New Modeling Methods for Multilevel Data

Tihomir Asparouhov and Bengt Muthén

May 27, 2011

Overview

- Two-level modeling for categorical data: weighted least squares estimation
- Three-level structural equation modeling: maximum-likelihood estimation
- Bayesian estimation for two and three level modeling
- Three-level modeling for categorical data: Bayesian estimation
- Cross classified structural equation models
- Two-level missing data imputation
- Two-level exploratory factor analysis
- Two-level multiple group modeling
- Plausible values for random effects

Two-level modeling for categorical data: weighted least squares estimation

Two-level WLS

- Asparouhov, T. & Muthén, B. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. Proceedings of the 2007 JSM.
- Extends the single level methodology originated by Muthén (1978) Contributions to factor analysis of dichotomous variables.
- Mplus Version 5 - 2007
- Advantages over ML: unlimited number of latent variables, fast computation, useful chi-square test of fit
- Disadvantages over ML: models only continuous and categorical variables, MAR missing data not supported, can yield non-positive definite variance/covariance for large number of random effects, only random intercepts.

Two-level WLS: Model

$$y_{pij} = k \Leftrightarrow \tau_{pk} < y^*_{pij} < \tau_{pk+1}. \tag{1}$$

$$y^*_{pij} = y_{wpij} + y_{bpj} \tag{2}$$

$$y_{wij} = \Lambda_w \eta_{wij} + \varepsilon_{wij} \tag{3}$$

$$\eta_{wij} = B_w \eta_{wij} + \Gamma_w x_{wij} + \xi_{wij}. \tag{4}$$

$$y_{bj} = \nu_b + \Lambda_b \eta_{bj} + \varepsilon_{bj} \tag{5}$$

$$\eta_{bj} = \alpha_b + B_b \eta_{bj} + \Gamma_b x_{bj} + \xi_{bj}. \tag{6}$$

Two-level WLS: Estimation

Part 1. Estimate the unrestricted model

$$y_{pij} = k \Leftrightarrow t_{pk} < y^*_{pij} < t_{pk+1}. \tag{7}$$

$$y^*_{pij} = y_{wpij} + y_{bpj}. \tag{8}$$

$$y_{wij} = \beta_w x_{wij} + \varepsilon_{wij} \tag{9}$$

$$y_{bj} = \mu_b + \beta_b x_{bj} + \varepsilon_{bj}. \tag{10}$$

- ML Univariate estimation - 1 dimension of numerical integration
- ML Bivariate estimation - 2 dimensions of numerical integration - estimation only for 2 correlation parameters

Two-level WLS: Estimation

Part 2. Estimate the structural model by minimizing

$$F = (s - s^*) W (s - s^*)^T \tag{11}$$

- s are the parameters of the unrestricted model
- s^* are the implied quantities from the structural model
- W is a weight matrix
- F is the base of the chi-square test of fit for the structural model v.s. the unrestricted model.

Two-level WLS: Simulation Study

- 6 polytomous observed variables with 5 categories
- 100 clusters of size 10
- 2 within and 2 between factors in a two-level CFA each measured by 3 variables
- ML uses 8 dimensional integration - montecarlo integration

Table: Two-level factor analysis model with categorical variables.

parameter	true value	WLSM bias	ML bias	WLSM coverage	ML coverage	Efficiency ratio
λ_{w2}	1.0	3%	2%	97%	100%	1.14
ψ_{w12}	0.4	2%	-14%	97%	89%	0.89
ψ_{w11}	0.7	2%	-23%	94%	75%	0.71
λ_{b2}	1.0	5%	4%	96%	94%	0.96
ψ_{b12}	0.2	-1%	-22%	94%	81%	0.91
ψ_{b11}	0.4	1%	-31%	93%	57%	0.77
τ_{11}	-0.3	-3%	-6%	96%	87%	1.17
τ_{12}	0.4	-1%	-14%	96%	81%	1.00
τ_{13}	1.2	0%	-11%	95%	55%	0.71
τ_{14}	1.8	0%	-10%	98%	47%	0.56
θ_{b1}	0.2	-2%	-55%	97%	32%	0.66

Three-level structural equation modeling: maximum-likelihood estimation

3-level SEM

- New feature in Mplus 7
- Three level multivariate data Y_{pijk}

$$Y_{pijk} = Y_{1pijk} + Y_{2pjk} + Y_{3pk} \tag{12}$$

- 3 sets of structural equations - one on each level

$$Y_{1ijk} = \Lambda_1 \eta_{ijk} + \varepsilon_{ijk} \tag{13}$$

$$\eta_{ijk} = B_1 \eta_{ijk} + \Gamma_1 x_{ijk} + \xi_{ijk}. \tag{14}$$

$$Y_{2jk} = \Lambda_2 \eta_{jk} + \varepsilon_{jk} \tag{15}$$

$$\eta_{jk} = B_2 \eta_{jk} + \Gamma_2 x_{jk} + \xi_{jk}. \tag{16}$$

$$Y_{3k} = \nu + \Lambda_3 \eta_k + \varepsilon_k \tag{17}$$

$$\eta_k = \alpha + B_3 \eta_k + \Gamma_3 x_k + \xi_k. \tag{18}$$

3-level SEM: Estimation

- Double EM-algorithm where the latent variables are Y_{2jk} and Y_{3k}
- E-step is based on $[Y_{3k}|*]$ and $[Y_{2jk}|Y_{3k},*]$
- M-step is a simple 3 group SEM maximization.
- Missing data: MAR support. Not a part of the EM-algorithm.
- Model accommodates easily variables defined on different levels: 7 types.
- Random coefficients for level 1 covariates and level 2 covariates

