

A variational Bayesian posterior approximation method for hidden Markov diagnosis classification models

Kazuhiro Yamaguchi¹

Alfonso J. Martinez²

¹University of Tsukuba,

²University of Iowa

E-mail: yamaguchi.kazuhir.ft@u.tsukubai.ac.jp

This work was supported by a JSPS Grant-in-Aid for JSPS Research Fellow 18J01312 and JSPS KAKANHI 19H00616, 20H01720, and 21H00936.

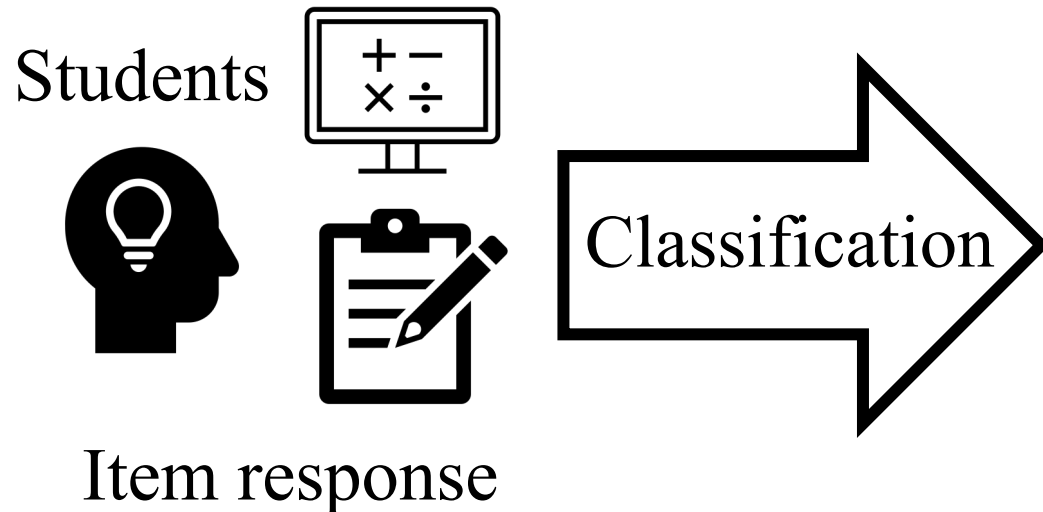
February 2nd, 2022 IASC-ARTS 2022

Table of Contents

1. Background
 1. Diagnostic classification models (DCMs)
 2. Longitudinal versions of DCMs and Bayesian estimation methods
 3. Objectives of this study
2. Formulation of Hidden Markov DCMs (HM-DCMs)
3. The VB method for HM-DCMs
4. Simulation study
5. Real data analysis example
6. Conclusion

Diagnostic Classification Models (DCMs)

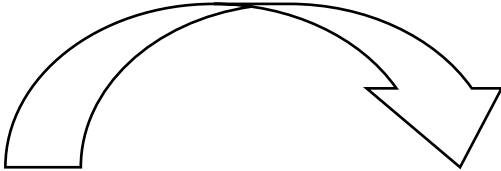
- DCMs are a special case of latent class models (Rupp & Templin, 2008).
 - The **cognitive components** are called “**attributes**.”
 - An attribute profile is a permutation of mastery and non-mastery of attributes.
- A main goal of DCMs is to classify students into one of attribute profiles.



Attribute profile		Attribute 1	
		Master	Non-master
Attribute 2	Master	✓	—
	Non-master	—	—

Tracking learning trajectories: Longitudinal DCMs

- We want to capture **changes in students' attribute mastery profile.**

Before learning  After learning

Attribute profile		Attribute 1	
		Master	Non-master
Attribute 2	Master	—	—
	Non-master	✓	—

Attribute profile		Attribute 1	
		Master	Non-master
Attribute 2	Master	✓	—
	Non-master	—	—

Longitudinal DCMs in the Literature

- Hidden Markov (HM; e.g., Beal, 2003) type models
 - Li et al. (2016): Latent transition model (LTA; Collins & Lanza, 2009) based DINA model
 - Kaya and Leite (2017): LTA-DINA and DINO (Templin & Henson, 2006)
 - Chen et al. (2017): Fully Bayesian estimation method for HM-DINA.
 - Madison and Bradshaw (2018): Transition version of LCDM (Henson et al., 2009) and ML estimation with Mplus
 - Wang (2021): A penalized EM for HM-DCMs
 - Wang et al. (2018a, b), Chen et al. (2017): HM-DINA for mental rotation skill test
 - Hung and Huang (2019): LTA-G-DINA
- Higher order proficiency type models
 - Huang (2017), Pan et al. (2020): A multilevel G-DINA (de la Torre, 2011)
 - Zhan et al. (2019a, b): Longitudinal version of DINA (Junker & Sijtsma, 2001)
 - Zhan (2021): Longitudinal version of probabilistic type DCM
 - Zhan et al. (2020): Higher order latent trait + attribute hierarchy situation
 - Lin et al. (2020): Growth curve approach for higher order trait
- In this study, **we focus on the HM type model**
 - **because DCMs are essentially a sub-model of the latent class model.**

