Analytical homework M&S – 1: The LISREL model

Ellen L. Hamaker
Utrecht University

LISREL (LInear Structural RELations) was developed by Karl Jöreskog and Dag Sörbom in 1970. It was the first program that allowed for the analysis of the covariance structure of a data set and as such it has become the standard of structural equation modeling (SEM). Although there are currently many diverse programs that allow for structural equation modeling (cf., Amos, Mplus, EQS, Mx, as well as functions in R), and many of these are more flexible, userfriendly (?!), or extended than LISREL, the latter remains an important player in the field of SEM. This dominant role is illustrated by for instance the TECH1 output option in Mplus, which results in the LISREL model matrices.

A SEM model can be represented in several ways: a) graphically, in a path diagram; b) as a set of equations; and c) in matrix notation. To really understand what you are doing when using SEM, it is good get familiar with all three representations.

For the matrix notation, there are several options: a) the full LISREL model; b) the y-model of the LISREL model; and c) McArdle's RAM notation. The latter can be thought of as a special case of the LISREL y-model. Note that a program like Mx allows you to specify your own covariance structure (rather than forcing you to use specific model matrices when specifying a model), which makes it extremely flexible (and it is for free!), but also requires the user to be more knowledgable.

The full LISREL model consists of two parts: the x-part which contains the exogenous variables, and the y-part which contains the endogenous variables. Both parts can contain observed and latent variables: Through relating the latent variables of the x- and y-part, the observed variables of these parts are also (indirectly) connected.

Although most discussion of the LISREL model contain these two parts, it is also possible to rewrite the x-y model into a y-model only. This has the advantage that the number of model matrices is reduced (while the sizes of these matrices typically increase). The TECH1 output option in Mplus is based on using the LISREL y-model.

Unless you are handling multiple group and/or longitudinal data, the means are typically not modeled. Therefore, we start with considering the y-model without means.

The observed and the modeled covariance matrices

To model the covariance structure of a set of observed variables, we compare the covariance matrix of the observed variables $\mathbf{S}$ with the modeled covariance matrix $\mathbf{\Sigma}$. When obtaining estimates of the model parameters, we try to somehow minimize the distance between $\mathbf{S}$ and $\mathbf{\Sigma}$ (although it is a bit more complex than just minimizing $\mathbf{S} - \mathbf{\Sigma}$).

The observed covariance matrix contains the variances and covariances of the observations, that is

$$\mathbf{S} = \begin{bmatrix} s_1^2 & & & \\ s_{21} & s_2^2 & & \\ \dots & & & \\ s_{p1} & s_{p2} & \dots & s_p^2 \end{bmatrix}, \tag{1}$$

where s_{hg} represents the covariance between variable h and variable g, and s_h^2 represents the variance of variable h. The modeled covariance matrix contains the modeled variances and covariances, that is

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & & & \\ \sigma_{21} & \sigma_2^2 & & \\ \dots & & & \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{bmatrix}. \tag{2}$$

These elements are functions of the model parameters, as will become clear below.

If we have observed p variables, the total number of unique elements in the observed covariance matrix is equal to $\{p \times (p+1)\}/2$. Hence, for our model to be identified, we should not estimate more than $\{p \times (p+1)\}/2$ parameters. If the number of parameters we estimate is equal to the number of observed statistics (i.e., unique elements in $\mathbf{S}$), the model is said to be *saturated*. The fit of such a model is perfect, and is really not informative. However, such a model can be of interest as it results in parameter estimates (with corresponding standard errors), such that we can make inferences based on it.

The measurement equation and the structural equation

To obtain analytical expressions for the elements of $\mathbf{\Sigma}$ in terms of the (unknown) model parameters, we need to specify the model using two equations: the measurement equation and the structural equation. In the *measurement equation* the observed variables $\boldsymbol{y}_i$ are related to the latent variables $\boldsymbol{\eta}_i$,

$$\boldsymbol{y}_i = \mathbf{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i \tag{3}$$

where

- $\boldsymbol{y}_i$ is a p-variate (column) vector with the observations of an individual,
- $\mathbf{\Lambda}$ is a $p \times q$ matrix with factor loadings (these are model parameters),
- $\boldsymbol{\eta}_i$ is a q-variate vector with latent variables (or factor scores) of the individual, and
- $\boldsymbol{\epsilon}_i$ is a p-variate vector with measurement errors for the individual.

