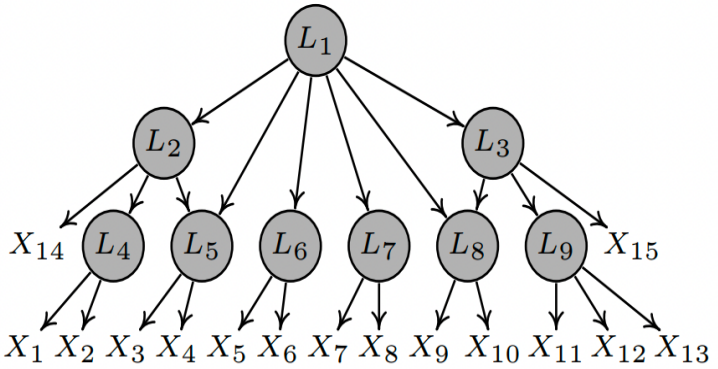


Identification of Linear Non-Gaussian Latent Hierarchical Structure

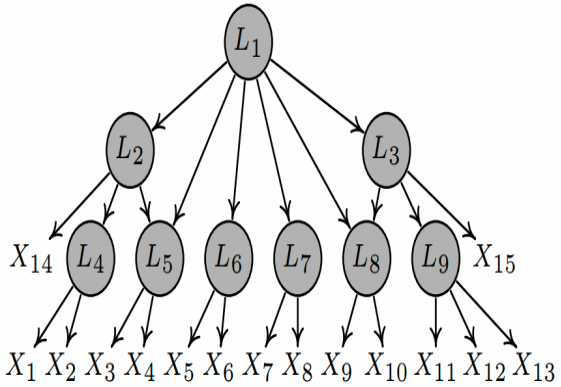
Feng Xie[1,2], Biwei Huang[3], Zhengming Chen[4], Yangbo He[1], Zhi Geng[2], Kun Zhang[3,5]

[1]Department of Probability and Statistics, Peking University, Beijing, China
[2]Department of Applied Statistics, Beijing Technology and Business University, Beijing, China
[3]Department of Philosophy, Carnegie Mellon University, Pittsburgh, USA
[4]School of Computer Science, Guangdong University of Technology, Guangzhou, China
[5]Machine Learning Department, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE.
Correspondence to: Feng Xie <xiefeng009@gmail.com>, Kun Zhang <kunz1@cmu.edu>

Problem Definition

	X_1	X_2	…	X_{15}
1	$X_{1,1}$	$X_{2,1}$	…	$X_{15,1}$
2	$X_{1,2}$	$X_{2,2}$	⋮	$X_{15,2}$
3	$X_{1,3}$	$X_{2,3}$	…	$X_{15,3}$
⋮	⋮	⋮	⋮	⋮
n	$X_{1,n}$	$X_{2,n}$	…	$X_{15,n}$

Observational dataset



Latent hierarchical structure

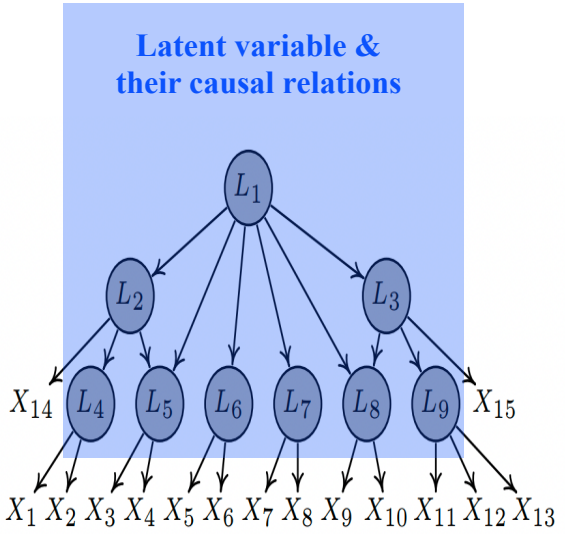
Is it possible to find latent variable L_i and their causal relations only from measured variables X_i?

