Assignment 2: Task 1 - CFA

Group 11

30.10.2018

0. Load Libraries

```
library(tidyverse)
library(lavaan)
library(semPlot)
library(lessR)
```

1. Problem Statement

We are given a covariance matrix and other descriptive statistics (mean, standard deviation, number of observations) of 8 variables that are assumed to measure abilities of professional basketball players. The task is to check whether a hypothesized model with two latent factors called “frontcourt skills” and “backcourt skills” fits the variance/covariance structure of the given data well, compare this model to another with three latent factors and finally perform a test for equality of two parameters. The goal is to find shared structures in the variances of the variables, i.e. factors that have an influence on groups of variables and can therefore be determined underlying unobservable factors that help explaining the variance/covariance structure of the data.

2. Descriptive Statistics

With only the variance matrix there is no possibility of checking for outliers in the data. One can see from looking at the complete table that the means of X3 and X8 and are substantially higher than the rest. This is not important for the analysis and can be safely ignored.

```
basketdf <- read.table("http://feb.kuleuven.be/martina.vandebroek/public/STATdata/basketball.txt",
                       header=T)
basket_cov <- basketdf %>% filter(X_TYPE_ == "COV") %>% select(2:ncol(basketdf))
rownames(basket_cov) <- colnames(basket_cov)
basket_cov <- as.matrix(basket_cov)
basketdf %>% filter(X_TYPE_ != "COV")
```

```
##   X_TYPE_         X1         X2         X3         X4         X5
## 1    MEAN   0.313544   0.757588   3.354687   1.137500   0.508394
## 2     STD   0.126765   0.101748   2.051235   0.420963   0.063846
## 3       N 320.000000 320.000000 320.000000 320.000000 320.000000
##           X6         X7         X8
## 1   0.755000   1.495000   5.056875
## 2   0.655414   1.178294   2.068820
## 3 320.000000 320.000000 320.000000
```

3. Assumptions

The most important condition to obtain meaningful results is that the number of inputs (unique values in variance/covariance matrix) is higher than the number of estimated parameters. Here we have 8 variables so our number of inputs is (8*(8+1))/2=36. For the first model we estimate 8 loadings, 8 error variances of the variables and 1 covariance between factors which lead to a total of 17 estimated parameters. The model is therefore well identified (as is the second model where 19 parameters are estimated). We confirm this with the inspect function where nonzero integers in the output are parameters that are to be estimated. A (approximate) multivariate normal distribution of the data also has to be assumed in order to estimate the paramters with maximum likelihood.

```
# Specify two-factor model
model1 <- '
# latent variables
backcourt =~ X1 + X2 + X3 + X4
frontcourt =~ X5 + X6 + X7 + X8
'
fit1 <- lavaan::sem(model1, sample.nobs = 320, sample.cov = basket_cov, orthogonal = F)

# Show number of estimated parameters
inspect(fit1)$psi
```

```
##            bckcrt frntcr
## backcourt  15
## frontcourt 17     16
```

4. Method and interpretation

Part A:

The model was already fit in section 3 with the sem function from the lavaan package. To assess the model fit we constructed a function that computes the goodness-of-fit (GFI) because this indicator is not given in the output of lavaan.

```
# Specify GFI function
GFI <- function(Si, Sobs){
  if( class(Si) == "list"){
    Si <- as.matrix(as.data.frame(Si$cov))
  }

  nominator <- sum(diag(solve(Si) %*% Sobs - diag(ncol(Sobs))))^2
  denominator <- sum(diag(solve(Si) %*% Sobs))^2
  return(1 - (nominator / denominator))
}
```

By using the summary (see output in Appendix), resid and modindices functions of the lavaan package the model are inspected. Overall the model does not fit the data well, for example the null hypothesis that the estimated covariance matrix is equal to the observed covariance matrix is rejected with a p-value of 0.

Overall the model does not fit the data well, the null hypothesis that the estimated covariance matrix is equal to the observed covariance matrix is rejected with a p-value of 0 and the comparative fit index (CFI) which compares the model with a baseline null model is 0.892. This value indicates that the model, while being better than the baseline model, can not be considered a good enough improvement to the baseline model. The GFI on the other hand is equal to 0.9999902. This tells us that the calculated covariance matrix (Si) is very similar to the observed one (Sobs). Finally, all estimated parameters are significant (the output of the summary function was deprecated, so that it does not show the parameter estimates because we do not interpret them in any way, the estimates were, however, checked and turned out to be significant).

```
# Display standardized residuals
resid(fit1, type = "standardized")
```

```
## $type
## [1] "standardized"
##
## $cov
##    X1     X2     X3     X4     X5     X6     X7     X8
## X1  0.000
## X2  3.282  0.000
## X3 -3.381 -1.090  0.000
## X4 -2.510 -1.897  6.089  0.000
## X5 -0.124  0.643 -0.648  0.859  0.000
## X6  0.583 -0.638 -1.506  1.465  2.924  0.000
## X7 -1.864  1.439 -2.563 -1.476 -1.325 -5.744  0.000
## X8  3.842  2.795  1.109 -0.764 -0.526  3.243  1.592  0.000
```

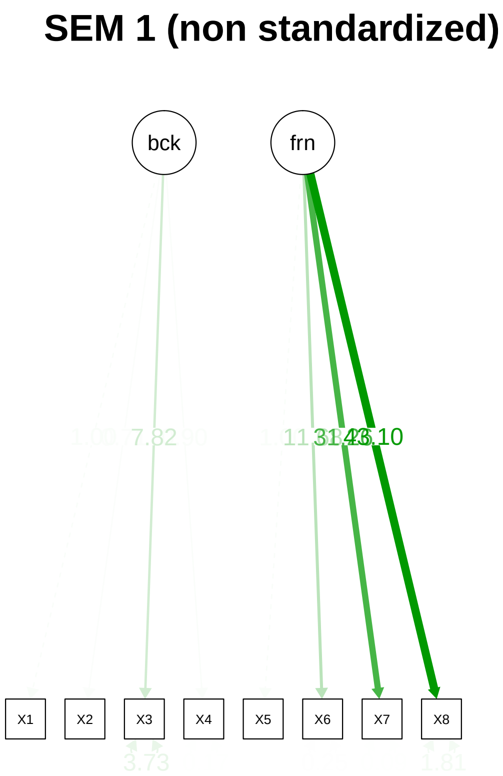
The resid function gives us the matrix of the standardized residuals. As a heuristic rule, standardized residuals over 1.96 (absolute value) are considered bad and indicate a bad fit. This is the case for the following 9 covariances: (X4,X3), (X7,X6), (X8,X1), (X3,X1), (X2,X1), (X8,X6), (X6,X5), (X8,X2) and (X7,X3).