Problems with Hidden Markov model in DCMs

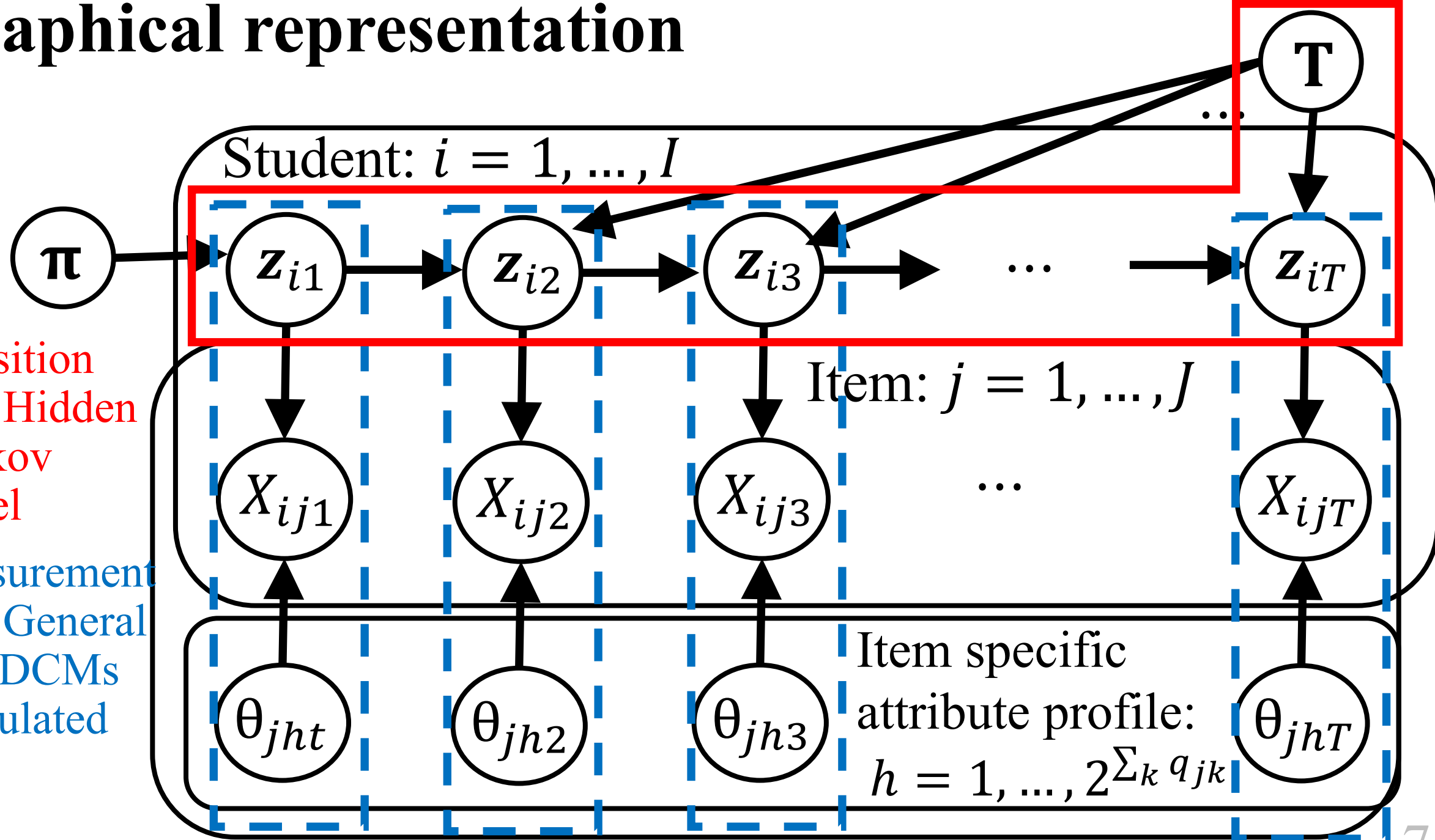
- HMM can become singular (Watanabe, 2018).
 - Maximum likelihood (ML) method is not appropriate for singular models.
 - Bayesian methods are appropriate for HMM.
- However, Bayesian estimation method for HMM with general DCMs have not been developed.
- Moreover, Markov chain Monte Carlo (MCMC) methods are computationally expensive in large sample and parameters settings.
- Therefore, a computationally efficient estimation method that utilizes a fully Bayesian framework is warranted.
- We employed variational Bayesian method (VB).
 - E.g., Yamaguchi & Okada (2020, 2021)

Objectives of this study:

- Constructing HM-DCMs in latent class formulation to derive a VB algorithm
- Developing VB algorithm for the HM-DCMs.
- Assessing parameter recovery of the proposed VB method.
- Applying this method for a real data example and compare MCMC method for more constrained model.

Graphical representation

- Transition part: Hidden Markov model
- Measurement part: General type DCMs formulated



Latent class formulation

- X_{ijt} : A random variable representing the j th ($j = 1, \dots, J_t$) item response of individual i ($i = 1, \dots, I$) at time t ($t = 1, \dots, T$) taking on a value of 1 for a correct response and 0 for an incorrect response
- x_{ijt} : A realization of X_{ijt}
- θ_{jlt} : The j th item correct response probability of an attribute profile l ($l = 1, \dots, L$) at time t
- z_{itl} : l th element of vector \mathbf{z}_{it} that is latent class indicator variable of individual i at time t

$$P(X_{ijt} = x_{ijt} | \mathbf{z}_{it}, \boldsymbol{\theta}_{jt}) = \prod_l \left\{ \theta_{jlt}^{x_{ijt}} (1 - \theta_{jlt})^{1-x_{ijt}} \right\}^{z_{itl}}$$

Q-matrix

- Q-matrix: $J \times K$ matrix with binary elements.
- The j th row of Q is $\mathbf{q}_j = [q_{j1}, \dots, q_{jK}]$ and its elements q_{jk} is 1 if item j measures attribute k and 0 otherwise.
- \rightarrow The \mathbf{q}_j s generate some constraints on the item response functions.

Item	\mathbf{q}_j	Attribute		
		1	2	3
1	\mathbf{q}_1	1	0	0
2	\mathbf{q}_2	0	1	0
3	\mathbf{q}_3	0	0	1
4	\mathbf{q}_4	1	1	0
5	\mathbf{q}_5	0	1	1
6	\mathbf{q}_6	1	0	1

Which patterns have the same probability?

Probability of correct response		k	1	0	1	0	0	1	1	0	1	Attribute profile α		
			2	0	0	1	0	1	0	1	1			
P($X = 1$)			3	0	0	0	1	0	1	1	1			
												The number of profile $2^3 = 8 = L$		
				1	2	3	4	5	6	7	8			
				1	2	3								
θ_{jh}	h	k	l	1	2	3								
				1	*	0	0	1	1	0	0	0	0	0
				2	*	0	1	0	0	1	0	0	0	0
				3	*	1	0	0	0	0	1	0	0	0
θ_{j4}	4	*	1	1	0	0	0	0	0	1	1	Restricting matrix G_j		
				0	1	1								
				The number of item specific profile $2^{\sum_k q_{jk}} = 2^2 = 4$				Item specific attribute profile (Distinguishable profile)						

