Learning or Forgetting? A Dynamic Approach for Tracking the Knowledge Proficiency of Students

ZHENYA HUANG, QI LIU, and YUYING CHEN, University of Science and Technology of China, China
LE WU, Hefei University of Technology, China and iFLYTEK Co., Ltd, China
KELI XIAO, Stony Brook University, USA
ENHONG CHEN, University of Science and Technology of China, China
HAIPING MA, Anhui University, China
GUOPING HU, iFLYTEK Research, China

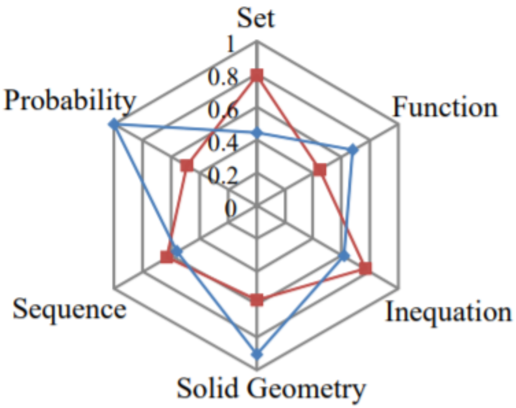
Reporter: Zhenya Huang

ACM Transactions on Information Systems, Vol. 38, No. 2, Article 19. Publication date: February 2020.

Outline

Background

- Cognitive diagnosis for knowledge proficiency
  - Domain: **Education**, Recruitment, Sports, Game, etc
  - Goal: Evaluating how much students learn about different knowledge concepts
    - Math subject: Function, Set, Inequality, etc
  - Fundamental task
    - Evaluation, Testing, Recommendation, etc

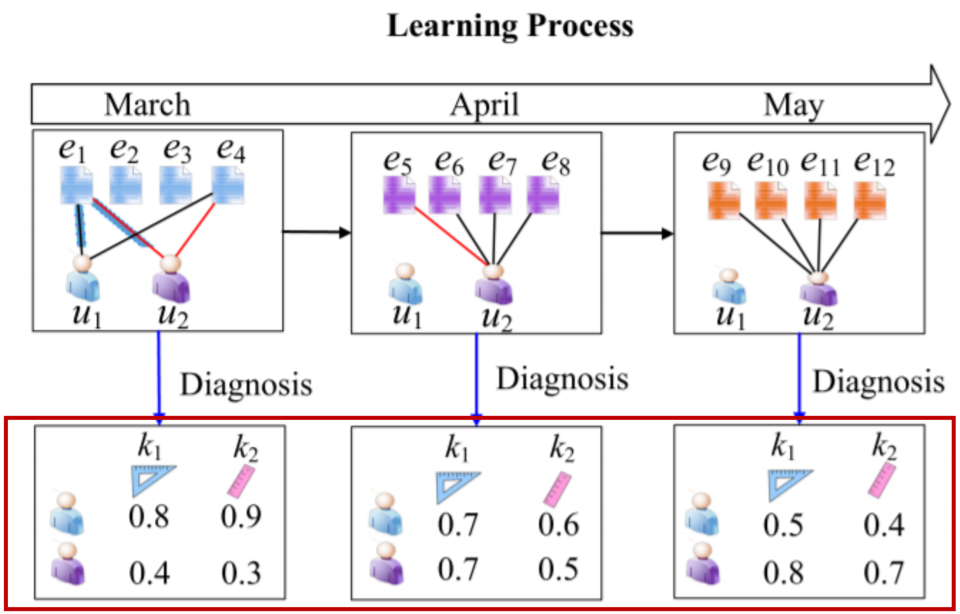


Background

- Learning activities
  - Taking courses, Practicing exercises, Taking Tests, etc
  - Classroom-based
    - Rely on expertise of teachers
    - Hard to record data
  - Online learning
    - Open environment with computer-aided technology
    - Learning data of students can be recorded
    - KhanAcademy, MOOC, etc

Background

- Cognitive diagnosis Problem



Q-matrix

	k_1(Function)	k_2(Inequality)
e_1	1	0
e_2	0	1
e_3	0	0
e_4	1	1
e_5	1	1
e_6	1	0
e_7	1	0
e_8	0	1
e_9	0	1
e_{10}	1	0
e_{11}	1	0
e_{12}	1	1

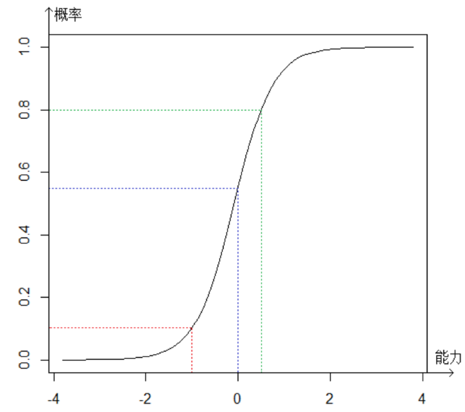
u_1, u_2: Student $e_1 \sim e_{12}$: Exercise k_1, k_2: Knowldge concept : Right answer : Wrong answer

Related work

- Static modeling
  - IRT: Item Response Theory

$$P(X_{ij}=1|\theta_j)=c_i+\frac{1-c_i}{1+\exp[-1.7a_i(\theta_j-b_i)]}$$

Latent trait



  - DINA:

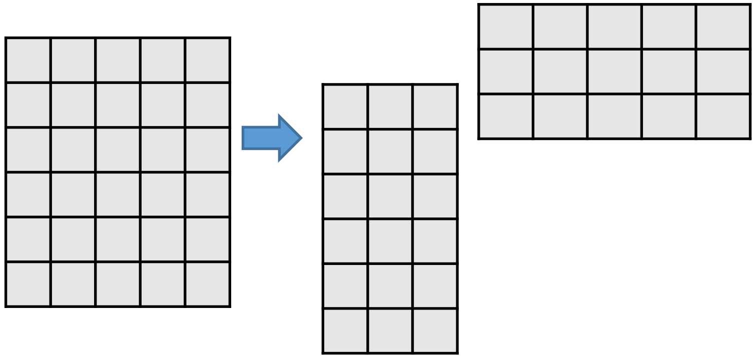
$$P_j(\boldsymbol{\alpha}_i)=P(X_{ij}=1|\boldsymbol{\alpha}_i)=g_j^{1-\eta_{ij}}(1-s_j)^{\eta_{ij}}.$$

