Using SEM Library in R software to Analyze Exploratory Structural Equation Models

Joan Guàrdia-Olmos[1], Maribel Peró-Cebollero[1,3], Sonia Benítez-Borrego[1], John Fox[2]
[1]University of Barcelona; Institute for Brain, Cognition and Behavior, Barcelona, SPAIN
[2]McMaster University, Toronto, CANADA
[3]Corresponding autor: Maribel Peró-Cebollero, e-mail: mpero@ub.edu

Abstract

Using R and its libraries is probably one of the most revolutionary operations in recent years regarding the use of statistical programs. Expanding their opportunities has far exceeded the first aspirations of promoters. The quantity and quality of R's libraries and the versatility of its procedures are key to its massive success both in programming and in the use of statistical methods. This increase is not only aimed at the most basic questions of statistics, but it is also essential for many applied researchers and even for teaching in Statistics.

Similarly, the Structural Equation Models (SEM) have become a commonly used technique in much of the applied research, especially in the domain of social sciences and R options for SEM, and they should be a piece to be considered and applied in parameter estimation.

Within the SEM, one of the latest contributions called Exploratory Structural Equation Models (ESEM) for a different perspective on exploratory factor structures. The aim of this paper is to show some of the feasible adaptations for parameter estimation through the **sem** library in the R Project.

Key Words: Exploratory Structural Equation Models, R Project.

1. Introduction

Since the appearance of the algorithms called Exploratory Structural Equation Models (ESEM), it was to be expected that this technique would be progressively applied in the study of the factorial structures of statistical techniques for dimension reduction. However, still few are the studies which have used the ESEM approximation to factorialize correlation structures. So far we have traced barely 15 studies within the 2009-2013 period in the style of Sánchez et al. (2012) or Marsh et al. (2010) as examples of good applications of this technique.

The statistical approximations to this type of studies are generally based on Confirmatory Factor Analyses (CFA) established from the algorithms of the measurement models of Structural Equation Models (SEM). That approximation has generally proved efficient enough to estimate construct validity and is so commonly used that much of the, for instance, psychometric structure of psychological assessment revolves around this type of techniques and their several derivatives. The use of Structural Equation Models (SEM) is a technique that has been extensively used for estimating factorial structures and, therefore, in psychometric studies of complex psychological phenomena. One of the advantages of this type of techniques, as indicated by Asparouhov and Muthén (2009), is that using CFA in the measurement model of SEM allows the researcher to propose a simple structure of the measurement model, since he is incorporating the previous substantive knowledge in the form of

certain restrictions in that measurement model, and that yields more parsimonious models. Nevertheless, unlike in the classic Exploratory Factor Analysis (EFA), setting forth a simple structure (where every item only loads in one factor and the remaining cross factor loadings are set to 0) can make the researcher specify a model more parsimonious than required by the data and, occasionally, that could lead to goodness-of-fit indexes that are not completely adequate.

In the line of what was mentioned above, it seems that the use of CFA in the measurement model of SEM may have contributed to calling into question the credibility and replicability of the proposed model in certain spheres where this technique is applied (Marsh et al., 2009). A modification of the CFA algorithm was recently put forward to provide SEM with an exploratory quality that was lacking in their original definition. These are the Exploratory Structural Equation Models (ESEM) and they are based on the new conception of the exploratory factor loading matrix which, in turn, is based on the use of parts of the model through EFA with rotated factor loading matrices in addition to, or instead of, parts of the measurement model through CFA. That is to say, the measurement model through EFA does not require strict restrictions in the cross factor loadings, unlike CFA. Furthermore, just like in the classic SEM approach, it grants access to the usual factor loadings parameters that allow us to make the common interpretations in terms of the saturation of each observable indicator in relation to the non-observable latent factor. The modification of the basic lines of SEM in order to convert it into ESEM may be followed in the paper by Asparouhov and Muthén (2009).

