Tracking Knowledge Proficiency of Students with Educational Priors

Yuying Chen[1], Qi Liu[1], Zhenya Huang[1], Le Wu[2], Enhong Chen[1*],
Runze Wu[1], Yu Su[3], Guoping Hu[4]
[1]Anhui Province Key Laboratory of Big Data Analysis and Application, School of Computer Science and Technology, University of Science and Technology of China
{cyy33222,huangzhy,wurz1990}@mail.ustc.edu.cn,{qiliuql,cheneh}@ustc.edu.cn
[2]Hefei University of Technology, Hefei, Anhui 230009, China, lewu@hfut.edu.cn
[3]School of Computer Science and Technology, Anhui University, yusu@iflytek.com
[4]iFLYTEK Research, gphu@iflytek.com

ABSTRACT

Diagnosing students' knowledge proficiency, i.e., the mastery degrees of a particular knowledge point in exercises, is a crucial issue for numerous educational applications, e.g., targeted knowledge training and exercise recommendation. Educational theories have converged that students learn and forget knowledge from time to time. Thus, it is necessary to track their mastery of knowledge over time. However, traditional methods in this area either ignored the explanatory power of the diagnosis results on knowledge points or relied on a static assumption. To this end, in this paper, we devise an *explanatory* probabilistic approach to *track* the knowledge proficiency of students over time by leveraging educational priors. Specifically, we first associate each exercise with a knowledge vector in which each element represents an explicit knowledge point by leveraging educational priors (i.e., *Q-matrix*). Correspondingly, each student is represented as a knowledge vector at each time in a same knowledge space. Second, given the student knowledge vector over time, we borrow two classical educational theories (i.e., *Learning curve* and *Forgetting curve*) as priors to capture the change of each student's proficiency over time. After that, we design a probabilistic matrix factorization framework by combining student and exercise priors for tracking student knowledge proficiency. Extensive experiments on three real-world datasets demonstrate both the effectiveness and explanatory power of our proposed model.

KEYWORDS

Knowledge Diagnosis; Dynamic Modeling; Educational Priors; Explanatory Power

*Corresponding author.

CIKM'17 , November 6–10, 2017, Singapore, Singapore

ACM ISBN 978-1-4503-4918-5/17/11…$15.00
https://doi.org/10.1145/3132847.3132929

1 INTRODUCTION

Real-world education service systems, such as intelligent tutoring systems, allow students to learn and do exercises individually. Their conveniences and rapid developments have caused an increasing attention of educators and publics [1, 5].

A key issue in such systems is the Knowledge Proficiency Diagnosis (KPD) of students, i.e., to discover the latent mastery degrees of students on each knowledge point [32]. Figure 1 shows a toy example for this KPD task. From the figure, there are two students (u_1 and u_2) doing different mathematical exercises from March to May 2016. Each exercise contains different knowledge points, which can be represented as a *Q-matrix* provided by educational experts [8]. Specifically, the number 1 in *Q-matrix* denotes that the corresponding exercise contains the knowledge point and 0 otherwise. As shown in this figure, exercise e_1 contains a knowledge point *Function*, and exercise e_9 is related to knowledge points *Function* and *Inequality*. The KPD task in educational area asks that: Given students' historical exercise records and the provided Q-matrix, how to diagnose the mastery degrees of students to the knowledge points (i.e., *Function* and *Inequality* in Figure 1)? In fact, as these diagnosis results are beneficial to numerous applications, such as targeted knowledge training [12] and personalized exercise recommendation [26], many efforts have been devoted to this KPD task. On one hand, cognitive diagnosis models from the educational psychology area usually characterize students' knowledge proficiency by a latent trait value [11] or a binary skill mastery vector [8]. On the other hand, by treating the KPD task as a data mining problem (i.e., student score prediction), matrix factorization techniques project each student in a latent space that depicts students' implicit knowledge states [16]. In summary, these two research directions usually model users' historical records without any temporal information, thus, they are good at predicting student's proficiency from a static perspective.

However, educational psychologists have long converged that the learning process of students is not static but evolves over time [29]. They claims that students learn and forget the knowledge from time to time. Two classical educational theories can well explain this dynamic change: the *Learning curve theory* argues that students can enhance their knowledge proficiency with constant trails or exercises [2] and the

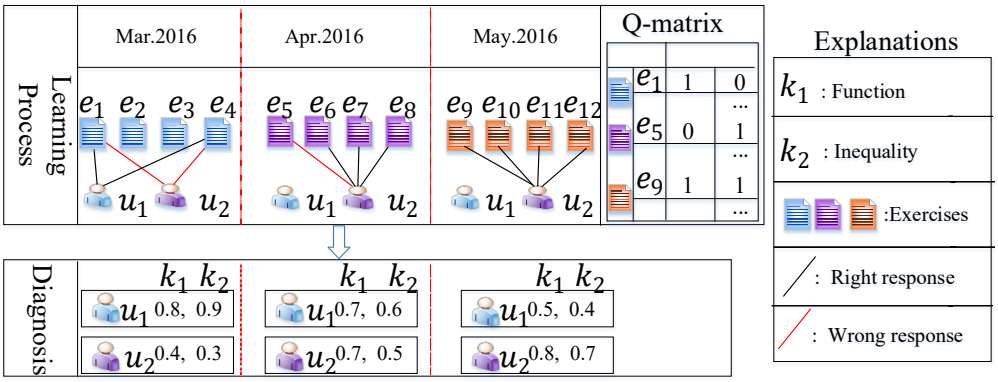


Figure 1: A showcase of KPD task on mathematical exercises related to the knowledge points of *Function* and *Inequality* from March to May, 2016. The left area contains two parts: the top part shows the student learning process with different exercises, and a related Q-matrix that depicts the knowledge points of the exercises. Specifically, each row of the Q-matrix denotes an exercise and each column stands for a knowledge point. The bottom part shows the corresponding diagnosis results of the two students to these two knowledge points over time.

Ebbinghaus forgetting curve theory indicates that students' knowledge proficiency may decline due to the forgetting factor [10]. Let us take the two students in Figure 1 as an example. As time goes on, student u_2 improved her proficiency on both two knowledge points with some exercises she took and learned. Therefore, she should focus on some exercises about new knowledge points. In contrast, student u_1's knowledge proficiency decreased with the possible reason of the forgetting factor. Thus, a timely review is necessary to reinforce these two knowledge points for u_1. Therefore, it is necessary to leverage these two educational theories to track students' knowledge proficiency over time.