Problem Definition

	X_1	X_2	…	X_{15}
1	$X_{1,1}$	$X_{2,1}$	…	$X_{15,1}$
2	$X_{1,2}$	$X_{2,2}$	⋮	$X_{15,2}$
3	$X_{1,3}$	$X_{2,3}$	…	$X_{15,3}$
⋮	⋮	⋮	⋮	⋮
n	$X_{1,n}$	$X_{2,n}$	…	$X_{15,n}$

Observational dataset

Discovery

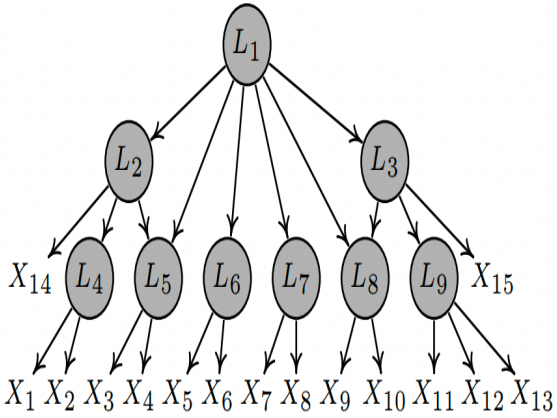


Latent hierarchical structure

Is it possible to find latent variable L_i and their causal relations only from measured variables X_i?

Linear, Non-Gaussian Latent Hierarchical Model

- **Measured variables** may not be directly causally related but were generated by causally related latent variables
- Some **latent variables** have only latent variables as children (i.e., no observed children)
- Assume variables were generated by the **Li**near, **N**on-**G**aussian **La**tent **H**ierarchical Model (LiNGLaH)



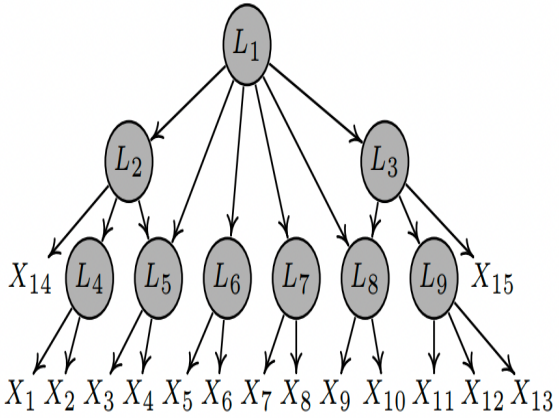
Find a sufficient graphical condition that renders the causal structure of a latent hierarchical model identifiable?

Linear, Non-Gaussian Latent Hierarchical Model

- **Measured variables** may not be directly causally related but were generated by causally related latent variables

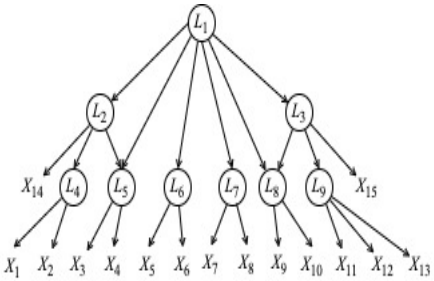
- Some **latent variables** have only latent variables as children (i.e., no observed children)

- Assume variables were generated by the **Li**near, **N**on-**G**aussian **La**tent **H**ierarchical Model (LiNGLaH)



+ Minimal Latent Hierarchical Structure Condition:
(1) each latent variable has at least three neighbors, and
(2) each latent variable has at least two pure children (which can be either latent or observed)

