Structural Equation Modeling for Social Relations: The R package **srm**

Alexander Robitzsch [1 2], Steffen Nestler [3], Oliver Lüdtke [1 2]

[1] IPN – Leibniz Institute for Science and Mathematics Education

[2] Centre for International Student Assessment (ZIB)

[3] University of Münster

Psychoco

Dortmund, February 2020

Content

1. Univariate and Multivariate Social Relations Model (SRM)
2. Structural Equation Models (SEM) for Multivariate Data
3. Social Relations Structural Equation Model (SR-SEM)
4. R package **srm**
5. Computational Aspects
6. Discussion

Univariate Social Relations Model (I)

- actor i rates partner j in dyad $d = (ij)$ on one variable y, e.g., ratings on
  - *I like person XX a lot.*
  - *I think that person XX is good at Mathematics.*
- social relations model (SRM)

$$y_{ij} = \mu + a_i + p_j + \varepsilon_{ij} \tag{1}$$

- actor effects a_i: *how much person i likes other persons*
- partner effects p_j: *how much person j is liked by other persons*
- relationship effects ε_{ij}: *specific effect that person i likes j*

Univariate Social Relations Model (II)

- social relations model (SRM)

$$y_{ij} = \mu + a_i + p_j + \varepsilon_{ij} \tag{1}$$

- model parameters at level of persons ($\mathbf{\Sigma}_u$) and dyads ($\mathbf{\Sigma}_r$)

$$\mathbf{\Sigma}_u = Var\begin{pmatrix} a_i \\ p_i \end{pmatrix} = \begin{pmatrix} \sigma_a^2 & \\ \sigma_{ap} & \sigma_p^2 \end{pmatrix} \tag{2}$$

$$\mathbf{\Sigma}_r = Var\begin{pmatrix} \varepsilon_{ij} \\ \varepsilon_{ji} \end{pmatrix} = \begin{pmatrix} \sigma_\varepsilon^2 & \\ \sigma_{\varepsilon\varepsilon} & \sigma_\varepsilon^2 \end{pmatrix} \tag{3}$$

Mixed Effects Representation of the SRM

- social relations model (SRM)

$$y_{ij} = \mu + a_i + p_j + \varepsilon_{ij} \tag{1}$$

- define vector of person effects for persons $i = 1, \ldots, I$: $\boldsymbol{u}_i = (a_i, p_i)$
- define vector of dyad effects for dyads $d = 1, \ldots, D$: $\boldsymbol{r}_d = (\varepsilon_{ij}, \varepsilon_{ji})$
- collect all observations in outcome $\boldsymbol{y} = (y_{ij})_{ij}$
- mixed effects model representation (see Nestler, 2016)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{i=1}^{I} \boldsymbol{Z}_i \boldsymbol{u}_i + \sum_{d=1}^{D} \boldsymbol{W}_d \boldsymbol{r}_d \tag{4}$$

  with design matrices $\boldsymbol{Z}_i$ and $\boldsymbol{W}_d$ (containing only zeros or ones)
- short form in mixed effects model notation: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{W}\boldsymbol{r}$

Multivariate Social Relations Model

- now consider V multiple outcomes $y_{1ij}, \dots, y_{Vij}$
- multiple (i.e., $2V$) actor and partner effects define person level variable $\boldsymbol{u}_i$
- relationship vector $\boldsymbol{r}_d$ can also be extended for multiple outcomes
- no general change in notation for mixed effects representation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{i=1}^{I} \boldsymbol{Z}_i \boldsymbol{u}_i + \sum_{d=1}^{D} \boldsymbol{W}_d \boldsymbol{r}_d \tag{4}$$

- in short: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{W}\boldsymbol{r}$
- estimation with ANOVA method (unweighted least squares) or (restricted) maximum likelihood

Structural Equation Models (SEM) for Multivariate Data

- model multivariate normally distributed outcome as a constrained model $\boldsymbol{y} \sim MVN(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ with a parameter vector $\boldsymbol{\theta}$
- ignore mean structure in the following for simplicity
- structural equation model (SEM)

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \\ \boldsymbol{\eta} &= \boldsymbol{B}\boldsymbol{\eta} + \boldsymbol{\xi} \end{aligned} \tag{5}$$

- model parameter vector $\boldsymbol{\theta}$ contains free parameters in $\boldsymbol{\Lambda}$, $\boldsymbol{B}$, $Var(\boldsymbol{\xi}) = \boldsymbol{\Phi}$, $Var(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$
- model implied covariance matrix

$$Var(\boldsymbol{y}) = \boldsymbol{\Sigma}_{\boldsymbol{y}} = \boldsymbol{\Sigma}_{\boldsymbol{y}}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Phi}((\boldsymbol{I} - \boldsymbol{B})^{-1})'\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \tag{6}$$

Maximum Likelihood Estimation in SEM

- maximum likelihood (ML) estimation maximizes

$$l(\boldsymbol{\theta}) = \text{const} - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{y}})'\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{y}}) \quad (7)$$

- gradient (score equation)

$$\frac{\partial l}{\partial \theta_h} = -\frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{y}}}{\partial \theta_h}\right) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{y}})'\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{y}}}{\partial \theta_h}\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{y}}) \quad (8)$$