Formulation of measurement part

- g_{jhtl} : Element of G-matrix of item j at time t .
 - An item specific attribute profile: α_{hj}^* and h ($h = 1 \dots, 2^{\sum_k q_{jk}}$).
 - For $\mathbf{q}_j = [1,1,0]$, four patterns of α_{hj}^* , $[0,0,*]$, $[0,1,*]$, $[1,0,*]$, and $[1,1,*]$, where “*” indicates the element can be ignored without loss of information.
 - Then, the element of the G-matrix becomes 1 if α_{hj}^* is the same with α_l except $\{k; q_{jk} = 0\}$ elements, otherwise 0.
- θ_{jht} : the j th item correct item response probability for item specific attribute profile h ($h = 1, \dots, 2^{\sum_k q_{jk}}$) at time t
- The IRF of the general DCM is:

$$P(X_{ijt} = x_{ijt} | \mathbf{z}_{it}, \boldsymbol{\theta}_{jt}) = \prod_l \prod_h \left\{ \left[\theta_{jht}^{x_{ijt}} (1 - \theta_{jht})^{1-x_{ijt}} \right]^{g_{jhtl}} \right\}^{z_{itl}}$$

Formulation of transition part: First order hidden Markov model

- Let $\tau_{ll'}$ be a transition probability from profile l to profile l'
($l, l' = 1, \dots, L$)
- Assumption: $\tau_{ll'} \geq 0$ and $\sum_{l'} \tau_{ll'} = 1$.

$$P(\mathbf{z}_{it} | \mathbf{z}_{i,t-1}, \mathbf{T}) = \prod_l \prod_{l'} \tau_{ll'}^{z_{i(t-1)} l z_{it} l'}$$

T		Attribute profile (l')				Sum
		Time t				
		00	01	10	11	
Attribute profile (l) Time $t - 1$	00	τ_{11}	τ_{12}	τ_{13}	τ_{14}	1
	01	τ_{21}	τ_{22}	τ_{23}	τ_{24}	1
	10	τ_{31}	τ_{32}	τ_{33}	τ_{34}	1
	11	τ_{41}	τ_{42}	τ_{43}	τ_{44}	1

Initial condition

- An initial attribute mastery profile for student i denoted $\mathbf{z}_{it=1}$:
- Categorical distribution with parameter $\boldsymbol{\pi} = [\pi_1, \dots, \pi_L]^\top$, where $\pi_l \geq 0$ and $\sum \pi_l = 1$

$$P(\mathbf{z}_{it=1} | \boldsymbol{\pi}) = \prod_l \pi_l^{z_{it=1}l}$$

Specification of prior distributions

- θ_{jht} : **Beta distribution** with non-negative hyper-parameters a_{jht}^0 and b_{jht}^0
 - $P(\theta_{jht} | a_{jht}^0, b_{jht}^0) \propto \theta_{jht}^{a_{jht}^0 - 1} (1 - \theta_{jht})^{b_{jht}^0 - 1},$
- $\boldsymbol{\pi}$: **Dirichlet distribution** with non-negative hyper-parameter $\boldsymbol{\delta}^0 = [\delta_1^0, \dots, \delta_L^0]^\top$
 - $P(\boldsymbol{\pi} | \boldsymbol{\delta}^0) \propto \prod_l \pi_l^{\delta_l^0 - 1}$
- $\boldsymbol{\tau}_l$: **Dirichlet distribution** with non-negative hyper-parameter $\boldsymbol{\omega}_l^0 = [\omega_1^0, \dots, \omega_L^0]^\top$
 - $P(\boldsymbol{\tau}_l | \boldsymbol{\omega}_l^0) \propto \prod_{l'} \tau_{ll'}^{\omega_{ll'}^0 - 1}$

VB method as an optimization problem

- A main goal of VB method: Finding approximate posterior distribution.
- Why?
 1. True posterior is difficult to express as an analytical form in many cases.
 2. MCMC is computationally very heavy in complex models.

- VB method is essentially an optimization problem with a constraint:

$$q = \underset{q}{\operatorname{argmin}} \operatorname{KL}(q(\boldsymbol{\Theta}, \boldsymbol{\pi}, \mathbf{T}) || P(\mathbf{Z}, \boldsymbol{\Theta}, \boldsymbol{\pi}, \mathbf{T} | \mathbf{X})) \text{ subject to}$$

$$q(\boldsymbol{\Theta}, \boldsymbol{\pi}, \mathbf{T}) = \left(\prod_t \prod_j \prod_h q(\theta_{jht}) \right) q(\boldsymbol{\pi}) \left(\prod_l q(\boldsymbol{\tau}_l) \right),$$

$$\text{where } \operatorname{KL}(f(\boldsymbol{\vartheta}) || g(\boldsymbol{\vartheta})) = \int f(\boldsymbol{\vartheta}) \log \frac{f(\boldsymbol{\vartheta})}{g(\boldsymbol{\vartheta})} d\boldsymbol{\vartheta}.$$

- Variational distributions are the closest ones to the true posterior in terms of KL divergence under independence of variational distributions.

VB posteriors and Expectations

- $q(\boldsymbol{\pi})$ is a **Dirichlet distribution** with parameters $\delta_l^* = \sum_i \mathbb{E}_{q(\mathbf{z}_i)}[z_{it=1l}] + \delta_l^0 - 1$
- $q(\theta_{ij t})$ is a **Beta distribution** with parameters
$$\begin{cases} a_{jht}^* = \sum_i \sum_l \mathbb{E}_{q(\mathbf{z}_i)}[z_{itl}] g_{jhlt} x_{ij t} + a_{jht}^0, \\ b_{jht}^* = \sum_i \sum_l \mathbb{E}_{q(\mathbf{z}_i)}[z_{itl}] g_{jhlt} (1 - x_{ij t}) + b_{jht}^0. \end{cases}$$
- $q(\boldsymbol{\tau}_l)$ is a **Dirichlet distribution** with $\omega_{ll'}^* = \sum_{t=2} \sum_i \mathbb{E}_{q(\mathbf{z}_i)}[z_{it-1,l} z_{itl'}] + \omega_{ll'}^0$
- These distributions are well-known and their means or variances can be calculated easily.
- $\mathbb{E}_{q(\mathbf{z}_i)}[z_{itl}]$ and $\mathbb{E}_{q(\mathbf{z}_i)}[z_{it-1,l} z_{itl'}]$ were calculated by effective forward-backward algorithm for each student.
- ELBO: Evidence of lower bound (lower bound of log marginal likelihood) is also calculated.