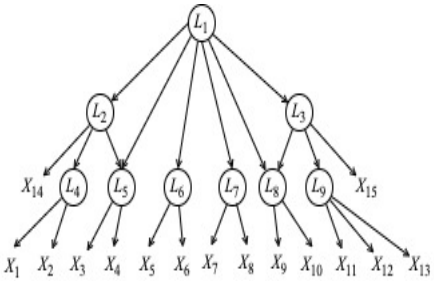
Model Estimation



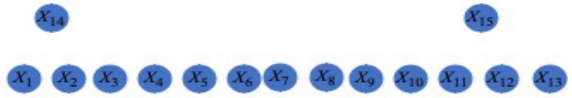
- **Step 1:** locate all latent variables
  - P1. Identify **causal clusters** from the active variable set
  - P2. Determine the number of **new latent variables** that need to be introduced for these clusters
  - P3. Update the active variable set

- **Step 2:** infer the causal structure among the identified latent variables
  - P1. identify the **causal order** among latent variables
  - P2. remove **redundant** edges

Model Estimation

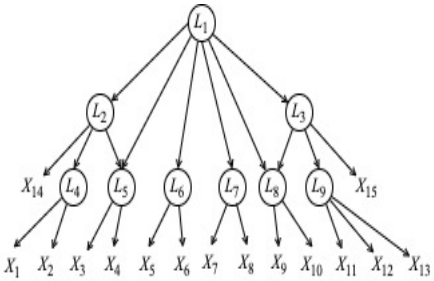


- **Step 1:** <u>locate all latent variables</u>
  - P1. Identify **causal clusters** from the active variable set
  - P2. Determine the number of **new latent variables** that need to be introduced for these clusters
  - P3. Update the active variable set

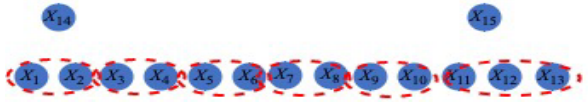


- **Step 2:** infer the causal structure among the identified latent variables
  - P1. identify the **causal order** among latent variables
  - P2. remove **redundant** edges

Model Estimation

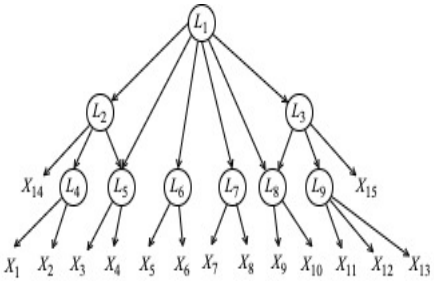


- **Step 1:** locate all latent variables
  - P1. <u>Identify **causal clusters** from the active variable set</u>
  - P2. Determine the number of **new latent variables** that need to be introduced for these clusters
  - P3. Update the active variable set

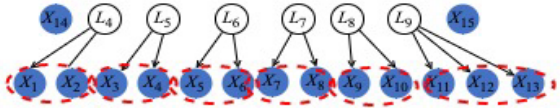


- **Step 2:** infer the causal structure among the identified latent variables
  - P1. identify the **causal order** among latent variables
  - P2. remove **redundant** edges

Model Estimation

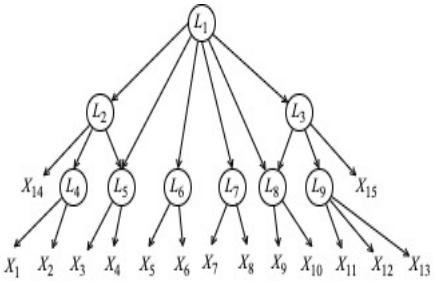


- **Step 1:** locate all latent variables
  - P1. Identify **causal clusters** from the active variable set
  - P2. Determine the number of **new latent variables** that need to be introduced for these clusters
  - P3. Update the active variable set

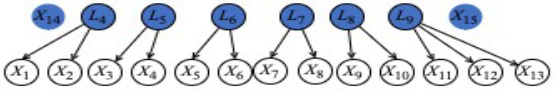


- **Step 2:** infer the causal structure among the identified latent variables
  - P1. identify the **causal order** among latent variables
  - P2. remove **redundant** edges