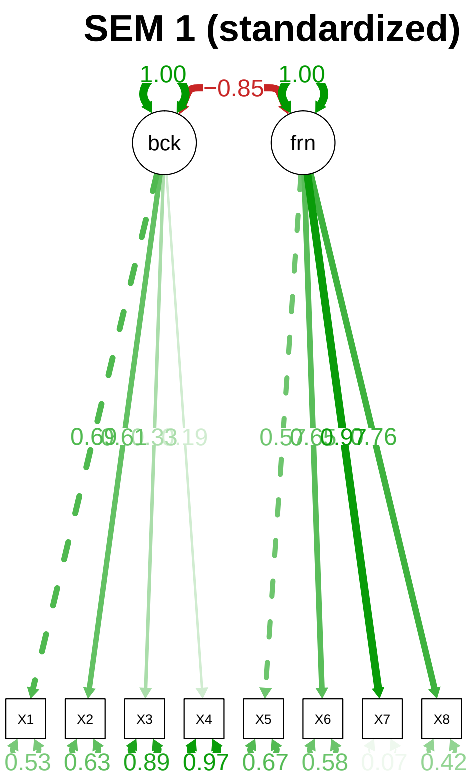
```
# Show modification indices greater than 10
modindices(fit1, sort. = T, minimum.value = 10)
```

```
##          lhs op rhs     mi     epc sepc.lv sepc.all sepc.nox
## 41        X3 ~~  X4 43.828   0.302   0.302    0.378    0.378
## 53        X6 ~~  X7 23.085  -0.136  -0.136   -0.898   -0.898
## 23 backcourt =~  X8 22.190  17.306   1.510    0.731    0.731
## 22 backcourt =~  X7 18.418 -10.325  -0.901   -0.766   -0.766
## 28        X1 ~~  X2 16.622   0.004   0.004    0.506    0.506
## 54        X6 ~~  X8 11.856   0.154   0.154    0.230    0.230
## 34        X1 ~~  X8 10.642   0.028   0.028    0.223    0.223
## 29        X1 ~~  X3 10.552  -0.040  -0.040   -0.223   -0.223
```

The modindices function outputs the modification indices that tell us the approximate decrease in χ^2 if that parameter would be estimated, mi is the decrease in the χ^2 statistic (LM Stat) in SAS) and epc is the expected change of the estimated parameter (Parm Change in SAS). Here, we see that for example the estimation of the covariance between X3 and X4 (or equivalenty lifting the constraint that this covariance is 0) would result in a decrease of the χ^2 statistic of approximately 43.828.

```
# Plot of standardized and unstandardized loading estimates
par(mfrow=c(1,2))
semPaths(fit1, "est",edge.label.cex = 1.4, curvePivot = TRUE)
title("SEM 1 (non standardized)", line = 3)
semPaths(fit1, "std",edge.label.cex = 1.4, curvePivot = TRUE)
title("SEM 1 (standardized)", line = 3)
```


SEM 1 (non standardized)


SEM 1 (standardized)

```r
par(mfrow=c(1,1))
```

Here, two path plots are shown to illustrate the underlying model specification and the estimated parameters. The left one uses non standardized parameters, which is why most of the paths are hardly visible because the parameters are low. In order to be able to see all paths, also the standarrdized plot, where the problem of paths fading away is greatly reduced, is shown.

Part B:

The model is respecified by splitting the backcourt factor into 2 factors: "shooting skills" and "neuromuscular coordination" while the frontcourt factor is equivalent to the first model. See again the output values of the summary function in the Appendix.

```r
# Specify three factor model
model2 <- '
# latent variables
shoot =~ X1 + X2
neuro =~ X3 + X4
frontcourt =~ X5 + X6 + X7 + X8
'
# Fitting the model
fit2 <- lavaan::sem(model2, sample.nobs = 320, sample.cov = basket_cov, orthogonal = F)
```

The p-value still indicates a good fit, but the CFI improved a lot in comparison to the first model (0.975 instead of 0.892), all other indicators are not significantly different to the first model (p-value still bad, GFI still good, parameters still significant). Because these indicators do not allow us to definitively say one model is better than another, both the AIC and BIC, which include a penalty for model complexity, are compared between the models. The second model has lower values on both of them (AIC: 2100.703 < 2156.084, BIC:

2172.301 < 2220.145), so one can safely assume that the second model fits the data better.

```
resid(fit2, type = "standardized")
```

```
## $type
## [1] "standardized"
##
## $cov
##     X1     X2     X3     X4     X5     X6     X7     X8
## X1  0.000
## X2  0.000  0.000
## X3 -0.553  1.247  0.000
## X4 -0.902 -0.666  0.000  0.000
## X5 -0.131  0.316 -0.154  1.266  0.000
## X6  0.566 -1.047 -0.949  1.981  2.952  0.000
## X7 -2.386 -1.501 -1.286  0.092 -1.439 -5.924  0.000
## X8  3.867  2.249  2.218 -0.234 -0.474  3.280  1.740  0.000
```