VB algorithm

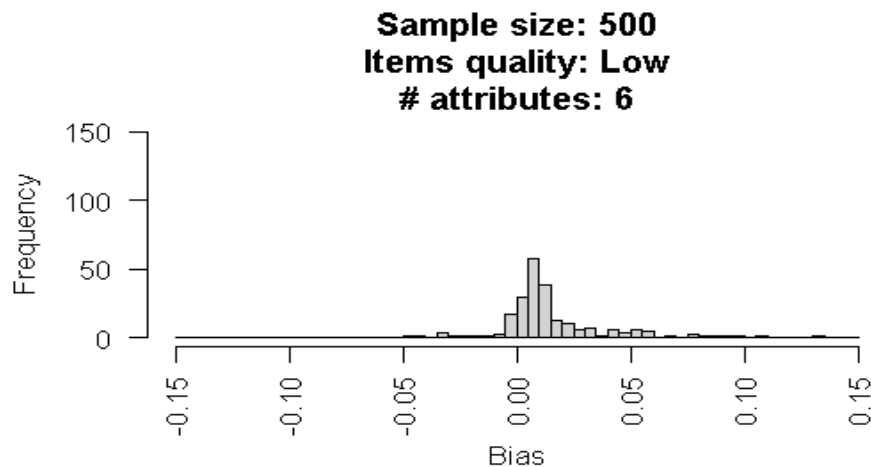
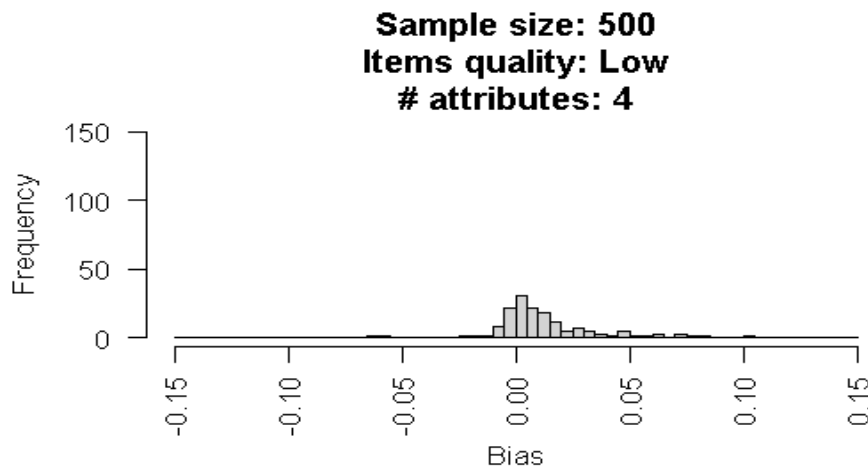
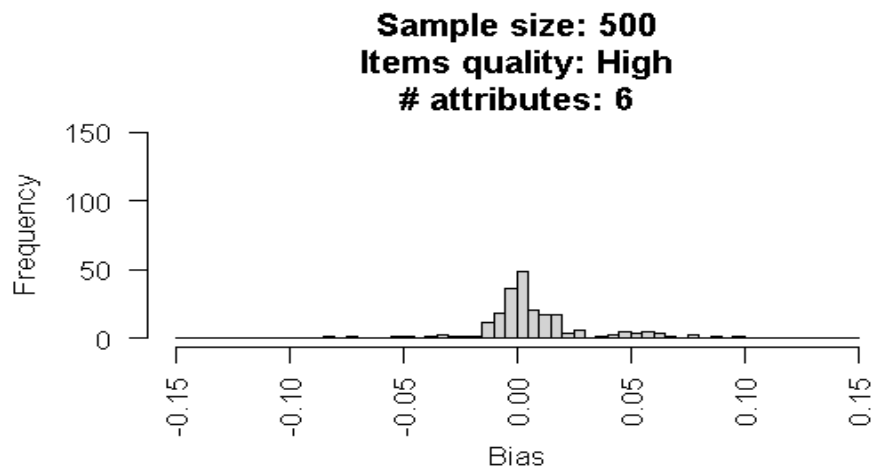
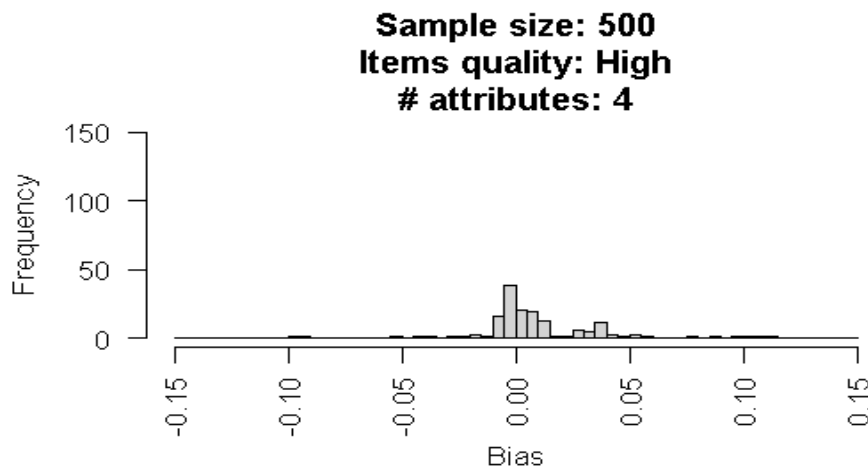
1. Set hyperparameters ($\mathbf{A}^0, \mathbf{B}^0, \boldsymbol{\delta}^0$, and $\boldsymbol{\Omega}^0$), and initialize variational parameters ($\mathbf{A}^*, \mathbf{B}^*, \boldsymbol{\delta}^*$, and $\boldsymbol{\Omega}^*$).
2. VE-step: Calculate all $\mathbb{E}_{q(\mathbf{z}_{it})}[z_{itl}]$ and $\mathbb{E}_{q(\mathbf{z}_i)}[z_{itl}'z_{it-1,l}]$ with forward-backward algorithm for each student.
3. VM-step: update variational parameters ($\mathbf{A}^*, \mathbf{B}^*, \boldsymbol{\delta}^*$, and $\boldsymbol{\Omega}^*$).
4. If the change of ELBO is smaller than the threshold, stop iteration and output mean and variance of variational posteriors, otherwise go back to the VE-step.

Simulation study

- Three factors were manipulated:
 1. **sample size**: small (500) or large (2000);
 2. **Q-matrix**: consisting of either four or six attributes
 - we assume the Q-matrix is the same across all time points
 3. **slip and guessing parameters**, representing high quality items (0.1) or low quality items (0.2).
 - Item parameters: $\theta_{jh'_at} = guess_j + a \frac{1 - slip_j - guess_j}{\sum_k q_{jk}}$
 - True item parameters were set to satisfy monotonicity conditions (e.g., Henson et al., 2009).
- Thus, a total of $2 \times 2 \times 2 = 8$ simulation conditions were evaluated.