Model Estimation

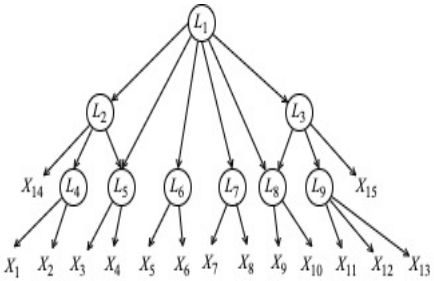


- **Step 1:** locate all latent variables
  - P1. Identify **causal clusters** from the active variable set
  - P2. Determine the number of **new latent variables** that need to be introduced for these clusters
  - P3. Update the active variable set

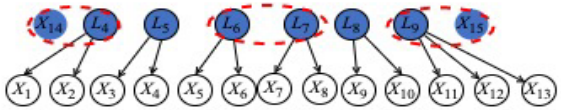


- **Step 2:** infer the causal structure among the identified latent variables
  - P1. identify the **causal order** among latent variables
  - P2. remove **redundant** edges

Model Estimation

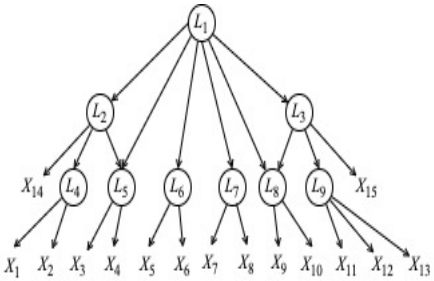


- **Step 1:** locate all latent variables
  - P1. Identify **causal clusters** from the active variable set
  - P2. Determine the number of **new latent variables** that need to be introduced for these clusters
  - P3. Update the active variable set

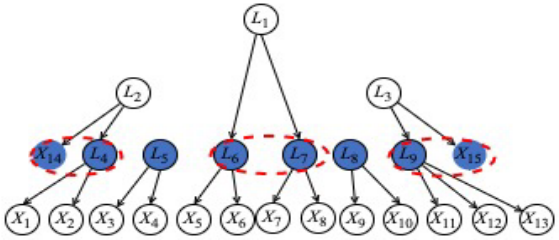


- **Step 2:** infer the causal structure among the identified latent variables
  - P1. identify the **causal order** among latent variables
  - P2. remove **redundant** edges

Model Estimation

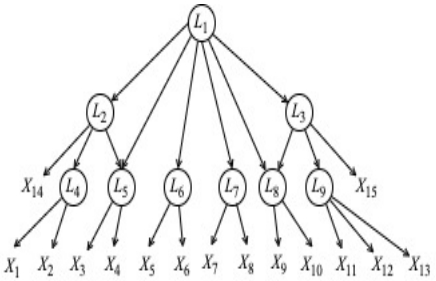


- **Step 1:** locate all latent variables
  - P1. Identify **causal clusters** from the active variable set
  - P2. Determine the number of **new latent variables** that need to be introduced for these clusters
  - P3. Update the active variable set

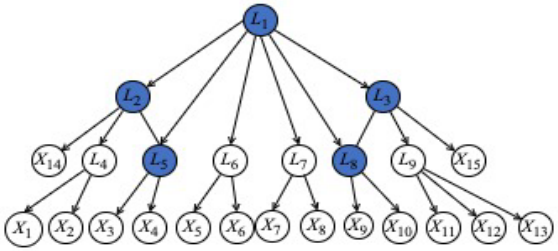


- **Step 2:** infer the causal structure among the identified latent variables
  - P1. identify the **causal order** among latent variables
  - P2. remove **redundant** edges