$$\Gamma_1 = \gamma_1 + \gamma_{1jk} + \gamma_{1k}$$

$$\Gamma_2 = \gamma_2 + \gamma_{2k}$$

  where γ_{1jk} is a level 2 latent variable while γ_{1k} and γ_{2k} are level 3 latent variables.
- No numerical integration
- Fast computation - all runs less than 1 min.
- Robust standard errors using sandwich estimator.

3-level SEM Example 1: Path Analysis / 3-level regression

$$Y_{ijk} = Y_{1ijk} + Y_{2jk} + Y_{3k}$$
$$Z_{ijk} = Z_{1ijk} + Z_{2jk} + Z_{3k}$$
$$Y_{1ijk} = \beta_1 Z_{1ijk} + \varepsilon_{ijk}$$
$$Y_{2jk} = \beta_2 Z_{2jk} + \varepsilon_{jk}$$
$$Y_{3k} = \alpha + \beta_3 Z_{3k} + \varepsilon_k$$

- Simulation: 80 level 3 clusters, with 15 level 2 clusters each of size 10.
- Extension of BFSPE, bias reduction (observed v.s. latent predictor), and Mediation modeling to 3 level
- Marsh et al. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling errors. MBR, 44, 764-802.
- Preacher et al. (2010). A general multilevel SEM framework for assessing multilevel mediation. Psychological Methods, 15, 209-233.

3-level SEM Example 1: Path Analysis Results

MODEL RESULTS

	Population	ESTIMATES Average	Std. Dev.	S. E. Average	M. S. E.	95% Cover	% Sig Coeff
• **Within Level**							
Y1 ON							
Y2	0.400	0.3987	0.0101	0.0094	0.0001	0.920	1.000
Variances							
Y2	1.200	1.1991	0.0152	0.0162	0.0002	0.960	1.000
Residual Variances							
Y1	1.200	1.1999	0.0136	0.0160	0.0002	0.980	1.000
• **Between CLUSTER1 Level**							
Y1 ON							
Y2	0.300	0.2933	0.0362	0.0352	0.0013	0.930	1.000
Variances							
Y2	0.700	0.7013	0.0362	0.0342	0.0013	0.940	1.000
Residual Variances							
Y1	0.700	0.7062	0.0375	0.0343	0.0014	0.920	1.000
• **Between CLUSTER2 Level**							
Y1 ON							
Y2	0.100	0.0692	0.1343	0.1271	0.0188	0.900	0.110
Means							
Y2	0.000	-0.0073	0.0736	0.0746	0.0054	0.950	0.050
Intercepts							
Y1	0.000	-0.0094	0.0689	0.0756	0.0048	0.960	0.040
Variances							
Y2	0.400	0.3938	0.0707	0.0703	0.0050	0.920	1.000
Residual Variances							
Y1	0.400	0.3920	0.0653	0.0680	0.0043	0.920	1.000

- Simulation: 1 factor on each level with 5 indicators.
- 3 simulations with different number of level 3 clusters: 80, 40, 20.
- Chi-square test of fit between the restricted model factor analysis model and the unrestricted variance covariance.
- DF=15, average chi-square value 15.156 / 15.571 / 17.707, rejection rates 5% / 9% / 19%

3-level SEM Example 2: Factor Analysis Results, 80 level 3 clusters

```
MODEL RESULTS

                              ESTIMATES                 S. E.     M. S. E.  95%  % Sig
                 Population   Average    Std. Dev.     Average              Cover Coeff
  • Within Level

 E        BY
  Y2                 1.000     0.9974     0.0200      0.0203      0.0004 0.950 1.000

 Variances
  E                  0.800     0.7999     0.0232      0.0246      0.0005 0.960 1.000

 Residual Variances
  Y1                 1.200     1.2031     0.0201      0.0204      0.0004 0.930 1.000

  • Between CLUSTER1 Level

 EB1      BY
  Y2                 1.000     1.0014     0.1620      0.1720      0.0260 0.940 1.000

 Variances
  EB1                0.200     0.2070     0.0477      0.0488      0.0023 0.970 1.000

 Residual Variances
  Y1                 0.900     0.8982     0.0528      0.0542      0.0028 0.940 1.000

  • Between CLUSTER2 Level

 EB2      BY
  Y2                 1.000     1.0205     0.1494      0.1413      0.0225 0.960 1.000

 Intercepts
  Y1                 1.000     0.9868     0.1082      0.1098      0.0118 0.960 1.000

 Variances
  EB2                0.600     0.5943     0.1556      0.1450      0.0240 0.920 1.000

 Residual Variances
```

