GLLAMM:

method, bayesian estimation, advantages, and applications to educational data.

José Manuel Rivera Espejo

Master of Science in Statistics and Data Science

KU Leuven

Leuven, August 30, 2021

Document Outline

The document is organized as follows:

1. Preliminary considerations
2. The GLLAMM for dichotomous outcomes
3. Bayesian estimation
4. Simulation study
5. Application
6. Conclusions

1. Preliminary consideration

IRT local independence

Comprised of two parts [1, 12]:

- local item independence
- local individual independence.

IRT models are **not robust** to the violation of local independence [29, 5, 13].

Educational data

often display **several** types of dependencies, violating the local item and/or individual independence, e.g.

- testlets [28];
- the measurement of multiple latent traits within individuals [25];
- cluster effects [24].

Proposed model

The GLLAMM follow a multilevel/hierarchical multidimensional approach to account for different dependencies.

- (**good**) control for dependencies in educational data
- (**important**) reach appropriate conclusion from the parameters

Computational implementation

Under the Bayesian framework, but the integration will be:

- Complex and highly dimensional (in parameters)
- On sparse binary data

but complex parametrizations introduce pathologies that prevent MCMC methods to achieve ergodicity [7, 8, 18, 19, 3], i.e. reach stationarity, convergence, and good mixing [16].

Computational implementation (cont.)

Four solutions are offered to solve the previous pathologies:

- Changing the settings of the MCMC method
  1. increasing the number of iterations per chain, with large burn-in and thinning processes
  2. designing model-specific MCMC algorithms.
- Readjusting the Bayesian model
  3. re-write the model in an alternative parametrization (**simple changes**)
  4. encode prior information through the prior distributions

Computational implementation (cont.)

However, still no simple rotation/rescaling of the parameter, or the amount of data, allow to visit the posterior distribution properly [3].

Multiple authors showed that changing the **posterior sampling geometries**, i.e. removing the dependence of the parameters on other sampled parameters, **improves** the performance of the MCMC methods [7, 8, 18, 19, 3]

2. The GLAMM for dichotomous outcomes

Model definition

Following Rabe-Hesketh et al. [21, 22], we define the GLLAMM in two parts:

1. the response model
2. the latent structure

Moreover, **the response model (1)** can be represented by a Generalized Linear Model (GLM) [17, 15] with:

1. a distributional
2. a systematic part

The response model

Conditional to all parameters $\mathbf{\Omega} = \{\boldsymbol{\beta}, \mathbf{\Lambda}, \mathbf{\Theta}, \mathbf{\Psi}, \mathbf{\Gamma}\}$; and the "stacked" vector of covariates $\mathbf{X}$ and $\mathbf{W}$; **the distributional part** is defined by:

$$f\left(y_{jkd} = 1 \mid \mathbf{X}, \mathbf{W}, \mathbf{\Omega}\right) = \pi_{jkd}^{n}(1 - \pi_{jkd})^{1-n} \qquad (1)$$

Furthermore, **the systematic part** is defined in the following form:

$$P\left(y_{jkd} = 1 \mid \mathbf{X}, \mathbf{W}, \mathbf{\Omega}\right) = \pi_{jkd} = h(\tau_k + \upsilon_{jkd}) \qquad (2)$$

where τ_k is k'th item threshold, assumed to be zero for the binary case [21].

The response model (cont.)

Moreover, the inverse-link function $h(\cdot)$ can be defined in three ways:

$$h(x) = \begin{cases} \exp(x)[1 + \exp(x)]^{-1} \\ \Phi(x) \\ \exp(-\exp(x)) \end{cases} \tag{3}$$

corresponding to the logistic, standard normal $\Phi(x)$, and Gumbel (extreme value type I) cumulative distributions, respectively.

The response model (cont.)

Finally, the linear predictor is defined by:

$$v_{jkd} = \sum_{p=1}^{P} x_{jp}\beta_p + \sum_{m=2}^{M+1}\sum_{k=1}^{K_{(m)}} \eta_k^{(m)} \alpha_k^{(m)} + \sum_{l=2}^{L+1}\sum_{d=1}^{D_{(l)}} \theta_{jd}^{(l)} \lambda_d^{(l)} \tag{4}$$

Which after the appropriate "stacking":

$$\begin{aligned} v_{jkd} &= \mathbf{X}_j \, \boldsymbol{\beta} + \boldsymbol{\eta} \, \boldsymbol{\alpha} \, \mathbf{A}_j + \boldsymbol{\theta} \, \boldsymbol{\lambda} \, \mathbf{B}_j \\ &= \mathbf{X}_j \, \boldsymbol{\beta} + \boldsymbol{\Theta} \, \boldsymbol{\Lambda} \, \mathbf{H}_j \end{aligned} \tag{5}$$

The latent structure

The structural model for the latent variables is represented in the following form:

$$\boldsymbol{\Theta} = \underset{(S\times S)}{\boldsymbol{\Psi}}\underset{(S\times 1)}{\boldsymbol{\Theta}} + \underset{(S\times Q)}{\boldsymbol{\Gamma}}\underset{(Q\times 1)}{\mathbf{W}} + \underset{(S\times 1)}{\boldsymbol{\zeta}} \tag{6}$$

where $S = K + D$, $K = \sum_m K_m$, and $D = \sum_l D_l$.