In fact, several research works from both the cognitive diagnosis area [6, 7, 20, 21, 34] and the data mining community [27, 33, 35] have already attempted to solve KPD task from a dynamic perspective. And the experimental results empirically showed the superiority of adding the temporal information for this task. However, there are still some questions to answer. Specifically, the data mining based methods, such as the TF model [33], only capture the latent factors of students over time, thus are hard to explain in practice. We argue that the explicit knowledge point explanation is especially important in the KPD task. With explicit knowledge point meaning, students can quantitatively measure the strengthes and weaknesses of themselves for self-improvement (e.g., targeted exercise training [12]). In addition, in the cognitive models, such as Baysian Knowlege Tracing based models [7, 20, 34], the learning and forgetting factors are viewed as additional parameters, which neglects that these two factors are closely related to the exercises they do at each time as suggested by educational experts. Thus, these cognitive models could not answer the question of how these educational theory could help to explain the evolution of students' knowledge proficiency over time. Therefore, few of the existing approaches can address the following challenge: how to embed the educational priors in the KPD modeling task to better *explain* and *predict* the KPD task of students to knowledge points?

To solve the challenge mentioned above, in this paper, we propose an explanatory probabilistic *Knowledge Proficiency Tracing* (KPT) model to track the KPD task of students over time by leveraging educational priors. Specifically, we first associate each exercise with a knowledge vector, in which each element represents an explicit knowledge point. The Q-matrix that is marked by educational experts to depict the relationship between knowledge points and exercises, is exploited as priors to generate exercise representations. To track students' knowledge proficiency, each student is represented as a knowledge vector at each time in the same knowledge space. Then we embed the classical educational theories (i.e., *Learning curve* and *Forgetting curve*) as priors to capture the change of each student's proficiency over time. After that, we design a probabilistic matrix factorization framework by combining student and exercise priors. Thus, the proposed model can *track* and *explain* students' knowledge proficiency over time. Finally, the experimental results clearly validate both the effectiveness and explanatory power of our proposed KPT model. To best of our knowledge, this is the first comprehensive attempt to incorporate three educational priors (*Q-matrix*, *Learning curve* and *Forgetting curve*) into a probabilistic matrix factorization framework for tracking KPD task with both precise and explanatory power.

2 RELATED WORK

Generally, we summarize the related work of our research as the following three categories.

The first category is student modeling [31, 35] in data mining area, with the goal to learn students' latent representations from their exercise. These learned representations could be applied to applications, such as score prediction [30]. Usually, we also regard the obtained representations of students as their implicit knowledge proficiency. There are two types of representative techniques: factorization models [16] and neural networks [22]. For instance, Thai-Nghe et al. [27] leveraged matrix factorization models to map each student into a latent vector that depicts students' implicit knowledge

Table 1: A toy example of exercise logs.

StudentId	ExerciseId	UpdateTime	Score
S_1	E_1	T_1	1
S_1	E_2	T_2	0.25
S_2	E_2	T_3	0
S_2	E_3	T_4	1
S_2	E_3	T_5	0.75
S_5	E_2	T_6	1
...	...	...	...

states. In order to track changes of student learning process, Thai-Nghe et al. [27] and Xiong et al. [33] proposed tensor factorization approaches by incorporating additional time dimensions for KPD over time. Recently, through establishing a bridge between knowledge points and neurons, Piech et al. [22] developed a recurrent neural network based approach to model student learning process, which improved the performance of score prediction task. Nevertheless, a common limitation of these works is that these models operate like a black box, thus the output of the student representations are hard to explain. That is to say, neither the latent vectors from factorization models nor the hidden layers from neural networks can correspond to any explicit knowledge point. In contrast, our model improves the traditional matrix factorization by leveraging educational priors (i.e., *Q-matrix*, *Learning curve* and *Forgetting curve*), which guarantees the explanatory power.

The second direction is KPD research in educational cognitive area, which aims at discovering the proficiency of students on defined knowledge points [9, 14]. Widely-used approaches could be divided into two aspects: one-dimensional models and multi-dimensional models. Among them, Item Response Theory (IRT), as a typical one-dimensional model, considered each student as a single proficiency variable (i.e., latent trait) [11]. Comparatively, multi-dimensional models, such as *Deterministic Inputs, Noisy-And gate model*, characterized students by a binary latent vector which described whether or not she mastered the knowledge points with the given Q-matrix prior [8]. Furthermore, Wu et al. [32] proposed FuzzyCDM to quantitatively diagnose student knowledge proficiency. However, to the best of our knowledge, all these methods rely on static assumption and ignore temporal factor for KPD task. In this work, we focus on the dynamic learning process of students and capture the change of each student's knowledge proficiency over time.

In order to explain the dynamic changes of students' knowledge proficiency during their learning process, educational psychologists have converged two classical theories: *Learning curve theory* argues that students can enhance their knowledge proficiency with constant trails or exercises [2] and *Ebbinghaus forgetting curve theory* indicates that students' knowledge proficiency may decline as time goes on [10]. Based on these two prior theories, researchers have attempted to develop a series of models for solving KPD task from an evolving perspective. For example, some IRT based models, such as Learning Factors Analysis [6] and Performance Factors Analysis [21], were proposed to improve traditional IRT, which assumed that students shared the same parameters of learning rate when exercising. Furthermore, Wang et al. [29] proposed a time-series IRT model to estimate a dynamic latent trait of each student. In addition, researchers proposed variations of Bayesian Knowledge Tracing (BKT) based models [7, 15, 20, 34] to capture the change of students' knowledge proficiency over time. Despite the importance of these efforts, there are still some limitations in practice: First, IRT based models only estimate a variable (e.g., latent trait) for each student so that they cannot discover students proficiency on multiple knowledge points simultaneously (i.e., KPD task for two knowledge points in Figure 1). Second, BKT based models focus on a simplified learning scenario where students are allowed to keep doing the same exercise while overlooking a more practical one in Figure 1 where students just do each exercise only once. In most cases, students seldom repeat doing the same exercises but seek more different exercises for learning. Last but not least, both IRT and BKT based models neglect the influence of exercises for the learning and forgetting factors directly, thus is hard to explain students' knowledge evolution over time.

Table 2: Some important notations.

Notation	Description
N	the total number of students
M	the total number of exercises
T	the total number of time windows
K	the total number of knowledge points
R_{ij}^t	the response of Student i on Exercise j in time window t
U_i^t	the knowledge proficiency of Student i in time window t
V_j	the correlation level of Exercise j on each knowledge point
b_j	the difficulty of Exercise j
α_i	the balance parameter of Student i

Based on the learning scenario that most students do each exercise only once, in this paper, we aim to *track* and *explain* students knowledge proficiency on multiple knowledge points leveraging by underlying theories (i.e., *Q-matrix*, *Learning curve* and *Forgetting curve*).

