An extension of the general 3-step ML approach to random effect EHA with multiple latent categorical predictors.

Yajing Zhu, Fiona Steele, Irini Moustaki

Department of Statistics
London School of Economics and Political Science, UK

RSS Glasgow, 5 Sept 2017

Research objective

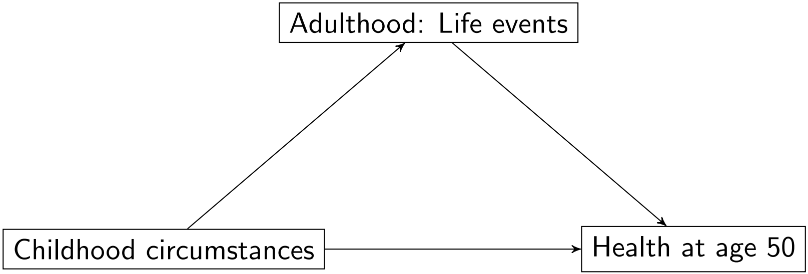

Figure 1: Develop a general joint modelling framework to explore the potential pathways between childhood circumstances, life events and health in mid-life.

Research objective

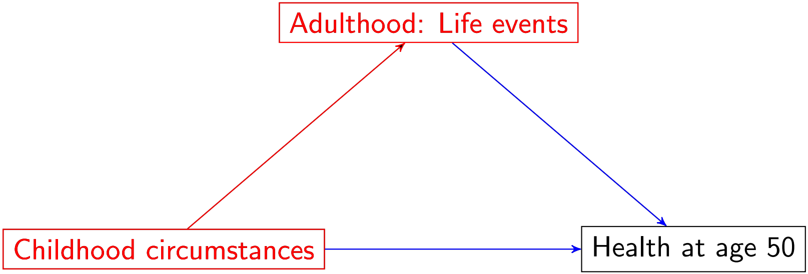

Figure 1: Develop a general joint modelling framework to explore the potential pathways between childhood circumstances, life events and health in mid-life.

Review of previous work

Main interest

How to include latent summaries of childhood SES as predictors of a distal outcome, in particular, a survival outcome?

- 1-step approach
  - Problem: unintended circular relationship.
- naive 3-step approaches (modal class, pseudo class)
  - Problem: misclassification, underestimated/overestimated standard errors.
- Advanced 3-step approaches (BCH, ML)
  - Problems with BCH: failure in estimating models with a categorical response and poor classification in the measurement model for the latent categorical predictor; severely underestimated standard errors.
  - Problems with ML: potential class shifts (can be monitored), explicit discussions of 1 LV only (but with genenalisablity).

Review: A 3-step ML approach

- 3-step approach with 1 LV: firstly proposed by Vermunt (2010); further developed by Asparouhov and Muthén (2014) to account for misclassification in the LCA step.

A 3-step ML approach

- Step 1: Estimate separate latent class models for categorical predictors.
- Step 2: Calculate misclassification probabilities.
- Step 3: Estimate models of interest, with categorical LVs as predictors.

Extension: A general 3-step ML approach I

Generalise the 3-step approach to ≥ 2 associated LVs by Zhu et al. (2017)

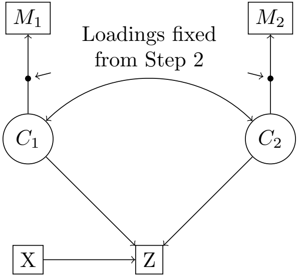

Figure 2: The 3-step approach with two latent categorical variables C_1 and C_2.

Extension: A general 3-step ML approach II

Assumption: $C_1 \perp\!\!\!\perp M_2|C_2$; $C_2 \perp\!\!\!\perp M_1|C_1$, $Z \perp\!\!\!\perp Us|C_1, C_2$. Log-likelihood function

$$l = \sum_{i=1}^{N} \log \sum_{C_2} \sum_{C_1} P(C_1, C_2) f(Z|X, C_1, C_2) P(M_1|C_1) P(M_2|C_2).$$

- Allows for a flexible association structure between LVs through a log-linear model.
- Tested the robustness to violation of model assumptions (findings: most sensitive to skewness, potential class shifts from Step 1 to Step 3).

Extension: random effects EHA I

Recall the substantive research question: does childhood circumstances influence life events (e.g. the hazard risk of partnership dissolution)? Z is now the event time, where events may be recurrent $\Rightarrow$ multilevel EHA with multiple categorical LVs.