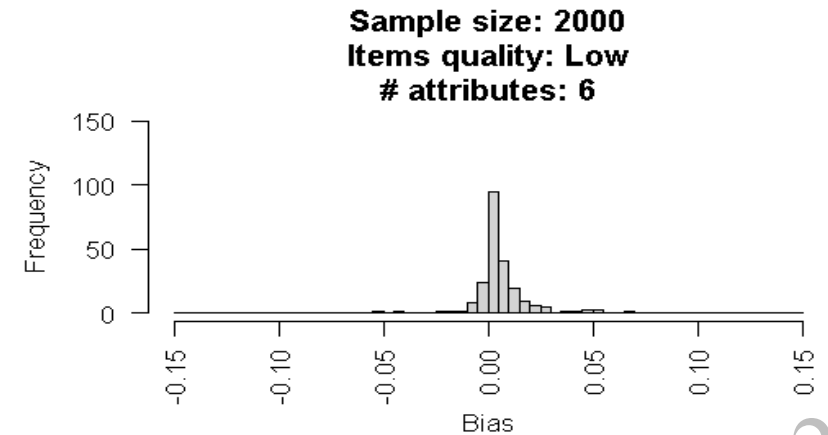
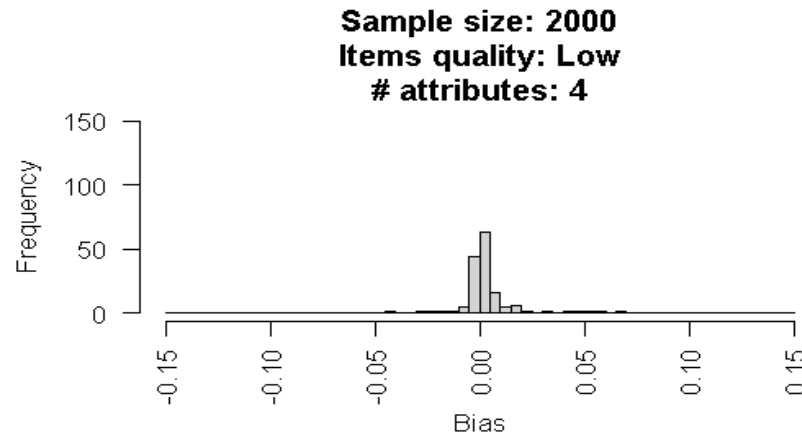
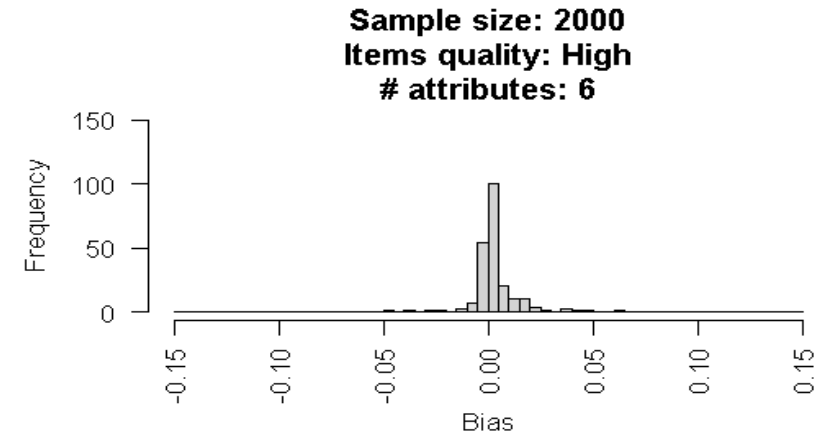
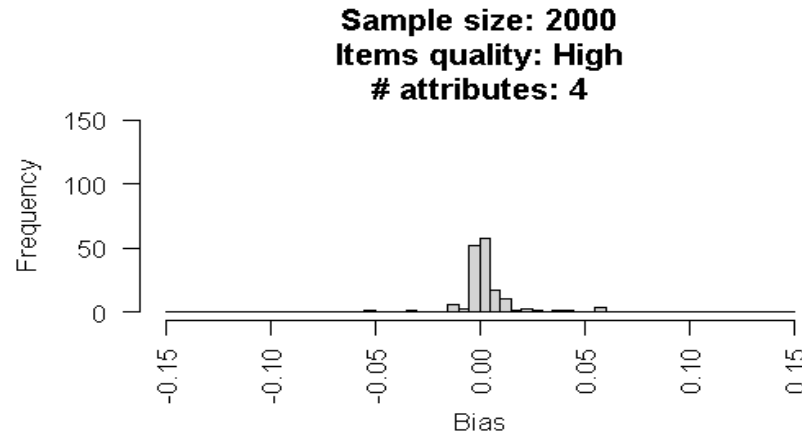
Histograms of biases of **correct item response probabilities** at the first time point in 500 sample size conditions

- Biases are distributed around zero.
- The VB method could recover appropriate parameter values.