3 KNOWLEDGE PROFICIENCY TRACKING MODEL

In this section, we first formally introduce the KPD task and our study overview. Then we introduce the technical details of our proposed model KPT. At last, we specify parameter learning and prediction of KPT.

3.1 Problem and Study Overview

Suppose there are N students, M exercises and K knowledge points in a learning system where students do exercises at different times recorded by students' exercise logs (as shown in Table 1). Specifically, the students' response logs can be represented as a response tensor $R \in \mathbb{R}^{N \times M \times T}$. If student i does exercise j at time t, R_{ij}^t denotes i's score of exercise j. In addition, we also have a Q-matrix provided by educational experts, which can be represented as a binary knowledge matrix $Q \in \mathbb{R}^{M \times K}$. If exercise j relates to knowledge point k, $Q_{jk} = 1$ and vice versa. It is worth mentioning that at different time, most students do the same exercises only once because they usually choose different exercises to learn a

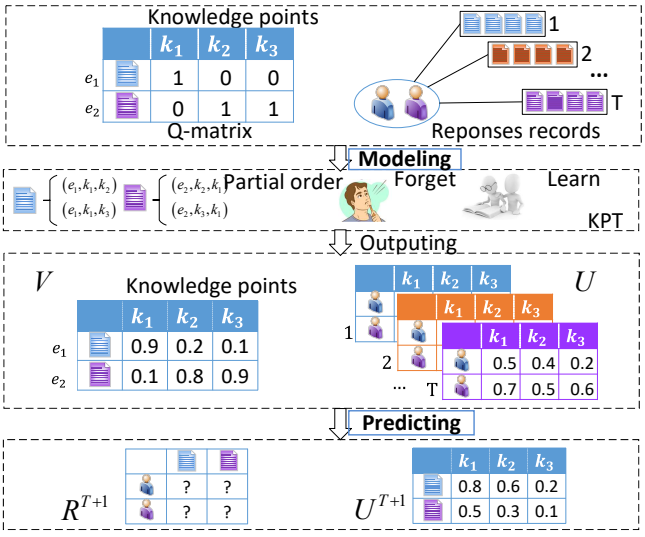


Figure 2: The framework of the KPT model.

specific knowledge point in general cases. Without loss of generality, the problem can be formulated as:

(PROBLEM FORMULATION) *Given the students' response tensor R and Q-matrix labeled by educational experts, our goal is two-fold: 1) modeling the change of students' knowledge proficiency from time 1 to T; 2) predicting students' knowledge proficiency and responses at time $T+1$.*

As shown in Figure 2, our solution is a two-stage framework, which contains a modeling stage and a predicting stage: 1) In modeling stage, given exercise response logs of students (Table 1) and Q-matrix labeled by experts, we first project each student's latent vector into a knowledge space with the help of the Q matrix prior provided by educational experts. Then, we propose KPT to address KPD of students over time by incorporating the *Learning* and *Forgetting curve* theories. After that, we can obtain students' knowledge proficiency U at different times and each exercise's knowledge vector V. 2) In predicting stage, KPT predicts students' responses (R^{T+1}) and knowledge proficiency (U^{T+1}) in the future.

In the following, we will specify the probabilistic modeling, parameter learning and prediction of KPT. For better illustration, some notations are summarized in Table 2.

3.2 Probabilistic Modeling with Priors

Inspired by many existing works [25, 27], for each student and each exercise, we model the response tensor R as:

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

where $\mathcal{N}(\mu,\sigma^2)$ is a Gaussian distribution with mean μ and variance σ^2. I is an indicator tensor and I_{ij}^t equals to 1 if student i does exercise j in time window t, and vice versa. $U_i^t \in \mathbb{R}^{K\times 1}$ is the knowledge proficiency of student i in time window t. $V \in \mathbb{R}^{M\times K}$ denotes the relationship between exercises and knowledge points. b_j is the difficulty bias of exercise j, which is widely adopted in KPT task modeling [11]. Given this likelihood function, in the following, we would detail how to incorporate the educational priors in the modeling process. We first explain how to embed the knowledge of the Q matrix to model V. Specifically, we incorporate Q-matrix prior to associate each exercise with a knowledge vector in which each element represents an explicit knowledge point. Then we model U by combining two educational theories (i.e., forgetting and learning) as priors to track students' dynamic learning process.

Modeling V with the Q-matrix prior. Traditional probabilistic matrix factorization models suffer from the interpretation problem as the learned latent dimensions are unexplainable. Comparatively, many efforts in educational field have been made to build an interpretative model by leveraging the prior knowledge based on Q-matrix. However, such traditional Q-matrix has two disadvantages: 1) inevitable error or subjective bias from manual labeling [18]; 2) the sparsity with the binary entries which does not fit probabilistic modeling well. To mitigate these existing issues, we refine and utilize a partial order [23] based on Q-matrix to reduce the subjective impact of experts and associate each exercise with sets of knowledge points. As for exercise j, the partial order $>_j^+$ can be defined as:

$$q >_j^+ p, \text{ if } Q_{jq} = 1 \text{ and } Q_{jp} = 0. \quad (2)$$

Specifically, for exercise j, if a knowledge point q is marked as 1, then we assume that q is more relevant to exercise j than all the other knowledge points with mark 0. Please note that we cannot infer comparability of knowledge points with the same mark. After that, we can transform the original Q-matrix into a set of comparability $D_T \in R^{M\times K\times K}$ by:

$$D_T = \{(j,q,p)|q >_j^+ p\}. \quad (3)$$

Thus, D_T is not as sparse as Q-matrix and can capture more accurate pairwise relationship between two knowledge points (q,p) based on an exercise j with a good interpretation. We learn the latent exercise matrix $V \in \mathbb{R}^{M\times K}$ by incorporating this prior partial order. The Bayesian formulation of finding the correct partial order for all pairs of knowledge points (q,p) turns to maximize the following posterior probability:

$$p(V|D_T) \propto p(D_T|V) \times p(V). \quad (4)$$

All exercises are presumed to be marked independently by educational experts. We also assume the ordering of each pair [23] of knowledge points (q,p) for a specific exercise is independent of the ordering of every other pair. Hence, the likelihood function $p(D_T|V)$ can be given as follows:

$$p(D_T|V) = \prod_{(j,q,p)\in D_T} p(q >_j^+ p|V_j). \quad (5)$$

In order to get a correct partial order relation on V, we define the probability that exercise j is more relevant to knowledge point q than knowledge point p as:

$$p(q >_j^+ p|V_j) = \frac{1}{1+\mathrm{e}^{-(V_{jq}-V_{jp})}}. \quad (6)$$

Besides, following the traditional Bayesian treatment, we also assume V follows a zero-mean Gaussian prior. Combining Eq. (4), (5) and (6), we can formulate the log posterior distribution over D_T on V as:

$$\ln p(V|D_T) = \ln \prod_{(j,q,p)\in D_T} p(>_j^+ |V)p(V)$$
$$= \sum_{j=1}^{M}\sum_{q=1}^{K}\sum_{p=1}^{K} I\left(q >_j^+ p\right) \ln \frac{1}{1+\mathrm{e}^{-\left(V_{jq}-V_{jp}\right)}}$$
$$- \frac{1}{2\sigma_V^2}||V||_F^2. \tag{7}$$