Model Estimation

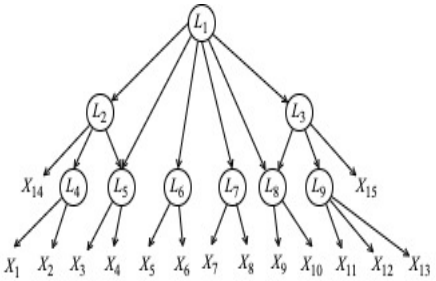


- **Step 1:** locate all latent variables
  - P1. Identify **causal clusters** from the active variable set
  - P2. Determine the number of **new latent variables** that need to be introduced for these clusters
  - P3. Update the active variable set

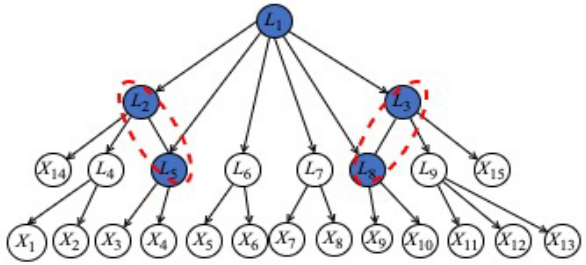


- **Step 2:** infer the causal structure among the identified latent variables
  - P1. identify the **causal order** among latent variables
  - P2. remove **redundant** edges

Model Estimation

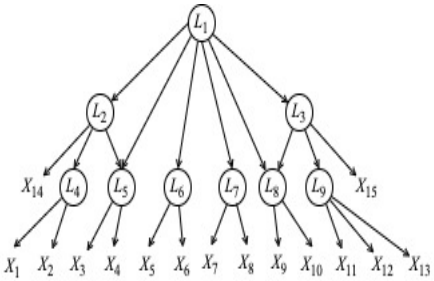


- **Step 1:** locate all latent variables
  - P1. Identify **causal clusters** from the active variable set
  - P2. Determine the number of **new latent variables** that need to be introduced for these clusters
  - P3. Update the active variable set

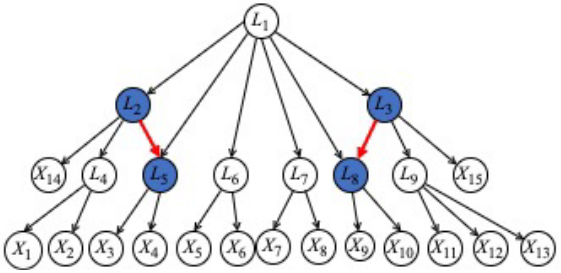


- **Step 2:** infer the causal structure among the identified latent variables
  - P1. identify the **causal order** among latent variables
  - P2. remove **redundant** edges

Model Estimation



- **Step 1:** locate all latent variables
  - P1. Identify **causal clusters** from the active variable set
  - P2. Determine the number of **new latent variables** that need to be introduced for these clusters
  - P3. Update the active variable set



- **Step 2:** <u>infer the causal structure among the identified latent variables</u>
  - P1. identify the **causal order** among latent variables
  - P2. remove **redundant** edges

Simulation

- 4 cases, with different latent structures, including tree-based and measurement-based structures
- Can we recover the ground-truth structure, including causal direction?
  - Structure Recovery Error rate
  - Error in the Number of Latent variable sets
  - Correct ordering rate

Table 1. Performance of LaHME, GIN, FOFC, BPC, CLRG and CLNJ on learning latent hierarchical structure.