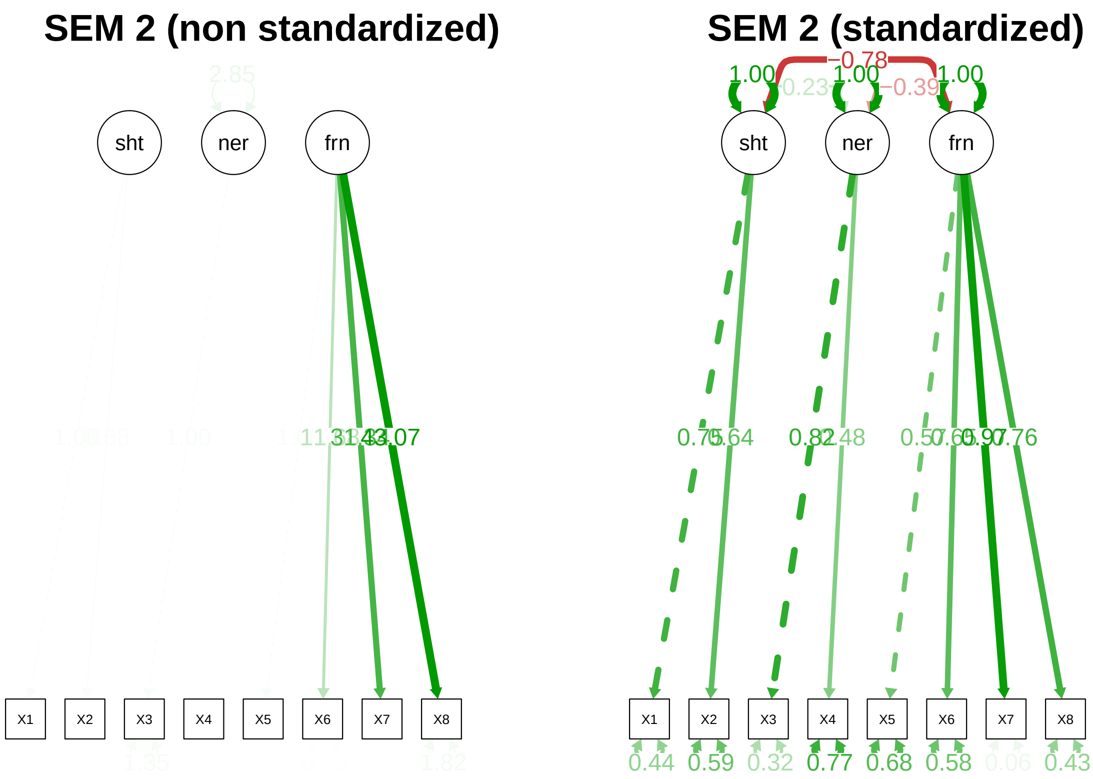
There are now less standardized residuals with a n absolute value over 1.96 (7 instead of 9) and also the values are smaller than those for the first model. From this one can conclude that the second model is an improvement to the first one.

```
# Show modification indices greater than 10
modindices(fit2, sort. = T, minimum.value = 10)
```

```
##      lhs op rhs     mi    epc sepc.lv sepc.all sepc.nox
## 64    X6 ~~  X7 24.343 -0.139  -0.139   -0.936   -0.936
## 28 shoot =~  X8 14.341  8.533   0.812    0.393    0.393
## 65    X6 ~~  X8 12.153  0.156   0.156    0.232    0.232
## 45    X1 ~~  X8 10.823  0.027   0.027    0.242    0.242
## 27 shoot =~  X7 10.078 -4.389  -0.417   -0.355   -0.355
```

The mi values are now lower, so it is more difficult than in the first case to improve the model further.

```
# Plot of standardized and unstandardized loading estimates
par(mfrow=c(1,2))
semPaths(fit2, "est",edge.label.cex = 1.4, curvePivot = TRUE)
title("SEM 2 (non standardized)", line = 3)
semPaths(fit2, "std",edge.label.cex = 1.4, curvePivot = TRUE)
title("SEM 2 (standardized)", line = 3)
```



```
par(mfrow=c(1,1))
```

The model could probably be improved by introducing new factor for X6 and X7 and/or a loading of the factor shooting skills on X8 (these two have the highest mi value in the modindices output).

Part C:

In order to be able to test whether the loadings of X7 and x8 are the same, a two sided t-test using the standardized loadings needs to be applied. The null hypothesis is that the loadings are equal, whereas the alternative hypothesis is that both values are not equal. The significance level is set to five percent and the remaining degrees of freedom are 301, as 19 variables were estimated.

```
# Extract standardized estimates and standard errors
estX7X8 <- standardizedSolution(fit2)$est.std[7:8]
seX7X8 <- standardizedSolution(fit2)$se[7:8]

# Set significance level
alpha <- 0.05

# Calculate t-statistic
tStat <- ((estX7X8[1] - estX7X8[2]) / seX7X8[1])

# Decide whether to reject H0 or not
# TRUE: Reject H0
# FALSE: Accept H0
ifelse(tStat < 0, tStat < -qt(p=1-(alpha/2), df=320-19),
                  tStat > qt(p=1-(alpha/2), df=320-19))
```

```
## [1] TRUE
```

The resulting test statistic is equal to 13.34606, which is much larger then the threshold value of approximately 1.94. Therefore the null hypothesis can be rejected, meaning that the loadings are significantly different from each other.

5. Alternative solutions

We want to test another two factor model, which consists offensive and defensive attributes. Therefore, X1, X2, X3 and X5 are measures of offensive attributes and the remaining observed variables are measures of the defensive abilities.

```
modelAlt <- '
# latent variables
offensive =~ X1 + X2 + X3 + X5
defensive =~ X4 + X6 + X7 + X8
'

# Fitting the model
fit3 <- lavaan::sem(modelAlt, sample.nobs = 320, sample.cov = basket_cov, orthogonal = F)
```

Analysing the fit criteria (see output in Appendix) results in the model being a bad one. The null hypothesis that the estimated covariance matrix is equal to the observed covariance matrix is rejected. The CFI value of 0.885 is lower than the desired value of 0.95 and the GFI value of approximately 0.2136 is far below the threshold value aswell. The AIC and BIC values are also higher than the values from the other two values.

```
# Show modification indices greater than 10
modindices(fit3, sort. = T, minimum.value = 10)
```

```
##          lhs op rhs     mi     epc sepc.lv sepc.all sepc.nox
## 42        X3 ~~  X4 43.595   0.298   0.298    0.372    0.372
## 23 offensive =~  X8 24.623  45.228   3.633    1.759    1.759
## 53        X6 ~~  X7 24.582  -0.146  -0.146   -1.043   -1.043
## 28        X1 ~~  X2 19.814   0.003   0.003    0.309    0.309
## 54        X6 ~~  X8 13.529   0.167   0.167    0.246    0.246
## 34        X1 ~~  X8 12.337   0.030   0.030    0.230    0.230
## 22 offensive =~  X7 11.219 -24.551  -1.972   -1.676   -1.676
```

One could improve this model by creating a new factor containing X3 and X4, which would result in the χ^2 statistic being improved by 43.828.