Knowledge vector

  - PMF: Probabilistic Matrix Factorization

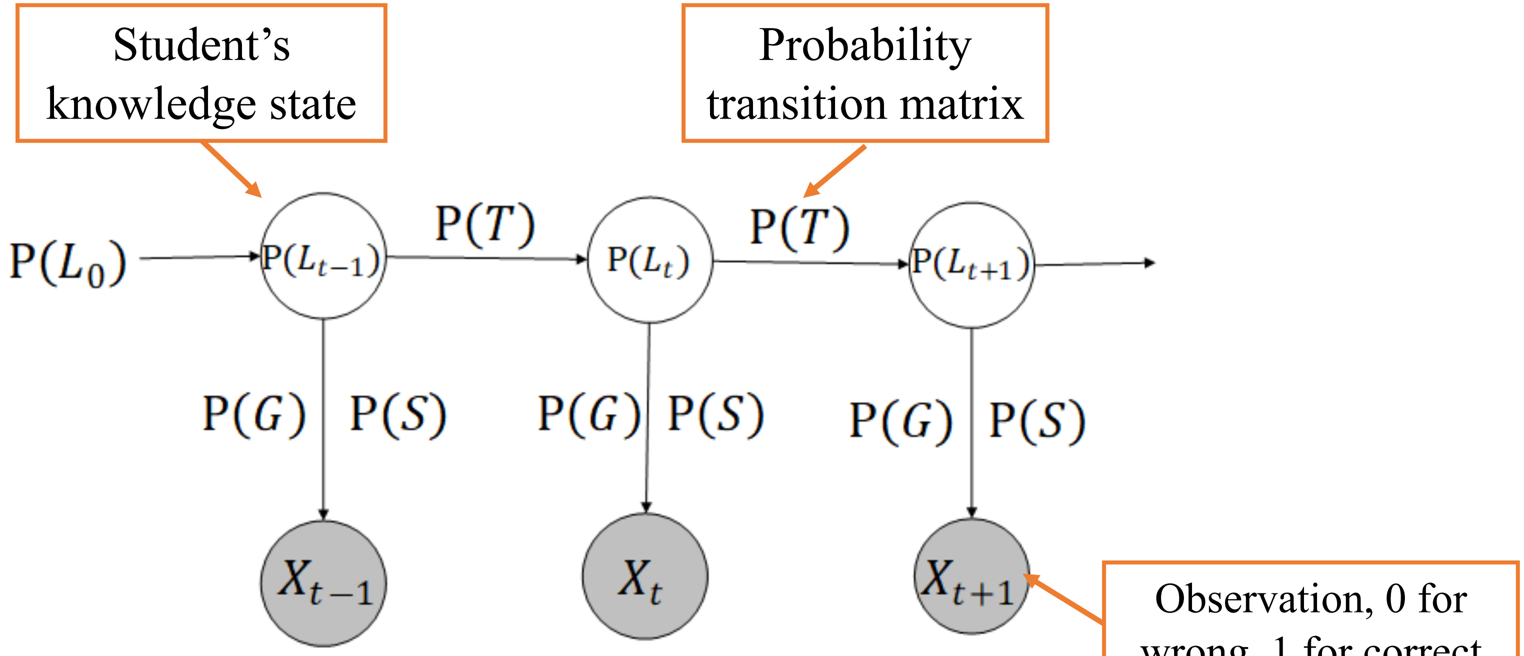
$$p(R|U,V,\sigma^2)=\prod_{i=1}^{N}\prod_{j=1}^{M}\left[\mathcal{N}(R_{ij}|U_i^TV_j,\sigma^2)\right]^{I_{ij}}$$

Latent vector



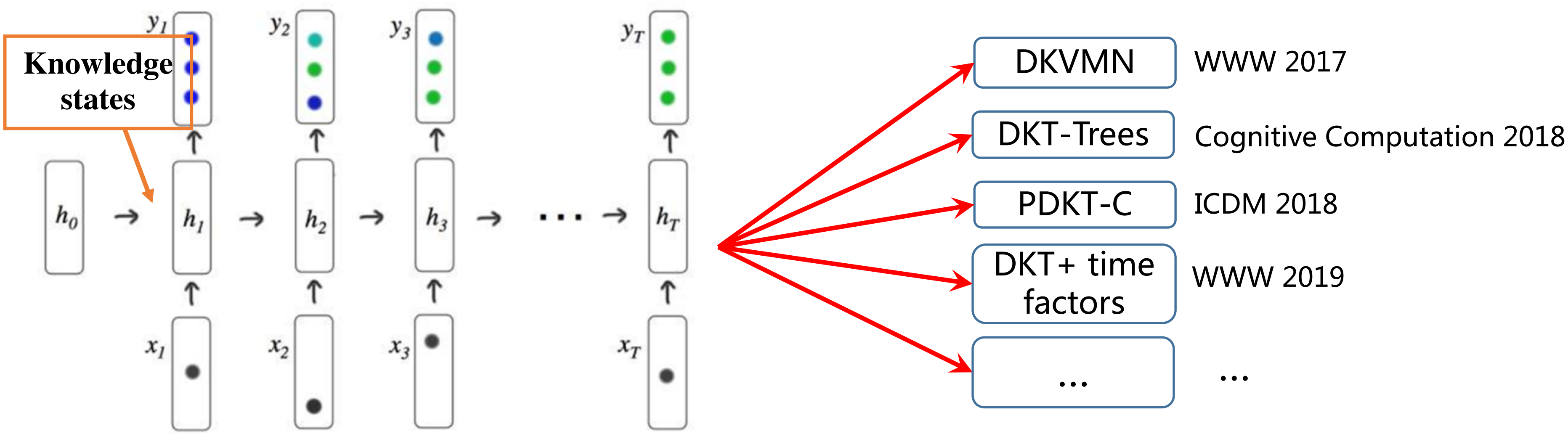
Related work

- Dynamic modeling
  - BKT: Baysian Knowledge Tracing
    - Hidden Markov Model
    - Tracing for single concept
    - Discrete results (mastered or non-mastered)