Notice equation (6) is the generalization of a single-level Structural Equation Models (SEM) to a multilevel setting.

Motivating example

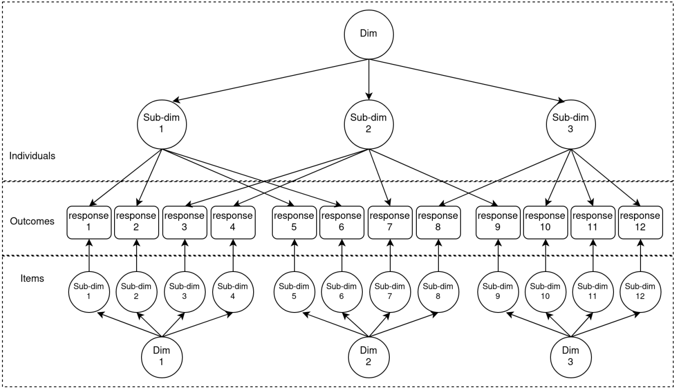

Figure: Path diagram of the dimensional structure for a hierarchical cross-classified IRT model.

Motivating example (cont.)

- Empty level-1 covariates matrix $\mathbf{X}_j$ $(P = 0)$
- $M = 2$ levels at the items block, with $K_2 = 12$ and $K_3 = 3$, i.e. $\boldsymbol{\eta}^{(2)} = [\eta_1^{(2)}, \ldots, \eta_{12}^{(2)}]^T$ and $\boldsymbol{\eta}^{(3)} = [\eta_1^{(3)}, \eta_2^{(3)}, \eta_3^{(3)}]^T$
- $L = 2$ levels in the individuals block, with $D_2 = 3$ and $D_3 = 1$, i.e. $\boldsymbol{\theta}^{(2)} = [\theta_1^{(2)}, \theta_2^{(2)}, \theta_3^{(2)}]^T$ and $\boldsymbol{\theta}^{(3)} = \theta_1^{(3)}$.
- Specific regression relationship among latents $\boldsymbol{\Psi}$, i.e. $\boldsymbol{\alpha}^{(3)} = [\alpha_{11}^{(3)}, \ldots, \alpha_{15}^{(3)}, \alpha_{21}^{(3)}, \ldots, \alpha_{25}^{(3)}, \alpha_{31}^{(3)}, \ldots, \alpha_{35}^{(3)}]^T$ and $\boldsymbol{\lambda}^{(3)} = [\lambda_1^{(3)}, \lambda_2^{(3)}, \lambda_3^{(3)}]^T$
- Empty structural covariates $\mathbf{W}$.

3. Bayesian Estimation

Bayesian GLLAMM for dichotomous outcomes

1. **Posterior distribution.** Given that $\mathbf{Y}$ is the observed data and $\boldsymbol{\Omega} = \{\beta, \boldsymbol{\Lambda}, \boldsymbol{\Theta}, \boldsymbol{\Psi}, \boldsymbol{\Gamma}\}$ the parameters:

$$P(\boldsymbol{\Omega} \mid \mathbf{Y}) = \frac{P(\mathbf{Y} \mid \boldsymbol{\Omega})\ P(\boldsymbol{\Omega})}{\int P(\mathbf{Y} \mid \boldsymbol{\Omega})\ P(\boldsymbol{\Omega})\ d\boldsymbol{\Omega}} \tag{7}$$

2. **Prior distributions.** Similar to Patz and Junker [20], we use an independent distributional structure for the joint priors:

$$\begin{aligned} P(\boldsymbol{\Omega}) =& P(\boldsymbol{\beta})\ [P(\boldsymbol{\alpha})\ P(\boldsymbol{\lambda})]\ [P(\boldsymbol{\eta})\ P(\boldsymbol{\theta})] \\ & [P(\boldsymbol{\Psi}_\eta)\ P(\boldsymbol{\Psi}_\theta)]\ [P(\boldsymbol{\Gamma}_\eta)\ P(\boldsymbol{\Gamma}_\theta)] \end{aligned} \tag{8}$$

Bayesian GLLAMM for dichotomous outcomes (cont.)