- expected information matrix for use in Fisher Scoring

$$E\left(\frac{\partial l^2}{\partial \theta_h \partial \theta_k}\right) = -\frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{y}}}{\partial \theta_h}\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{y}}}{\partial \theta_k}\right) \quad (9)$$

- update equation in Fisher scoring

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \left[E\left(\frac{\partial l^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right)\right]^{-1}\frac{\partial l}{\partial \boldsymbol{\theta}} \quad (10)$$

Social Relations Structural Equation Model (SR-SEM)

- multivariate SRM $\Rightarrow$ covariance structure of person effects $\boldsymbol{\Sigma}_u$ and dyad effects $\boldsymbol{\Sigma}_r$
- consider restricted models $\boldsymbol{\Sigma}_u = \boldsymbol{\Sigma}_u(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_r = \boldsymbol{\Sigma}_r(\boldsymbol{\theta})$, e.g. models with factor structures or relationship among several constructs $\Rightarrow$ social relations structural equation model (SR-SEM)
- SEM at level of persons: $\boldsymbol{\theta}_u = (\boldsymbol{\Lambda}_u, \boldsymbol{B}_u, \boldsymbol{\Phi}_u, \boldsymbol{\Psi}_u)$
- SEM at level of dyads $\boldsymbol{\theta}_r = (\boldsymbol{\Lambda}_r, \boldsymbol{B}_r, \boldsymbol{\Phi}_r, \boldsymbol{\Psi}_r)$
- or pose some equality constraints among both levels (e.g., invariance of factor loadings)

ML Estimation in SR-SEM

- stack all observations (dyads, variables) of a round robin design in outcome vector $\boldsymbol{y}$
- $\boldsymbol{y}$ is multivariate normally distributed if all effects of the SRM are normally distribution
- ML estimation of $\boldsymbol{\theta}$ needs $\boldsymbol{\Sigma}_{\boldsymbol{y}}$ and $\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{y}}}{\partial \theta_h}$ (see normal theory based ML)
- multivariate SRM has mixed effects representation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{i=1}^{I} \boldsymbol{Z}_i \boldsymbol{u}_i + \sum_{d=1}^{D} \boldsymbol{W}_d \boldsymbol{r}_d \qquad (4)$$

$\Rightarrow$

$$\boldsymbol{\Sigma}_{\boldsymbol{y}} = Var(\boldsymbol{y}) = \sum_{i=1}^{I} \boldsymbol{Z}_i \boldsymbol{\Sigma}_u \boldsymbol{Z}_i' + \sum_{d=1}^{D} \boldsymbol{W}_d \boldsymbol{\Sigma}_r \boldsymbol{W}_d' \qquad (11)$$

$$\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{y}}}{\partial \theta_h} = \sum_{i=1}^{I} \boldsymbol{Z}_i \frac{\partial \boldsymbol{\Sigma}_u}{\partial \theta_h} \boldsymbol{Z}_i' + \sum_{d=1}^{D} \boldsymbol{W}_d \frac{\partial \boldsymbol{\Sigma}_r}{\partial \theta_h} \boldsymbol{W}_d' \qquad (12)$$

R Package **srm**

- R package **srm** on CRAN
- covers SEM at both levels (persons and dyads)
- satisfactory computation time (computational shortcuts, use of Rcpp)
- ML estimation using Fisher scoring and quasi-Newton approach using observed information matrix
- Fisher scoring relatively stable, at least more stable than Quasi-Newton algorithms with observed information matrix

srm Package: Model Syntax

- inspired by multilevel syntax of **lavaan** (level identifiers `%person` and `%dyad`)
- SRM decomposition $Y_{ij} = \mu + a_i + p_j + \varepsilon_{ij}$ plainly translates to `V1=V1@A+V1@P+V1@AP`
- Example syntax for unidimensional factor model

```
\%Person
f1@A=~Wert1@A+Wert2@A+Wert3@A
f1@P=~Wert1@P+Wert2@P+Wert3@P

\%Dyad
f1@AP=~Wert1@AP+Wert2@AP+Wert3@AP

# define some constraints
Wert1@AP ~~ 0*Wert1@PA
Wert3@AP ~~ 0*Wert3@PA
```