The measurement errors are assumed to be uncorrelated with the latent variables, and have a mean vector of zeros, and covariance matrix $\mathbf{\Theta}$ (containing model parameters).

The *structural equation* allows us to specify structural relationships between the latent variables, and can be expressed as

$$\boldsymbol{\eta}_i = \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\zeta}_i \tag{4}$$

where

- $\mathbf{B}$ is a $q \times q$ matrix with the structural relationships (these are model parameters), and
- $\boldsymbol{\zeta}_i$ is a q-variate vector with residuals of the individual.

The latter have a mean vector of zeros and covariance matrix $\boldsymbol{\Psi}$ (which contains model parameters).

2nd order factor model Suppose we have observed twelve variables ($p = 12$), for which we consider the following second-order factor model in Mplus:
MODEL: f1 BY y1-y3;
f2 BY y4-y6;
f3 BY y7-y9;
f4 BY y10-y12;
f5 BY f1-f4;

This implies we have

$$\boldsymbol{y}_i = \begin{bmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \\ y_{4i} \\ y_{5i} \\ y_{6i} \\ y_{7i} \\ y_{8i} \\ y_{9i} \\ y_{10i} \\ y_{11i} \\ y_{12i} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\eta}_i = \begin{bmatrix} \eta_{1i} \\ \eta_{2i} \\ \eta_{3i} \\ \eta_{4i} \\ \eta_{5i} \end{bmatrix} \tag{5}$$

The measurement equation relates the observed variables to the latent variables, that is

$$\begin{bmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \\ y_{4i} \\ y_{5i} \\ y_{6i} \\ y_{7i} \\ y_{8i} \\ y_{9i} \\ y_{10i} \\ y_{11i} \\ y_{12i} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \lambda_{21} & 0 & 0 & 0 & 0 \\ \lambda_{31} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & \lambda_{52} & 0 & 0 & 0 \\ 0 & \lambda_{62} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \lambda_{83} & 0 & 0 \\ 0 & 0 & \lambda_{93} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \lambda_{11,4} & 0 \\ 0 & 0 & 0 & \lambda_{12,4} & 0 \end{bmatrix} \begin{bmatrix} \eta_{1i} \\ \eta_{2i} \\ \eta_{3i} \\ \eta_{4i} \\ \eta_{5i} \end{bmatrix} + \begin{bmatrix} \epsilon_{1i} \\ \epsilon_{2i} \\ \epsilon_{3i} \\ \epsilon_{4i} \\ \epsilon_{5i} \\ \epsilon_{6i} \\ \epsilon_{7i} \\ \epsilon_{8i} \\ \epsilon_{9i} \\ \epsilon_{10i} \\ \epsilon_{11i} \\ \epsilon_{12i} \end{bmatrix} \tag{6}$$

EXERCISE 1 Write out the equation above, such that you get analytical expressions for each observed variable in terms of λ's, η's and ϵ's. Which part of the model is defined by this?

The structural equation relates the latent variables to each other, and for this model it is

$$\begin{bmatrix} \eta_{1i} \\ \eta_{2i} \\ \eta_{3i} \\ \eta_{4i} \\ \eta_{5i} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \beta_{25} \\ 0 & 0 & 0 & 0 & \beta_{35} \\ 0 & 0 & 0 & 0 & \beta_{45} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \eta_{1i} \\ \eta_{2i} \\ \eta_{3i} \\ \eta_{4i} \\ \eta_{5i} \end{bmatrix} + \begin{bmatrix} \zeta_{1i} \\ \zeta_{2i} \\ \zeta_{3i} \\ \zeta_{4i} \\ \zeta_{5i} \end{bmatrix} \tag{7}$$

EXERCISE 2 Write out the equation above, such that you get analytical expressions for each latent variable η in terms of ζ's and β's. What part of the model is this?