Discrete-time event history data

- Denote by y_{ij} the duration of episode j of individual i, which is fully observed if an event occurs ($\delta_{ij} = 1$) and right-censored if not ($\delta_{ij} = 0$).
- Data restructuring: convert the observed data (y_{ij}, δ_{ij}) to a sequence of binary responses (y_{tij}), indicating whether an event has occurred in time interval $[t, t+1)$.
- Discrete-time hazard function: $h_{tij} = Pr(y_{tij} = 1 | y_{t-1,ij} = 0)$

Extension: random effects EHA II

Recall a general 3-step approach: Step 1& 2 estimates separate latent class models for C_1, C_2; computes misclassification probabilities $P(M_1|C_1)$, $P(M_2|C_2)$.

Step 3 is a random effects logit model, allowing for a log-linear structure between LVs.

$$\log\left(\frac{h_{tij}}{1-h_{tij}}\right) = \alpha_t + \beta^T \mathbf{X}_{tij} + \sum_{k_1=1}^{K_1-1} \tau_{k_1}^{C_1} I(C_{1i} = k_1) + \sum_{k_2=1}^{K_2-1} \tau_{k_2}^{C_2} I(C_{2i} = k_2) + u_i$$

- α_t is the baseline hazard function
- $\mathbf{X}_{tij}$ is the vector of time-varying and time-invariant predictors
- $\tau_{k_1}^{C_1}$ and $\tau_{k_2}^{C_2}$ are the class-specific coefficients of LVs
- $u_i \sim N(0, \sigma_u^2)$ is the time-invariant individual-specific unobservable

Estimation I

Denote all LVs (both continuous and discrete) by a vector $\xi_i = (u_i, C_{1i}, C_{2i})$, if they are all observed, the complete data log-likelihood is written as:

$$\begin{aligned} l &= \sum_{i=1}^{N} \log f\left(\mathbf{y}_{tij}, M_{1i}, M_{2i}, \xi_i | \mathbf{X}_{tij}\right) \\ &= \sum_{i=1}^{N} \left[\log f_1(\mathbf{y}_{tij}, M_{1i}, M_{2i} | \mathbf{X}_{tij}, \xi_i) + \log \phi(\xi_i)\right], \end{aligned} \tag{1}$$

where $\phi(\xi_i)$ is the joint distribution of LVs.

Estimation II

Assumptions: conditional independence of manifest variables and the distal outcome ($\mathbf{y}_{tij}$); conditional on u_i, durations in a risk set for a given individual are independent:

$$f_1(\mathbf{y}_{tij}, M_{1i}, M_{2i}|\mathbf{X}_{tij}, \xi_i) = f\left(\mathbf{y}_{tij}|\mathbf{X}_{tij}, \xi_i\right) P(M_{1i}|C_{1i})P(M_{2i}|C_{2i}),$$

such that (1) can be decomposed into five terms, i.e.

$$\begin{aligned} l = \sum_{i=1}^{N} \Bigg[& \log f(\mathbf{y}_{tij}|\mathbf{X}_{tij}, \xi_i) + \log P(M_{1i}|C_{1i}) + \log P(M_{2i}|C_{2i}) \\ & + \log P(C_{1i}, C_{2i}) + \log \phi(u_i) \Bigg], \end{aligned} \tag{2}$$

where $\phi(\cdot)$ is the normal density of u_i and we assume independence of (C_{1i}, C_{2i}) and u_i.

Estimation III

Note that the distributions of observed quantities are conditional on the LVs, use EM algorithm.

- E-step: expected score function is computed where the expectation is taken with respect to the posterior distribution of ξ_i given all observed data, i.e.
$$g(\xi_i | \mathbf{y}_{tij}, \mathbf{X}_{tij}, M_{1i}, M_{2i}).$$
- M-step: update parameters by using root-finding algorithms to solve the functions in E-step.

Simulation study: Data generation I

Target: generate discrete time-to-event data for repeatable events. After an event occurs, the origin is reset to zero.

Simplification: 2 binary LVs (ref=category 2), linear baseline hazard.

After generating latent classes C_1 and C_2 from the log-linear model, the manifest variable Us, we generate event times from

$$\log\left(\frac{h_{tij}}{1-h_{tij}}\right) = \alpha_t + \beta^T \mathbf{X}_{tij} + \tau^{C_1} I(C_{1i}=1) + \tau^{C_2} I(C_{2i}=1) + u_i.$$

Simulation study: Data generation II

Examples of the event histories of three individuals from the generated datasets.