3-level SEM Example 2: Factor Analysis Results, 40 level 3 clusters

MODEL RESULTS

	Population	ESTIMATES Average	Std. Dev.	S. E. Average	M. S. E.	95% Cover	% Sig Coeff
Within Level							
E BY							
Y2	1.000	0.9983	0.0273	0.0281	0.0007	0.950	1.000
Variances							
E	0.800	0.7966	0.0312	0.0347	0.0010	0.970	1.000
Residual Variances							
Y1	1.200	1.2012	0.0273	0.0284	0.0007	0.950	1.000
Between CLUSTER1 Level							
EB1 BY							
Y2	1.000	1.0305	0.2954	0.2631	0.0873	0.930	1.000
Variances							
EB1	0.200	0.2146	0.0819	0.0693	0.0069	0.900	0.910
Residual Variances							
Y1	0.900	0.8847	0.0742	0.0765	0.0057	0.960	1.000
Between CLUSTER2 Level							
EB2 BY							
Y2	1.000	1.0198	0.1996	0.2088	0.0398	0.950	1.000
Intercepts							
Y1	1.000	0.9838	0.1579	0.1535	0.0249	0.950	1.000
Variances							
EB2	0.600	0.5848	0.2036	0.2001	0.0413	0.920	0.950
Residual Variances							
Y1	0.300	0.2819	0.1076	0.0943	0.0118	0.810	0.900

3-level SEM Example 2: Factor Analysis Results, 20 level 3 clusters

MODEL RESULTS

	Population	ESTIMATES Average	Std. Dev.	S. E. Average	M. S. E.	95% Cover	% Sig Coeff
Within Level							
E BY							
Y2	1.000	0.9948	0.0388	0.0387	0.0015	0.950	1.000
Variances							
E	0.800	0.7974	0.0455	0.0482	0.0021	0.920	1.000
Residual Variances							
Y1	1.200	1.1977	0.0382	0.0388	0.0015	0.970	1.000
Between CLUSTER1 Level							
EB1 BY							
Y2	1.000	1.1296	0.5277	0.4536	0.2925	0.940	0.830
Variances							
EB1	0.200	0.2172	0.1165	0.1002	0.0137	0.920	0.530
Residual Variances							
Y1	0.900	0.8722	0.1221	0.1104	0.0155	0.930	1.000
Between CLUSTER2 Level							
EB2 BY							
Y2	1.000	1.0622	0.3626	0.3714	0.1340	0.900	0.890
Intercepts							
Y1	1.000	0.9687	0.2032	0.2147	0.0418	0.950	0.990
Variances							
EB2	0.600	0.5824	0.2963	0.2640	0.0872	0.830	0.630
Residual Variances							
Y1	0.300	0.2742	0.1450	0.1292	0.0215	0.880	0.610

3-level SEM Example 2: Factor Analysis with random coefficients

- Simulation: 5 observed indicator variables Y_{pijk} and one covariate X_{ijk}.

$$Y_{pijk} = \mu_p + \lambda_p \eta_{ijk} + \varepsilon_{pijk}$$

$$\eta_{ijk} = \eta_{0ijk} + s_{jk} x_{ijk}$$

$$\eta_{0ijk} = \eta_{1ijk} + \eta_{2jk} + \eta_{3k}$$

$$s_{jk} = s_{2jk} + s_{3k}$$

- 5 indicators variables measuring one factor regressed on a single predictor with random intercept and slope varying over level 2 and level 3
- 100 level 3 clusters, each with 20 level 2 clusters, each with 5 observations with 5 indicator variables

3-level SEM Example 2: Factor Analysis with random coefficients results

MODEL RESULTS

	Population	ESTIMATES Average	Std. Dev.	S. E. Average	M. S. E.	95% Cover	% Sig Coeff
Within Level							
F BY							
Y2	1.000	0.9998	0.0165	0.0143	0.0003	0.900	1.000
Residual Variances							
Y1	2.000	2.0027	0.0516	0.0480	0.0026	0.930	1.000
F	1.000	0.9989	0.0414	0.0396	0.0017	0.950	1.000
Between CLUSTER1 Level							
S WITH							
F1	0.300	0.2972	0.0329	0.0359	0.0011	0.960	1.000
Variances							
F1	0.700	0.6971	0.0516	0.0492	0.0026	0.930	1.000
S	0.600	0.5926	0.0460	0.0517	0.0022	0.960	1.000
Between CLUSTER2 Level							
S WITH							
F2	0.300	0.2934	0.0601	0.0630	0.0036	0.950	1.000
Means							
S	1.300	1.2985	0.0705	0.0720	0.0049	0.940	1.000
Intercepts							
Y1	0.000	-0.0015	0.0769	0.0787	0.0059	0.950	0.050
Variances							
F2	0.500	0.4871	0.0784	0.0822	0.0062	0.950	1.000
S	0.400	0.3981	0.0691	0.0717	0.0047	0.960	1.000

