Multilevel SEM for Discrete Data in the 'Wide Format' Approach

Mariska Barendse and Yves Rosseel

SEM working group
Amsterdam
16 March 2018

goal of this study

- study the 'wide format' for discrete multilevel data instead of the 'long format' approach
  - latent variables
  - covariates
  - fit indices
- small simulation study (MML; WLSMV; PL)
- real data example
  - student-Teacher Relationship Scale (STRS; Koomen, Verschueren & Pianta, 2007)
  - 1047 students from 559 primary school teachers
  - average cluster size of 1.873

SEM with discrete data: estimation

- full information approach: marginal maximum likelihood (MML: e.g., Bock & Lieberman, 1970)

$$L_i(\boldsymbol{\theta}) = f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}) = \int_{D(\boldsymbol{\eta})} f(\mathbf{y}_i|\boldsymbol{\eta},\mathbf{x}_i;\boldsymbol{\theta}) f(\boldsymbol{\eta}|\mathbf{x}_i;\boldsymbol{\theta}) d\boldsymbol{\eta}$$

  - computationally intensive (numerical integration)

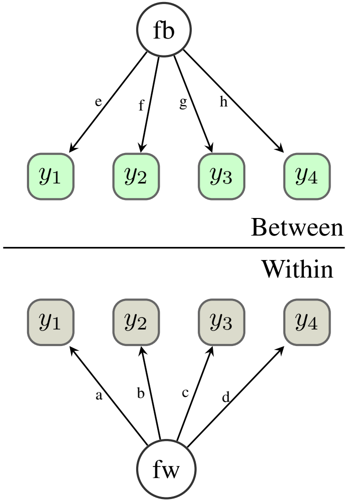
- limited information approach:
  - WLSMV (robust weighted least squares: three steps; Muthén, 1978 ; Muthén, Du Toit & Spisic, 1997)
  - pairwise likelihood estimation (Jöreskog& Moustaki, 2001)

$$pl(\boldsymbol{\theta}) = \sum_{k<l} \ln L(\theta;(\mathbf{y}_i,\mathbf{y}_j)) = \sum_{i<j} \sum_{\mathbf{y}_i=1}^{m_i} \sum_{\mathbf{y}_j=1}^{m_j} p_{\mathbf{y}_i,\mathbf{y}_j} \ln[p_{\mathbf{y}_i,\mathbf{y}_j}/\pi_{\mathbf{y}_i,\mathbf{y}_j}(\theta)]$$

multilevel SEM with discrete data: estimation

- full information: multilevel MML (e.g., Hedeker & Gibbons, 1994)
- limited information: multilevel WLS (Asparouhov & Muthén, 2007)
- limited information: PL (+ random effects)
  - generalized linear mixed models
    - (crossed) random effects: e.g., Bellio & Varin, 2005; Tibaldi, Verbeke, Molenberghs, Renard, Van den Noortgate, & De Boeck, 2007; Choa & Rabe-Hesketh, 2011
    - weighted version: Renard, Molenberghs & Geys, 2004
  - longitudinal models
    - e.g., Albert & Shih, 2010; Fu , Tao , Shi , Zhang & Lin, 2011; Cagnone, Moustaki & Vasdekis (2009; 2012)
    - weighted version: Vasdekis, Rizopoulos & Moustaki, 2014

two-level model – long




```
model <- '
  level: 1
    fw =~ a*y1 + b*y2 + c*y3 + d*y4
  level: 2
    fb =~ e*y1 + f*y2 + g*y3 + h*y4
'
fit <- sem(myModel, myData,
           cluster = "school")
```

```
Parameter Estimates:

Level 1 [within]:

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  fw =~
    y1        (a)     1.000
    y2        (b)     0.875    0.074   11.879    0.000
    y3        (c)     0.943    0.076   12.450    0.000
    y4        (d)     1.078    0.091   11.898    0.000

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .y1                0.967    0.078   12.433    0.000
   .y2                0.976    0.075   12.998    0.000
   .y3                0.989    0.076   13.097    0.000
   .y4                0.962    0.089   10.765    0.000
    fw                0.979    0.131    7.453    0.000

Level 2 [clus]:

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  fb =~
    y1        (e)     1.000
    y2        (f)     1.117    0.166    6.722    0.000
    y3        (g)     1.028    0.138    7.439    0.000
```