		Structure Recovery Error Rate ↓						Error in Hidden Variables ↓						Correct-Ordering Rate ↑					
Algorithm		LaHME	GIN	FOFC	BPC	CLRG	CLNJ	LaHME	GIN	FOFC	BPC	CLRG	CLNJ	LaHME	GIN	FOFC	BPC	CLRG	CLNJ
Case 1	1k	0.1	0.2	1.0	1.0	1.0	1.0	0.1	0.1	0.5	0.6	2.0	2.0	0.96	0.92	-	-	-	-
	5k	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.1	2.0	2.0	1.0	1.0	-	-	-	-
	10k	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	2.0	2.0	1.0	1.0	-	-	-	-
Case 2	1k	0.2	1.0	1.0	1.0	1.0	1.0	0.2	3.2	3.8	3.9	4.0	4.0	0.9	0.08	-	-	-	-
	5k	0.1	1.0	1.0	1.0	1.0	1.0	0.1	3.0	3.6	3.8	4.0	4.0	0.96	0.1	-	-	-	-
	10k	0.0	1.0	1.0	1.0	1.0	1.0	0.0	3.0	3.5	3.8	4.0	4.0	1.0	0.1	-	-	-	-
Case 3	1k	0.1	1.0	1.0	1.0	1.0	1.0	0.2	1.3	3.0	3.1	3.0	3.0	0.92	0.0	-	-	-	-
	5k	0.0	1.0	1.0	1.0	1.0	1.0	0.0	1.2	3.0	3.2	3.0	3.0	1.0	0.0	-	-	-	-
	10k	0.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	3.2	3.4	3.0	3.0	1.0	0.0	-	-	-	-
Case 4	1k	0.3	1.0	1.0	1.0	1.0	1.0	0.4	3.4	7.0	7.2	8.0	8.0	0.9	0.0	-	-	-	-
	5k	0.2	1.0	1.0	1.0	1.0	1.0	0.2	3.2	6.6	6.9	8.0	8.0	0.94	0.0	-	-	-	-
	10k	0.0	1.0	1.0	1.0	1.0	1.0	0.0	3.1	5.8	6.7	8.0	8.0	1.0	0.0	-	-	-	-

Simulation

- 4 cases, with different latent structures, including tree-based and measurement-based structures
- Can we recover the ground-truth structure, including causal direction?
  - Structure Recovery Error rate
  - Error in the Number of Latent variable sets
  - Correct ordering rate

Table 1. Performance of LaHME, GIN, FOFC, BPC, CLRG and CLNJ on learning latent hierarchical structure.