ESEM parameter estimation has been enormously facilitated by the presence of MPlus and that alone should be a reason to activate its use. However, as has been said, the technique is not much in use yet. Accordingly, we think that the possibility to conduct ESEM through the R Project's routines would be a step further toward the normalization of the use of ESEM. Therefore, the goal of this paper is to present a simple programming choice in R by using the **`sem`** library to estimate the ESEM parameters in the factorialization of a correlation matrix.

2. Example

Asparouhov and Muthén (2009) present a simple example in the factorialization of a correlation matrix (R) from simulations in order to show the fit of a structure of the Λ_x factor loading matrix. The population factor loading matrix is as follows:

$$\Lambda_x = \begin{vmatrix} 0.8 & 0 \\ 0.8 & 0 \\ 0.8 & 0 \\ 0.8 & 0.25 \\ 0.8 & 0.25 \\ 0 & 0.8 \\ 0 & 0.8 \\ 0 & 0.8 \\ 0 & 0.8 \\ 0 & 0.8 \end{vmatrix}$$

where this factor loading matrix is the result of the factorialization of the following original R matrix

$$
R = \begin{vmatrix}
1 & & & & & & & & & & \\
.686 & 1 & & & & & & & & & \\
.684 & .682 & 1 & & & & & & & & \\
.712 & .714 & .704 & 1 & & & & & & & \\
.710 & .712 & .704 & .780 & 1 & & & & & & \\
.457 & .454 & .454 & .615 & .620 & 1 & & & & & \\
.464 & .468 & .461 & .622 & .628 & .781 & 1 & & & & \\
.466 & .463 & .455 & .616 & .621 & .781 & .782 & 1 & & & \\
.469 & .463 & .458 & .621 & .624 & .780 & .782 & .784 & 1 & & \\
.460 & .463 & .455 & .616 & .621 & .777 & .779 & .776 & .780 & 1 & \\
.373 & .378 & .364 & .502 & .497 & .626 & .620 & .622 & .623 & .624 & 1
\end{vmatrix}
$$

3. General Model

This is an 11x11 matrix, wherein the first ten vectors correspond to the 10 variables (Y_i) to factorialize and the last vector corresponds to an exogenous variable (x) necessary to formulate ESEM in accordance with the algorithm by Asparouhov and Muthén (2009), since the starting SEM model is as follows:

$$Y = \nu + \Lambda\eta + Kx + \varepsilon,$$

wherein we consider k endogenous variables Y, q exogenous variables x, and m latent variables η. Obviously, the SEM model follows exactly the general scheme in LISREL notation:

$$\eta = \alpha + \beta\eta + \Gamma\xi + \zeta.$$

The assumptions of the SEM model in the ESEM case are the usual ones, and so ε and ζ are normally distributed [$N(0, \theta, \Psi$ respectively)]; Var(ξ_i) = 1. The problem with this formulation is that, in Jöreskog's words (1978), the Λ_x matrix proposes all the λ_{ij} parameters as free, and therefore we get an unidentified model because the maximum number of parameters to estimate fits $p(p+1)/2$, where p is the number of observable variables. In fact, to the $(p \cdot m)$ factor loadings we should add the p variances, which makes the system undetermined. A simple explanation to this problem can be found in Steiger (2013). The solution given by Asparouhov and Muthén (2009) is to modify the Λ_x matrix so that a new m-dimension H matrix is generated and to substitute the η vectors for Hη in the model, so that Λ_x is replaced by $\Lambda_x H^{-1}$ and the model appears modified:

$$H\eta = H\alpha + (H\beta H^{-1)}H\eta + H\Gamma\xi + H\zeta H^{T}.$$

In the case of orthogonal factors, we put forth that $HH^{-T} = I$, so that the $Var(H\eta) = I$. In the case of oblique rotations, the restriction is somewhat more complex, since $diag(H\Psi H^{T} - I) = 0$ or simpler still $diag(H\Psi H^{T}) = 1$. This makes the $(p \cdot m)$ elements of Λ_x not have specification restrictions anymore. One question pending analysis is the effect of the possible oblique rotation solutions in case that is the model.