Three-level structural equation modeling for categorical and continuous variable: Bayesian estimation

Bayesian Estimation

- The same 3-level structural model as with ML
- MCMC estimation: slopes / variances / latent variables / missing data / between level data
- The key is the likelihood posteriors $[Y_{3k}|*]$ and $[Y_{2jk}|Y_{3k},*]$ - the same as in the EM algorithm, followed by multiple group analysis
- Algorithm uses largest blocks possible for efficient mixing
- Advantages over ML: accommodates priors, not reliant on asymptotic theory, posterior for random effects and more accurate estimates, flexible modeling in structural part
- Easily extends modeling to categorical variables
- New feature in Mplus 7

3-level with categorical data: univariate regression

- y_{ijk} categorical variable with 3 possible outcomes
- $x_{1ijk}, x_{2jk}, x_{3k}$ covariates at level 1, 2 and 3.

$$y_{ijk} = l \Leftrightarrow \tau_l < y^*_{ijk} < \tau_{l+1}$$

$$y^*_{ijk} = \alpha_{jk} + \alpha_k + \beta_1 x_{1ijk} + \beta_2 x_{2jk} + \beta_3 x_{3k} + \varepsilon_{ijk}$$

$$\beta_1 = \beta_{10} + \beta_{1jk} + \beta_{1k}$$

$$\beta_2 = \beta_{20} + \beta_{2k}$$

- Variance of ε_{ijk} is 1 as in probit regression
- ML would use 5 dimensional integration - very difficult
- Bayesian estimation $\leq$ 1 min
- Default of weakly informative priors: IW(I,p+1), N(0,5). Key for small sample problems.

3-level with categorical data: univariate regression results

```
Between CLUSTER1 Level

 S          ON
  X2              0.400      0.3938     0.0635     0.0550     0.0040 0.880 1.000

 Y         WITH
  S               1.000      0.9732     0.1201     0.1151     0.0150 0.880 1.000

 Residual Variances
  Y               1.500      1.4758     0.1676     0.1478     0.0284 0.910 1.000
  S               1.600      1.5260     0.1850     0.1740     0.0393 0.920 1.000

Between CLUSTER2 Level

 S          ON
  X3              0.550      0.5183     0.2106     0.2246     0.0449 0.950 0.620

 S2         ON
  X3              0.550      0.5646     0.2540     0.2135     0.0641 0.870 0.710

 Y          ON
  X3              1.300      1.2793     0.1815     0.1795     0.0330 0.920 1.000

 S2        WITH
  S               0.300      0.2596     0.2505     0.2412     0.0637 0.930 0.220
  Y              -0.200     -0.2188     0.1844     0.1908     0.0340 0.920 0.180

 Y         WITH
  S               0.700      0.6413     0.2068     0.2182     0.0458 0.950 0.940

 Intercepts
  S               1.700      1.6872     0.1582     0.1573     0.0249 0.940 1.000
  S2              0.300      0.2971     0.1497     0.1487     0.0222 0.930 0.540

 Thresholds
  Y$1            -2.100     -2.0968     0.1380     0.1152     0.0189 0.860 1.000
  Y$2             2.100      2.0661     0.1600     0.1179     0.0265 0.820 1.000

 Residual Variances
  Y               1.300      1.2368     0.2444     0.2315     0.0631 0.910 1.000
  S               2.200      2.0684     0.3702     0.3756     0.1530 0.930 1.000
  S2              2.000      1.9354     0.3586     0.3428     0.1314 0.950 1.000
```

3-level SEM with categorical data: Factor model

- y_{pijk} 5 categorical variable with 3 possible outcomes
- 1 factor on each level
- ML would use 13 dimensions of integration

$$y_{pijk} = l \Leftrightarrow \tau_{pl} < y^*_{pijk} < \tau_{pl+1}$$

$$y^*_{pijk} = y_{pjk} + y_{pk} + \lambda_{1p} f_{1ijk} + \lambda_{2p} f_{2jk} + \lambda_{3p} f_{3k} + \varepsilon_{pijk}$$

- Variances of $\varepsilon_{pijk}, f_{1ijk}, f_{2jk}, f_{3k}$ are 1

3-level SEM with categorical data: Factor model results

```
MODEL RESULTS

                          ESTIMATES              S. E.     M. S. E.  95%  % Sig
              Population  Average   Std. Dev.  Average             Cover Coeff
Within Level

 E1       BY
  Y1             1.300     1.3130     0.0566     0.0496     0.0033 0.890 1.000

Between CLUSTER1 Level

 E2       BY
  Y1             1.000     1.0094     0.0639     0.0564     0.0041 0.920 1.000

 Residual Variances
  Y1             0.400     0.4042     0.0607     0.0573     0.0037 0.930 1.000

Between CLUSTER2 Level

 E3       BY
  Y1             0.600     0.6032     0.1056     0.1072     0.0111 0.940 1.000

 Thresholds
  Y1$1          -1.000    -0.9994     0.1077     0.0866     0.0115 0.890 1.000
  Y1$2           1.000     1.0180     0.1002     0.0876     0.0103 0.890 1.000

 Residual Variances
  Y1             0.300     0.3139     0.0750     0.0772     0.0058 0.960 1.000
```