6. Conclusion

The main purpose of CFA is to identify a certain factor model, which fits the data well based on several criteria, as mentioned above. After respecifying the first factor model, we came out with a better solution in terms of model fit. Nonetheless, this model could still not be completely validated, as the null hypothesis of the χ^2 test is rejected. The proposed alternative model turns out to be worse than all the other models.

Appendix

```
# Output of model 1 fit measures
summary(fit1,fit.measures = TRUE, estimates = F)
```

```
## lavaan 0.6-3 ended normally after 88 iterations
##
##   Optimization method                            NLMINB
##   Number of free parameters                          17
##
##   Number of observations                            320
##
##   Estimator                                          ML
##   Model Fit Test Statistic                      113.897
##   Degrees of freedom                                 19
##   P-value (Chi-square)                            0.000
##
## Model test baseline model:
##
##   Minimum Function Test Statistic               909.437
##   Degrees of freedom                                 28
##   P-value                                         0.000
##
## User model versus baseline model:
##
##   Comparative Fit Index (CFI)                     0.892
##   Tucker-Lewis Index (TLI)                        0.841
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)               -1061.042
##   Loglikelihood unrestricted model (H1)       -1004.093
##
##   Number of free parameters                          17
##   Akaike (AIC)                                 2156.084
##   Bayesian (BIC)                               2220.145
##   Sample-size adjusted Bayesian (BIC)          2166.224
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                           0.125
##   90 Percent Confidence Interval          0.103  0.148
##   P-value RMSEA <= 0.05                           0.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                            0.073
```

```
GFI(Si = fitted(fit1), Sobs = basket_cov)
```

```
## [1] 0.9999902
```

```r
# Output of model 2 fit measures
summary(fit2,fit.measures = TRUE, estimates = F)
```

```
## lavaan 0.6-3 ended normally after 96 iterations
##
##   Optimization method                           NLMINB
##   Number of free parameters                         19
##
##   Number of observations                           320
##
##   Estimator                                         ML
##   Model Fit Test Statistic                      54.516
##   Degrees of freedom                                17
##   P-value (Chi-square)                           0.000
##
## Model test baseline model:
##
##   Minimum Function Test Statistic              909.437
##   Degrees of freedom                                28
##   P-value                                        0.000
##
## User model versus baseline model:
##
##   Comparative Fit Index (CFI)                    0.957
##   Tucker-Lewis Index (TLI)                       0.930
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -1031.351
##   Loglikelihood unrestricted model (H1)      -1004.093
##
##   Number of free parameters                         19
##   Akaike (AIC)                                2100.703
##   Bayesian (BIC)                              2172.301
##   Sample-size adjusted Bayesian (BIC)         2112.036
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.083
##   90 Percent Confidence Interval          0.059  0.108
##   P-value RMSEA <= 0.05                          0.014
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.038
```

```r
GFI(Si = fitted(fit2), Sobs = basket_cov)
```

```
## [1] 0.9999902
```

```r
# Output of model 3 fit measures
summary(fit3,fit.measures = TRUE, estimates = F)
```

```
## lavaan 0.6-3 ended normally after 90 iterations
##
##   Optimization method                           NLMINB
##   Number of free parameters                         17
##
##   Number of observations                           320
##
##   Estimator                                         ML
##   Model Fit Test Statistic                     120.647
##   Degrees of freedom                                19
##   P-value (Chi-square)                           0.000
##
## Model test baseline model:
##
##   Minimum Function Test Statistic              909.437
##   Degrees of freedom                                28
##   P-value                                        0.000
##
## User model versus baseline model:
##
##   Comparative Fit Index (CFI)                    0.885
##   Tucker-Lewis Index (TLI)                       0.830
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -1064.417
##   Loglikelihood unrestricted model (H1)      -1004.093
##
##   Number of free parameters                         17
##   Akaike (AIC)                                2162.834
##   Bayesian (BIC)                              2226.895
##   Sample-size adjusted Bayesian (BIC)         2172.974
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.129
##   90 Percent Confidence Interval          0.108  0.152
##   P-value RMSEA <= 0.05                          0.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.075
```

```r
GFI(Si = fitted(fit3), Sobs = basket_cov)
```

```
## [1] 0.2136608
```

Assignment 2: Task 2 - SEM

Group 11

30.10.2018

0. Load Libraries

```
library(lavaan)
library(plotrix)
library(lessR)
library(semPlot)
```

1. Problem Statement

We are given a dataset containing the covariance matrix on some basic statistical measures and two advanced measurements of 320 basketball players during the NBA 2017-2018 season. There are also two advanced measurements (Y1, Y2) about offensive performance of the players. Additionally, we are given a hypothesized model of the causal relationships between the observed constructs, which should be tested for the statistics of the fit.