Related work

- Dynamic modeling
  - DKT: Deep Knowledge Tracing
    - Apply RNNs (LSTM) to model student knowledge over time
    - Tracing **all concepts** together
    - Hidden states can represent the latent knowledge states



Background

- Limitation
  - Ignoring the dynamic memory factors
    - How can we learn and remember knowledge?
    - Why do we forget what we have learned ?
  - Lack of interpretability
    - Don’t know the meaning of latent vectors/ hidden states
  - Learning records are sparse
    - Students practice very few exercises

Outline

Problem & Overview

- Given
  - Exercising logs as a score tensor: $R \in \mathbb{R}^{N \times M \times T}$
  - Q-matrix representing exercise-knowledge relation: $Q \in \mathbb{R}^{M \times K}$

- Goal
  - Tracking the change of knowledge proficiency of students from time 1 to T
  - Predicting her proficiency on K concepts and performance scores on specific exercises at time T + 1

(a) Exercising log example

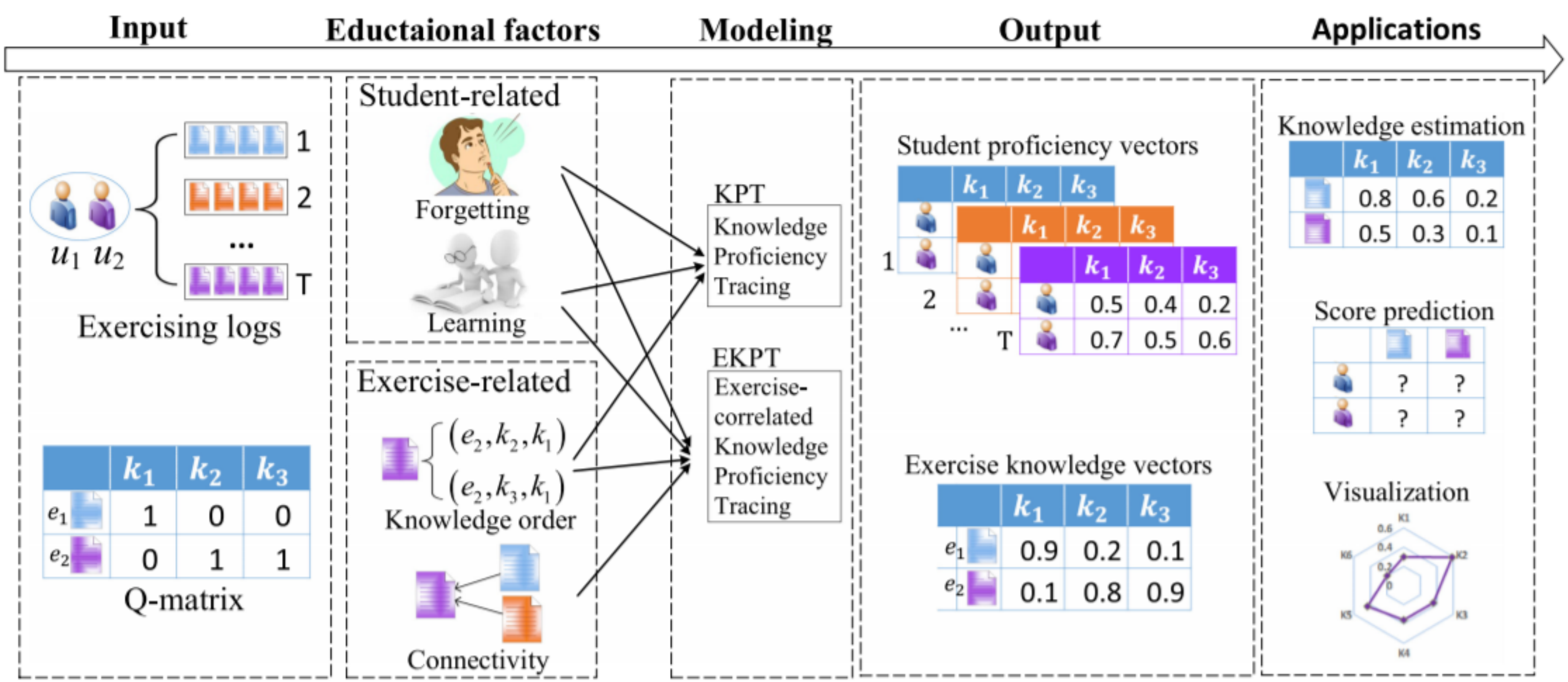
Student	Exercise	Time	Score
u_1	e_1	t_1	0
u_1	e_5	t_2	0.25
u_2	e_2	t_1	0
u_2	e_3	t_3	1
u_2	e_1	t_3	0.75
u_3	e_4	t_4	1
...	...	...	...

(b) Q-matrix example

Exercise	Knowledge concepts				
	k_1	k_2	k_3	k_4	k_5
e_1	1	0	0	0	0
e_2	0	0	1	0	0
e_3	0	0	0	1	1
e_4	0	1	0	0	0
e_5	1	0	0	0	0
...	...	...	...	...	...

Problem & Overview

- Model overview
  - KPT: Knowledge Proficiency Tracing model
  - EKPT: Exercise-correlated Knowledge Proficiency model



	k_1	k_2	k_3
e_1	1	0	0
e_2	0	1	1

	k_1	k_2	k_3
	0.5	0.4	0.2
	0.7	0.5	0.6

	k_1	k_2	k_3
e_1	0.9	0.2	0.1
e_2	0.1	0.8	0.9

	k_1	k_2	k_3
	0.8	0.6	0.2
	0.5	0.3	0.1

	?	?
	?	?