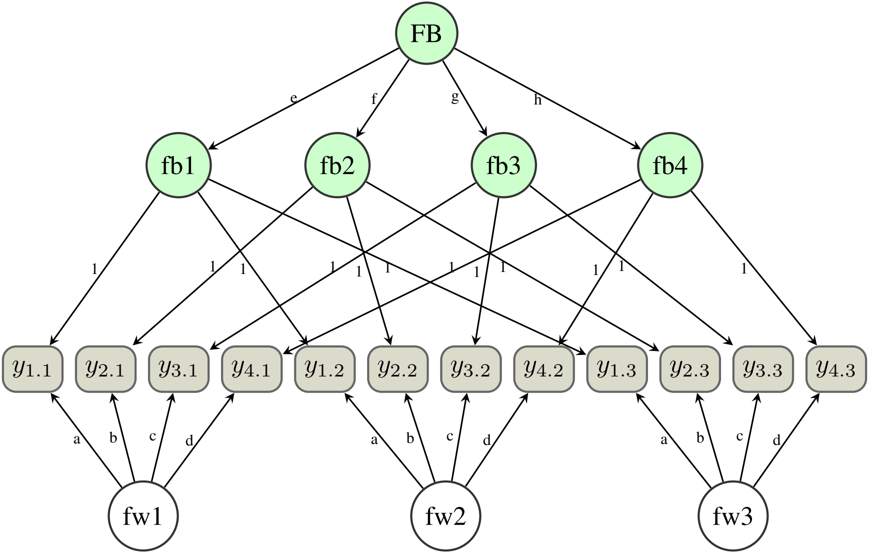
```
    y4          (h)    0.749    0.138    5.424    0.000

 Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .y1                0.000
   .y2                0.000
   .y3                0.000
   .y4                0.000
    fb                0.360    0.106    3.404    0.001
```

wide versus long data

- each row corresponds to a single cluster
- rows are independent
- unbalanced data can be handled by filling in missing values for the smaller clusters

	y1.1	y2.1	y3.1	y4.1	y5.1	y6.1	sk.1	sl.1	y1.2	y2.2	y3.2	y4.2	y5.2	y6.2	sk.2	sl.2
1	2	1	1	1	1	3	0	1	1	1	1	1	1	3	1	1
2	2	2	4	2	2	3	1	1	NA	NA	NA	NA	NA	NA	NA	NA
3	1	3	1	1	2	1	0	1	NA	NA	NA	NA	NA	NA	NA	NA
4	1	1	2	2	1	2	0	1	1	1	2	1	1	2	1	1
5	2	3	4	3	3	3	0	0	NA	NA	NA	NA	NA	NA	NA	NA
6	1	1	1	1	1	1	1	0	1	1	4	1	1	1	0	0

two-level model – wide

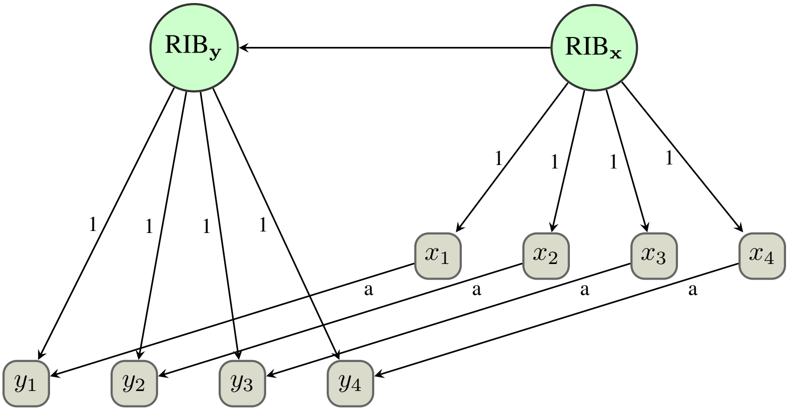


```
Parameter Estimates:

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  fw1 =~
    y1.1      (a)     1.000
    y2.1      (b)     0.875    0.074   11.879    0.000
    y3.1      (c)     0.943    0.078   12.100    0.000
    y4.1      (d)     1.078    0.090   11.946    0.000
  fw2 =~
    y1.2      (a)     1.000
    y2.2      (b)     0.875    0.074   11.879    0.000
    y3.2      (c)     0.943    0.078   12.100    0.000
    y4.2      (d)     1.078    0.090   11.946    0.000
  fw3 =~
    y1.3      (a)     1.000
    y2.3      (b)     0.875    0.074   11.879    0.000
    y3.3      (c)     0.943    0.078   12.100    0.000
    y4.3      (d)     1.078    0.090   11.946    0.000
  fb1 =~
    y1.1              1.000
    y1.2              1.000
    y1.3              1.000
  fb2 =~
    y2.1              1.000
    y2.2              1.000
    y2.3              1.000
```