Modeling U with two dynamic learning theories. Now we specify the modeling of students' latent tensor U. As mentioned before, during students' dynamic learning process, there are two widely accepted theories in educational psychology that could guide us in the modeling process: 1) *Learning curve.* [2] depicts the knowledge we learned can be enhanced with several exercises. 2) *Ebbinghaus forgetting curve* [28] hypothesizes the knowledge we learned will be gradually forgotten over time.

Combining the two theories as priors, we assume a student's current knowledge proficiency is mainly influenced by two underlying reasons: 1) The more exercises she does, the higher level of related knowledge proficiency she will get. 2) The longer the time passes, the more knowledge she will forget. Formally, we model two effects of each student's knowledge proficiency at time window $t = 2, 3, ..., T$ as:

$$p(U_i^t) = \mathcal{N}(U_i^t|\bar{U}_i^t, \sigma_U^2\mathbf{I}), \text{where} \quad \bar{U}_i^t = \left\{\bar{U}_{i1}^t, \bar{U}_{i2}^t ... \bar{U}_{iK}^t\right\}$$
$$\bar{U}_{ik}^t = \alpha_i l^t(*) + (1-\alpha_i) f^t(*), s.t. \quad 0 \le \alpha_i \le 1, \tag{8}$$

where $U_i^t \in \mathbb{R}^{K\times 1}$, the knowledge proficiency of student i in time window t, follows a Gaussian distribution with mean $\bar{U}_i^t$ and variance $\sigma_U^2\mathbf{I}$. U_{ik}^t is student i's knowledge proficiency on knowledge point k at time t. $l^t(*)$ is the learning factor which means the learned knowledge at time t after several exercises and $f^t(*)$ is the forgetting factor which indicates the remaining knowledge at time t. α_i balances the two factors to capture the students' learning characteristics. Intuitively, if student i has a large a_i, she may be diligent. Thus $l^t(*)$, instead of $f^t(*)$, affects her future knowledge proficiency more significantly, and vice versa. In the following, we formally define $l^t(*)$ and $f^t(*)$.

$l^t(*)$ captures the growth of knowledge with exercises:

$$l^t(*) = U_{ik}^{t-1}\frac{D * f_k^t}{f_k^t + r}, \tag{9}$$

where f_k^t denotes the frequency of knowledge k examined in time window t. r and D are two hyper-parameters, which control the magnitude and multiplier of growth respectively.

$f^t(*)$ depicts the decline of knowledge over time:

$$f^t(*) = U_{ik}^{t-1}\mathrm{e}^{-\frac{\Delta t}{S}}, \tag{10}$$

where Δt is the time interval between time window $t-1$ and time window t, S is a hyper-parameter that denotes the strength of memory.

At the initial time $t = 1$, we do not know the initial level of each student. Therefore, we assume a zero-mean Gaussian distribution of student's knowledge proficiency at that time. Then we summarize the prior over user latent tensor as:

$$p(U|\sigma_U^2, \sigma_{U1}^2) = \prod_{i=1}^{N}\mathcal{N}(U_i^1|0, \sigma_{U1}^2\mathbf{I}) \prod_{t=2}^{T}\mathcal{N}(U_i^t|\overline{U_i^t}, \sigma_U^2\mathbf{I}). \tag{11}$$

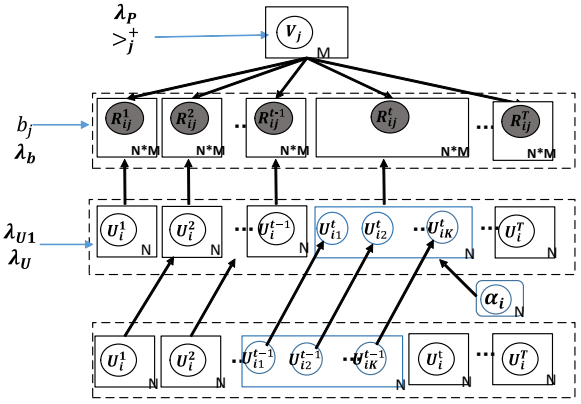


Figure 3: Graphical representation of KPT.

3.3 Model Learning and Prediction

We summarize the graphical representation of the proposed latent model in Figure 3, where the shaded and unshaded variables indicate the observed and latent variables. Given students' response tensor R and partial order $>_j^+$ based on Q-matrix, our goal is to learn the parameters $\Phi = [U, V, \alpha, b]$, $\alpha = [\alpha_i]_{i=1}^N$. Particularly, combining Eq. (1), (4) and(11), the posterior distribution over Φ is:

$$p(U, V, \alpha, b|R, D_T) \propto p(R|U, V, \alpha, b) \times p(U|\sigma_U^2, \sigma_{U1}^2) \times p(V|D_T). \tag{12}$$

Maximizing the log posterior of the above equation is equivalent to minimizing the following objective:

$$\min_{\Phi} \mathcal{E}(\Phi) = \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}^t[\hat{R}_{ij}^t - R_{ij}^t]^2$$
$$- \lambda_P \sum_{j=1}^{M}\sum_{q=1}^{K}\sum_{p=1}^{K} I\left(q >_j^+ p\right) \ln \frac{1}{1+\mathrm{e}^{-\left(V_{jq}-V_{jp}\right)}} + \frac{\lambda_V}{2}\sum_{i=1}^{M}||V_i||_F^2$$
$$+ \frac{\lambda_U}{2}\sum_{t=2}^{T}\sum_{i=1}^{N}||\overline{U_i^t} - U_i^t||_F^2 + \frac{\lambda_{U1}}{2}\sum_{i=1}^{N}||U_i^1||_F^2, \tag{13}$$

where $\lambda_P = \sigma_R^2$, $\lambda_U = \frac{\sigma_R^2}{\sigma_U^2}$, $\lambda_{U1} = \frac{\sigma_R^2}{\sigma_{U1}^2}$ and $\lambda_V = \frac{\sigma_R^2}{\sigma_V^2}$. Among them, λ_P is a tradeoff coefficient between the responses prediction loss and partial order loss, and λ_U is a coefficient that measures how student's knowledge proficiency changes over time. λ_{U1} and λ_V are regularization parameters for students' knowledge proficiency at time 1 and the exercise-knowledge correlation matrix.