Outline

KPT model

- Probabilistic modeling
  - For each student and exercise, modeling the responses as:

$$p(R|U,V,b) = \prod_{t=1}^{T}\prod_{i=1}^{N}\prod_{j=1}^{M}\left[\mathcal{N}\left(R_{ij}^t \middle| \left\langle \boxed{U_i^t, V_j} \right\rangle - b_j, \sigma_R^2\right)\right]^{I_{ij}^t},$$

    - $U_i^t \in \mathbb{R}^{K\times 1}$: proficiency vector of student i, representing how much students learn on K concepts at time t
    - $V_j \in \mathbb{R}^{K\times 1}$: knowledge vector of exercise j, denoting the latent correlation between exercise j and K concepts
  - How to establish the corresponding relationship among **students**, **exercises** and **knowledge concepts**?

KPT model

- Modeling V with Q-matrix prior
  - Goal: project exercise into knowledge space, enhancing interpretability
  - Traditional Q-matrix
    - Denoting exercise-knowledge correlation
    - Binary entries: do not fit for probabilistic modeling
  - Our work assumption
    - If Qjq = 1, then this concept q is more relevant to exercise j than all other concepts with mark 0

$$\forall p, q \in K, p \neq q, \text{if } Q_{jq} = 1 \text{ and } Q_{jp} = 0 \Rightarrow q >_j^+ p,$$
$$\forall p, q \in K, p \neq q, \text{if } Q_{jq} = 1 \text{ and } Q_{jp} = 1 \Rightarrow q \not>_j^+ p,$$
$$\forall p, q \in K, p \neq q, \text{if } Q_{jq} = 0 \text{ and } Q_{jp} = 0 \Rightarrow q \not>_j^+ p.$$

Exercise	Knowledge concepts				
	k_1	k_2	k_3	k_4	k_5
e_1	1	0	0	0	0
e_2	0	0	1	0	0
e_3	0	0	0	1	1
e_4	0	1	0	0	0
e_5	1	0	0	0	0
...	...	...	...	...	...

$$\left\{\begin{array}{l} \text{e1, k1, k2} \\ \text{e1, k1, k3} \\ \text{e2, k3, k1} \\ \dots \\ \text{e5, k1, k2} \\ \text{e5, k1, k4} \end{array}\right.$$

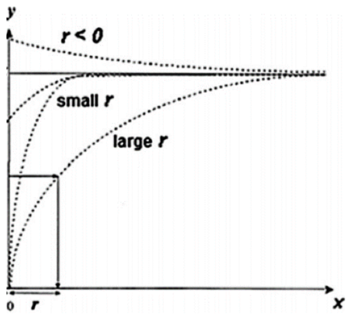
KPT model

- Modeling U with learning theories
  - Goal: explain the dynamic factors in the learning process

$$p\left(U_i^t\right) = \mathcal{N}\left(U_i^t \middle| \bar{U_i^t}, \sigma_U^2 \mathbf{I}\right), \quad \text{where } \bar{U_i^t} = \left\{\bar{U_{i1}^t}, \bar{U_{i2}^t}, \ldots, \bar{U_{iK}^t}\right\},$$

$$\bar{U_{ik}^t} = \alpha_i \boxed{L_{ik}^t(*)} + (1-\alpha_i)\boxed{F_{ik}^t(*)}, \quad s.t.\ 0 \le \alpha_i \le 1,$$

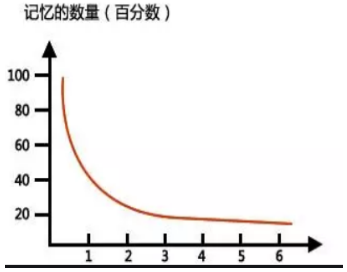
  - Two learning theories
    - **Learning curve**: The **more exercises she does**, the higher level of proficiency on the related knowledge she will get



$$L_{ik}^t(*) = U_{ik}^{t-1} \frac{D \boxed{f_{ik}^t}}{f_{ik}^t + r},$$

Number of practice times

    - **Forgetting curve**: The **longer the time passes**, the more knowledge she will forget

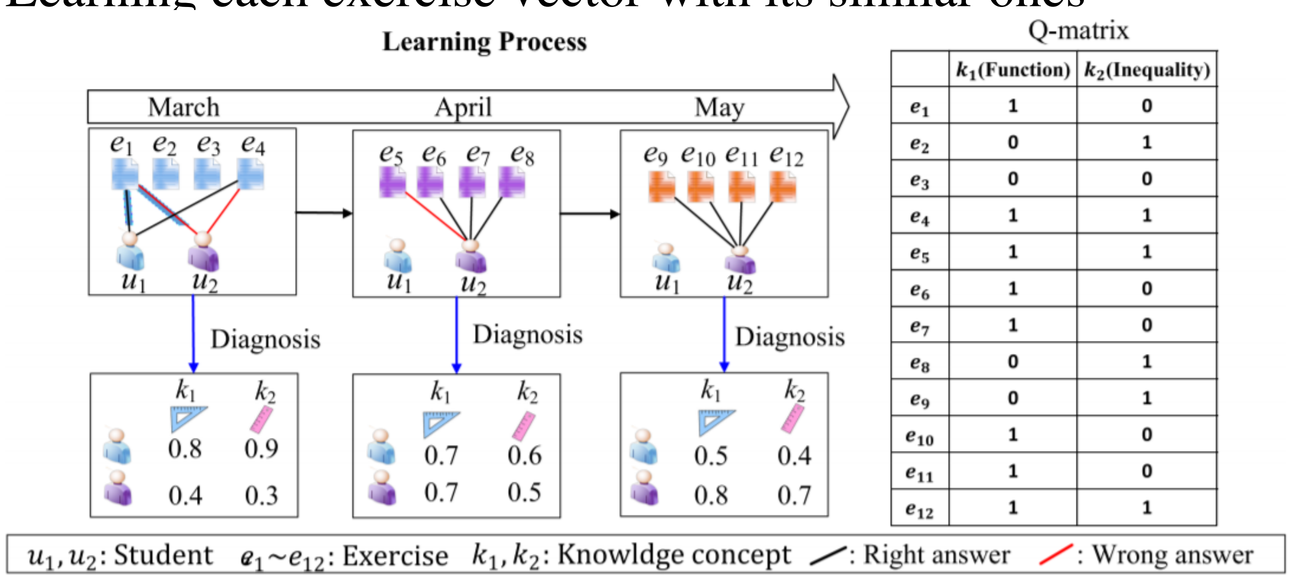


$$F_{ik}^t(*) = U_{ik}^{t-1} e^{-\frac{\boxed{\Delta t}}{S}},$$

Time interval

EKPT model

- Sparsity problem
  - Students practice very few exercises compared with the huge exercise space
  - Inaccurate if students just practices few exercises at each time
- EKPT model
  - Exercise connectivity assumption
    - Students may get consistent scores on these knowledge-based exercises
    - Learning each exercise vector with its similar ones



	k_1(Function)	k_2(Inequality)
e_1	1	0
e_2	0	1
e_3	0	0
e_4	1	1
e_5	1	1
e_6	1	0
e_7	1	0
e_8	0	1
e_9	0	1
e_{10}	1	0
e_{11}	1	0
e_{12}	1	1