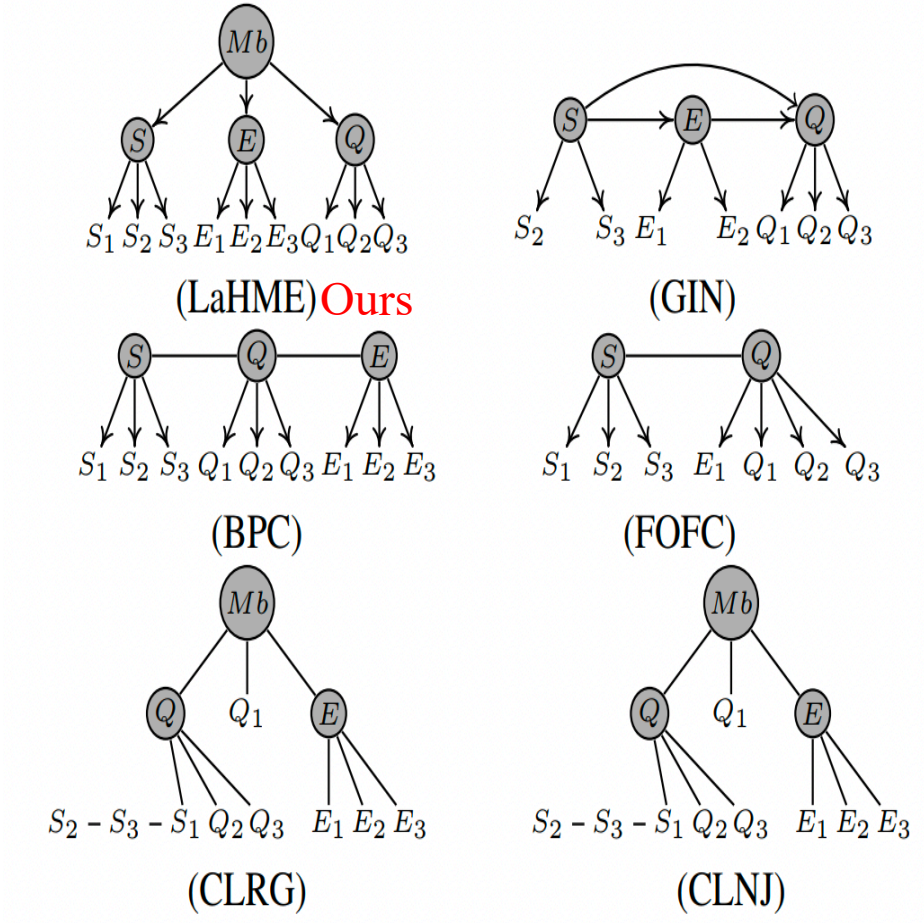
		Structure Recovery Error Rate ↓						Error in Hidden Variables ↓						Correct-Ordering Rate ↑					
Algorithm		LaHME	GIN	FOFC	BPC	CLRG	CLNJ	LaHME	GIN	FOFC	BPC	CLRG	CLNJ	LaHME	GIN	FOFC	BPC	CLRG	CLNJ
Case 1	1k	0.1	0.2	1.0	1.0	1.0	1.0	0.1	0.1	0.5	0.6	2.0	2.0	0.96	0.92	-	-	-	-
	5k	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.1	2.0	2.0	1.0	1.0	-	-	-	-
	10k	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	2.0	2.0	1.0	1.0	-	-	-	-
Case 2	1k	0.2	1.0	1.0	1.0	1.0	1.0	0.2	3.2	3.8	3.9	4.0	4.0	0.9	0.08	-	-	-	-
	5k	0.1	1.0	1.0	1.0	1.0	1.0	0.1	3.0	3.6	3.8	4.0	4.0	0.96	0.1	-	-	-	-
	10k	0.0	1.0	1.0	1.0	1.0	1.0	0.0	3.0	3.5	3.8	4.0	4.0	1.0	0.1	-	-	-	-
Case 3	1k	0.1	1.0	1.0	1.0	1.0	1.0	0.2	1.3	3.0	3.1	3.0	3.0	0.92	0.0	-	-	-	-
	5k	0.0	1.0	1.0	1.0	1.0	1.0	0.0	1.2	3.0	3.2	3.0	3.0	1.0	0.0	-	-	-	-
	10k	0.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	3.2	3.4	3.0	3.0	1.0	0.0	-	-	-	-
Case 4	1k	0.3	1.0	1.0	1.0	1.0	1.0	0.4	3.4	7.0	7.2	8.0	8.0	0.9	0.0	-	-	-	-
	5k	0.2	1.0	1.0	1.0	1.0	1.0	0.2	3.2	6.6	6.9	8.0	8.0	0.94	0.0	-	-	-	-
	10k	0.0	1.0	1.0	1.0	1.0	1.0	0.0	3.1	5.8	6.7	8.0	8.0	1.0	0.0	-	-	-	-

Application to multitasking behavior Data

- The data set consists of 202 samples

Latent Factors	Children (Indicators)
Speed (S)	Correctly marked Numbers (S1), Correctly marked Latters (S2), and Correctly marked Figures (S3)
Error (E)	Errors marking Numbers (E1), Errors marking Latters (E2), and Errors marking Figures (E3)
Question (Q)	Correctly answered Questions Par.1 (Q1), Correctly answered Questions Par.2 (Q2), and Correctly answered Questions Par.3 (Q3)
Multitasking behavior (Mb)	Speed, Error, and Question

- Consistent with the hypothesized model given in Himi et al., 2019



Conclusion

- Essential to learn the linear latent hierarchical structure
- Provide sufficient conditions for structural identifiability
- Future work: n-factor model, nonlinear hierarchical structure…