Specifically, the derivative of each parameter are:

$$\nabla_{U_{ik}^t} = \sum_{j=1}^{M} I_{ij}^t(\hat{R}_{ij}^t - R_{ij}^t)V_{jk} + \mathcal{I}[t=1]\lambda_{U1}U_{ik}^1$$
$$+ \mathcal{I}[t \ge 2]\lambda_U(\overline{U_{ik}^t} - U_{ik}^t) \tag{14}$$
$$+ \lambda_U(\overline{U_{ik}^{(t+1)}} - U_{ik}^{(t+1)})((1-\alpha_i)\mathrm{e}^{-\frac{\Delta t}{S}} + \alpha_i\frac{Df_k^t}{f_k^t + r}),$$

$$\nabla_{V_{jk}} = \sum_{t=1}^{T}\sum_{i=1}^{N} I_{ij}^{t}(\hat{R}_{ij}^{t} - R_{ij}^{t})U_{ik}^{t} + \lambda_V V_{jk} - \lambda_P \sum_{p=1}^{K} I\left(k >_j^+ p\right)\frac{\mathrm{e}^{-\left(V_{jk}-V_{jp}\right)}}{1+\mathrm{e}^{-\left(V_{jk}-V_{jp}\right)}} - \lambda_P \sum_{q=1}^{K} I\left(q >_j^+ k\right)\frac{-\mathrm{e}^{-\left(V_{jq}-V_{jk}\right)}}{1+\mathrm{e}^{-\left(V_{jq}-V_{jk}\right)}}, \quad (15)$$

$$\nabla_{\alpha_i} = \lambda_U \sum_{t=2}^{T}\sum_{k=1}^{K}(\overline{U_{ik}^{t}} - U_{ik}^{t})(U_{ik}^{t}(\frac{Df_k^t}{f_k^t + r} - \mathrm{e}^{-\frac{\Delta t}{S}})), \quad (16)$$

$$\nabla_{b_j} = \sum_{i=1}^{M} I_{ij}^{t}(\hat{R}_{ij}^{t} - R_{ij}^{t}), \quad (17)$$

here $\mathcal{I}[x]$ is an indicator function that equals to 1 if x is true.

We can update U, V and b directly by using Stochastic Gradient Descent (SGD) method [4]. With the bound constraints of α_i, a local minimum can be found by the Projected Gradient (PG) method [17]. Specifically, for each $\alpha_i \in [0,1]$ the PG method updates the current solution α_i^k in k-th iteration to α_i^{k+1} by the following rule:

$$\alpha_i^{k+1} = P[\alpha_i^k - \eta\nabla_{\alpha_i}], P(\alpha_i) = \begin{cases} \alpha_i & \text{if} \quad 0 \le \alpha_i \le 1, \\ 0 & \text{if} \quad \alpha_i < 0, \\ 1 & \text{if} \quad \alpha_i > 1. \end{cases} \quad (18)$$

With students' knowledge proficiency $U^1, U^2, ..., U^T$ and related parameters, students' responses and knowledge proficiency at time T+1 can be predicted as:

$$U_i^{(T+1)} = \left\{U_{i1}^{(T+1)}, U_{i2}^{(T+1)}, ..., U_{iK}^{(T+1)}\right\},$$

$$U_{ik}^{(T+1)} \approx (1-\alpha_i)U_{ik}^{T}\mathrm{e}^{-\frac{\Delta(T+1)}{S}} + \alpha_i U_{ik}^{T}\frac{Mf_k^{T+1}}{f_k^{T+1}+r},$$

$$\hat{R}_{ij}^{(T+1)} \approx \langle U_i^{(T+1)}, V_j\rangle - b_j. \quad (19)$$

After obtaining $\hat{R}^{(T+1)}$ and $U^{(T+1)}$ at time T+1, we can recommend relevant exercises with high probability to get wrong response or forget for student i. In summary, we give the training algorithm of KPT in Algorithm 1.

Algorithm 1: Parameter Learning of the KPT Model

```
Initialize U, V, α and b ;
while not converged do
    for i = 1, 2, ...N do
        for t = 1, 2, ..., T do
            for k = 1, 2, ..., K do
                Fix V, α, b, update U_ik^t using SGD;
        Fix U, V, b, update α_i using PG ;
    for i = 1, 2, ..., M do
        for k = 1, 2, ..., K do
            Fix U, α, b, update V_jk using SGD;
        Fix U, V, α, update b using SGD;
Return U, V, α and b ;
```

Time Complexity. KPT costs most of time in computing the knowledge proficiency of each student and balancing parameters. Suppose there are r non-empty entries in response tensor R, then the average response records of each student in each time window are $t_r = \frac{r}{N\times T}$. In each iteration, the time complexity is $O(N \times T \times K \times t_r = O(K \times r))$ for U, $O(K \times r)$ for V, and $O(r)$ for the balance parameters. Thus the total complexity of parameter learning in each iteration is $O(K \times r)$, which is linear with the records and time windows.

4 EXPERIMENTS

In this section, we first introduce our experimental datasets and setups. Then, we report experimental results from the following four aspects: (1) the predictive performance of our KPT model; (2) the effectiveness on KPD task; (3) the influence of parameter settings in KPT; (4) the explanatory power of the diagnosis results of KPT by a case study.

4.1 Datasets

In the experiments, we use three real-world datasets, i.e., Math1, Math2 and ASSIST, respectively. Among them, Math1 and Math2 are two private datasets which are collected from daily exercise records of high school students (Table 1) for mathematics problems. ASSIST is a public dataset Assistments[1] 2009-2010 "Non-skill builder" [20], which records the student mathematics exercise logs in an on-line tutor.

In Math1 and Math2, each dataset contains responses of students with time record and a given Q-matrix (an example is shown in Table 3) by educational experts. We treat each month as a time window, and thus there are 4 (10) time points in dataset Math1 (Math2). In data splitting process, we use the data till time T for model training, i.e., T=3 (T=9) in Math1 (Math2), and the records of the last time window are for testing.