ID1	>1 event, censored									
Calendar Time	1	2	3	4	5	6	7	8	9	10
Gap-time t	1	2	3	1	2	3	1	2	3	1
Episode j	1	1	1	2	2	2	3	3	3	4
D_i	1	1	1	1	1	1	1	1	1	1
y_{tij}	0	.	.	.	.	.	.	.	.	.
ID2	>1 event, not censored									
Calendar Time	1	2	3	4	5	6	7	8	9	10
Gap-time t	1	1	2	1	1	1	1	1	2	1
Episode j	1	2	2	3	4	5	6	7	7	8
D_i	10	10	10	10	10	10	10	10	10	10
y_{tij}	1	0	1	1	1	1	1	0	1	1
ID3	no event, censored									
Calendar Time	1	2	3	4	5	6	7	8	9	10
Gap-time t	1	2	3	4	5	6	7	8	9	10
Episode j	1	1	1	1	1	1	1	1	1	1
D_i	3	3	3	3	3	3	3	3	3	3
y_{tij}	0	0	0	.	.	.	.	.	.	.

Simulation study: Results (High entropy, N=500)

Models are estimated LatentGOLD 5.1 in settings with combinations of sample sizes (N=500, 2000) and entropy values (0.8 and 0.4) of the measurement models.

Entropy=0.8, N=500	TRUE	Bias (%)	SE	SD	Coverage
β_0	−2.00	−0.04	0.24	0.23	0.96
$\beta_1(t)$	1.50	0.30	0.11	0.11	0.95
$\beta_2(x)$	1.50	0.54	0.10	0.11	0.95
$\beta_3(x_t)$	−0.50	0.56	0.06	0.06	0.95
$\tau(C_1 = 1)$	2.50	0.18	0.19	0.19	0.94
$\tau(C_2 = 1)$	−1.00	1.30	0.18	0.18	0.96
σ_u^2	1.00	−0.57	0.23	0.24	0.94
ω_1	0.70	−0.67	0.16	0.17	0.95
ω_2	0.40	1.16	0.17	0.18	0.94
ω_{12}	−0.50	−0.46	0.23	0.22	0.97

Simulation study: Results (High entropy, N=2000)

Entropy=0.8, N=2000	TRUE	Bias (%)	SE	SD	Coverage
β_0	-2.00	-0.54	0.12	0.12	0.94
$\beta_1(t)$	1.50	-0.15	0.06	0.06	0.94
$\beta_2(x)$	1.50	-0.11	0.05	0.05	0.94
$\beta_3(x_t)$	-0.50	-0.20	0.03	0.03	0.96
$\tau(C_1 = 1)$	2.50	-0.15	0.09	0.09	0.95
$\tau(C_2 = 1)$	-1.00	0.41	0.09	0.09	0.95
σ_u^2	1.00	-0.11	0.11	0.11	0.95
ω_1	0.70	-0.77	0.08	0.09	0.94
ω_2	0.40	-0.28	0.09	0.09	0.95
ω_{12}	-0.50	-1.24	0.11	0.12	0.93

Simulation study: Results (Low entropy, N=500)

Entropy=0.4, N=500	TRUE	Bias (%)	SE	SD	Coverage
β_0	-2.00	-2.83	0.32	0.43	0.85
$\beta_1(t)$	1.50	0.82	0.11	0.12	0.95
$\beta_2(x)$	1.50	0.48	0.11	0.12	0.94
$\beta_3(x_t)$	-0.50	0.99	0.06	0.06	0.96
$\tau(C_1 = 1)$	2.50	-4.76	0.27	0.32	0.88
$\tau(C_2 = 1)$	-1.00	-1.73	0.34	0.34	0.94
σ_u^2	1.00	27.07	0.36	0.43	0.86
ω_1	0.70	-4.09	0.33	0.55	0.77
ω_2	0.40	-9.30	0.36	0.60	0.78
ω_{12}	-0.50	-11.42	0.52	0.51	0.97

Simulation study: Results (Low entropy, N=2000)