Cross-classified structural equation modeling

Cross-classified data

- Y_{pijk} is the $p-$th observation for person i belonging to level 2 cluster j and level 3 cluster k.
- Level 2 clusters are not nested within level 3 clusters
- Examples:
  - To model income: individuals are nested within the same geographical location and are nested within occupation clusters
  - Students are nested within schools and nested within neighborhoods
  - Student performance scores are nested within students and within teachers
  - Studies where observations are nested within persons and treatments/situations
  - Studies where observations are nested within neighborhoods and interviewer
  - Generalizability theory (Brennan, 2001; Cronbach, Rajaratnam, & Gleser, 1963), Items are considered a random sample from a population of items.

Cross-classified modeling

- Why do we need to model both sets of clustering?
- A model that ignores the dependence between observations in the same cluster yields misspecifications, underestimates SE, fails to discover the true predictor/explanotory effect stemming from the clusters
- Why do we need random effects and not fixed effects?
- Fixed effects possible: Tucker3
- Fixed effect for one set of clusters and random for the other: dummy variables for one set of clusters
- If the number of clusters is more than 10 - modeling with fixed effects may lead to too many parameters and less accurate model.

Cross-classified modeling

Gonzalez, De Boeck, Tuerlinckx (2008) A Double-Structure Structural Equation Model for Three-Mode Data. Psychological Methods, 337 - 353.

Table 1
Three Approaches for Modeling Three-Mode Data

Variable	Tucker3 (a)	Tucker3 (b)	SEM–MTMM	2sSEM
Interactions	[PSR]	[PSR]	[PS] [PR]	[PR] [SR]
Parameters				
Persons	fixed	random	random	random
Situations	fixed	fixed	fixed	random
Responses	fixed	fixed	fixed	fixed
Dependence structure				
Dependent	irrelevant	pairs (S, R)	pairs (S, R)	R
Independent	irrelevant	P	P	nonoverlapping pairs (P, S)

Note. Variables that interact are grouped together inside a set of brackets; variables that are independent of one another are placed in separate brackets. SEM–MTMM = structural equation model for multitrait–multimethod data; 2sSEM = double-structure structural equation model; P = persons; S = situations; R = responses.

Cross-classified model

- Univariate model
- Both cluster levels contribute with random effects

$$Y_{ijk} = Y_{1ijk} + Y_{2j} + Y_{3k}$$

- Y_{2j} and Y_{3k} are random effects for the two cluster levels

Cross-classified model

- General SEM model

$$Y_{pijk} = Y_{1pijk} + Y_{2pj} + Y_{3pk}$$

- 3 sets of structural equations - one on each level

$$Y_{1ijk} = \nu + \Lambda_1 \eta_{ijk} + \varepsilon_{ijk}$$

$$\eta_{ijk} = \alpha + B_1 \eta_{ijk} + \Gamma_1 x_{ijk} + \xi_{ijk}$$

$$Y_{2j} = \Lambda_2 \eta_j + \varepsilon_j$$

$$\eta_j = B_2 \eta_j + \Gamma_2 x_j + \xi_j$$

$$Y_{3k} = \Lambda_3 \eta_k + \varepsilon_k$$

$$\eta_k = B_3 \eta_k + \Gamma_3 x_k + \xi_k$$

Cross - classified model estimation

- The parameter ν and α can be used on any level
- Bayesian MCMC estimation
- The two sets of random effects Y_{2pj}, Y_{3pk} are treated separately
- The two sets of random effects $[Y_{2pj}|*, Y_{3pk}]$ and $[Y_{3pk}|*, Y_{2pj}]$ are level 2 random effect
- Easily extends to categorical variables
- New feature in Mplus 7

Cross-classified model, example 1: Factor model

- 1 factor at the individual level and 1 factor at each of the clustering levels, 5 indicator variables on the individual level

$$y_{pijk} = \mu_p + y_{pj} + y_{pk} + \lambda_{1p} f_{1ijk} + \lambda_{2p} f_{2j} + \lambda_{3p} f_{3k} + \varepsilon_{pijk}$$

- Variances of $f_{1ijk}, f_{2jk}, f_{3k}$ are 1, all loadings are estimated.
- 50 level 2 clusters, 50 level 3 clusters, 1 unit within each cluster intersection
- Total sample size 2500
- 1 unit on the within level would not work if the clusters were nested
- Estimation takes less than 1 min per replication