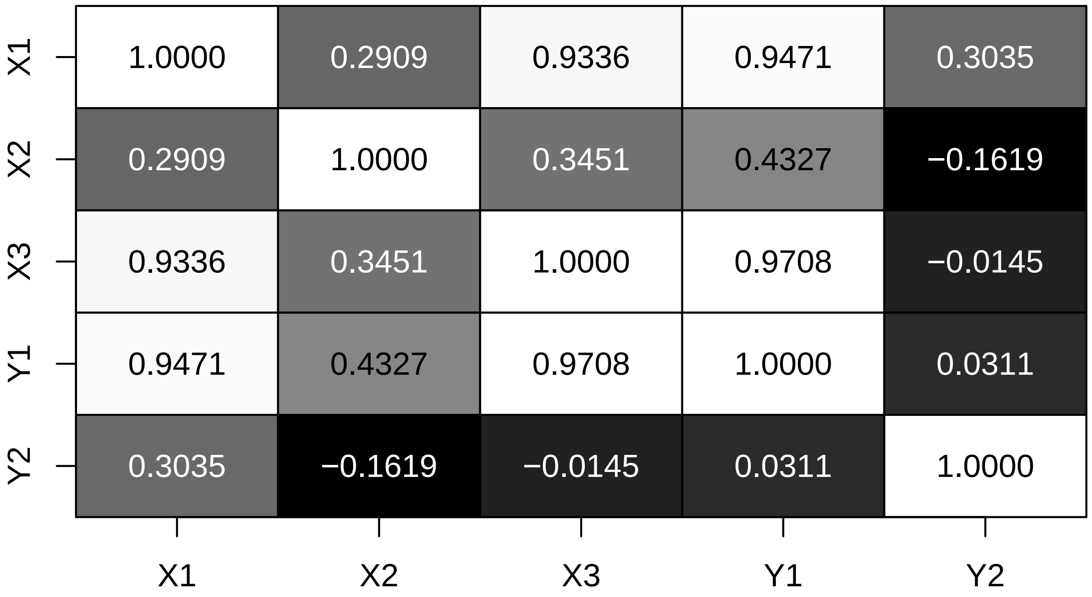
2. Descriptive Statistics

In the raw data, we are given the covariance matrix of 320 observations measured by 8 variables. For interpretability we converted the covariance matrix into a correlation matrix and visualized it in a heatmap. The correlations are either very high (>0.9) or very low (<0.2) with some mid ranged values in between a range of [0.29, 0.45]. Within the basic statistical measurements only X1 and X3 are highly correlated. Both X1 and X3 are strongly correlated with one of the advanced measures, Y1.

```
cov <-read.delim("../data/advanced_basketball.txt", header = TRUE, sep="",dec = ".", skipNul = FALSE)
cov <- cov[-1]
cov <- cov[-c(1,2,3),]
rownames(cov) = c("X1","X2","X3","Y1","Y2")
colnames(cov) = c("X1","X2","X3","Y1","Y2")
cov <- as.matrix(cov)
cov
```

```
##           X1        X2        X3        Y1        Y2
## X1 20.723150  2.716768 14.430665 22.819525  0.065662
## X2  2.716768  4.207564  2.403467  4.697706 -0.015785
## X3 14.430665  2.403467 11.529588 17.447140 -0.002344
## Y1 22.819525  4.697706 17.447140 28.013290  0.007814
## Y2  0.065662 -0.015785 -0.002344  0.007814  0.002258
```

```
color2D.matplot(cov2cor(cov),show.values=4,axes=FALSE,
  xlab="",ylab="")

axis(1,at=0.5:4.5,labels=rownames(cov))
axis(2,at=4.5:0.5,labels=colnames(cov))
```



	X1	X2	X3	Y1	Y2
X1	1.0000	0.2909	0.9336	0.9471	0.3035
X2	0.2909	1.0000	0.3451	0.4327	−0.1619
X3	0.9336	0.3451	1.0000	0.9708	−0.0145
Y1	0.9471	0.4327	0.9708	1.0000	0.0311
Y2	0.3035	−0.1619	−0.0145	0.0311	1.0000

3. Assumptions

First, we assume multivariate normal distribution. When the maximum likelihood method is used the data is required to be multivariate normally distributed, as small changes in multivariate normality can lead to a large difference in the chi-square test. Additionally, equations must be greater than the estimated parameters or models should be over identified or exact identified. Under-identified models are not considered. Specifically, we have 15 inputs and 11 parameters to be estimated, which satisfies the condition of $p \leq inputs$.

4. Method

The objective of structural equation models is to explain the covariances of the observed variables in terms of the relationships of these variables to the assumed underlying latent variables and the relationships postulated between the latent variables themselves. SEM involves estimating a number of model parameters from the observed covariance matrix S^{obs} to minimize the difference between this matrix and a matrix S^{i} implied by the fitted model. The most commonly used method of estimation for SEM is maximum likelihood under the assumption that the observed data have a multivariate normal distribution. Y1 and Y2 represent endogenous variables, which means that their variance is considered to be explained in part by other variables in the model. X1, X2 and X3 are exgoneous variables whose variance is assumed to be caused by variables not in the causal model.

First, we construct the hypothesized model in lavaan syntax and execute the sem function with a Maximum Likelihood estimation. There are other estimation procedures, but ML is the most efficient when the dataset is large, the data are multivariate normal distributed and the input is a covariance matrix. All these assumptions are satisfied for our task.

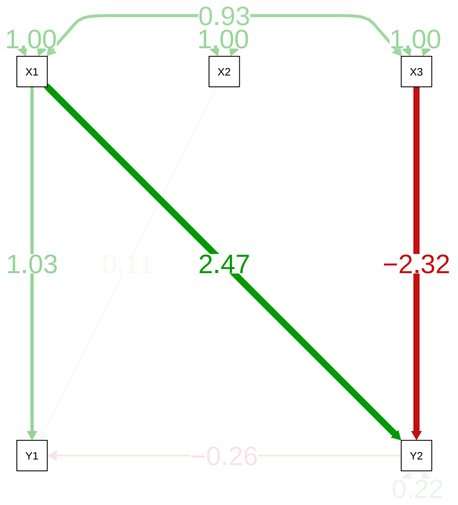
```
model <- '
# regressions
Y1 ~ X1 + X2 + Y2
Y2 ~ X1 + X3
# residual covariance
X1 ~~ X3
'

GFI <- function(Si, Sobs){
  if( class(Si) == "list"){
    Si <- as.matrix(as.data.frame(Si$cov))
  }

  nominator <- sum(diag(solve(Si) %*% Sobs - diag(ncol(Sobs))))^2
  denominator <- sum(diag(solve(Si) %*% Sobs))^2
  return(1 - (nominator / denominator))
}


fit <- lavaan::sem(model,sample.nobs = 320,sample.cov=cov, fixed.x=F)
# see appendix
#summary(fit, fit.measures=TRUE)
Si2 <- as.matrix(as.data.frame(fitted(fit)$cov))
Si2 <- corReorder(R = Si2, vars = c(X1, X2, X3, Y1, Y2), heat.map = F)
```

```
GFI(Si = Si2, Sobs = cov)
```

```
## [1] 0.9999902
```

```
semPaths(fit, "std",edge.label.cex = 2.0, cex=1.5, curvePivot = TRUE)
```

```
## Warning in qgraph(Edgelist, labels = nLab, bidirectional = Bidir, directed
## = Directed, : The following arguments are not documented and likely not
## arguments of qgraph and thus ignored: cex
```