srm Package: Model Output

```
   index group      lhs op     rhs   mat fixed   est    se lower
1     NA     1     F1@A =~ Wert1@A LAM_U     1 1.000    NA  -Inf
2     NA     1     F1@P =~ Wert1@P LAM_U     1 1.000    NA  -Inf
3      1     1     F1@A ~~    F1@A PHI_U    NA 0.322 0.071  -Inf
4      2     1     F1@A ~~    F1@P PHI_U    NA 0.098 0.043  -Inf
5      3     1     F1@P ~~    F1@P PHI_U    NA 0.160 0.049  -Inf
6     NA     1  Wert1@A ~~ Wert1@A PSI_U     0 0.000    NA  -Inf
7     NA     1  Wert1@A ~~ Wert1@P PSI_U     0 0.000    NA  -Inf
8     NA     1  Wert1@P ~~ Wert1@P PSI_U     0 0.000    NA  -Inf
9     NA     1     F1@A ~1    F1@A  MU_U     0 0.000    NA  -Inf
10    NA     1     F1@P ~1    F1@P  MU_U     0 0.000    NA  -Inf
11     4     1  Wert1@A ~1 Wert1@A  BETA    NA 0.150 0.093  -Inf
12    NA     1    F1@AP =~ Wert1@AP LAM_D    1 1.000    NA  -Inf
13    NA     1    F1@PA =~ Wert1@PA LAM_D    1 1.000    NA  -Inf
14     6     1    F1@AP ~~   F1@AP PHI_D    NA 1.531 0.081  -Inf
15     5     1    F1@AP ~~   F1@PA PHI_D    NA 0.069 0.081  -Inf
16     6     1    F1@PA ~~   F1@PA PHI_D    NA 1.531 0.081  -Inf
17    NA     1 Wert1@AP ~~ Wert1@AP PSI_D    0 0.000    NA  -Inf
18    NA     1 Wert1@AP ~~ Wert1@PA PSI_D    0 0.000    NA  -Inf
19    NA     1 Wert1@PA ~~ Wert1@PA PSI_D    0 0.000    NA  -Inf
```

Computational Aspects

- matrices of derivatives $\frac{\partial \boldsymbol{\Sigma}_u}{\partial \theta_h}$ and $\frac{\partial \boldsymbol{\Sigma}_r}{\partial \theta_h}$ have known forms (known from single-level SEMs)
- inverse matrix $\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}$ computationally demanding because its dimension is $D(D-1)V$
- total likelihood based on sum of independent likelihoods corresponding to different round robin groups

$\Rightarrow$ $\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}$ must only be computed for round robin designs with same number of persons (without missing data)

Faster Computation of $\Sigma_{\boldsymbol{y}}^{-1}$: Woodbury Identity

- tip from Yves Rosseel (June 2019)
- observations in the SR-SEM are of the form $\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$, where $\boldsymbol{U} = Var(\boldsymbol{u})$ and $\boldsymbol{E} = Var(\boldsymbol{e})$ are block diagonal matrices of functions of $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_r$, respectively
- $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_r$ computationally inexpensive to invert (because of lower dimension), and, therefore, also block diagonal matrices $\boldsymbol{U}$ and $\boldsymbol{E}$
- it holds that

$$Var(\boldsymbol{y}) = \boldsymbol{\Sigma}_{\boldsymbol{y}} = \boldsymbol{Z}\boldsymbol{U}\boldsymbol{Z}^T + \boldsymbol{E} \tag{13}$$

- use Woodbury identity for inversion

$$\left(\boldsymbol{Z}\boldsymbol{U}\boldsymbol{Z}^T + \boldsymbol{E}\right)^{-1} = \boldsymbol{E}^{-1} - \boldsymbol{E}^{-1}\boldsymbol{Z}\left(\boldsymbol{U}^{-1} + \boldsymbol{Z}^T\boldsymbol{E}^{-1}\boldsymbol{Z}\right)\boldsymbol{Z}^T\boldsymbol{E}^{-1} \tag{14}$$

Skipping Zero Entries in Matrix Computations

- in computation of the first and second derivative, matrix multiplications $\mathbf{\Sigma}_{\boldsymbol{y}}^{-1}\frac{\partial \mathbf{\Sigma}_{\boldsymbol{y}}}{\partial \theta_h}$ for all parameters θ_h have to be computed
- many entries in $\frac{\partial \mathbf{\Sigma}_{\boldsymbol{y}}}{\partial \theta_h}$ are zero (e.g., derivative with respect to a particular item loading)
- skip these computations in matrix computations by *hard coding sparse matrix multiplications* in **Rcpp**

$\Rightarrow$ skipping redundant computations led to most important speed improvement

More Advanced Models and Extensions

- multiple group models (e.g., round robin designs in different age groups or different school tracks)
- discrete moderators x (e.g., gender) of model parameters $\boldsymbol{\theta} = \boldsymbol{\theta}(x)$ can be handled by including pseudo variables (original variable $\times$ dummy variables for moderator values)
- generic variables at person level (self ratings) are round robin variables with constraints: $y_{ij} = \mu + 0 \cdot a_i + 1 \cdot p_j + 0 \cdot \varepsilon_{ij}$
- level-specific fit indices for assessing differences between multivariate saturated SRM and SRM-SEM

Alternative Estimators

- least squares estimation (Bond & Malloy, 2018)
- composite likelihood methods (pairwise likelihood estimation), particularly attractive for high-dimensional models and categorical data
- MCMC techniques (Hoff, 2005; Gill & Swartz, 2001)
- maximum a posterior (MAP) estimation using prior distributions (penalized maximum likelihood estmation)
- plausible value imputation: estimate a saturated multivariate SRM at first, then plugin the PVs into a standard single-level SEM
- two-step methods: estimation of „factor scores“, then plug-in factor scores into path models (with some unreliablity correction)

Many thanks!

Alexander Robitzsch
robitzsch@ipn.uni-kiel.de