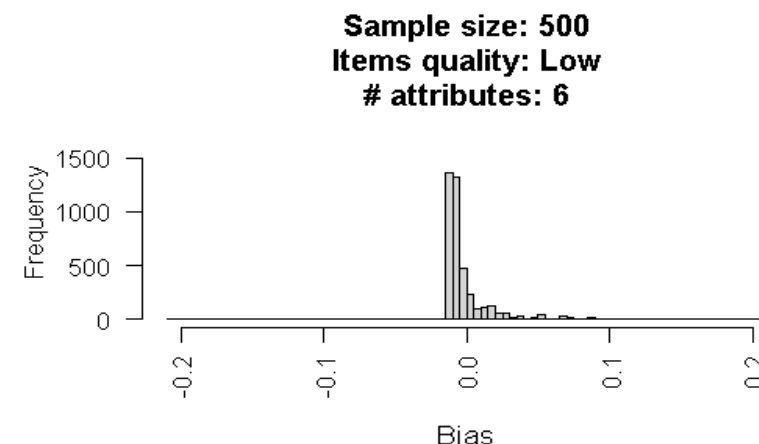
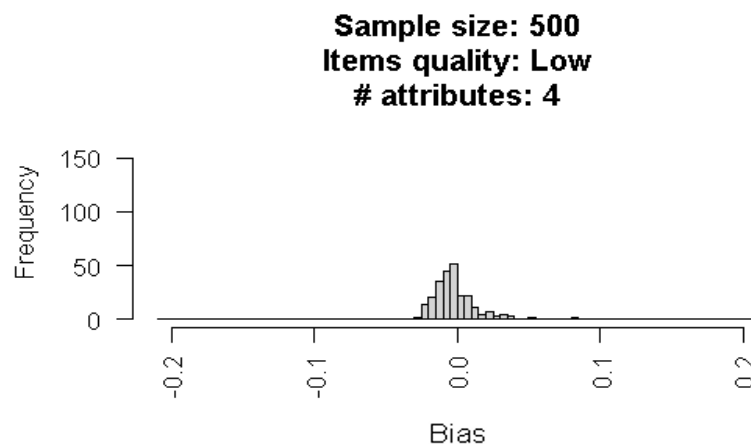
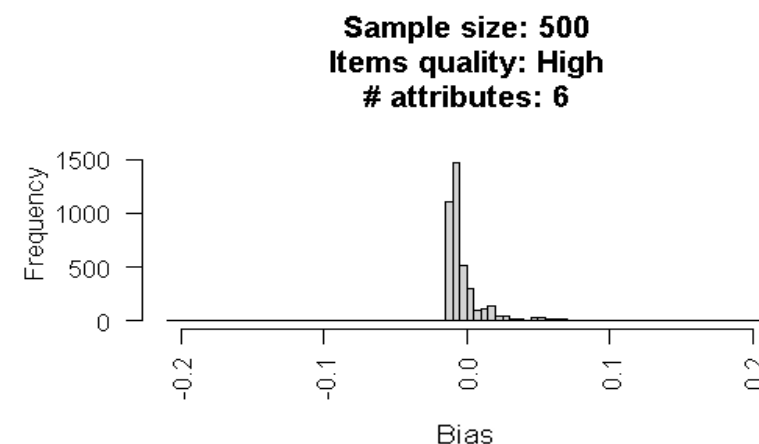
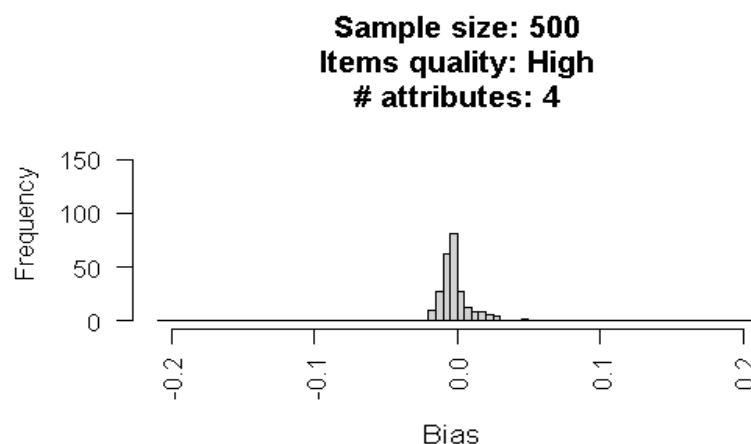
Histograms of biases of **correct item response probabilities** at the first time point in 2000 sample size conditions

- Biases became smaller than 500 sample size conditions.



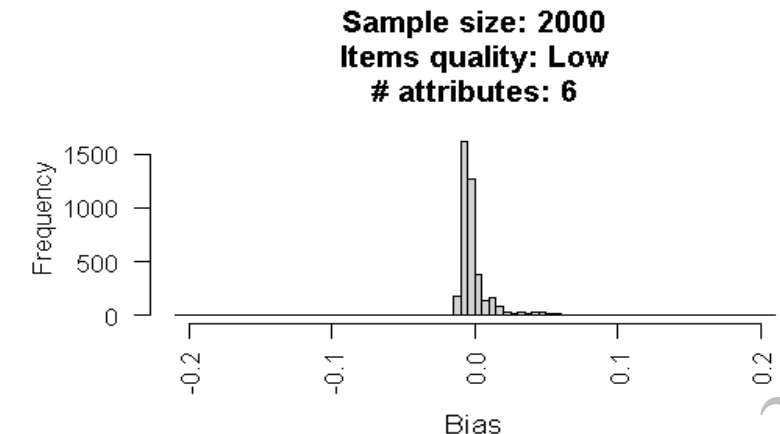
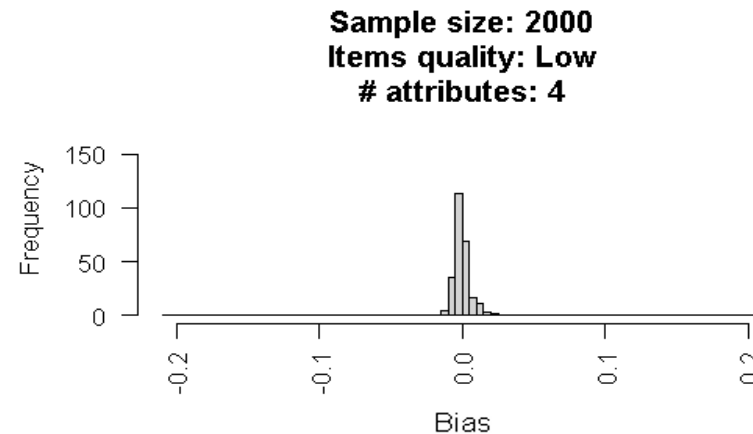
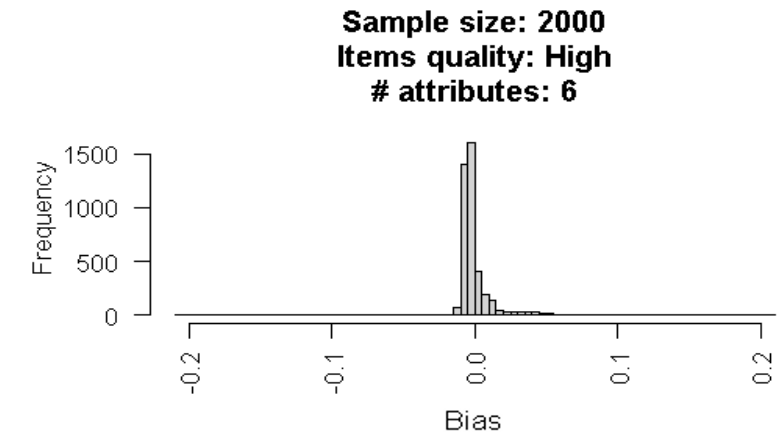
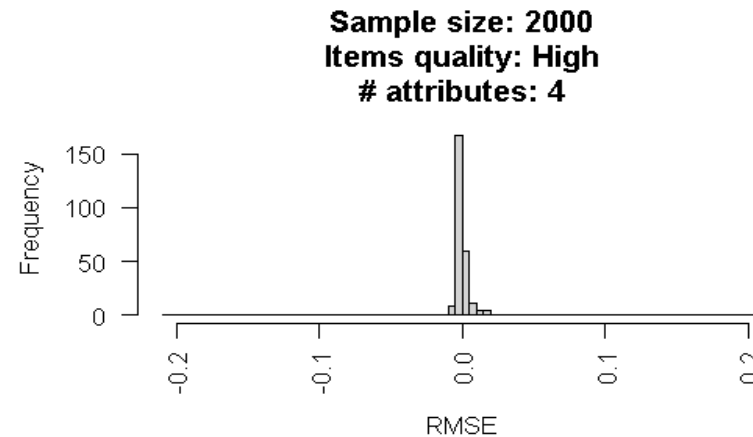
Histograms of biases of **transition parameters** in 500 sample size conditions