EXERCISE 3 Note that you could enter the expression obtained in Exercise 2 into the expression obtained in Exercise 1, such that the observed variables are written as a function of the ζ's only (and not of the η's). Do this.

The covariance structure

From the above, it is clear that there are four model matrices containing parameters we wish to estimate in the y-model without means, that is:

- $\mathbf{\Lambda}$, the $p \times q$ matrix with factor loadings
- $\mathbf{\Theta}$, the $p \times p$ covariance matrix of the measurement errors ϵ
- $\mathbf{B}$, the $q \times q$ matrix with structural relationships
- $\mathbf{\Psi}$, the $q \times q$ covariance matrix of the latent residuals

Hence, to specify a model, we need to specify these four model matrices. From these matrices, we can derive the modeled covariance matrix $\mathbf{\Sigma}$, as is shown below.

First, we determine the covariance matrix of the latent variables $\boldsymbol{\eta}$. If we model the covariance structure only (as we are doing here), we can assume that all the latent variables and the observed variables have a mean of zero. Hence, since $\mathrm{E}[\boldsymbol{\eta}] = \mathbf{0}$, the covariance matrix of $\boldsymbol{\eta}$ can be written as

$$\begin{aligned} \mathrm{Var}(\boldsymbol{\eta}_i) = \mathrm{E}[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T] &= \mathrm{E}\left[\{\mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\zeta}_i\}\{\mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\zeta}_i\}^T\right] \\ &= \mathrm{E}\left[\{\mathbf{B}\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T\mathbf{B}^T + \boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^{\boldsymbol{T}} + \mathbf{B}\boldsymbol{\eta}_i\boldsymbol{\zeta}_i^T + \boldsymbol{\zeta}_i\boldsymbol{\eta}_i^T\mathbf{B}^T\right] \\ &= \mathbf{B}\mathrm{E}\big[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T\big]\mathbf{B}^T + \mathrm{E}\big[\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^T\big] + \mathbf{B}\mathrm{E}\big[\boldsymbol{\eta}_i\boldsymbol{\zeta}_i^T\big] + \mathrm{E}\big[\boldsymbol{\zeta}_i\boldsymbol{\eta}_i^T\big]\mathbf{B}^T \end{aligned}$$

Note that we can take $\mathbf{B}$ outside the expectations, because it contains only constants. However, we are stuck with $\mathrm{E}[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T]$ on the right-hand side as well, while we were trying to find an expression for this. Hence, this approach is not so helpful.

To obtain an expression for the covariance matrix of $\boldsymbol{\eta}$ in terms of model parameters only, we first have to find an expression of $\boldsymbol{\eta}$ which is not based on the recursion that is present in Equation 4. This is done as follows:

$$\begin{aligned}\boldsymbol{\eta}_i &= \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\zeta}_i \\ \boldsymbol{\eta}_i - \mathbf{B}\boldsymbol{\eta}_i &= \boldsymbol{\zeta}_i \\ \big(\mathbf{I} - \mathbf{B}\big)\boldsymbol{\eta}_i &= \boldsymbol{\zeta}_i \\ \boldsymbol{\eta}_i &= \big(\mathbf{I} - \mathbf{B}\big)^{-1}\boldsymbol{\zeta}_i,\end{aligned} \tag{8}$$

where $\mathbf{I}$ is a $q \times q$ identity matrix.

Using this alternative expression for $\boldsymbol{\eta}$ (which does not itself depend on $\boldsymbol{\eta}$), we can define the covariance matrix as

$$\begin{aligned}\mathrm{E}[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T] &= \mathrm{E}\Big[\big\{\big(\mathbf{I} - \mathbf{B}\big)^{-1}\boldsymbol{\zeta}_i\big\}\big\{\big(\mathbf{I} - \mathbf{B}\big)^{-1}\boldsymbol{\zeta}_i\big\}^T\Big] \\ &= \mathrm{E}\Big[\big(\mathbf{I} - \mathbf{B}\big)^{-1}\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^T\big(\mathbf{I} - \mathbf{B}\big)^{-1T}\Big] \\ &= \big(\mathbf{I} - \mathbf{B}\big)^{-1}\mathrm{E}\big[\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^T\big]\big(\mathbf{I} - \mathbf{B}\big)^{-1T} \\ &= \big(\mathbf{I} - \mathbf{B}\big)^{-1}\boldsymbol{\Psi}\big(\mathbf{I} - \mathbf{B}\big)^{-1T},\end{aligned} \tag{9}$$

which shows that the variances and covariances of $\boldsymbol{\eta}_i$ can be written as functions of the elements of $\mathbf{B}$ and $\boldsymbol{\Psi}$.