Cross-classified model example 1: Factor model results

```
MODEL RESULTS

                            ESTIMATES              S. E.     M. S. E.  95%  % Sig
              Population   Average   Std. Dev.   Average              Cover Coeff
Within Level

 E         BY
  Y1               1.500    1.5018     0.0286     0.0324     0.0008 0.970 1.000

 Residual Variances
  Y1               1.200    1.1999     0.0400     0.0437     0.0016 0.990 1.000

Between CLUSTER1 Level

 E2        BY
  Y1               1.000    1.0000     0.2822     0.2865     0.0788 0.920 0.940

 Residual Variances
  Y1               1.500    1.6146     0.3755     0.4754     0.1527 0.990 1.000

Between CLUSTER2 Level

 E3        BY
  Y1               0.800    0.8606     0.1538     0.1698     0.0271 0.950 1.000

 Intercepts
  Y1               2.200    2.1561     0.2477     0.2900     0.0627 0.990 1.000

 Residual Variances
  Y1               0.500    0.5517     0.1582     0.1647     0.0274 0.940 1.000
```

Cross-classified model, example 2: Gonzalez's example

- Observations are nested within individual and several specific situations, the dependent variables are 4 binary outcomes
- 1 observation for each pair of clustering units

$$y^*_{pjk} = y_{pj} + y_{pk} + \varepsilon_{pjk}$$

- Variances of ε_{pjk} is fixed to 1
- 100 level 2 clusters, 100 level 3 clusters
- Identical structural model for the two cluster random effects

$$y_{1j} = \beta_1 y_{3j} + \beta_2 y_{4j} + \varepsilon_{1j}$$

$$y_{2j} = \beta_3 y_{3j} + \beta_4 y_{4j} + \varepsilon_{2j}$$

$$y_{1k} = \beta_1 y_{3k} + \beta_2 y_{4k} + \varepsilon_{1k}$$

$$y_{2k} = \beta_3 y_{3k} + \beta_4 y_{4k} + \varepsilon_{2k}$$

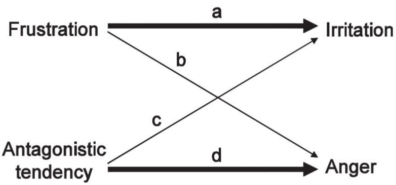


Figure 3. Graphical representation of the research questions. *a*, *b*, *c*, and *d* are effect parameters.

Identical structural model is estimated for the two-sets of random effects.

Cross-classified model, example 2: Gonzalez's example results

Between CLUSTER1 Level

Y1 ON							
Z1	0.300	0.3059	0.0891	0.0932	0.0079	0.940	0.910
Z2	-0.300	-0.2932	0.0870	0.0900	0.0075	0.970	0.940
Y2 ON							
Z1	0.500	0.5046	0.0834	0.0877	0.0069	0.970	1.000
Z2	-0.500	-0.5045	0.1004	0.0939	0.0100	0.960	1.000
Y1 WITH							
Y2	0.700	0.6823	0.1553	0.1443	0.0242	0.940	1.000
Z1 WITH							
Z2	0.700	0.7129	0.1394	0.1458	0.0194	0.970	1.000
Variances							
Z1	1.500	1.5237	0.2282	0.2370	0.0521	0.960	1.000
Z2	0.900	0.9117	0.1251	0.1408	0.0156	0.960	1.000
Residual Variances							
Y1	1.500	1.4707	0.2335	0.2317	0.0548	0.950	1.000
Y2	0.900	0.8946	0.1421	0.1427	0.0200	0.970	1.000

Cross-classified model, example 2: Gonzalez's example results

Between CLUSTER2 Level

Y1 ON							
Z1	0.300	0.3059	0.0891	0.0932	0.0079	0.940	0.910
Z2	-0.300	-0.2932	0.0870	0.0900	0.0075	0.970	0.940
Y2 ON							
Z1	0.500	0.5046	0.0834	0.0877	0.0069	0.970	1.000
Z2	-0.500	-0.5045	0.1004	0.0939	0.0100	0.960	1.000
Y1 WITH							
Y2	0.400	0.3929	0.0788	0.0814	0.0062	0.960	1.000
Z1 WITH							
Z2	0.400	0.4056	0.0805	0.0816	0.0064	0.950	1.000
Thresholds							
Y1$1	0.200	0.1998	0.1400	0.1461	0.0194	0.970	0.310
Y2$1	-0.500	-0.4932	0.1400	0.1453	0.0194	0.950	0.910
Z1$1	0.200	0.2263	0.1517	0.1398	0.0235	0.930	0.350
Z2$1	-0.500	-0.4675	0.1503	0.1285	0.0234	0.880	0.910
Variances							
Z1	0.500	0.5208	0.0767	0.0811	0.0063	0.950	1.000
Z2	0.800	0.8133	0.1342	0.1255	0.0180	0.960	1.000
Residual Variances							
Y1	0.500	0.5053	0.0748	0.0802	0.0056	0.950	1.000
Y2	0.800	0.7995	0.1277	0.1265	0.0161	0.960	1.000