5. Interpretation

The Goodness-of-fit (GFI) value of 0.99 (>0.95) and Comparative Fit Index (CFI) 0.977 (>0.95) imply that the data is a good fit. Moreover, AIC (Akaike's Information Criterion) estimates the quality of each model, and it is better to obtain a smaller AIC. In our model it is at 3515.273, but this information is only useful when comparing it with a modified version of the model. The Standardized Root Mean Square Residual (SRMR) is defined as the standardized difference between the observed correlation and the predicted correlation and anything below 0.8 is considered a good fit, while we have a value of 0.152 in this model. The overall p- value (0.000) of the Chi-Square test suggests to reject H0, indicating that we cannot assume S^i and S^{obs} to be equal. This test is very sensitive to sample size and given the large sample of 320 observations might be cause for this. We attach more weight to the GFI index, which represents the proportion of the variances and covariances explained by the model. CFI measures whether the model fits the data better than a more restricted baseline model and we therefore conclude, that the model fits the data. The full summary of all the values can be seen in the appendix.

Nevertheless, we attempted to improve the model by adjusting the hypothesized structure. We noticed, that the residual covariances of the model are not too small indicating that there are some related causes which have not been identified yet. This holds especially for X2 and Y2 and X3 and X2, and X1 and X2.

```
modindices(fit, sort=T)
```

```
##    lhs op rhs     mi    epc sepc.lv sepc.all sepc.nox
## 24  X2  ~  Y1 38.835  0.142   0.142    0.355    0.355
## 27  X2  ~  X3 38.105  0.208   0.208    0.345    0.208
## 26  X2  ~  X1 27.087  0.131   0.131    0.291    0.291
## 28  X3  ~  Y1 18.814  0.446   0.446    0.671    0.671
## 31  X3  ~  X2 13.445  0.122   0.122    0.073    0.036
## 19  Y2  ~  X2  9.414 -0.002  -0.002   -0.080   -0.039
## 25  X2  ~  Y2  8.392 -6.991  -6.991   -0.162   -0.162
## 20  X1  ~  Y1  6.661 -0.364  -0.364   -0.408   -0.408
## 13  Y1 ~~  X1  5.414 -0.362  -0.362   -0.102   -0.102
## 14  Y1 ~~  X3  5.414  0.252   0.252    0.095    0.095
## 12  Y1 ~~  Y2  5.414  0.003   0.003    0.149    0.149
## 17  Y1  ~  X3  5.414  0.171   0.171    0.113    0.069
## 22  X1  ~  X2  2.428 -0.069  -0.069   -0.031   -0.015
## 18  Y2  ~  Y1  0.007  0.000   0.000   -0.013   -0.013
```

```
resid(fit)$cov
```

```
##    Y1     Y2     X1     X2     X3
## Y1  1.886
## Y2 -0.004  0.000
## X1  0.714  0.000  0.000
## X2  3.577 -0.016  2.708  0.000
## X3  0.692  0.000  0.000  2.396  0.000
```

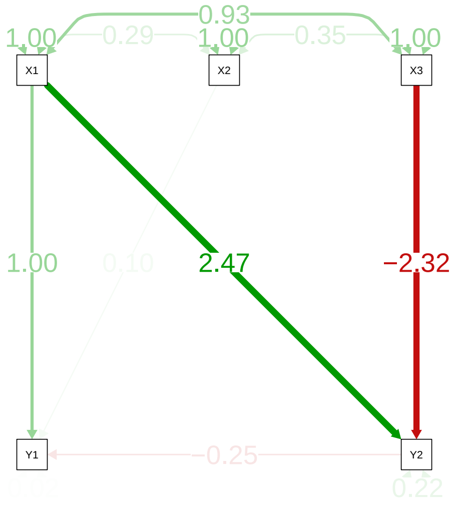
Modification indices state how model fit would change if you added new parameters to the model. We sorted the modification indices by mi which is an estimate of how much the model fit would improve if each parameter were added. The first suggestion is a regression of Y1, "an average estimate of how many offensive plays the player was involved in, per game", on X2, "average number of assists per game". However, this would not make any sense, since X2 is a basic measurement that an advanced measure should not regress on. The suggested modifications do not take the logic behind the model into account, as SEM is not an exploratory technique, but rather used to validate a theoretically solid model. It is important to consider the overall objective of CFA is to confirm, not to explore, which is why following all modification indices can be dangerous, as it can lead to over-fitting the data and decreasing the generalizability of the results.

We adjusted the second modification suggestion and added a residual covariance between X1, "average points per game” and X2, “average number of assists per game”. This resulted in an improved CFI (0.983) and an improved AIC (3502.237) and SRMR (0.131), but almost no change in all the other evaluation measures. The p-value still indicates a rejection of H0. We then also added a covariance between X1 and X2 to the model which further improved the CFI (0.994) and AIC (3475.934) also decreased. The SRMR greatly decreased to a value of 0.021, surpassing the lower threshold of 0.8 which is considered to be an indicator of a good fit. The remaining modification indices and residuals imply that further improvement could be yielded by adding the regression of X2 on Y2. Theoretically, this would again not make sense, because shooting efficiency (Y2) is not dependant on assists (X2). The structure is visualized below.

```
model2 <- '
# regressions
Y1 ~ X1 + X2 + Y2
Y2 ~ X1 + X3
# residual covariance
X1 ~~ X3
X2 ~~ X3
X1 ~~ X2
'

fit2 <- lavaan::sem(model2,sample.nobs = 320,sample.cov=cov)
```

```
semPaths(fit2, "std",edge.label.cex = 2.0, cex=1.5, curvePivot = TRUE)
```

```
## Warning in qgraph(Edgelist, labels = nLab, bidirectional = Bidir, directed
## = Directed, : The following arguments are not documented and likely not
## arguments of qgraph and thus ignored: cex
```



```
#summary(fit2, fit.measures=T)
```

It is also possible, to identify relationships, which are not needed. “Shot attempts” (X3) has an indirect effect on “offensive plays” (Y1) through “shooting efficiency” (Y2). As Y1 is already regressed on by one of the variables determining shooting efficiency (X1), and deleting the regression of Y2 on Y1 did not indicate a large drop in Chi-Square from the modification indices, we tested a model with a direct link between X3 and Y1 and no regression of Y2 on Y1. This model had slightly improved statistical measures (AIC= 3472.458, p-value= 0.001, GIF and SMRS stayed equal) in comparison to the previous ‘model2’.