Entropy=0.4, N=2000					
	TRUE	Bias (%)	SE	SD	Coverage
β_0	-2.00	-1.82	0.16	0.19	0.87
$\beta_1(t)$	1.50	0.06	0.06	0.05	0.95
$\beta_2(x)$	1.50	-0.05	0.06	0.06	0.94
$\beta_3(x_t)$	-0.50	0.29	0.03	0.03	0.95
$\tau(C_1 = 1)$	2.50	-1.27	0.12	0.13	0.94
$\tau(C_2 = 1)$	-1.00	1.50	0.16	0.18	0.94
σ_u^2	1.00	5.07	0.17	0.19	0.93
ω_1	0.70	-3.01	0.16	0.25	0.79
ω_2	0.40	0.03	0.17	0.25	0.83
ω_{12}	-0.50	-3.39	0.26	0.26	0.96

A real data example

Comparison of the general 3-step approach with the modal class approach: effects of four dimensions of childhood socio-economic situations on the risk of partnership dissolution.

Table 1: Raw effects of LVs on the log-hazard of partnership dissolution

Categorical LVs	Modal class Est.	(SE)	3-step Est.	(SE)
Fathers social class (ref.=high)				
Low	−0.06	(0.082)	0.13	(0.153)
Medium	0.01	(0.065)	0.04	(0.083)
Financial difficulty (ref. =low)				
High	0.03	(0.077)	0.04	(0.182)
Material hardship (ref.=low)				
Medium	−0.08	(0.062)	−0.04	(0.094)
High	−0.16**	(0.069)	−0.05	(0.083)
Unstable family structure (ref. =stable)				
Unstable	0.29**	(0.083)	0.19*	(0.110)

Next step I: Extension of the 3-step approach to structural equation models

Recall the substantive research question:

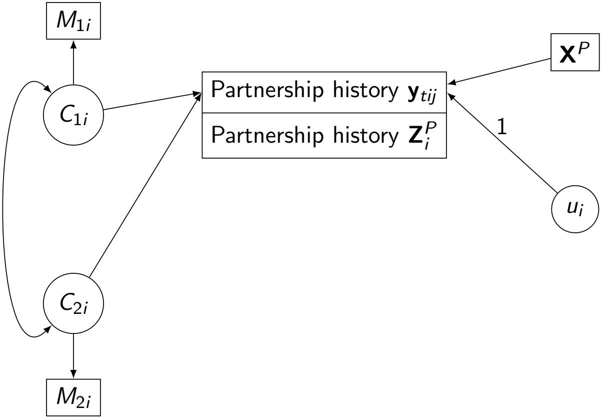

Figure 3: A general path diagram of a multilevel SEM with factorised individual-level random effects.

Next step I: Extension of the 3-step approach to structural equation models

Recall the substantive research question:

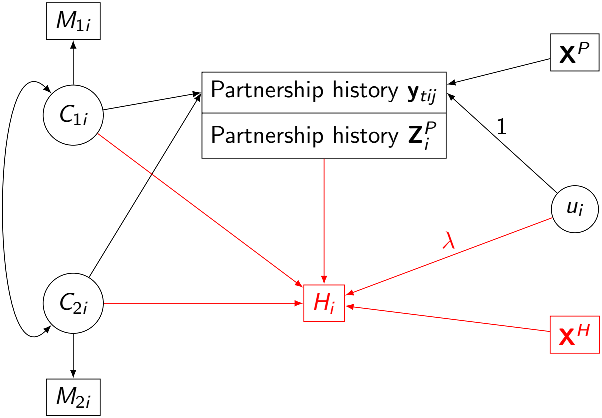

Figure 3: A general path diagram of a multilevel SEM with factorised individual-level random effects.

Next step I: Model specification

Joint modelling of the health and partnership dissolution risk.
Methodology: joint modelling of level-1 and level-2 distal outcomes of mixed types that are predicted by latent categorical variables.

$$\begin{aligned} \text{logit}\big[h_{tij}\big] &= \alpha_t + \alpha_1 C_{1i} + \alpha_2 C_{2i} + \alpha^T \mathbf{X}_{tij}^{(P)} + u_i, \\ \text{logit}\big[P(H_i = 1)\big] &= \beta_0 + \beta_1 C_{1i} + \beta_2 C_{2i} + \beta_3^T \mathbf{X}_i^{(H)} + \beta_4 Z_i^{(P)} + \lambda u_i. \end{aligned} \tag{3}$$

- H_i is health status (binary or ordered), 1=poor health.
- $\mathbf{X}_i^{(H)}$ is a vector of health-relevant covariates.
- $\mathbf{X}_{tij}^{(P)}$ is a vector of predictors of separation hazard.
- $Z_i^{(P)}$ is a summary indicator of partnership stability derived from the partnership history (e.g. the total number of partners during ages 16-50).