3. **Likelihood.** Following Rabe-Hesketh et al. [21], the likelihood function is build in a recursive way.

$$f\left(y_{jkd}=1 \mid \mathbf{X},\mathbf{W},\boldsymbol{\Omega}\right)=\pi_{jkd}^{n}(1-\pi_{jkd})^{1-n} \tag{9}$$

$$f\left(\mathbf{y}=\mathbf{1} \mid \mathbf{X},\mathbf{W},\boldsymbol{\Omega}\right)=\prod_{j=1}^{J}\prod_{d=1}^{D}\prod_{k=1}^{K} f\left(y_{jkd}=1 \mid \mathbf{X},\mathbf{W},\boldsymbol{\Omega}\right) \tag{10}$$

$$f_{(m)}^{(l)}\left(\mathbf{y}=\mathbf{1} \mid \mathbf{X},\mathbf{W},\boldsymbol{\Omega}\right)=\int\left[\prod f_{(m-1)}^{(l-1)}\left(\mathbf{y}=\mathbf{1} \mid \mathbf{X},\mathbf{W},\boldsymbol{\Omega}\right)\right] P(\boldsymbol{\Theta}_{(m)}^{(l)} \tag{11}$$

$$\mathcal{L}(\mathbf{X},\mathbf{W},\boldsymbol{\Omega})=\prod_{m=2}^{M+1}\prod_{l=2}^{L+1} f_{(m)}^{(l)}\left(\mathbf{y}=\mathbf{1} \mid \mathbf{X},\mathbf{W},\boldsymbol{\Omega}\right) \tag{12}$$

$$\ell(\mathbf{X},\mathbf{W},\boldsymbol{\Omega})=\log\mathcal{L}(\mathbf{X},\mathbf{W},\boldsymbol{\Omega}) \tag{13}$$

Computational implementation

1. With Hamiltonian Monte Carlo (HMC) and `Stan` [27].
2. **No burn-in and thinning**. However, a warm-up phase is required to "tune-up" the number of steps (`leapfrogs`), and the `step size` [27].
3. We will use a total $3,000$ **effective iterations**, coming from 3 chains of $2,000$ iterations each, where $1,000$ of them will be spend on warm-up.
4. **Initial starts** sampled from the priors defined in the model.
5. Prior distributions selected based on **prior predictive simulations**.

To center or not to center

Even the most simple hierarchical models present formidable pathologies, that no simple rotation/rescaling of the parameter can be performed to visit the posterior distribution properly [3].

Example, the devil's funnel [16]:

$$\begin{aligned} v &\sim N(0,3) \\ \theta &\sim N(0,\exp(v)) \end{aligned} \tag{14}$$

Equation (14) describes a centered parametrization (CP)

To center or not to center (cont.)

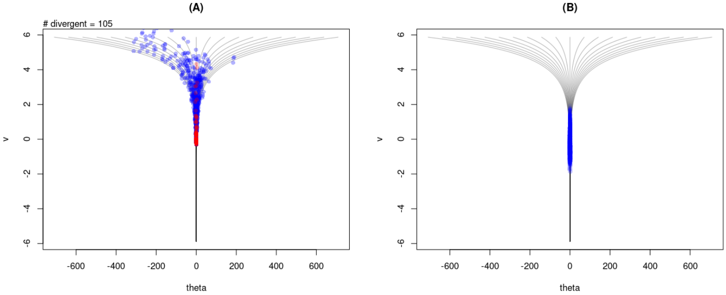

Figure: Posterior sampling geometry. Centered Parametrization.

To center or not to center (cont.)

How can we solve this?:

1. Adapt HMC warm-up (`adapt_delta`$= 0.99$).
2. Use regularizing priors
3. Use the **non-centered parametrization**.

Non-centered parametrization (NCP)

Changing the posterior sampling geometry means to modify equation (14) in the following way:

$$
\begin{aligned}
v &\sim N(0,3) \\
z &\sim N(0,1) \\
\theta &= \exp(v)\ z
\end{aligned}
\tag{15}
$$

Non-centered parametrization (cont.)

Changing the posterior sampling geometry means to modify equation (14) in the following way:

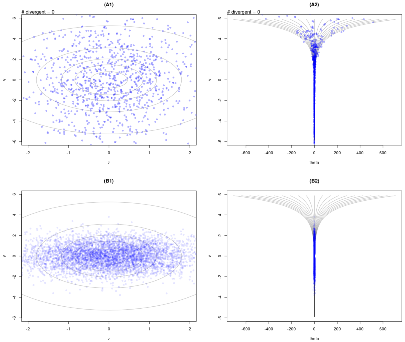

Figure: Posterior sampling geometry. Non-Centered Parametrization.

4. Simulation studies

Objectives

Designed to assess three attributes of the bayesian implementation of the GLLAMM for dichotomous outcomes:

1. **Performance.** In terms of achieving ergodicity, under the CP and NCP,
2. **Recovery capacity.** Capacity to recover the parameters of interest, especially the structural regression parameters.
3. **Retrodictive accuracy.** Capacity to retrodict the data of interest, according to a set of aggregating dimensions.

Evaluation criteria

1. **Performance.** Trace, trank and ACF plots with support of `Rhat` and `n_eff` statistics developed by Gelman et al. [9] (pp. $284 - 287$).
2. **Recovery capacity.** we used the between replica root mean squared error (RMSE_B).
3. **Retrodictive accuracy.** we used the average within $\overline{\text{RMSE}}_W$ and between prediction root mean squared error RMSE_B, of the responses' predictive proportion $\hat{p}$, versus the observed proportion p.

Results (Performance)

1. **Performance.** In terms of achieving ergodicity, under the CP and NCP,
2. **Recovery capacity.** Capacity to recover the parameters of interest, especially the structural regression parameters.
3. **Retrodictive accuracy.** Capacity to retrodict the data of interest, according to a set of aggregating dimensions.