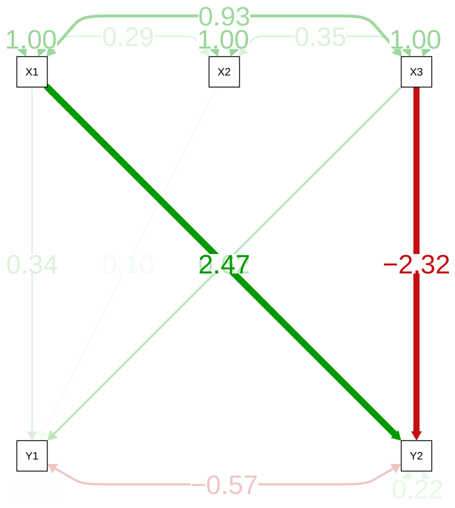
```
model3 <- '
# regressions
Y1 ~ X1 + X2 + X3
Y2 ~ X1 + X3
# residual covariance
X1 ~~ X3
X2 ~~ X3
X1 ~~ X2
'

fit3 <- lavaan::sem(model3,sample.nobs = 320,sample.cov=cov)

#summary(fit3, fit.measures=T)
Si2 <- as.matrix(as.data.frame(fitted(fit3)$cov))
Si2 <- corReorder(R = Si2, vars = c(X1, X2, X3, Y1, Y2), heat.map = F)
```

```
GFI(Si = Si2, Sobs = cov)
```

```
## [1] 0.9999902
```

```
semPaths(fit3, "std",edge.label.cex = 2.0, cex=1.5, curvePivot = TRUE)
```

```
## Warning in qgraph(Edgelist, labels = nLab, bidirectional = Bidir, directed
## = Directed, : The following arguments are not documented and likely not
## arguments of qgraph and thus ignored: cex
```



We conducted an anova to compare the modified models which confirmed the significant improvement of 'model3' over 'model2'.

```
anova(fit3,fit2)
```

```
## Chi Square Difference Test
##
##      Df    AIC    BIC  Chisq Chisq diff Df diff Pr(>Chisq)
## fit3  1 3472.5 3525.2 10.965
## fit2  2 3475.9 3524.9 16.441     5.4764       1    0.01927
```

6. Alternatives

In our attempt at improving the model, we incrementally added variables, as described before and yielded another model, as seen below, which achieves the statistically best values, with a p-value of 0.019, CFI of 0.998, RMSEA of 0.118, AIC of 3467.0 and GFI of 0.999. However, we do not consider the structure useful, as it makes the model overly complex. It also introduces a regression of "number of assists" (X2) on "shooting efficiency" (Y2) which is not logically consistent. We therefore discarded this modification.

```
model4 <- '
# regressions
Y1 ~ X1 + X2 + Y2
Y2 ~ X1 + X3 + X2
# residual covariance
X1 ~~ X3
X2 ~~ X3
X1 ~~ X2
'
fit4 <- lavaan::sem(model4,sample.nobs = 320,sample.cov=cov)
```

7. Conclusion

We evaluated the fit of a path model using different evaluation measures and concluded that the model is a good representation of the data. We attempted to improve the model and assessed multiple structural changes. We compared the fit of the improved models with each other, which revealed that 'model3' fits the data even better, while keeping the underlying theoretical concepts intact.