Next, we will determine the modeled covariance matrix of the observed variables, that is,

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathrm{E}[\mathbf{y}_i\mathbf{y}_i^T] \\ &= \mathrm{E}\Big[\big\{\boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i\big\}\big\{\boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i\big\}^T\Big] \\ &= \mathrm{E}\Big[\boldsymbol{\Lambda}\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T\boldsymbol{\Lambda}^T + \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T + \boldsymbol{\Lambda}\boldsymbol{\eta}_i\boldsymbol{\epsilon}^T + \boldsymbol{\epsilon}_i\boldsymbol{\eta}_i^T\boldsymbol{\Lambda}^T\Big] \\ &= \boldsymbol{\Lambda}\mathrm{E}\big[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T\big]\boldsymbol{\Lambda}^T + \mathrm{E}\big[\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T\big] + \boldsymbol{\Lambda}\mathrm{E}\big[\boldsymbol{\eta}_i\boldsymbol{\epsilon}_i^T\big] + \mathrm{E}\big[\boldsymbol{\epsilon}_i\boldsymbol{\eta}_i^T\big]\boldsymbol{\Lambda}^T\end{aligned}$$

where the latter two terms are zero (because – by definition – the latent variables are uncorrelated with the measurement errors), and the first term contains the covariance matrix of the latent variables, such that we can write,

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\big(\mathbf{I} - \mathbf{B}\big)^{-1}\boldsymbol{\Psi}\big(\mathbf{I} - \mathbf{B}\big)^{-1T}\boldsymbol{\Lambda}^T + \boldsymbol{\Theta}. \tag{10}$$

This is the *general expression* for the covariance structure of the y-model, which shows that the modeled covariance matrix is a function of the four model matrices, that is $\boldsymbol{\Lambda}$, $\mathbf{B}$, $\boldsymbol{\Psi}$ and $\boldsymbol{\Theta}$.

Example: 2nd order factor model (continued) We can now determine the modeled covariance matrix of the observations for the second-order factor model that was defined before, using Equation 10. The matrices $\boldsymbol{\Lambda}$ and $\mathbf{B}$ were already presented explicitly above. The matrices $\boldsymbol{\Psi}$ and $\boldsymbol{\Theta}$ still need to be defined. As we have not included any covariances between the residuals of the latent variables, $\mathbf{\Psi}$ is a diagonal matrix, that is, $\mathbf{\Psi} = \text{diag}[\psi_1^2, \psi_2^2, \psi_3^2, \psi_4^2, \psi_5^2]$. Similarly, as we have not included any covariances between the measurement errors, $\mathbf{\Theta}$ is also a diagonal matrix, that is, $\mathbf{\Theta} = \text{diag}[\theta_1^2, \theta_2^2, \ldots, \theta_{12}^2]$. Using these matrices in Equation 10 would result in expressing the elements of $\mathbf{\Sigma}$ in terms of the unknown model parameters.

However, it is difficult to determine $\mathbf{\Sigma}$ by hand, because of the inverse of $\mathbf{I} - \mathbf{B}$ that we need to take (which requires us to obtain the determinant of this matrix first, which is really not easy). Alternatively, we can determine the expression of each element of $\mathbf{\Sigma}$ separately, using the analytical expressions obtained above in Exercise 3 in the appropriate expectations.

EXERCISE 4 Determine the expressions for the variance of y_1 in terms of the unknown model parameters.

EXERCISE 5 Determine the expression for the covariance between y_2 and y_{12} in terms of the unknown model parameters.