Two-level factor model with random loadings

$$y_{pij} = y_{pj} + \lambda_{pj}\eta_{ij} + \varepsilon_{pjk}$$

Bayesian estimation, based on two explicit posterior distributions

$$[y_{pj}, \lambda_{pj} | *, \eta_{ij}]$$

$$[\eta_{ij} | *, y_{pj}, \lambda_{pj}]$$

Multiple imputations for two-level data

Multiple imputations for two-level data

- Bayesian estimation of unrestricted H1 model

$$y_{pij} = y_{wpij} + y_{bpj}$$

- Unrestricted variance covariance matrix for y_{wpij}
- Unrestricted means and variance covariance matrix for y_{bpj}
- Imputation of data based on the MCMC estimation
- Categorical data imputation based on y^*_{pij} and unrestricted correlation matrices
- PX methodology for estimating correlation matrices: Mplus 6.1

Multiple imputations for two-level categorical data example

- Asparouhov and Muthén (2010) Multiple Imputation with Mplus
- P categorical variables with 4 categories, M clusters of size 30
- Two-level factor model with 1 factor at each level

$$Y^*_{pij} = \lambda_{wp}\eta_{wij} + \lambda_{bp}\eta_{bj} + \varepsilon_{bpj} + \varepsilon_{wpij}$$

- Variance of ε_{wpij} is fixed to 1
- Generate MAR missing data using

$$P(Y_j \text{ is missing}) = \frac{1}{1 + Exp(-1.5 + 0.15\sum_{k=26}^{30} Y_k)}.$$

- Analyze the data with WLSMV directly and WLSMV after imputation

Multiple imputations for two-level categorical data results

Table: *MSE* for the threshold parameters for two-level imputation with categorical variables.

Number of Clusters	Number of Variables	Direct WLSMV	H1 Imputed WLSMV
50	10	0.352	0.268
200	10	0.204	0.077
50	30	0.418	0.273
200	30	0.235	0.074

Imputation based results are more accurate. WLSMV supports MCAR, MARX but not MAR.

Two-level exploratory factor analysis

Two-level exploratory factor analysis

- Mplus 5
- Based on ML estimation for all continuous variables and WLS for continuous and categorical
- Estimate an unrotated solution

$$Y_{ij} = \mu + \Lambda_{0w}\eta_{0w} + \Lambda_{0b}\eta_{0b} + \varepsilon_{ij}$$

- Λ_{0w} and Λ_{0b} have fixed 0 above the diagonal
- η_{0w} and η_{0b} are standard normal
- Independent rotation of Λ_{0w} and Λ_{0b} based on Jennrich's gradient projection algorithm (GPA)
- GPA minimizes $f(\Lambda_{0w}H_w^{-1})$ and $f(\Lambda_{0b}H_b^{-1})$ over orthogonal rotations H_w and H_b where f is the rotation criteria

Two-level exploratory factor analysis

- Mplus uses as a default the Geomin rotation criteria

$$f(\Lambda) = \sum_{i=1}^{p} \left(\prod_{j=1}^{m} (\lambda_{ij}^2 + \varepsilon) \right)^{1/m}$$

- Chi-square test of fit for evaluating the model
- As a preliminary step in deciding the number of factors estimate within level EFA or between level EFA
- Within level EFA estimates unrestricted variance covariance on the between level and factor model on the within level only
- Between level EFA estimates estimates unrestricted variance covariance on the within level and a factor model on the between level only

Two-level exploratory factor analysis example

Exploratory Factor Analysis Of Aggression Items

Item Distributions for Cohort 3: Fall 1st Grade (n=362 males in 27 classrooms)

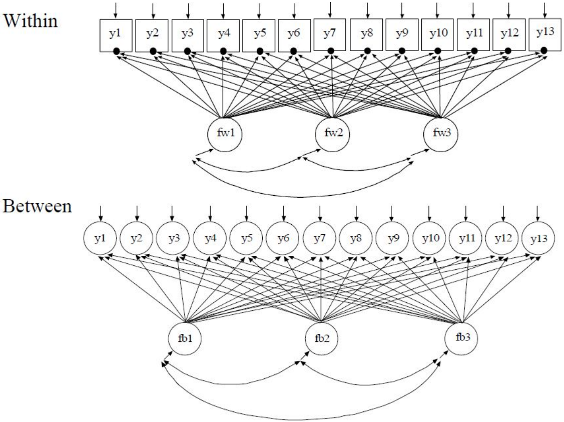
	Almost Never (*scored as 1*)	*Rarely* (*scored as 2*)	*Sometimes* (*scored as 3*)	*Often* (*scored as 4*)	*Very Often* (*scored as 5*)	*Almost Always* (*scored as 6*)
Stubborn	42.5	21.3	18.5	7.2	6.4	4.1
Breaks Rules	37.6	16.0	22.7	7.5	8.3	8.0
Harms Others	69.3	12.4	9.40	3.9	2.5	2.5
Breaks Things	79.8	6.60	5.20	3.9	3.6	0.8
Yells at Others	61.9	14.1	11.9	5.8	4.1	2.2
Takes Others' Property	72.9	9.70	10.8	2.5	2.2	1.9
Fights	60.5	13.8	13.5	5.5	3.0	3.6
Harms Property	74.9	9.90	9.10	2.8	2.8	0.6
Lies	72.4	12.4	8.00	2.8	3.3	1.1
Talks Back to Adults	79.6	9.70	7.80	1.4	0.8	1.4
Teases Classmates	55.0	14.4	17.7	7.2	4.4	1.4
Fights With Classmates	67.4	12.4	10.2	5.0	3.3	1.7
Loses Temper	61.6	15.5	13.8	4.7	3.0	1.4