Next step II: Advantages

- Joint modelling handles endogeneity of $Z_i^{(P)}$ in the health model.
- Allow for differential effects (λ) of a common set of individual-specific unobservables (u_i) on the hazard of separation and health.
- Ease interpretation: $\lambda > 0 \Rightarrow$ people with certain unobserved time-invariant characteristics that put them in the higher-than-average risk group of divorce also tend to have poor health at age 50.
- Generalisability: can handle data with complex structures (e.g. multilevel, longitudinal, mixed response types); multivariate health outcome $\Rightarrow$ better identification of σ_u^2 (factor model).

Next step III: Limitations & Future work

- Conditional independence assumption on multiple distal outcomes.
- Causal interpretation: Not yet!
- Borrow strength from both the event history literature (e.g. multi-process models, competing-risk models) and the latent variable modelling literature (e.g. different specifications of the structural model).

References

Tihomir Asparouhov and Bengt Muthén. Auxiliary variables in mixture modeling: Three-step approaches using mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(3):329–341, 2014.

David J Bartholomew, Martin Knott, and Irini Moustaki. *Latent variable models and factor analysis: A unified approach*, volume 904. John Wiley & Sons, 2011.

Mary Dupuis Sammel, Louise M Ryan, and Julie M Legler. Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):667–678, 1997.

Jeroen K. Vermunt. Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18(4):450–469, 2010.

Yajing Zhu, Fiona Steele, and Irini Moustaki. A general 3-step maximum likelihood approach to estimate the effects of multiple latent categorical variables on a distal outcome. *Structural Equation Modeling: A Multidisciplinary Journal*, pages 1–14, 2017.

Appendix: More on estimation I

1) Estimation for ω parameters in the log-linear model

Denote by parameter vector $\theta_1 = (\omega_0, \omega_{k_1}^{C_1}, \omega_{k_2}^{C_2}, \omega_{k_1 k_2}^{C_1 C_2})$. The individual contribution to the expected score function of θ_1 can be written:

$$E\left[S_i(\theta_1)\right] = \sum_{k_2=1}^{K_2} \sum_{k_1=1}^{K_1} \int_u S_i(\theta_1) g(\xi_i | \mathbf{y}_{tij}, \mathbf{X}_{tij}, M_{1i}, M_{2i}) du. \tag{4}$$

In the M-step, we need to solve $\sum_{i=1}^{N} E\left[S_i(\theta_1)\right] = 0$. Integrals in (4) can be approximated by, e.g. Monte Carlo methods (Sammel et al., 1997) or Gaussian-Hermite quadratures which replaces the integral with a weighted summation over u_i. 2) Estimation for parameters in the survival model

Appendix: More on estimation II

Denote by parameter vector $\theta_2 = (\alpha_t, \beta, \tau_{k_1}, \tau_{k_2}, \sigma_u)$, the individual contribution to the expected score function of θ_2 is

$$E\left[S_i(\theta_2)\right] = \sum_{k_2=1}^{K_2} \sum_{k_1=1}^{K_1} \int_u S_i(\theta_2) g(\xi_i | \mathbf{y}_{tij}, \mathbf{X}_{tij}, M_{1i}, M_{2i}) du. \qquad (5)$$

Similar to earlier practices, solving $\sum_{i=1}^{N} E\left[S_i(\theta_2)\right] = 0$ requires the approximation of the integral in (5). Higher dimensions of the latent variables (either discrete or continuous) can be computationally expensive. Summary of estimation in Step 3:

Appendix: More on estimation III

1. Generate initial estimates for all parameters (θ_1, θ_2).
2. E-step: compute $E\left[S_i(\theta_1)\right]$ and $E\left[S_i(\theta_2)\right]$ given in (4) and (5).
3. M-step: solve for $\sum_{i=1}^{N} E\left[S_i(\theta_1)\right] = 0$ and $\sum_{i=1}^{N} E\left[S_i(\theta_2)\right] = 0$, update parameter estimates.
4. Repeat steps 2 and 3 until convergence is reached.

Standard errors:
Denote by vector $\theta = (\theta_1, \theta_2)$. To obtain asymptotic standard errors: compute the information matrix $I(\theta)$ using maximum likelihood estimates; take diagonal elements of the inverse of $I(\hat{\theta})$. An alternative: use parametric bootstrap methods that are available in many software packages (Bartholomew et al., 2011).