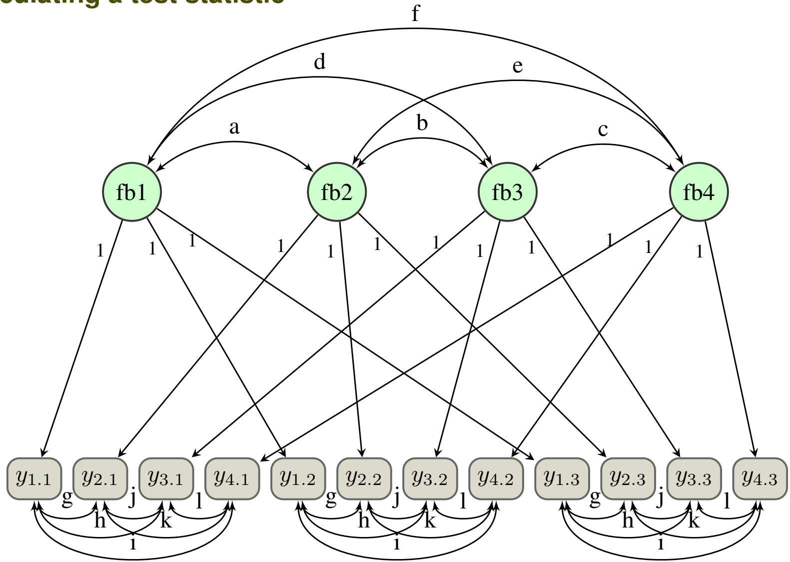
```
  fb3 =~
    y3.1              1.000
    y3.2              1.000
    y3.3              1.000
  fb4 =~
    y4.1              1.000
    y4.2              1.000
    y4.3              1.000
  fbb =~
    fb1        (e)    1.000
    fb2        (f)    1.117    0.157    7.102    0.000
    fb3        (g)    1.028    0.148    6.960    0.000
    fb4        (h)    0.749    0.135    5.563    0.000

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
    fw1       (wv)    0.979    0.130    7.512    0.000
    fw2       (wv)    0.979    0.130    7.512    0.000
    fw3       (wv)    0.979    0.130    7.512    0.000
    fb1               0.000
    fb2               0.000
    fb3               0.000
    fb4               0.000
    fbb               0.360    0.104    3.472    0.001
```

including a covariate

- covariates are modeled jointly with the other variables.



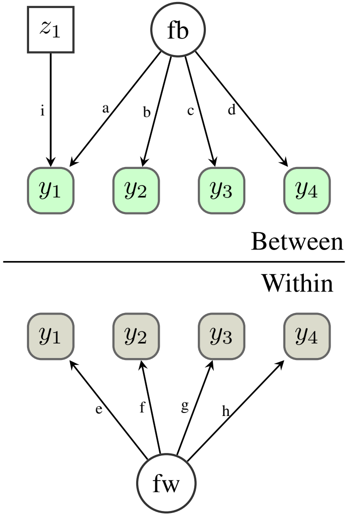
calculating a test statistic



‘wide format’ with discrete data

- single level estimation methods can be used to estimate multilevel models in the ‘wide format’
  - least squares estimation methods (e.g., WLSMV)
  - PL estimation methods
  - MML (only path models)
- unequal number of observations per cluster
  - continuous: fiml
  - PL: ’pairwise missing’, ’available cases’, . . .
  - least squares estimation methods: ’pairwise missing’, ’available cases’, . . .

small simulation study



data generation details:

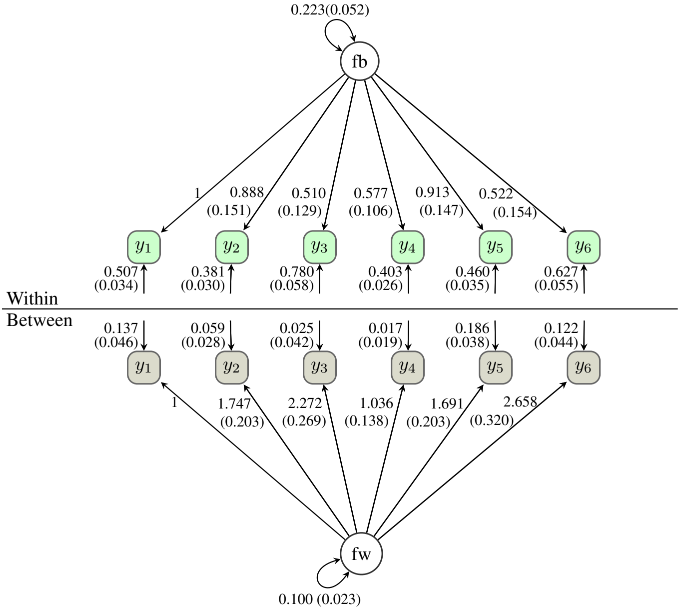
-observations per clusters: 200
-responses in cluster: 3
-total sample size: 600
-replications: 500
-%bias $= ((\hat{\theta} - \theta) = /\theta) * 100$

conclusion:

-results MML, PL, and WLSMV are close
-correctly specified model
*parameter est.(-0.800–5.360% bias)
*efficiency(1.602–13.330% bias)
-misspecified model
*parameter est.(-3.560–5.360% bias)
*efficiency(-3.560–12.400% bias)

example data:

- student-Teacher Relationship Scale (STRS; Koomen, Verschueren & Pianta, 2007)
  - six questions regarding dependent child behaviour filled in by primary school teachers
  - five-point scale ranging from 1 (definitely does not apply) to 5 (definitely does apply)
- selected classes where teachers filled in a questionnaire for maximum three children
  - 1047 students from 559 primary school teachers
  - average cluster size of 1.873
  - ICC: 0.088–0.365


0.223(0.052)
fb
1
0.888 (0.151)
0.510 (0.129)
0.577 (0.106)
0.913 (0.147)
0.522 (0.154)
y_1 y_2 y_3 y_4 y_5 y_6
0.507 (0.034)
0.381 (0.030)
0.780 (0.058)
0.403 (0.026)
0.460 (0.035)
0.627 (0.055)
Within
Between
0.137 (0.046)
0.059 (0.028)
0.025 (0.042)
0.017 (0.019)
0.186 (0.038)
0.122 (0.044)
y_1 y_2 y_3 y_4 y_5 y_6
1
1.747 (0.203)
2.272 (0.269)
1.036 (0.138)
1.691 (0.203)
2.658 (0.320)
fw
0.100 (0.023)

between level parameter estimates for discrete estimation methods

	MML (long)		WLSMV (wide)		PL (wide)	
pop.	$\overline{\lambda}$	SE	$\overline{\lambda}$	SE	$\overline{\lambda}$	SE
$\lambda_{1,1}$	1.000	-	1.000	-	1.000	-
$\lambda_{1,2}$	0.990	0.147	1.145	0.109	1.133	0.200
$\lambda_{1,3}$	0.426	0.084	0.520	0.055	0.516	0.114
$\lambda_{1,4}$	0.726	0.115	0.819	0.079	0.817	0.157
$\lambda_{1,5}$	0.838	0.124	0.885	0.080	0.929	0.140
$\lambda_{1,6}$	0.417	0.104	0.572	0.059	0.574	0.129
VAR(fb)	0.798	0.156	0.715	0.084	0.640	0.143

within level parameter estimates for discrete estimation methods

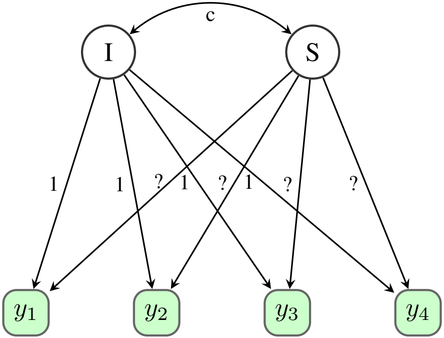
	MML (long)		WLSMV (wide)		PL (wide)	
pop.	$\overline{\lambda}$	SE	$\overline{\lambda}$	SE	$\overline{\lambda}$	SE
$\lambda_{1,1}$	1.000	-	1.000	-	1.000	-
$\lambda_{1,2}$	2.522	0.367	2.551	0.355	2.531	0.438
$\lambda_{1,3}$	2.200	0.345	2.462	0.359	2.337	0.423
$\lambda_{1,4}$	1.680	0.264	1.779	0.252	1.732	0.302
$\lambda_{1,5}$	1.758	0.253	1.731	0.234	1.732	0.322
$\lambda_{1,6}$	3.324	0.567	3.075	0.488	3.087	0.685
VAR(fw)	0.167	0.048	0.140	0.032	0.142	0.049

conclusion and discussion

- ‘wide format’ is a useful approach to study multilevel SEM with discrete data
  - with covariates
  - can handle unbalanced data
  - MML/PL is a single step procedure that can also handle random slopes
- advantages:
  - single level software can be used to estimate multilevel models
  - can handle many latent variables (WLSMV long format?)
  - does not need as many restrictions as the long format
  - all restrictions can be tested
- disadvantages:
  - tedious to specify (we will offer an better way in lavaan)
  - large clusters can be problematic

end

random slope



- balanced: same x for all clusters (growth model)
- unbalanced: different x for all clusters (casewise estimation)