As for ASSIST, we preprocess the original dataset for our KPD task as follows: (1) we select 71 exercises with 7 frequent knowledge points from ASSIST in the experiments because the knowledge points at different time require a high coverage. (2) Since the system allows students to repeat doing the same exercises, we just take the first-attempt responses of them to each exercise as records for fairness. (3) ASSIST only records the order (no explicit time information) of student exercise history, we divided each student logs into four parts according to their sequencial order, thus we have 4 time windows in ASSIST. Specifically, we use the first 3 order logs of student for model training, and the remain one is for testing.

In summary, Figure 4 shows the preview of three Q-matrices (we only show subsets of 28 exercises for better illustration), where each row of each subfigure denotes an exercise and each column stands for a knowledge point. The white one means the exercise is related to the knowledge point, while the black one indicates the exercise does not contain the knowledge point. Moreover, a better explanation about Q-matrix is shown in Table 3, which contains 5 exercises and the related knowledge points in Math1. From the Figure 4 and Table 3, we can see that most of the exercises are less than two knowledge points, which indicates that Q-matrix is very sparse. Table 4 summarizes the basic statistics of three datasets.

4.2 Experimental Setup

KPT Setting. We first introduce the parameter settings of $l^t(*)$ and $f^t(*)$, i.e., *Learning curve* and *Ebbinghaus forgetting curve*, respectively. Specifically, for $l^t(*)$, we set $D = 2$ to control the multiplier of growth and the average frequency

[1] https://sites.google.com/site/assistmentsdata/home/assistment-2009-2010-data

Table 3: A practical example of Q-matrix in Math1.

ExerciseId	Knowledge Points
E1	Function
E2	Function, Set
E3	Function, Derivative, Inequality
E4	Solid geometry, Trigonometric function
E5	Propositional logic

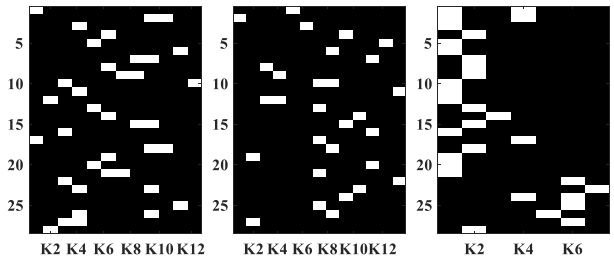


Figure 4: Q-matrix of two DataSets.

among all knowledge points r as 4, 9, 6 in Math1, Math2, ASSIST respectively; For $f^t(*)$, we set Δt as 1 for all time intervals between time window $t-1$ and t. And memory strength S as 5 to fit forgetting curve. Then, as for several regularization parameters in KPT model, we set $\lambda_{U1}=\lambda_V=0.01$. λ_U is set to be 3, 1 and 2 in Math1, Math2 and ASSIST, and λ_P is set to be 1.5, 1 and 2 in Math1, Math2 and ASSIST respectively(we will discuss the sensitivity of parameters in the next subsection).

Baseline Approaches. To compare the performance of our proposed KPT model, we borrow some baselines from various perspectives. The details of them are as follows:

- *IRT*: a cognitive diagnosis method modeling students' latent trait and exercises' parameters by a logistic-like function [3] .
- *DINA*: a cognitive diagnosis method modeling each student's knowledge proficiency by a binary vector with Q-matrix [8].
- *PMF*: a probabilistic matrix factorization method that projects students and exercises into low-rank latent factors [24].
- *BKT*: a kind of Hidden Markov Model (HMM) which models students' latent knowledge state as a set of binary variables and determines when a knowledge point has been learned [7].
- *LFA*: an improved IRT model that assumes students share the same parameters of learning rate during their learning process [6].
- *QMIRT*: QMIRT is a variant of basic IRT model, where we extend the latent trait value of each student in IRT to a multi-dimension knowledge proficiency vector with our proposed partial order prior of Q-matrix.
- *QPMF*: QPMF is a variant of basic PMF model, where we incorporate our proposed partial order prior of Q-matrix into PMF to improve the explanatory power. Particularly, QPMF is also a simplified model of KPT that does not consider priors of forgetting and learning.

Table 4: The statistics of the three datasets.

Dataset	Math1	Math2	ASSIST
Training scores logs	521,248	347,424	13,443
Testing scores logs	74,464	18,312	1,822
Students	9,308	1,306	215
Exercises	64	280	71
Time windows	4	10	4
Knowledge points	12	13	7
Average knowledge points of each exercise	1.15	1.3215	1.02

Concretely, the chosen baselines are all widely-used in the educational psychology area (IRT, DINA, BKT, LFA) and data mining community (PMF), and the two variants (QMIRT, QPMF) are adopted to highlight the effectiveness of our proposed partial order Q-matrix prior. Also, all these baselines can be categorized into static diagnostic models (IRT, DINA, PMF, QMIRT, QPMF) and the dynamic ones (LFA, BKT). For better illustration, we summarize the characteristics of these models in Table 5.

In the following experiments, both KPT and baselines are implemented by Python and all experiments are run on a Linux server with four 2.0GHz Intel Xeon E5-2620 CPUs and 100G memory. For fairness, all parameters in these baselines are tuned to have the best performances.

4.3 Experimental Results

Students' Responses Prediction. To demonstrate the predictive performance of KPT model, we first conduct experiments on the task of predicting the responses of students (i.e., whether or not a student get the correct answer to a specific exercise) at time window $T+1$ (the first task in predicting stage of our framework in Figure 2). In this task, we adopt all the baselines mentioned above for comparison, and use the widely-used *mean absolute error* (MAE) and *root mean square error* (RMSE) as the evaluation metrics [24].

Figure 5 shows the overall results of all models in students' score prediction task. There are several observations: First, our proposed model KPT performs best on all three datasets. Second, QMIRT and QPMF outperform traditional IRT and PMF, which indicates the effectiveness of incorporating the partial order based Q-matrix prior. Third, KPT and LFA, as dynamic models, perform better than those with static assumption (IRT, DINA, PMF), which demonstrates that it is more effective to diagnose students' knowledge proficiency from an evolving perspective. However, BKT does not perform well on this task. We guess a possible reason is that BKT focuses on the scenario that students keep doing the same exercises. But in our data, most students just do a specific exercise only once, thus the exercise sequence lengths of students are not enough for BKT. In summary, these evidences demonstrate the rationality of three priors (i.e., *Q-matrix*, *learning curve* and *forgetting curve*).