EKPT model

- EKPT model
  - Modeling V with exercise connectivity
    - For exercise j, we define a neighbor set

$$N_{V_j} = \{l | k \in j \cap l, l \in V, k \in K\}$$

    - The knowledge vector of exercise j is influenced by the set:

$$V_j = \sum_{l \in N_{V_j}} \boxed{w(j,l)} \times V_l + \theta_V, \theta_V \sim \mathcal{N}\left(0, \sigma_V^2\right).$$

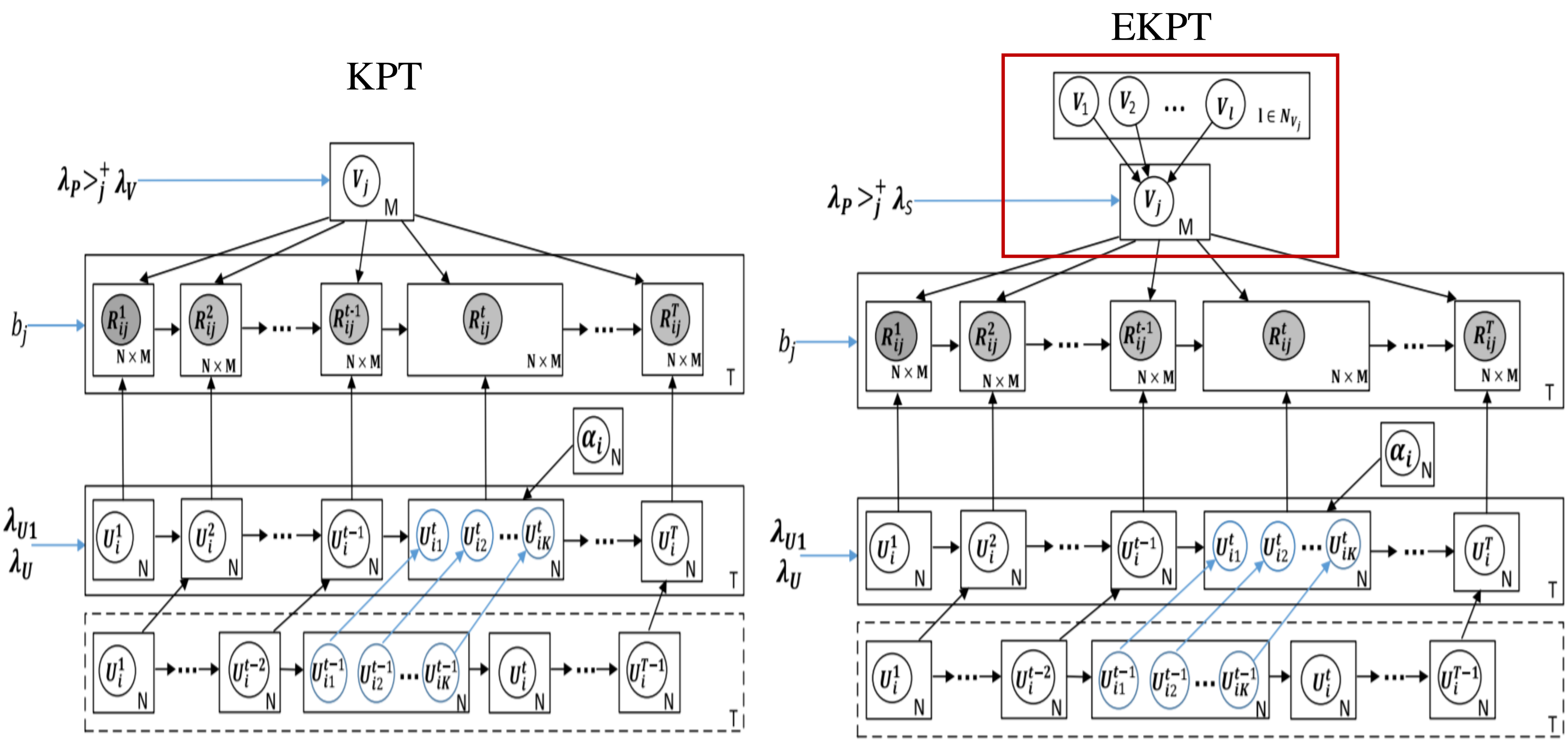
    - $w(j,l)$ is the weight influence, which can be any weight function, like

$$V_j = \boxed{\frac{1}{|N_{V_j}|}} \sum_{l \in N_{V_j}} V_l + \theta_V, \theta_V \sim \mathcal{N}\left(0, \sigma_V^2\right).$$

Equal contribution for all neighbor exercicses

Model

➢ Model Comparasion



Model

➢ Model Learning

KPT

$$\min_{\Phi} \mathcal{E}(\Phi) = \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}^{t}\left(\hat{R}_{ij}^{t} - R_{ij}^{t}\right)^2 - \lambda_P \sum_{j=1}^{M}\sum_{q=1}^{K}\sum_{p=1}^{K} I\left(q >_j^+ p\right) \ln \frac{1}{1+e^{-(V_{jq}-V_{jp})}} + \frac{\lambda_V}{2}\sum_{j=1}^{M} ||V_j||_F^2 + \frac{\lambda_U}{2}\sum_{t=2}^{T}\sum_{i=1}^{N} ||\overline{U_i^t} - U_i^t||_F^2 + \frac{\lambda_{U1}}{2}\sum_{i=1}^{N} ||U_i^1||_F^2,$$