Two-level exploratory factor analysis example

Hypothesized Aggressiveness Factors

- Verbal aggression
  - Yells at others
  - Talks back to adults
  - Loses temper
  - Stubborn
- Property aggression
  - Breaks things
  - Harms property
  - Takes others' property
  - Harms others
- Person aggression
  - Fights
  - Fights with classmates
  - Teases classmates

Two-level exploratory factor analysis example



Two-level exploratory factor analysis example

```
Number of clusters                    27

Average cluster size       13.407

Estimated Intraclass Correlations for the Y Variables
```

	Intraclass		Intraclass		Intraclass
Variable	Correlation	Variable	Correlation	Variable	Correlation
U1	0.110	U2	0.121	U3	0.208
U4	0.378	U5	0.213	U6	0.250
U7	0.161	U8	0.315	U9	0.208
U10	0.140	U11	0.178	U12	0.162
U13	0.172				

Two-level exploratory factor analysis example

Within-level Factors	Between-level Factors	Df	Chi-Square	CFI	RMSEA
unrestricted	1	65	66(p=0.43)	1.000	0.007
1	1	130	670	0.991	0.107
2	1	118	430	0.995	0.084
3	1	107	258	0.997	0.062
4*	1	97	193	0.998	0.052

*4th factor has no significant loadings

Two-level exploratory factor analysis example

	Within-Level Loadings			Between-Level Loadings
	Property	**Verbal**	**Person**	**General**
Stubborn	0.00	**0.78***	0.01	**0.65***
Breaks Rules	0.31*	0.25*	0.32*	**0.61***
Harms Others and Property	**0.64***	0.12	0.25*	**0.68***
Breaks Things	**0.98***	0.08	-0.12*	**0.98***
Yells At Others	0.11	**0.67***	0.10	**0.93***
Takes Others' Property	**0.73***	-0.15*	0.31*	**0.80***
Fights	0.10	0.03	**0.86***	**0.79***
Harms Property	**0.81***	0.12	0.05	**0.86***
Lies	**0.60***	0.25*	0.10	**0.86***
Talks Back To Adults	0.09	**0.78***	0.05	**0.81***
Teases Classmates	0.12	0.16*	**0.59***	**0.83***
Fights With Classmates	-0.02	0.13	**0.88***	**0.84***
Loses Temper	-0.02	**0.85***	0.05	**0.87***

Two-level multiple group modeling

Two-level multiple group modeling

- Two types of grouping variable: between (private v.s. public schools), within (female v.s. males)
- Between grouping variable: straight forward, likelihood is sum of independent groups
- Within grouping variable: if you treat it as if the grouping variable is between the two groups within the cluster are independent - misspecification
- Modeling options: the between random effect is the same across groups OR more flexible the between random effects is different across groups
- Is teacher's effect on student performance the same for females and males?
- Are there two correlated random effects (teachers ability to relate to different gender) or is it one random effect (teachers ability) that may affect students performance differently?

Two-level multiple group modeling

- Within level grouping variable - one random effect (assuming fixed variance of 1): twolevel mixture(with known class variable)

$$Y_{ijg} = \mu_g + \beta_g Y_j + \varepsilon_{ijg}$$

- Within level grouping variable - two random effects (assuming different variance across groups): twolevel mixture(with known class variable)

$$Y_{ijg} = \mu_g + Y_{gj} + \varepsilon_{ijg}$$

- Both use numerical integration currently
- Using 3 level modeling and treating the grouping variable as level 2 variable (Y_{gj} are correlated via Y_j) : no numerical integration

$$Y_{ijg} = \mu_g + Y_{gj} + Y_j + \varepsilon_{ijg}$$

- LRT to decide between the three models - if nested

Plausible Values

Plausible Values

- With Baysian estimation we can obtain posterior distribution for the random effects and plausible valeus as draws from the posterior
- Plausible values are more accurate than ML factor scores especially for small sample size
- Plausible values variability accounts for the variability in the parameter estimates
- Can be used for further analysis - comparing the random effects of different clusters, estimate difference, can be used for further modeling
- In 3 level modeling plausible values on level 2 accurately reproduce the correlation between level 2 random effects, not factor scores.