Results
(Recovery capacity)

1. **Performance.** In terms of achieving ergodicity, under the CP and NCP,
2. **Recovery capacity.** Capacity to recover the parameters of interest, especially the structural regression parameters.
3. **Retrodictive accuracy.** Capacity to retrodict the data of interest, according to a set of aggregating dimensions.

Results (Retrodictive accuracy)

1. **Performance.** In terms of achieving ergodicity, under the CP and NCP,
2. **Recovery capacity.** Capacity to recover the parameters of interest, especially the structural regression parameters.
3. **Retrodictive accuracy.** Capacity to retrodict the data of interest, according to a set of aggregating dimensions.

5. Application

Specific goals

The research have two main goals:

- to describe the method, estimation procedures, and advantages of the GLLAMM framework, and
- to tests the policy implications of the method and its results, in a data composed of large repeated teacher's standardized educational assessments from Peru.

6. Conclusions and further development

Specific goals

The research have two main goals:

- to describe the method, estimation procedures, and advantages of the GLLAMM framework, and
- to tests the policy implications of the method and its results, in a data composed of large repeated teacher's standardized educational assessments from Peru.

Appendix

Cluster effects

Individual clustering involves the addition of more random effects to the linear predictor defined in equation (4):

$$\begin{aligned} \upsilon_{jkdc} &= \upsilon_{jkd} + \sum_{c=1}^{C} \delta_c \\ &= \upsilon_{jkd} + \boldsymbol{\delta}\mathbf{Z}_j \end{aligned} \tag{16}$$

Example (restrictions for IRT)

we could set the restriction $\boldsymbol{\alpha}^{(2)} = -\boldsymbol{\lambda}^{(2)}$ where $\boldsymbol{\lambda}^{(2)} > \mathbf{0}$. In that case we get a multidimensional generalization of the linear predictor observed in the archetypical Rasch [23], or 2PL [14] models, i.e. $\lambda_d^{(2)}(\theta_{jd}^{(2)} - \eta_k^{(2)})$

In addition,
$\boldsymbol{\alpha}^{(3)} = [\alpha_{11}^{(3)}, \ldots, \alpha_{15}^{(3)}, \alpha_{21}^{(3)}, \ldots, \alpha_{25}^{(3)}, , \alpha_{31}^{(3)}, \ldots, \alpha_{35}^{(3)}]^T$
$= [1, \ldots, 1]^T$, indicating texts difficulties explain directly the items difficulties at the lower level.

Example (restrictions for IRT)

Moreover, notice that because in the IRT framework $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ should be orthogonal to each other by design, we can further decompose equation (6) in the following form:

$$\boldsymbol{\eta} = \underset{(K\times K)}{\boldsymbol{\Psi}_\eta}\underset{(K\times 1)}{\boldsymbol{\eta}} + \underset{(K\times Q)}{\boldsymbol{\Gamma}_\eta}\underset{(Q\times 1)}{\mathbf{W}_\eta} + \underset{(K\times 1)}{\boldsymbol{\zeta}_\eta} \tag{17}$$

$$\boldsymbol{\theta} = \underset{(D\times S)}{\boldsymbol{\Psi}_\theta}\underset{(D\times 1)}{\boldsymbol{\theta}} + \underset{(D\times Q)}{\boldsymbol{\Gamma}_\theta}\underset{(Q\times 1)}{\mathbf{W}_\theta} + \underset{(D\times 1)}{\boldsymbol{\zeta}_\theta} \tag{18}$$

Model Assumptions

Following Skrondal and Rabe-Hesketh [26], the framework has two main assumptions:

(M1) **Complete latent space.**[11] In the GLLAMM representation, the space is complete if we consider all latent variables $\boldsymbol{\Theta}$ at levels $l > 1$ and $m > 1$.

(M2) **Local Independence.** It assumes independence conditional on all the latent dimensions and covariates, at different hierarchical levels; effectively modeling all the observed dependencies.

Model Assumptions (cont.)

Local Independence is defined as follows:

$$f\left(\mathbf{y}=\mathbf{1} \mid \mathbf{X}, \mathbf{W}, \boldsymbol{\Omega}\right)=\prod_{j=1}^{J} \prod_{d=1}^{D} \prod_{k=1}^{K} f\left(y_{jkd}=1 \mid \mathbf{X}, \mathbf{W}, \boldsymbol{\Omega}\right) \quad (19)$$

Model Assumptions (cont.)

but comes from:

1. **Local item independence,**

$$f\left(y_{j..}=1 \mid \mathbf{X}, \mathbf{W}, \boldsymbol{\Omega}\right)=\prod_{d=1}^{D} \prod_{k=1}^{K} f\left(y_{jkd}=1 \mid \mathbf{X}, \mathbf{W}, \boldsymbol{\Omega}\right) \tag{20}$$

2. **Local individual independence,**

$$f\left(y_{.kd}=1 \mid \mathbf{X}, \mathbf{W}, \boldsymbol{\Omega}\right)=\prod_{j=1}^{J} f\left(y_{jkd}=1 \mid \mathbf{X}, \mathbf{W}, \boldsymbol{\Omega}\right) \tag{21}$$

Why bayesian?