ALGORITHM 1: Parameter Learning of the KPT Model

Initialize U, V, α and b;
while *not converged* **do**
  for $i = 1, 2, \ldots N$ **do**
    for $t = 1, 2, \ldots, T$ **do**
      for $k = 1, 2, \ldots, K$ **do**
        Fix V, α, b, update U_{ik}^t by Equation (15) using SGD;
    Fix U, V, b, update α_i by Equation (17) and Equation (19) using PG;
  for $j = 1, 2, \ldots, M$ **do**
    for $k = 1, 2, \ldots, K$ **do**
      Fix U, α, b, update V_{jk} by Equation (16) using SGD;
    Fix U, V, α, update b by Equation (18) using SGD;
Return U, V, α and b;

EKPT

$$\min_{\Phi} \mathcal{E}(\Phi) = \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}^{t}\left(\hat{R}_{ij}^{t} - R_{ij}^{t}\right)^2 - \lambda_P \sum_{j=1}^{M}\sum_{q=1}^{K}\sum_{p=1}^{K} I\left(q >_j^+ p\right) \ln \frac{1}{1+e^{-(V_{jq}-V_{jp})}} + \frac{\lambda_S}{2}\sum_{j=1}^{M} ||V_j - \frac{1}{|N_{V_j}|}\sum_{l \in N_{V_j}} V_l||_F^2 + \frac{\lambda_U}{2}\sum_{t=2}^{T}\sum_{i=1}^{N} ||\overline{U_i^t} - U_i^t||_F^2 + \frac{\lambda_{U1}}{2}\sum_{i=1}^{N} ||U_i^1||_F^2,$$

ALGORITHM 2: Parameter Learning of the EKPT Model

Initialize U, V, α and b;
while *not converged* **do**
  for $i = 1, 2, \ldots N$ **do**
    for $t = 1, 2, \ldots, T$ **do**
      for $k = 1, 2, \ldots, K$ **do**
        Fix V, α, b, update U_{ik}^t by Equation (15) using SGD;
    Fix U, V, b, update α_i by Equation (17) and Equation (19) using PG ;
  for $j = 1, 2, \ldots, M$ **do**
    for $k = 1, 2, \ldots, K$ **do**
      Fix U, α, b, update V_{jk} by Equation (25) using SGD;
    Fix U, V, α, update b by Equation (18) using SGD;
Return U, V, α and b;

Model

- Application
  - Knowledge Proficiency Estimation

$$\hat{U}_i^{(T+1)} = \left\{\hat{U}_{i1}^{(T+1)}, \hat{U}_{i2}^{(T+1)}, \ldots, \hat{U}_{iK}^{(T+1)}\right\},$$

$$\hat{U}_{ik}^{(T+1)} \approx \alpha_i U_{ik}^T \frac{D f_{ik}^{T+1}}{f_{ik}^{T+1} + r} + (1-\alpha_i) U_{ik}^T \mathrm{e}^{-\frac{\Delta(T+1)}{S}},$$

  - Student Performance Prediction

$$\hat{R}_{ij}^{(T+1)} \approx \left\langle U_i^{(T+1)}, V_j \right\rangle - b_j. \qquad \hat{R}_{ij}^{(T+1)} = \begin{cases} \hat{R}_{ij}^{(T+1)} & \text{if} \quad 0 \le \hat{R}_{ij}^{(T+1)} \le 1, \\ 0 & \text{if} \quad \hat{R}_{ij}^{(T+1)} < 0, \\ 1 & \text{if} \quad \hat{R}_{ij}^{(T+1)} > 1. \end{cases}$$

  - Diagnosis results explanation and visualization

Outline

Experiment

➢ Dataset

sparse

Dataset	Math1	Math2	Assist	Adaptive
Training logs	521,248	347,424	263,327	229,848
Testing logs	74,464	18,312	43,888	38,308
# of students	9,308	1,306	7197	3,217
# of exercises	64	280	3211	411
# of time windows	4	10	7	7
# of knowledge concepts	12	13	20	12
Avg. knowledge concepts per exercise	1.15	1.3215	1.5073	1.06

➢ Baseline

	Model	Data Source				Application			Dynamic Explanation?
		Q-matrix	Multi-Skill	Repeating	Time	Knowledge Estimation	Score Prediction	Visualization	
Static models	IRT [17]	×	×	×	×	×	√	×	×
	DINA [15]	√	√	×	×	√	√	√	×
	PMF [63]	×	×	×	×	×	√	×	×
Dynamic models	BKT [31]	√	×	√	√	√	√	√	√
	LFA [9]	√	√	√	√	×	√	×	√
	DKT [52]	×	√	√	√	×	√	×	√
Variants	QMIRT	√	√	×	×	√	√	√	×
	QPMF	√	√	×	×	√	√	√	×
Ours	**KPT**	√	√	×	√	√	√	√	√
	EKPT	√	√	×	√	√	√	√	√

Experiment

- Knowledge Proficiency Estimation
  - **DOA**: if **a** masters better than **b** on a **concept k** at time T, then **a** will have a higher probability to get **correct** answers to the **exercises** related to **concept k** than **b** at time **T**

$$DOA(k) = \sum_{j=1}^{M} I_{jk} \sum_{a=1}^{N} \sum_{b=1}^{N} \frac{\delta\left(U_{ak}^{T+1}, U_{bk}^{T+1}\right) \cap \delta\left(R_{aj}^{T+1}, R_{bj}^{T+1}\right)}{\delta\left(U_{ak}^{T+1}, U_{bk}^{T+1}\right)},$$

(a) Math1

K	Models					
	EKPT	KPT	QPMF	QMIRT	DINA	BKT
K1	**0.807**	0.798	0.565	0.595	0.524	0.558
K2	**0.751**	0.733	0.576	0.621	0.473	0.623
K3	**0.830**	0.827	0.614	0.629	0.497	0.523
K4	**0.769**	0.752	0.581	0.675	0.486	0.565
K5	**0.799**	0.791	0.559	0.723	0.476	0.578
K6	**0.844**	0.838	0.730	0.766	0.485	0.628
K7	**0.851**	0.842	0.697	0.634	0.520	0.697
K8	**0.799**	0.784	0.699	0.657	0.498	0.617
K9	**0.796**	0.771	0.609	0.712	0.501	0.645
K10	0.813	**0.834**	0.597	0.515	0.489	0.503
K11	**0.796**	0.786	0.608	0.631	0.478	0.617
K12	0.811	**0.842**	0.532	0.641	0.523	0.645
Avg	**0.806**	0.799	0.614	0.650	0.496	0.601