Knowledge Proficiency Diagnosis. As mentioned before, the second task in predicting stage of our framework in Figure 2 is predicting students' knowledge proficiency in the future. In order to validate the effectiveness of this prediction

Table 5: Characteristics of the Baselines and KPT.

Model	Data Source				Prediction		Dynamic
	Q-matrix	Multi-Skill Question	Repeating Answer Question	Time	Multi-Knowledge Proficiency	Response	Explanation?
IRT[3]	×	×	×	×	×	√	×
DINA[8]	√	√	×	×	√	√	×
PMF[24]	×	×	×	×	×	√	×
BKT[7]	√	×	√	√	√	√	√
LFA [6]	√	√	√	√	×	√	√
QMIRT	√	√	×	×	√	√	×
QPMF	√	√	×	×	√	√	×
KPT	√	√	×	√	√	√	√

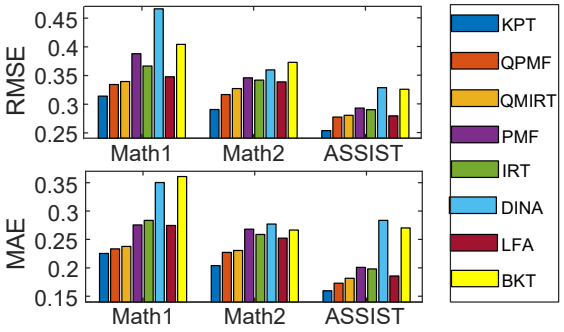


Figure 5: Responses prediction task performance.

(i.e., whether or not the diagnosis results of students are good), we also conduct several experiments.

Intuitively, if student a masters better than student b on a specific knowledge point at time $T+1$ (calculated by Eq. 19), she will have a higher probability to get correct answers to the related exercises than student b at time $T+1$. We adopt *Degree of Agreement* (DOA) [13, 19] metric to evaluate this ranking performance. Particularly, for a specific knowledge k, the DOA result on k is defined as:

$$DOA(k) = \sum_{j=1}^{M} I_{jk} \sum_{a=1}^{N} \sum_{b=1}^{N} \frac{\delta\left(U_{ak}^{T+1}, U_{bk}^{T+1}\right) \cap \delta\left(R_{aj}^{T+1}, R_{bj}^{T+1}\right)}{\delta\left(U_{ak}^{T+1}, U_{bk}^{T+1}\right)} \quad (20)$$

where U_{ak}^{T+1} is knowledge proficiency of student a on knowledge point k at time $T+1$. R_{aj}^{T+1} (denoted in Table 2) is student a's response on exercise j at time $T+1$. $\delta(x,y)$ is an indicator function, where $\delta(x,y)=1$ if $x>y$. I_{jk} is an another indicator function, where $I_{jk}=1$ if exercise j contains knowledge point k. Then DOA value ranges from 0 to 1 and the larger the better. Furthermore, we also average DOA(k) of all knowledge points for measuring the whole effectiveness on KPD task, which is denoted as DOA-Avg.

For model comparisons, we choose DINA, QMIRT, QPMF and BKT for this KPD task as baselines because all other latent factor models mentioned before are unexplainable for the diagnosis, i.e., each dimension of student latent vectors cannot correspond to any explicit knowledge point.

Figure 6 illustrates the whole effectiveness results of all models on KPD task and Table 6 shows the results of each specific knowledge point in all three datasets. Specifically, for all datasets, KPT performs best on KPD task for all knowledge points, followed by QPMF and QIRT, which indicates that the educational prior of *Q-matrix* does effectively. Besides, we also observe that traditional cognitive diagnosis model DINA does not perform well, indicating that the static model is unsuitable for solving the KPD task over time. Last but not least, we can see that BKT, as a dynamic model, does not perform as well as KPT. This observation demonstrates the effectiveness of incorporating both priors of *Learning curve* and *Forgetting curve*.

Sensitivity of Parameters. In our KPT model, there are four parameters playing crucial roles: λ_{U1}, λ_V, λ_U and λ_P. Among them, λ_{U1} and λ_V are the regularization parameters of students' vectors of knowledge proficiency at time $T=1$ and exercises' vectors of knowledge related, respectively. Since λ_{U1} and λ_V have a similar form to PMF model, we tune them on PMF and set them under the setting of the best performance on PMF. In the following, we report the setting parameters λ_U and λ_P with the evaluation metrics of RMSE and DOA-Avg on both two tasks mentioned above.

λ_U regularizes that students learn and forget knowledges from time to time, Figure 7(a), Figure 7(c) and Figure 7(e) visualizes the performance with the increasing values of λ_U from 1, 0.1, 1 to 10, 5, 5 in datasets Math1, Math2, ASSIST respectively. As we can see from the figure, as λ_U increases, the performances of KPT firstly increase but decrease when λ_U surpasses 3, 1, 2 in datasets Math1 ,Math2, ASSIST. Therefore, we set $\lambda_U=3$, 1, 2 in Math1, Math2, ASSIST for obtaining the best results.

Also, as shown in Figure 7(b), Figure 7(d) and Figure 7(f), parameter λ_P, which controls how much the KPT model is restricted by the partial order Q-matrix prior, has the similar property to λ_U. As a result, we set $\lambda_P=1.5$, 1, 2 in Math1, Math2, ASSIST respectively because the performance of KPT achieves the best when it reaches the corresponding value.

Case Study. We argue that KPT can track KPD of students in an interpretable way. Figure 8 visualizes the diagnosis results of a student on six knowledge points at three particular time in Math2 (we only show six knowledge points for better illustration). From the figure, she makes progress on knowledge point "Function" from Mar (0.08) to May (0.36), 2016 with possible learning factor (she does from 2 to 7 exercises about *Function* from March to May). In contrast, her proficiency on knowledge point *Analytic geometry* declines (from 0.65 to 0.36) over time with possible forgetting factor because she practices less than 2 relevant exercises at each time. These observations imply that she

Table 6: KPD task performance for each knowledge point.