- It is built on a simulation-based estimation method, therefore, it can handle all kinds of priors and data-generating processes [6].
- While the likelihood for the data and priors for the parameters are used to define the posterior sampling distributions, they can also be used in a generative way [16].
- Because the procedure integrates prior knowledge about the parameters, it can produce results even in scenarios where the Maximum Likelihood methods (ML) have issues of non-convergence or improper estimation [26, 6, 16]

There is nothing wrong?

- It exposes the user to somewhat-arbitrary decisions about the running of the chains, in order to ensure a proper performance (**solution:** Hamiltonian Monte Carlo (HMC) [3]).
- The user can include all type of information through the priors distributions, making their elicitation convenient for manipulation (**solution:** prior predictive simulations and/or sensitivity analysis).
- Visual evaluation of performance, making it hard to assess if a proper posterior investigation have been made [10] (**solution:** help of `Rhat`, `n_eff`, and change of posterior sampling geometry).

There is nothing wrong? (cont.)

- The procedure makes it hard to discover parameters' lack of identification [26] (**solution:** regularizing priors).
- Oftentimes the posterior sampling geometry of the model makes it hard to find proper solutions for the parameter space [3] (**solution:** change the posterior sampling geometry).
- The greater the complexity of the model, the harder it is to communicate/share and takes more time (**solution:** No solution, but it is a small "price" to pay).

There is nothing wrong? (cont.)

- The procedure makes it hard to discover parameters' lack of identification [26] (**solution:** regularizing priors).
- Oftentimes the posterior sampling geometry of the model makes it hard to find proper solutions for the parameter space [3] (**solution:** change the posterior sampling geometry).
- The greater the complexity of the model, the harder it is to communicate/share and takes more time (**solution:** No solution, but it is a small "price" to pay).

Prior elicitation

Un-informative priors are not the solution:

1. Uninformative / weakly-informative latent prior:

$$\begin{aligned} \theta &\sim N(0,100) & \theta &\sim N(0,1) \\ \text{logit}(p) &= \theta & \text{logit}(p) &= \theta \end{aligned}$$

2. Uninformative / weakly-informative hierarchical latent prior:

$$\begin{aligned} v &\sim \log N(0,3) & v &\sim \log N(0,0.5) \\ \theta &\sim N(0,v) & \theta &\sim N(0,v) \\ \text{logit}(p) &= \theta & \text{logit}(p) &= \theta \end{aligned}$$

Prior elicitation (cont.)

Un-informative priors are not the solution:

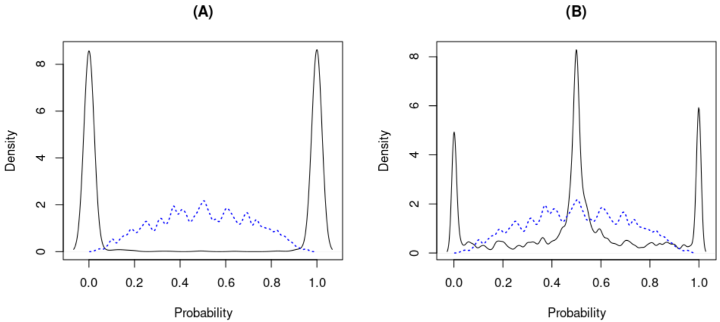

Figure: Prior predictive simulation. Examples of uninformative and mildly informative priors.

The benefits of using regularizing priors

Consider the use of a regularizing prior on equation (14):

$$\begin{aligned} v &\sim N(0,1) \\ \theta &\sim N(0, \exp(v)) \end{aligned} \tag{22}$$

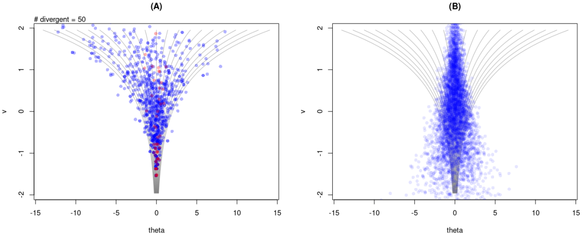

Figure: Posterior sampling geometry. Centered Parametrization with mildly informative priors.

Simulation study design

A $3 \times 2 \times 2$ fractional factorial design:

- Three different samples sizes to generate the data under analysis: 500, 250, and 100.
- Two parametrization of the models: CP and NCP.
- Two models of interest: the first- and second-order latent variable model.

Ten (10) data sets were generated for each study condition. Each data set resembled responses to 25 binary scored items, conforming to the SOLV model defined in figure 7. The model was motivated by the hypothesized structure of the reading comprehension sub-test, from the Peruvian public teaching career national assessment

Simulation study design (cont.)