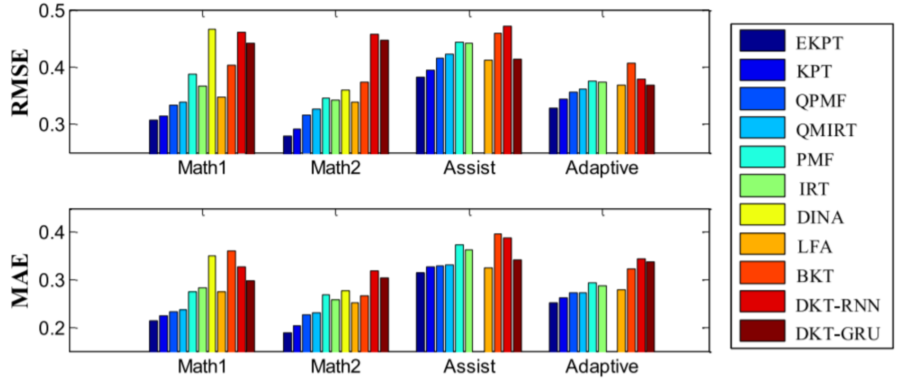
(d) Adaptive

K	Models				
	EKPT	KPT	QPMF	QMIRT	BKT
K1	**0.742**	0.732	0.656	0.645	0.578
K2	**0.799**	0.780	0.756	0.740	0.609
K3	**0.796**	0.793	0.752	0.736	0.592
K4	**0.804**	0.802	0.737	0.638	0.679
K5	**0.812**	0.808	0.597	0.632	0.552
K6	**0.818**	0.812	0.659	0.648	0.547
K7	**0.821**	0.815	0.587	0.668	0.687
K8	**0.824**	0.818	0.624	0.591	0.532
K9	**0.824**	0.809	0.704	0.692	0.645
K10	**0.823**	0.819	0.730	0.776	0.732
K11	**0.830**	0.820	0.658	0.685	0.702
K12	**0.809**	0.792	0.709	0.693	0.690
Avg	**0.809**	0.801	0.681	0.679	0.629

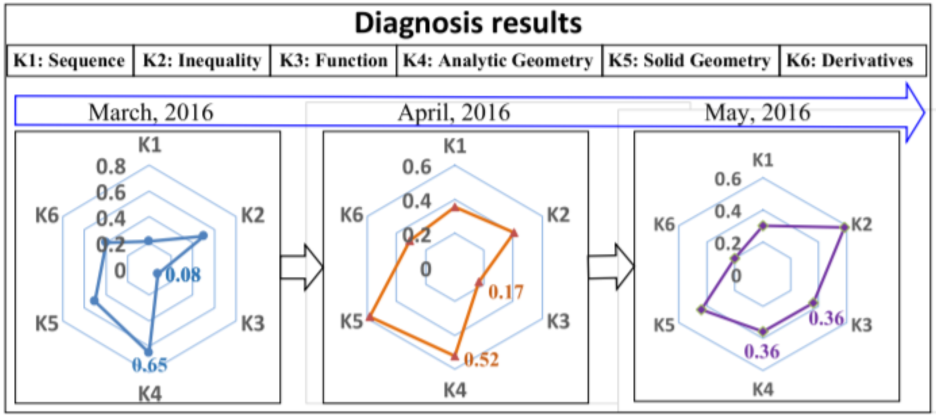
  - Our models perform better than baselines
  - EKPT is better than KPT on sparse dataset

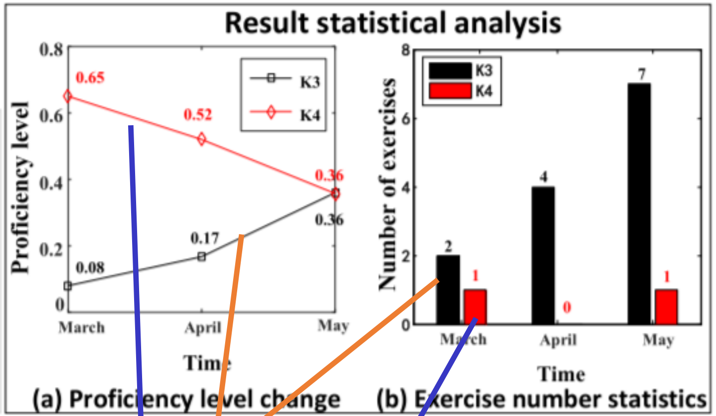
Experiment

- Student Performance Prediction
  - **MAE, RMSE**



  - Dynamic models are better than static ones
  - Deep learning based models (DKT) perform not very good
    - Possible: Time is not longer enough, Data volume may not support
- Diagnosis results visualization





  - The student practices many times on K3, knowledge proficiency increases
  - The student practices very few exercises on K4, she may forget what she have learned

Experiment

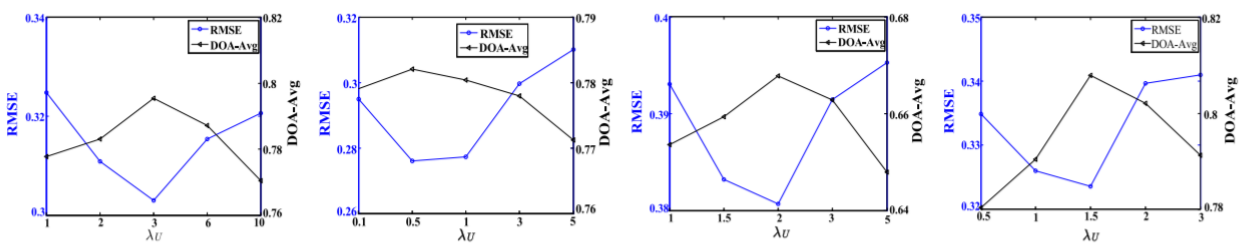
➢ Model Analysis

➢ Computational Performance

- ➢ Though our model needs more time for training, they are competitive compared with DKT (deep learning based ones)