4. `sem` Library in the R Project

When this formulation is used based on the **`sem`** library in R (Fox et al., 2013), translating Asparouhov's and Muthén's proposal (2009) would be unviable because the system is infraestimated, as has been mentioned, and the result in R would be:

```
> sem.esem1 <- sem(Model.esem1, R.esem1, N=1000)
Error en solve.default(C[ind, ind]) :
  Lapack routine dgesv: system is exactly singular: U[2,2] = 0
```

This is evidence of the infraidentification problem ESEM tries to solve with the proposed modification. According to Steiger (2013), the R-based solution to this matter would be to modify the procedure and follow the next programming pattern:

4.1. Estimating factor loadings based on the Exploratory Factor Analysis

For this step, we must define the loading matrix derived from two factors under the AFE model. If we name R.esem1 the R matrix of correlations between the 10 (Y_i) variables observed, that is, without the exogenous variable (x), the instruction would be the following:

```
> R.Expl1 <- factanal(R.esem1, factors =2)
```

Obviously, for this estimation, we need a solid hypothesis to determine *a priori* the number of factors to extract or use previous statistical criteria to determine the *m* latent variables. To extract the non-rotated factor loading matrix, we can run the following routine:

```
F <- matrix(R.Expl1$loadings[1:20],10,2)
```

which generates the non-rotated factor loading matrix, which we named F:

```
> F
       [,1]    [,2]
 [1,] 0.7284 -0.4214
 [2,] 0.7280 -0.4230
 [3,] 0.7217 -0.4266
 [4,] 0.8435 -0.2316
 [5,] 0.8461 -0.2236
 [6,] 0.8175  0.3254
 [7,] 0.8233  0.3138
 [8,] 0.8204  0.3193
 [9,] 0.8225  0.3164
[10,] 0.8183  0.3185
```

The highest factor loadings are the ones in $\lambda_{(51)}$ = .8461 and $\lambda_{(32)}$ = -.4266 for each of the two factors assumed.

4.2. Study of the rotation

In order to do this, we will select the F1 submatrix out of the F matrix involving those two factor loadings:

```
> F1 <- rbind(F[5,], F[3,])
> F1
     [,1]  [,2]
[1,] 0.84 -0.22
[2,] 0.72 -0.42
```

Based on this submatrix, we can construct the new rotated factor loading matrix which allows us to have initial evidence in order to better specify the ESEM proposal. In the case of R, we can put forth the following:

```
> H <- solve(F1) %*% diag (sqrt(diag(F1 %*% t(F1))))
> Rotated.F <- zapsmall (F %*% H)
> Rotated.F
           [,1]       [,2]
 [1,] 0.8411111  0.0000000
 [2,] 0.8323212  0.0000000
 [3,] 0.8331221  0.0000000
 [4,] 0.8361713  0.2360174
 [5,] 0.8683317  0.2477112
 [6,] 0.0000000  0.8166427
 [7,] 0.0000000  0.8100584
 [8,] 0.0000000  0.8222133
 [9,] 0.0000000  0.8122114
[10,] 0.0000000  0.8067221
```

Likewise, we can estimate the matrix of correlations between factors in order to show the structure to be defined in SEM. Simply,

```
> solve(H) %*% t(solve(H))
          [,1]      [,2]
[1,] 1.0000000 0.9632562
[2,] 0.9632562 1.0000000
```

which implies an intense correlation between the two factors and it would consequently prevent us from defining the matrix $\phi = I$, so $\phi_{21} \neq 0$. From the analysis of the Rotated.F matrix, we can infer that some loadings should not undergo a SEM analysis, since they are clearly null loadings, like the value $\lambda_{61} = \lambda_{12} = 0$ among others.

Therefore, with that prevention, it would be feasible to obtain a SEM model within the **sem** library with similar features to those put forward by Asparouhov and Muthén (2009).