- Transition parameters were also correctly recovered even in 500 sample size conditions.



Histograms of biases of **transition parameters** in 2000 sample size conditions

- Transition parameters were precisely recovered with proposed VB method.



Real data example: Spatial responding test

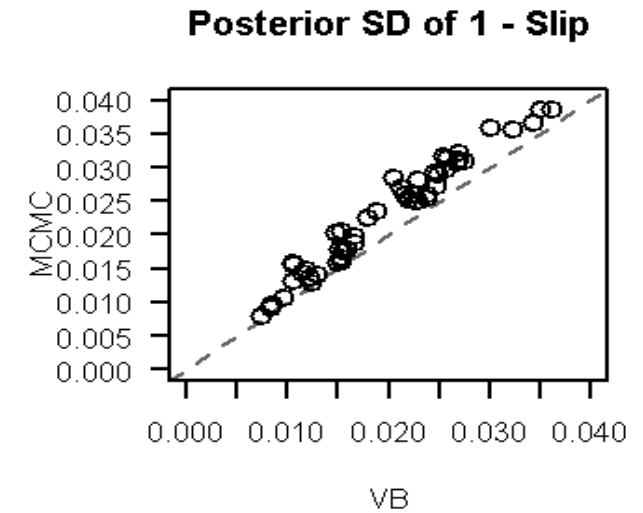
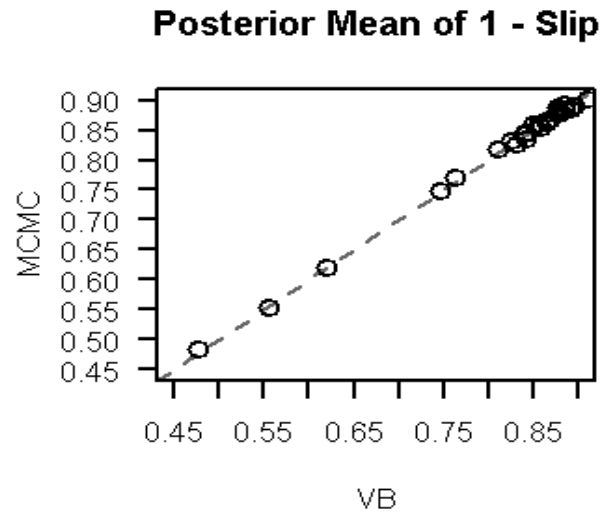
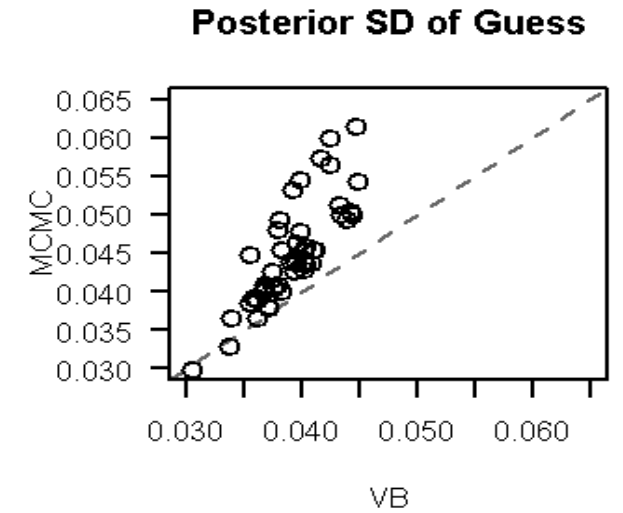
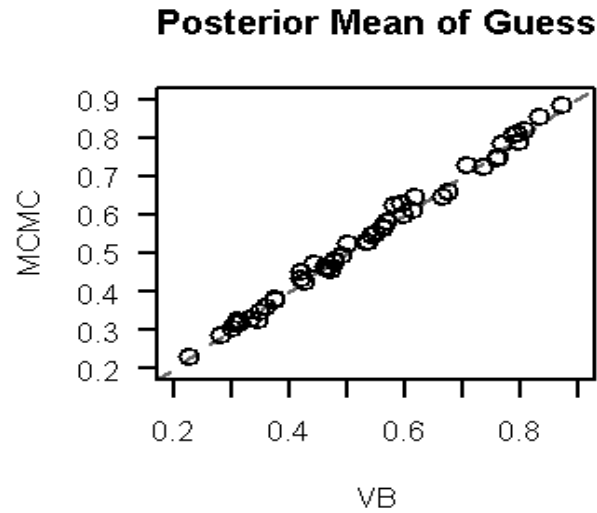
- **Spatial responding test** (e.g., Chen et al., 2017) had **five time points** corresponding to five test blocks.
 - This test consisted of **50 items** that were divided five blocks.
- The number of **subjects were 351** and the **four attributes** (1. 90° x-axis, 2. 90° y-axis, 3. 180° x-axis, and 4. 180° y-axis) were assumed.
- The data and MCMC estimation code can be found in the “hmcdm” package (Zhang, Wang, & Chen, 2020)
- The G-matrix was modified to represent the DINA model: $g_{j1l} = 1 - \prod_k \alpha_{lk}^{q_{jk}}$ and $g_{j2l} = \prod_k \alpha_{lk}^{q_{jk}}$. In this setting, guessing and slip parameters corresponded to the θ parameters as $\theta_{j1} = guess_j$ and $\theta_{j2} = 1 - slip_j$.
 - In addition, we also modified algorithm to deal with nondecreasing attributes constraints (Chen et al., 2017).