Dataset	Time	Stastic Models			Dynamic Models				Variants		Our Models	
		IRT	DINA	PMF	BKT	LFA	DKT (RNN)	DKT (GRU)	QMIRT	QPMF	KPT	EKPT
Math1	Each	0.022	0.316	0.023	/	0.024	0.403	0.479	0.036	0.025	0.083	0.101
	Total	1.960	18.05	1.833	1.516	2.483	22.867	195.375	3.647	2.535	8.334	11.66
Math2	Each	0.011	0.616	0.021	/	0.012	0.122	0.157	0.016	0.012	0.067	0.073
	Total	1.051	57.28	1.283	0.581	1.152	7.720	10.435	1.603	1.589	7.334	7.738
Assist	Each	0.015	/	0.033	/	0.026	1.594	3.207	0.283	0.265	0.467	0.735
	Total	2.320	/	4.951	1.275	2.991	73.324	147.522	26.38	29.94	47.13	77.15
Adaptive	Each	0.013	/	0.029	/	0.015	0.273	0.338	0.105	0.110	0.233	0.453
	Total	2.154	/	3.466	1.017	1.942	11.734	12.522	8.412	10.45	24.73	48.92

➢ Parameter sensitivity

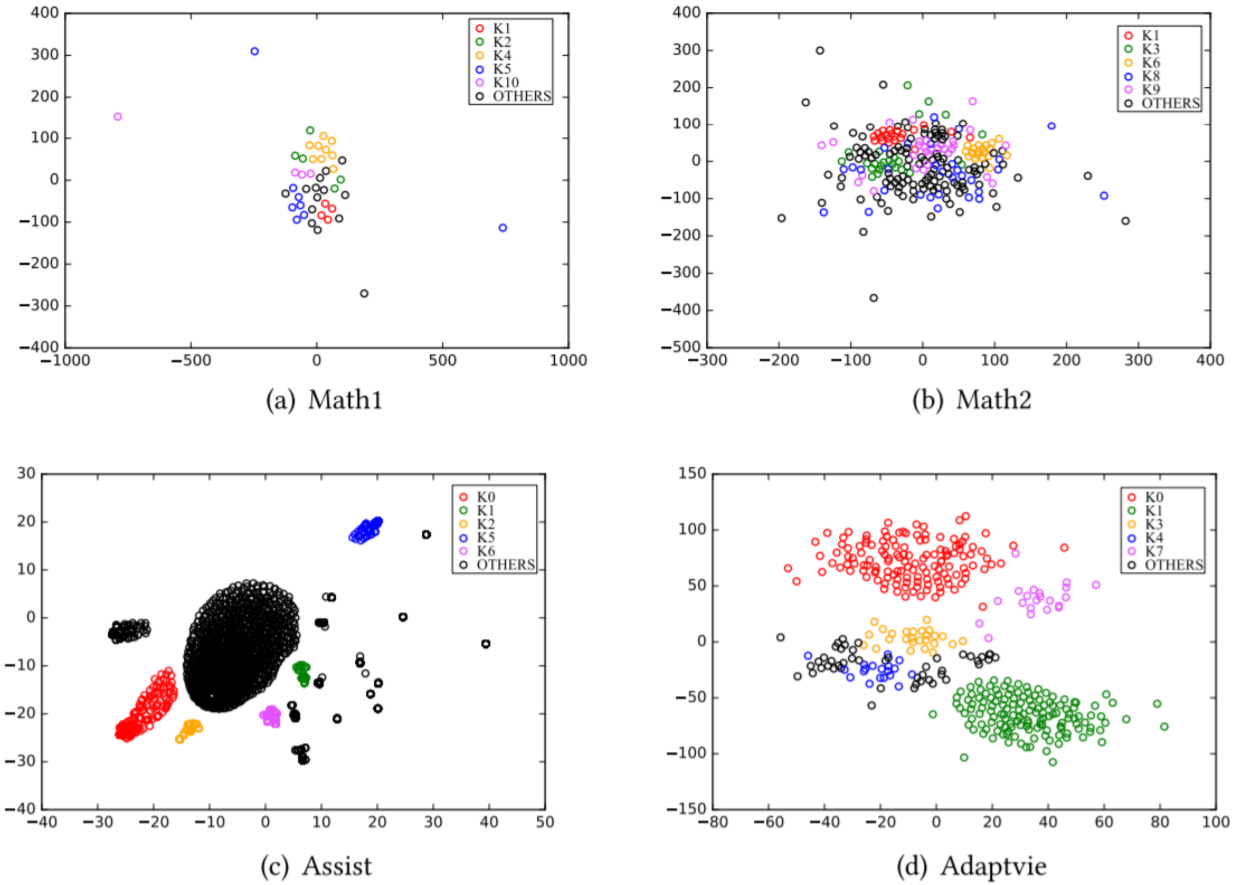


(a) Math1 (b) Math2 (c) Assist (d) Adaptive

Fig. 8. The impact of λ_U on four datasets.

Experiment

- Model Analysis
  - Exercise relationship
    - Exercise with same concepts are grouped together



(a) Math1 (b) Math2

(c) Assist (d) Adaptvie

Outline

Conclusion & Future work

- Conclusion
  - A focused study on tracking the knowledge proficiency of students
  - Two explanatory probabilistic models considering different educational factors
    - Incorporating learning theories for explaining the knowledge change
    - Incorporating Q-matrix for improving the interpretability
    - Incorporating exercise connectivity property to address sparsity problem
  - Experiments on different datasets show the both effectiveness and explanatory power of our models

- Future work
  - Consider different specific modeling for learning and forgetting factors
  - Consider student behaviors and social connections for more precise diagnosis
  - Consider different learning scenarios
    - Game
    - Multiple-attempt response
    - Repeated learning

Thanks for your listening!

huangzhy@mail.ustc.edu.cn