4.3 Use of the sem library.

In terms of what ESEM proposes and based on the original R matrix, we could formulate that:

```
# ESPECIFICATION MODEL
> Model.esem1<-specifyModel()
F1-> Y1,lam11,.8
F1-> Y2,lam21,.8
F1-> Y3,lam31,.8
F1-> Y4,lam41,.8
F1-> Y5,lam51,.8
##F1-> Y6,lam61,0
##F1-> Y7,lam71,0
##F1-> Y8,lam81,0
##F1-> Y9,lam91,0
##F1-> Y10,lam101,0
##F1-> X,lam111,0
##F2-> Y1,lam12,0
##F2-> Y2,lam22,0
##F2-> Y3,lam32,0
F2-> Y4,lam42,.25
F2-> Y5,lam52,.25
F2-> Y6,lam62,.8
F2-> Y7,lam72,.8
F2-> Y8,lam82,.8
F2-> Y9,lam92,.8
F2-> Y10,lam102,.8
F2-> X,lam112,0
F3-> Y1,NA,0
F3-> Y2,NA,0
F3-> Y3,NA,0
F3-> Y4,NA,0
F3-> Y5,NA,0
F3-> Y6,NA,0
F3-> Y7,NA,0
F3-> Y8,NA,0
F3-> Y9,NA,0
F3-> Y10,NA,0
F3-> X,lam113,1
F1 <-> F2, NA,.9
F1 <-> F1, NA,1
F2 <-> F2, NA,1
F3 <-> F3, NA,1
F1 <- F3, NA,.5
F2 <- F3, NA,1
```

thus obtaining the following result:

```
> sem.esem1 <- sem(Model.esem1, R.esem1, N=10000)
> summary(sem.esem1)

 Model Chisquare =  30.511   Df =  42 Pr(>Chisq) = 0.90588
 Chisquare (null model) =  96839   Df =  55
 Goodness-of-fit index =  0.99945
 Adjusted goodness-of-fit index =  0.99913
 RMSEA index =  0   90% CI: (NA, 0.00288)
 Bentler-Bonnett NFI =  0.99968
 Tucker-Lewis NNFI =  1.0002
 Bentler CFI =  1
 SRMR =  0.0027821
 AIC =  78.511
 AICc =  30.632
 BIC =  251.56
 CAIC =  -398.32

 Normalized Residuals
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.5910 -0.0977  0.0499  0.0626  0.1480  0.7130

 R-square for Endogenous Variables
    F1     Y1     Y2     Y3     Y4     Y5     Y6     Y7     Y8     Y9    Y10      X     F2
0.2000 0.6868 0.6886 0.6754 0.7791 0.7794 0.7793 0.7825 0.7807 0.7835 0.7751 0.9979 0.5000


 Parameter Estimates
       Estimate Std Error z value   Pr(>|z|)
lam11  0.740956 0.0072844 101.71757  0.0000e+00 Y1 <--- F1
lam21  0.741945 0.0072785 101.93664  0.0000e+00 Y2 <--- F1
lam31  0.734803 0.0073213 100.36525  0.0000e+00 Y3 <--- F1
lam41  0.619217 0.0076279  81.17735  0.0000e+00 Y4 <--- F1
lam51  0.612312 0.0075920  80.65195  0.0000e+00 Y5 <--- F1
lam42  0.186392 0.0054915  33.94201 1.6004e-252 Y4 <--- F2
lam52  0.193207 0.0054816  35.24656 3.8741e-272 Y5 <--- F2
lam62  0.624253 0.0049406 126.35293  0.0000e+00 Y6 <--- F2
lam72  0.625552 0.0049288 126.91674  0.0000e+00 Y7 <--- F2
lam82  0.624821 0.0049354 126.59886  0.0000e+00 Y8 <--- F2
lam92  0.625953 0.0049252 127.09124  0.0000e+00 Y9 <--- F2
lam102 0.622585 0.0049555 125.63465  0.0000e+00 Y10 <--- F2
```

```
phi21   0.494268 0.0088950  55.56701  0.0000e+00 F2 <--> F1
V[Y1]   0.313014 0.0055230  56.67504  0.0000e+00 Y1 <--> Y1
V[Y2]   0.311178 0.0055026  56.55077  0.0000e+00 Y2 <--> Y2
V[Y3]   0.324376 0.0056500  57.41202  0.0000e+00 Y3 <--> Y3
V[Y4]   0.220629 0.0039621  55.68498  0.0000e+00 Y4 <--> Y4
V[Y5]   0.220342 0.0039391  55.93749  0.0000e+00 Y5 <--> Y5
V[Y6]   0.220776 0.0038060  58.00788  0.0000e+00 Y6 <--> Y6
V[Y7]   0.217529 0.0037660  57.76142  0.0000e+00 Y7 <--> Y7
V[Y8]   0.219358 0.0037885  57.90113  0.0000e+00 Y8 <--> Y8
V[Y9]   0.216527 0.0037537  57.68389  0.0000e+00 Y9 <--> Y9
V[Y10] 0.224936 0.0038574  58.31320  0.0000e+00 Y10 <--> Y10
V[X]    0.002149 0.0074636   0.28793  7.7340e-01 X <--> X

 Iterations =  23
```