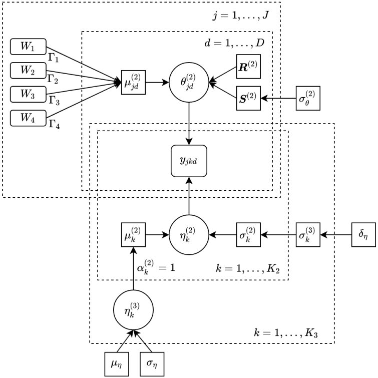

Figure: Directed Acyclig Graph (DAG). First-order latent variable model (SOLV).

Simulation study design (cont.)

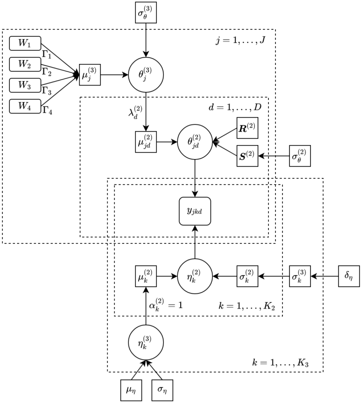

Figure: Directed Acyclig Graph (DAG). Second-order latent variable model (SOLV).

Likelihood, priors and hyper-priors (centered parametrization)

$$y_{jkd} \sim \text{Bernoulli}(\pi_{jkd}) \tag{23}$$

$$\text{logit}(\pi_{jkd}) = \upsilon_{jkd} \tag{24}$$

$$\upsilon_{jkd} = \theta_{jd}^{(2)} - \eta_{k}^{(2)} \tag{25}$$

$$\boldsymbol{\theta}_{j}^{(2)} = \left[\theta_{j1}^{(2)}, \theta_{j2}^{(2)}, \theta_{j3}^{(2)} \right] \tag{26}$$

$$\boldsymbol{\theta}_{j}^{(2)} \sim \text{MVNormal}\left(\boldsymbol{\mu}_{j}^{(2)}, \boldsymbol{\Sigma}^{(2)} \right) \tag{27}$$

$$\boldsymbol{\Sigma}^{(2)} = \boldsymbol{S}^{(2)} \cdot \boldsymbol{R}^{(2)} \cdot \boldsymbol{S}^{(2)} \tag{28}$$

$$\boldsymbol{S}^{(2)} = \boldsymbol{\sigma}_{\theta}^{(2)} \mathbf{I} \tag{29}$$

Likelihood, priors and hyper-priors (centered parametrization, cont.)

For the FOLV model:

$$\boldsymbol{\mu}_j^{(2)} = \left[\mu_{j1}^{(2)},\ \mu_{j2}^{(2)}, \mu_{j3}^{(2)}\right] \tag{30}$$

$$\mu_{jd}^{(2)} = \Gamma_0 + \Gamma_1 W_{1j} + \Gamma_2 (W_{2j} - W_{2\text{min}}) + \Gamma_3 W_{3j} + \Gamma_4 W_{4j} \tag{31}$$

Likelihood, priors and hyper-priors (centered parametrization, cont.)

For the SOLV model:

$$\boldsymbol{\mu}_j^{(2)} = \left[\mu_{j1}^{(2)},\ \mu_{j2}^{(2)}, \mu_{j3}^{(2)}\right] \tag{32}$$

$$\boldsymbol{\lambda}^{(2)} = \left[\lambda_1^{(2)},\ \lambda_2^{(2)}, \lambda_3^{(2)}\right] \tag{33}$$

$$\mu_{jd}^{(2)} = \lambda_d^{(2)}\ \theta_j^{(3)} \tag{34}$$

$$\theta_j^{(3)} \sim \text{Normal}\left(\mu_j^{(3)}, \sigma_\theta^{(3)}\right) \tag{35}$$

$$\mu_j^{(3)} = \Gamma_0 + \Gamma_1 W_{1j} + \Gamma_2(W_{2j} - W_{2\text{min}}) + \Gamma_3 W_{3j} + \Gamma_4 W_{4j} \tag{36}$$

Likelihood, priors and hyper-priors (centered parametrization, cont.)

For the items:

$$\eta_k^{(2)} \sim \text{Normal}\left(\mu_k^{(2)}, \sigma_k^{(2)}\right) \quad (37)$$

$$\mu_k^{(2)} = \boldsymbol{\eta}^{(3)}\mathbf{A} \quad (38)$$

$$\sigma_k^{(2)} = \boldsymbol{\sigma}^{(3)}\mathbf{A} \quad (39)$$

$$\boldsymbol{\eta}^{(3)} = [\eta_1^{(3)}, \eta_2^{(3)}, \eta_3^{(3)}, \eta_4^{(3)}, \eta_5^{(3)}] \quad (40)$$

$$\boldsymbol{\sigma}^{(3)} = [\sigma_1^{(3)}, \sigma_2^{(3)}, \sigma_3^{(3)}, \sigma_4^{(3)}, \sigma_5^{(3)}] \quad (41)$$

$$\eta_k^{(3)} \sim \text{Normal}\left(\mu_\eta, \sigma_\eta\right) \quad (42)$$

$$\sigma_k^{(3)} \sim \text{Exponential}\left(\delta_\eta\right) \quad (43)$$

Likelihood, priors and hyper-priors (centered parametrization, cont.)