Other analysis settings

- The maximum number of iterations of the VB algorithm was 500 and convergence criteria was 10^{-4} .
- The number of MCMC sampling was 40,000 and the first 20,000 iterations were discarded as burn-in period.
- The number of MCMC chains were four.
- Other settings were the default of the hmcdm package.
- Convergence criteria was Gelman-Rubin's $\hat{R} \leq 1.1$ (Gelman & Rubin, 1992) for each parameter.

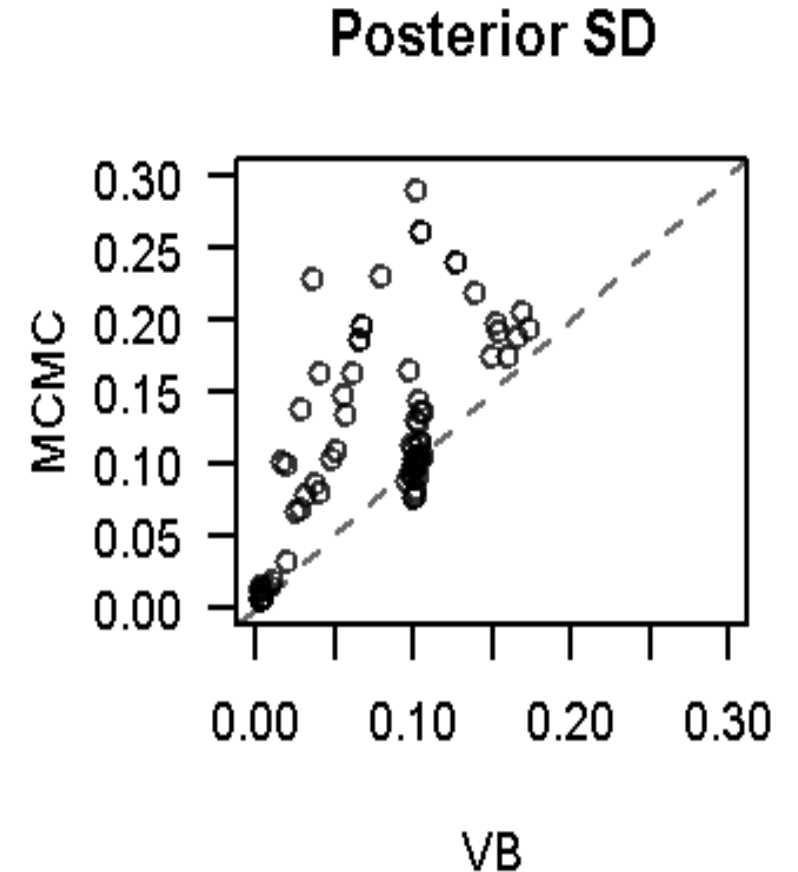
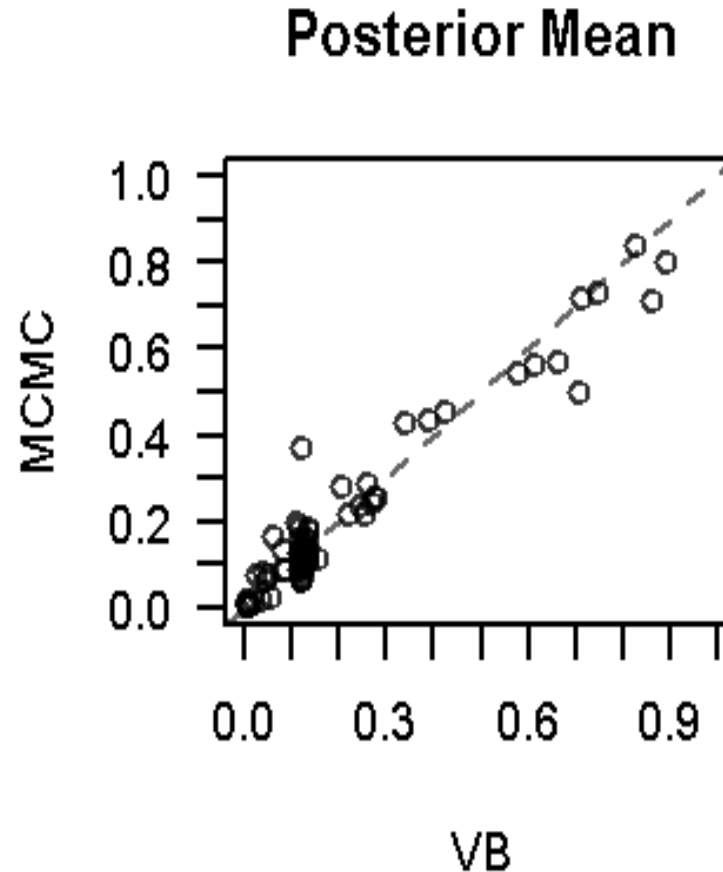
Result of 1-slip and guessing

- VB was more than **34 times** faster than MCMC in this condition.
 - VB took **42.286 sec** and MCMC took **1454.312 sec** for estimation.
- The parameter estimates of the VB results were very similar to the MCMC method.
 - Correlations between two methods were greater than **0.99**.
- The posterior SDs of VB method were slightly smaller than the ones of the MCMC.



Result of transition probability

- The **correlation was 0.969** – this value was satisfactory high.
- SD result was slightly different.
 - SDs of VB method underestimate posterior SDs.



Result of attribute mastery profile

- The mastery agreement results of “90° x-axis”, “90° y-axis”, and “180° y-axis” were **all greater than 0.95**.
- The agreement results of “180° x-axis” were slightly worse but they were in the range from **0.926 (at the third time point) to 0.951 (at the first time point)**, which is still a relatively high agreement result.
- The whole mastery profile agreements were ranged from **0.877 at the third time point to the 0.914 at the fifth time point**.

Time point	Attribute				All
	90° x-axis	90° y-axis	180° x-axis	180° y-axis	
1	.951	.974	.951	.977	.891
2	.974	.966	.943	.986	.903
3	.957	.963	.926	.983	.877
4	.957	.969	.931	.986	.886
5	.960	.983	.943	.991	.914

Conclusion

- We developed a VB algorithm for the longitudinal type DCMs, namely HM-DCMs.
- The VB algorithm showed similar parameter estimates to the MCMC method in real data example in much faster estimation time.
- To speed up estimation, C++ version of estimation code may be required.