5. Conclusions

In light of the procedure proposed, the matter could be solved by a somewhat long, unfriendly process, though efficient, by establishing an ESEM approximation based on the **sem** library in R. Several aspects would still need to be tackled: namely the estimation of the discrepancy function and Bartlett's multiplier [`> sem.esem1$criterion`] to further purify the model; as well as the study of the modification indices for a more efficient fit [`> modIndices(sem.esem1)`]. Furthermore, the application of **sem** to ESEM estimation procedure require a very hard modification of **sem** package in order to generate a simple way to estimate ESEM parameters. The use of actual **sem** possibilities is not adjusted to ESEM configuration. Obviously, this is a first approximation and we must carry on a simulation study about this possibility [`bootsem`].

Acknowledgements: This research was made possible by the PSI2010-21214-C02-01 project and was carried out by members of the Generalitat de Catalunya's SGR 388 Consolidated Research Group.

References

Asparouhov, T. & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*(3), 397–438.

Fox, J., Zhenghua, N., Byrnes, J., Culbertson, M., DebRoy, S., Friendly, M., Jones, R.H., Kramer, A., Monette, G. and R-Core (2013). *Package* **sem**. Document dowloaded on 26 April, 2013 from http://cran.r-project.org/web/packages/sem/sem.pdf.

Jöreskog, K.G. (1978). Structural analysis of covariance and correlation matrices. Psychometrika, 43(4), 445-477.

Marsh, H.W., Muthén, B., Morin, A., Lüdtke, O., Asparaouhov, T., Trautwein, U., & Nagengast, B. (2010). A new look at the Big Five Factor structure through Exploratory Structural Equation Modeling. *Psychological Assessment, 22*(3), 471–491.

Marsh, H.W., Muthén, B., Asparaouhov, T., Lüdtke, O., Robitzsch, A., Morin, A., & Trautwein, U. (2009). Exploratory structural equation modeling, interpretating CFA and EFA: Application to student's evaluations of university teaching. *Structural Equation Modeling, 16*(3), 439–476.

Sánchez, D., Barrada, J.R., López, G., Fauquet, J., Almenara, C.A., & Trepat, E. (2012). Analysis of the factor structure of the Sociocultural Attitudes towards Appearance Questionnaire (SATAG-3) in Spanish secondary-school students through exploratory structural equation modeling. *Body Image*, *9*(1), 163-171.

Steiger (2013).*Confirmatory Factor Analysis with R.* Document downloaded from: http://ebookbrowse.com/confirmatory-factor-analysis-with-r-pdf-d23298127. 15, April, 2013.