Remaining priors and hyper-priors

$$\boldsymbol{R}^{(2)} \sim \text{LkjCorrelation}(2) \tag{44}$$
$$\Gamma_{1c} \sim \text{Normal}(0, 0.5) \tag{45}$$
$$\Gamma_{2} \sim \text{Normal}(0, 0.5) \tag{46}$$
$$\Gamma_{3c} \sim \text{Normal}(0, 1) \tag{47}$$
$$\Gamma_{4c} \sim \text{Normal}(0, 0.5) \tag{48}$$

Likelihood, priors and hyper-priors (Non-centered parametrization)

Under the NCP, equation (37) was re-defined as follows:

$$\eta_k^{(2)} = \mu_k^{(2)} + \sigma_k^{(2)} z_k^{(2)} \tag{49}$$

$$z_k^{(2)} \sim \text{Normal}(0,1) \tag{50}$$

Equation (27) was re-defined as follows:

$$\boldsymbol{\theta}_j^{(2)} = \boldsymbol{\mu}_j^{(2)} + \boldsymbol{S}^{(2)} \cdot \boldsymbol{L}_\Sigma^{(2)} \cdot (\mathbf{z}_j \mathbf{I}) \tag{51}$$

$$\mathbf{z}_j = [z_{j1}, \ldots, z_{jd}]^T \tag{52}$$

$$z_{jd} \sim \text{Normal}(0,1) \tag{53}$$

$$\boldsymbol{L}_\Sigma^{(2)} \sim \text{LKJCorrelationCholesky}(2) \tag{54}$$

Likelihood, priors and hyper-priors (Non-centered parametrization, cont.)

Finally, equation (35) was re-defined as follows:

$$\theta_j^{(3)} = \mu_j^{(3)} + \sigma_\theta^{(3)} z_j \tag{55}$$

$$z_j \sim \text{Normal}(0, 1) \tag{56}$$

Identification

We used the unit variance identification scheme (UVI), that is, to set the scale of the higher-order dimension and sub-dimensions to one:

- $\sigma_\theta^{(3)} = 1$
- $\mathbf{S}^{(2)} = \boldsymbol{\sigma}_\theta^{(2)}\mathbf{I}$ with $\boldsymbol{\sigma}_\theta^{(2)} = [1, 1, 1]^T$
- $\mu_\eta = 0,\ \sigma_\eta = 1,\ \delta_\eta = 2$

The first two turned the covariance matrix into a correlation, i.e. $\boldsymbol{\Sigma}^{(2)} = \boldsymbol{R}^{(2)}$.

Prior predictive investigation

From two perspectives:

- the IRT perspective
- the outcome perspective

Prior predictive investigation (cont.)

From the IRT perspective:

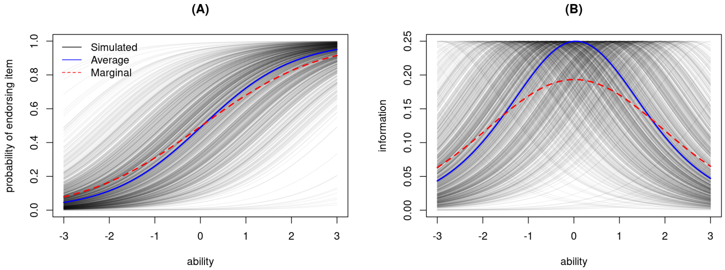

Figure: First-order latent variable model (FOLV). (A) Item Characteristics Curve, ICC. (B) Item Information Function, IIF.

From the outcome perspective:

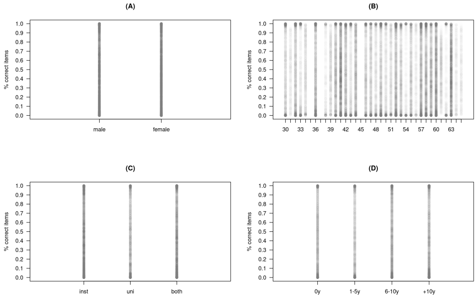

Figure: First-order latent variable model (FOLV). Aggregated endorsement rate per simulated covariate: (A) gender, (B) age, (C) education, and (D) experience.

References I

[1] Baker, F. [2001]. The basic of item response theory, *Technical report*, ERIC Clearinghouse on Assessment and Evaluation.

[2] Baker, F. and Kim, S. [1992]. *Item Response Theory: Parameter Estimation Techniques*, Statistics for the Social and Behavioral Sciences, CRC Press, Taylor and Francis Group. **doi:** https://www.doi.org/10.1201/9781482276725.

[3] Betancourt, M. and Girolami, M. [2012]. Hamiltonian monte carlo for hierarchical models. **url:** arxiv.org/abs/1312.0906v1.

References II

[4] Bock, R. [1972]. Estimating item parameters and latent ability when responses are scored in two or more nominal categories, *Psychometrika* **37**(1). **doi:** https://doi.org/10.1007/BF02291411.

[5] Chen, W. and Thissen, D. [1997]. Local dependence indexes for item pairs using item response theory, *Journal of Educational and Behavioral Statistics* **22**(3): 265–289. **doi:** https://doi.org/10.3102/10769986022003265.

