Assignment 1: Task 1 - CFA

Group 11

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

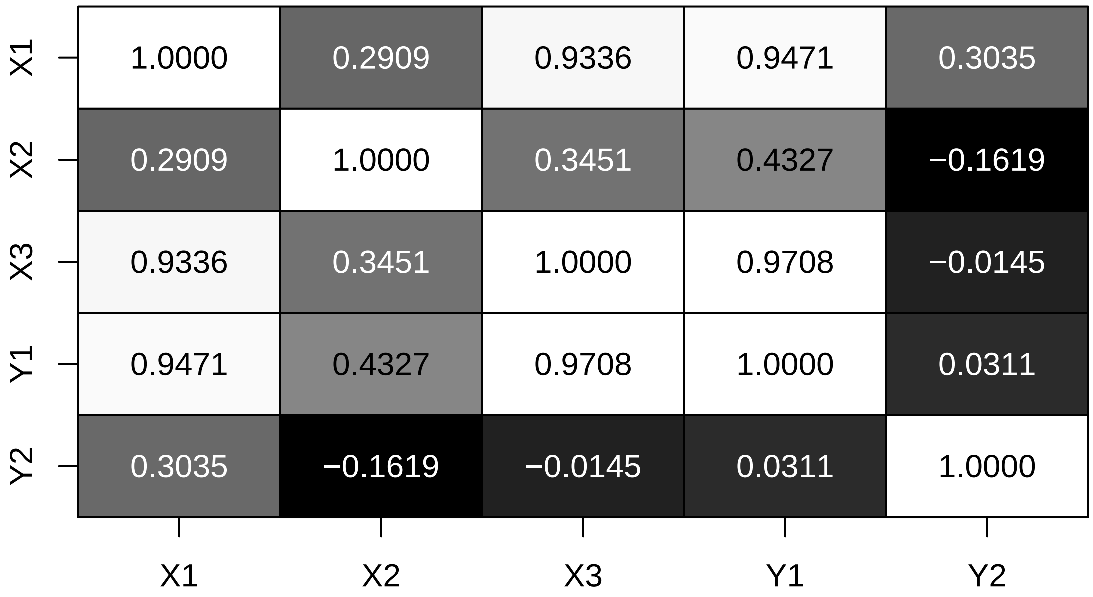
In the raw data, we are given the covariance matrix of 320 observations measured by 8 variables. For interpretability we converted the covariance matrix into a correlation matrix and visualized it in a heatmap. The correlations are either very high (>0.9) or very low (<0.2) with some mid ranged values 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

Multivariate normal distribution: The maximum likelihood method is used and assumed for multivariate normal distribution. Small changes in multivariate normality can lead to a large difference in the chi-square test. Linearity: A linear relationship is assumed between endogenous and exogenous variables. Model identification: Equations must be greater than the estimated parameters or models should be over identified or exact identified. Under identified models are not considered.

Uncorrelated error terms: Error terms are assumed uncorrelated with other variable error terms.

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. As all the variables in our dataset are measureable constructs, we will conduct a path analysis, which is essentially a SEM without latent factors. First, we construct the hypothesized model in lavaan syntax and execute the sem function.

```
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))
```

```r
  }

  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)
```

```
## Warning in lavaan::lavaan(model = model, sample.cov = cov, sample.nobs
## = 320, : lavaan WARNING: syntax contains parameters involving exogenous
## covariates; switching to fixed.x = FALSE
```

```r
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
```

```
#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)
```

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. RMSEA…. However, the overall p- value (0.000) of the Chi-Square test suggests to reject H0, indicating that we cannot assume S^1 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.

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 the model is almost under-identified and 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. (The GFI stays the same, Chi-Square p-value also, but smaller residuals, RMSEA and better CFI (.977, before .994))

We adjusted the second suggestion and added a residual covariance between X1, "average points per game" and X3, "average number of assists per game".
This resulted in an improved CFI (0.983) 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 X3 to the model which further improved the CFI (0.994) and also the RMSEA (0.15) decreased. The 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).

Since the CFA model should not be exploratory (i.e. you should know what parameters you want to include in the model before you begin), modification indices can be dangerous. If you make the changes they suggest, you run a serious risk of over-fitting your data and reducing the generalizability of your results.

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

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

```
modindices(fit2)
```

```
##    lhs op rhs     mi      epc  sepc.lv sepc.all sepc.nox
## 14  Y1 ~~  Y2  5.476    0.003    0.003    0.151    0.151
## 15  Y1 ~~  X1  5.476   -0.358   -0.358   -0.101   -0.101
## 16  Y1 ~~  X2  5.476   -1.831   -1.831   -1.147   -1.147
## 17  Y1 ~~  X3  5.476    0.243    0.243    0.092    0.092
## 18  Y2 ~~  X1 10.779   -0.071   -0.071   -0.706   -0.706
## 19  Y2 ~~  X2 10.779   -0.008   -0.008   -0.172   -0.172
## 20  Y2 ~~  X3 10.779    0.022    0.022    0.300    0.300
## 21  Y1  ~  X3  5.476    0.173    0.173    0.111    0.111
## 22  Y2  ~  Y1  0.008    0.000    0.000   -0.014   -0.014
## 23  Y2  ~  X2 10.779   -0.002   -0.002   -0.091   -0.091
## 24  X1  ~  Y1  3.193   -0.445   -0.445   -0.516   -0.516
## 25  X1  ~  Y2 10.779 -145.201 -145.201   -1.516   -1.516
## 28  X2  ~  Y1  8.683    0.500    0.500    1.286    1.286
## 29  X2  ~  Y2 10.779  -15.901  -15.901   -0.368   -0.368
## 32  X3  ~  Y1  1.296    0.184    0.184    0.286    0.286
## 33  X3  ~  Y2 10.779   46.020   46.020    0.644    0.644
```

```r
Si2 <- as.matrix(as.data.frame(fitted(fit2)$cov))
Si2 <- corReorder(R = Si2, vars = c(X1, X2, X3, Y1, Y2), heat.map = F)
```

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

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

Alternatives

Conclusion

We evaluated the fit of a path model consisting of manifest variables using different criteria and concluded that the model is a good representation of the data. We attempted to improve the model and assessed multiple structural changes but eventually decided, that the initial model should not be altered, as the relevant measures had already exceeded the respective threshold to be considered good and the remaining one - p value of Chi-Square - could not be improved significantly. Additionally, none of the alterations changed the GFI value, which is one of the most important measures for the model fit.