Appendix

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

```
## lavaan 0.6-3 ended normally after 88 iterations
##
##   Optimization method                           NLMINB
##   Number of free parameters                         11
##
##   Number of observations                           320
##
##   Estimator                                         ML
##   Model Fit Test Statistic                      59.781
##   Degrees of freedom                                 4
##   P-value (Chi-square)                           0.000
##
## Model test baseline model:
##
##   Minimum Function Test Statistic             2430.889
##   Degrees of freedom                                10
##   P-value                                        0.000
##
## User model versus baseline model:
##
```

```
##   Comparative Fit Index (CFI)                    0.977
##   Tucker-Lewis Index (TLI)                       0.942
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -1746.637
##   Loglikelihood unrestricted model (H1)      -1716.746
##
##   Number of free parameters                         11
##   Akaike (AIC)                                3515.273
##   Bayesian (BIC)                              3556.725
##   Sample-size adjusted Bayesian (BIC)         3521.835
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.209
##   90 Percent Confidence Interval          0.164  0.257
##   P-value RMSEA <= 0.05                          0.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.152
##
## Parameter Estimates:
##
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##   Standard Errors                             Standard
##
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   Y1 ~
##     X1                1.156    0.010  114.968    0.000
##     X2                0.264    0.021   12.394    0.000
##     Y2              -28.323    0.964  -29.394    0.000
##   Y2 ~
##     X1                0.026    0.001   33.995    0.000
##     X3               -0.032    0.001  -31.937    0.000
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   X1 ~~
##     X3               14.386    1.178   12.207    0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .Y1                0.607    0.048   12.649    0.000
##    .Y2                0.000    0.000   12.649    0.000
##     X1               20.658    1.633   12.649    0.000
##     X3               11.494    0.909   12.649    0.000
##     X2                4.194    0.332   12.649    0.000
```

```r
summary(fit2, fit.measures=TRUE)
```

```
## lavaan 0.6-3 ended normally after 98 iterations
##
##   Optimization method                           NLMINB
##   Number of free parameters                         13
##
##   Number of observations                           320
##
##   Estimator                                         ML
##   Model Fit Test Statistic                      16.441
##   Degrees of freedom                                 2
##   P-value (Chi-square)                           0.000
##
## Model test baseline model:
##
##   Minimum Function Test Statistic             2430.889
##   Degrees of freedom                                10
##   P-value                                        0.000
##
## User model versus baseline model:
##
##   Comparative Fit Index (CFI)                    0.994
##   Tucker-Lewis Index (TLI)                       0.970
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -1724.967
##   Loglikelihood unrestricted model (H1)      -1716.746
##
##   Number of free parameters                         13
##   Akaike (AIC)                                3475.934
##   Bayesian (BIC)                              3524.922
##   Sample-size adjusted Bayesian (BIC)         3483.688
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.150
##   90 Percent Confidence Interval          0.089  0.221
##   P-value RMSEA <= 0.05                          0.005
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.021
##
## Parameter Estimates:
##
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##   Standard Errors                             Standard
##
## Regressions:
##                     Estimate  Std.Err  z-value  P(>|z|)
##   Y1 ~
##     X1                1.156    0.011  108.422    0.000
##     X2                0.264    0.023   11.649    0.000
##     Y2              -28.323    0.981  -28.876    0.000
##   Y2 ~
##     X1                0.026    0.001   33.995    0.000
##     X3               -0.032    0.001  -31.937    0.000
##
## Covariances:
##                      Estimate  Std.Err  z-value  P(>|z|)
##   X1 ~~
##     X3               14.386    1.178   12.207    0.000
##   X2 ~~
##     X3                2.396    0.411    5.835    0.000
##   X1 ~~
##     X2                2.708    0.542    4.997    0.000
##
## Variances:
##                      Estimate  Std.Err  z-value  P(>|z|)
##    .Y1                0.607    0.048   12.649    0.000
##    .Y2                0.000    0.000   12.649    0.000
##     X1               20.658    1.633   12.649    0.000
##     X2                4.194    0.332   12.649    0.000
##     X3               11.494    0.909   12.649    0.000
```

```r
summary(fit3, fit.measures=TRUE)
```

```
## lavaan 0.6-3 ended normally after 85 iterations
##
##   Optimization method                           NLMINB
##   Number of free parameters                         14
##
##   Number of observations                           320
##
##   Estimator                                         ML
##   Model Fit Test Statistic                      10.965
##   Degrees of freedom                                 1
##   P-value (Chi-square)                           0.001
##
## Model test baseline model:
##
##   Minimum Function Test Statistic             2430.889
##   Degrees of freedom                                10
##   P-value                                        0.000
##
## User model versus baseline model:
##
##   Comparative Fit Index (CFI)                    0.996
##   Tucker-Lewis Index (TLI)                       0.959
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -1722.229
##   Loglikelihood unrestricted model (H1)      -1716.746
##
##   Number of free parameters                         14
##   Akaike (AIC)                                3472.458
##   Bayesian (BIC)                              3525.214
##   Sample-size adjusted Bayesian (BIC)         3480.808
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.176
##   90 Percent Confidence Interval          0.093  0.277
##   P-value RMSEA <= 0.05                          0.008
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.021
##
## Parameter Estimates:
##
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##   Standard Errors                             Standard
##
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   Y1 ~
##     X1                 0.398    0.032   12.304    0.000
##     X2                 0.266    0.023   11.811    0.000
##     X3                 0.959    0.044   21.816    0.000
##   Y2 ~
##     X1                 0.026    0.001   33.995    0.000
##     X3                -0.032    0.001  -31.937    0.000
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   X1 ~~
##     X3                14.386    1.178   12.207    0.000
##   X2 ~~
##     X3                 2.396    0.411    5.835    0.000
##   X1 ~~
##     X2                 2.708    0.542    4.997    0.000
##  .Y1 ~~
##    .Y2                -0.012    0.001   -8.861    0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .Y1                 0.884    0.070   12.649    0.000
##    .Y2                 0.000    0.000   12.649    0.000
##     X1                20.658    1.633   12.649    0.000
##     X2                 4.194    0.332   12.649    0.000
##     X3                11.494    0.909   12.649    0.000
```

```r
summary(fit4, fit.measures=TRUE)
```

```
## lavaan 0.6-3 ended normally after 100 iterations
##
##   Optimization method                           NLMINB
##   Number of free parameters                         14
```

```
##
##   Number of observations                           320
##
##   Estimator                                         ML
##   Model Fit Test Statistic                       5.476
##   Degrees of freedom                                 1
##   P-value (Chi-square)                           0.019
##
## Model test baseline model:
##
##   Minimum Function Test Statistic             2430.889
##   Degrees of freedom                                10
##   P-value                                        0.000
##
## User model versus baseline model:
##
##   Comparative Fit Index (CFI)                    0.998
##   Tucker-Lewis Index (TLI)                       0.982
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -1719.484
##   Loglikelihood unrestricted model (H1)      -1716.746
##
##   Number of free parameters                         14
##   Akaike (AIC)                                3466.969
##   Bayesian (BIC)                              3519.725
##   Sample-size adjusted Bayesian (BIC)         3475.320
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.118
##   90 Percent Confidence Interval          0.038  0.223
##   P-value RMSEA <= 0.05                          0.075
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.001
##
## Parameter Estimates:
##
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##   Standard Errors                             Standard
##
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   Y1 ~
##     X1                1.156    0.011  107.184    0.000
##     X2                0.264    0.023   11.402    0.000
##     Y2              -28.323    1.002  -28.265    0.000
##   Y2 ~
##     X1                0.026    0.001   34.124    0.000
##     X3               -0.032    0.001  -31.019    0.000
##     X2                -0.002    0.001   -3.340    0.001
##
## Covariances:
##                   Estimate  Std.Err  z-value  P(>|z|)
##   X1 ~~
##     X3                14.386    1.178   12.207    0.000
##   X2 ~~
##     X3                 2.396    0.411    5.835    0.000
##   X1 ~~
##     X2                 2.708    0.542    4.997    0.000
##
## Variances:
##                   Estimate  Std.Err  z-value  P(>|z|)
##    .Y1                 0.607    0.048   12.649    0.000
##    .Y2                 0.000    0.000   12.649    0.000
##     X1                20.658    1.633   12.649    0.000
##     X2                 4.194    0.332   12.649    0.000
##     X3                11.494    0.909   12.649    0.000
```