[6] Fox, J. [2010]. *Bayesian Item Response Modeling, Theory and Applications*, Statistics for Social and Behavioral Sciences, fienberg, s. and van der linden, w. edn, Springer Science+Business Media, LLC.

References III

[7] Gelfand, A., Sahu, S. and Carlin, B. [1995]. Efficient parametrisations for normal linear mixed models, *Biometrika* **82**(3): 479–488.
doi: https://doi.org/10.1093/biomet/82.3.479.

[8] Gelfand, A., Sahu, S. and Carlin, B. [1996]. Efficient parameterizations for generalised linear models (with discussion), *in* J. Bernardo, J. Berger, A. Dawid and a. Smith (eds), *Bayesian Statistics*, Vol. 5, pp. 165–180.

[9] Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. [2014]. *Bayesian Data Analysis*, Texts in Statistical Science, third edn, Chapman and Hall/CRC.

References IV

[10] Gelman, A. and Rubin, D. [1996]. Markov chain monte carlo methods in biostatistics, *Statistical Methods in Medical Research* **5**(4): 339–355. **doi:** https://doi.org/10.1177/096228029600500402.

[11] Hambleton, R. and Swaminathan, H. [1991]. *Item Response Theory*, Evaluation in Education and Human Services series, Springer Science+Business Media, LLC.

[12] Hambleton, R., Swaminathan, H. and Rogers, H. [1991]. *Fundamentals of Item Response Theory*, SAGE Publications Inc.

References V

[13] Jiao, H., Kamata, A., Wang, S. and Jin, Y. [2012]. A multilevel testlet model for dual local dependence, *Journal of Educational Measurement* **49**(1): 82–100. **doi:** https://doi.org/10.1111/j.1745-3984.2011.00161.x.

[14] Lord, F. and Novik, M. [2008]. *Statistical Theories of Mental Test Scores*, Information Age Publishing.

[15] McCullagh, P. and Nelder, J. [1989]. *Generalized Linear Models*, Monographs on Statistics Applied Probability, Chapman Hall/CRC Press.

References VI

[16] McElreath, R. [2020]. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, Texts in Statistical Science, 2 edn, Chapman and Hall/CRC.
doi: https://doi.org/10.1201/9780429029608.

[17] Nelder, J. and Wedderburn, W. [1972]. Generalized linear models, *Royal Statistical Society* **135**(3): 370–384.
doi: https://doi.org/10.2307/2344614.
url: https://www.jstor.org/stable/2344614.

References VII

[18] Papaspiliopoulos, O., Roberts, G. and Skold, M. [2003]. Non-centered parameterisations for hierarchical models and data augmentation, *Bayesian Statistics* **7**: 307–326.
url: http://econ.upf.edu/ omiros/papers/val7.pdf.

[19] Papaspiliopoulos, O., Roberts, G. and Skold, M. [2007]. A general framework for the parametrization of hierarchical models, *Statistical Science* **22**(1): 59–73.
doi: https://www.doi.org/10.1214/088342307000000014.

References VIII

[20] Patz, R. J. and Junker, B. W. [1999]. A straightforward approach to markov chain monte carlo methods for item response models, *Journal of Educational and Behavioral Statistics* **24**(2): 146–178. **doi:** 10.3102/10769986024002146.

[21] Rabe-Hesketh, S., Skrondal, A. and Pickles, A. [2004a]. Generalized multilevel structural equation modeling, *Psychometrika* **69**(2): 167–190. **doi:** https://www.doi.org/10.1007/BF02295939.

References IX

[22] Rabe-Hesketh, S., Skrondal, A. and Pickles, A. [2004b]. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects, *Journal of Econometrics* **128**(2): 301–323.
doi: https://www.doi.org/10.1016/j.jeconom.2004.08.017.
url: http://www.sciencedirect.com/science/article/pii/S030

[23] Rasch, G. [1980]. *Probabilistic Models for Some Intelligence and Attainment Tests*, University of Chicago Press.

References X

[24] Raudenbush, S. and Bryk, A. [2002]. *Hierarchical linear models: Applications and data analysis methods (Vol. 1)*, Advanced Quantitative Techniques in the Social Sciences, SAGE Publications Inc.

[25] Reckase, M. [2009]. *Multidimensional Item Response Theory*, Statistics for Social and Behavioral Sciences, Springer Science+Business Media, LLC.

References XI

[26] Skrondal, A. and Rabe-Hesketh, S. [2004]. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Interdisciplinary Statistics, Chapman Hall/CRC Press.

[27] Stan Development Team. [2021]. *Stan Modeling Language Users Guide and Reference Manual, version 2.26*, Vienna, Austria.
url: https://mc-stan.org.

[28] Wainer, H., Bradlow, E. and Wang, X. [2007]. *Testlet response theory and its applications*, Cambridge University Press.

References XII

[29] Yen, W. [1984]. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model, *Applied Psychological Measurement* **8**(2): 125–145. **doi:** https://doi.org/10.1177/014662168400800201.