(a) Math1

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

(b) Math2

K	Baselines				
	KPT	QPMF	QMIRT	DINA	BKT
K1	**0.804**	0.743	0.754	0.517	0.568
K2	**0.757**	0.632	0.659	0.534	0.753
K3	**0.818**	0.761	0.723	0.510	0.669
K4	0.688	0.733	**0.734**	0.534	0.711
K5	**0.891**	0.703	0.668	0.474	0.553
K6	**0.699**	0.547	0.653	0.489	0.644
K7	**0.791**	0.677	0.722	0.483	0.730
K8	**0.726**	0.722	0.659	0.523	0.668
K9	**0.736**	0.558	0.541	0.507	0.567
K10	**0.652**	0.639	0.650	0.511	0.614
K11	**0.888**	0.836	0.692	0.522	0.630
K12	**0.798**	0.737	0.794	0.498	0.528
K13	**0.813**	0.797	0.804	0.453	0.633

(c) ASSIST

K	Baselines				
	KPT	QPMF	QMIRT	DINA	BKT
K1	**0.793**	0.747	0.716	0.605	0.592
K2	**0.823**	0.653	0.673	0.593	0.672
K3	**0.887**	0.852	0.671	0.631	0.577
K4	**0.792**	0.598	0.755	0.525	0.569
K5	**0.891**	0.576	0.672	0.511	0.624
K6	**0.871**	0.647	0.657	0.628	0.604
K7	**0.901**	0.793	0.654	0.573	0.796

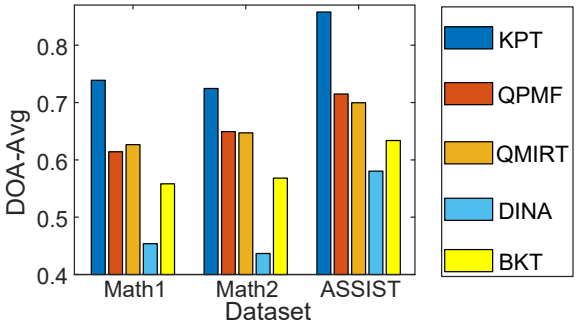


Figure 6: KPD task performance for all knowledge points.

needs a timely review on *Analytic geometry*. Therefore, these evidences could lead to more personalized training for her.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we designed an explanatory probabilistic KPT model for solving the KPD task of students over time by leveraging educational priors. Specifically, we associated each exercise with a knowledge vector with the *Q-matrix* prior. And each student was also represented as a knowledge vector at each time in the same knowledge space. Then we embedded the classical educational theories (i.e., *Learning curve* and *Forgetting curve*) as priors to capture the change of each student's proficiency over time. After that, we designed a probabilistic matrix factorization framework by combining student and exercise priors. Extensive experiments on three real-world datasets clearly demonstrated the effectiveness and explanatory power of our proposed model.

In the future, there are some directions for further studies. First, we will consider to combine more kinds' of users' behaviors (e.g., reading records) for the KPD task. Second, as students may learn difficult knowledge points (e.g., *Function*) after some basic ones (e.g., *Set*), it is interesting to take this kind of knowledge relationship into account for KPD task.

6 ACKNOWLEDGEMENT

The authors thank Defu Lian for the valuable suggestions. This research was partially funded by the National High Technology Research and Development Program (863 Program) of China (No. 2015AA015409), the National Natural Science Foundation of China (Grants No. 61672483, U1605251, 61325010, 61602147), the Youth Innovation Promotion Association of CAS (No. 2014299), and the Anhui Provincial Natural Science Foundataion(Grant No. 1708085QF155).

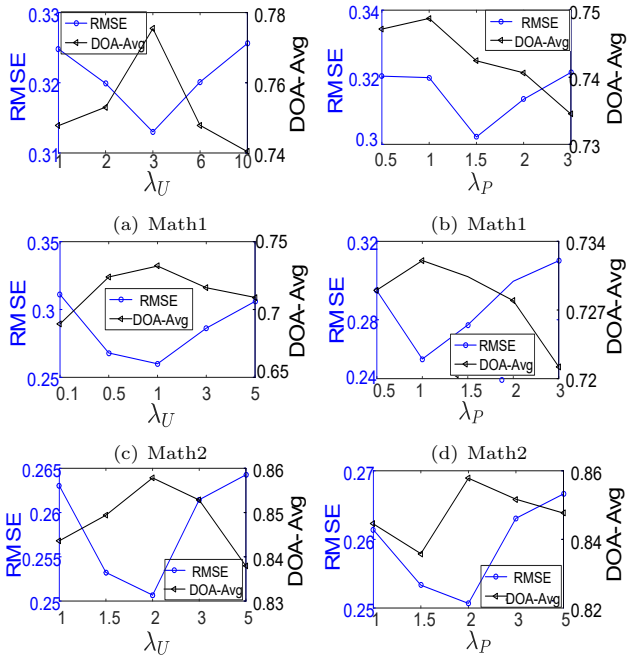


Figure 7: The impact of λ_U and λ_P .

REFERENCES

[1] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2014. Engaging with massive online courses. In *Proceedings of the 23rd international conference on World wide web*. ACM, 687–698.

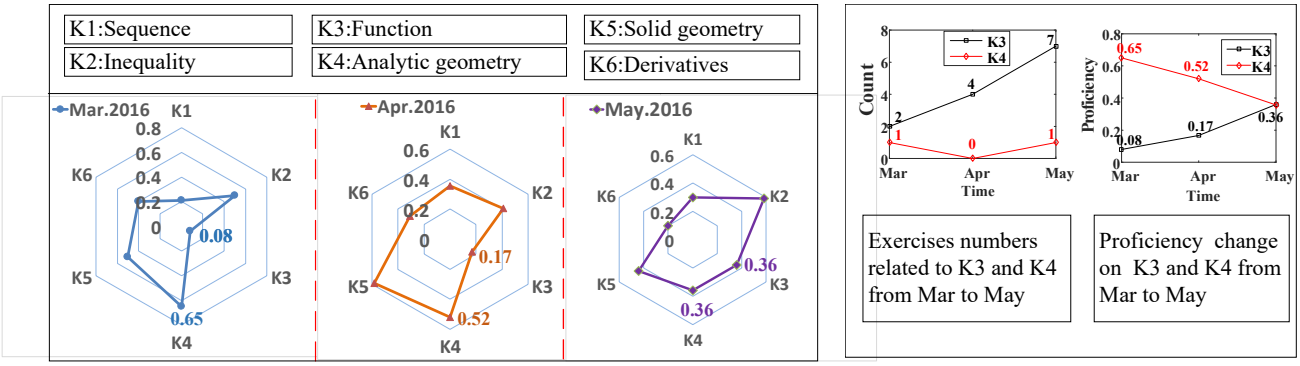
[2] Michel Jose Anzanello and Flavio Sanson Fogliatto. 2011. Learning curve models and applications: Literature review and research directions. *International Journal of Industrial Ergonomics* 41, 5 (2011), 573–583.

[3] Allan Birnbaum. 1968. Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores* (1968).

[4] Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer, 177–186.

[5] Hugh Burns, Carol A Luckhardt, James W Parlett, and Carol L Redfield. 2014. *Intelligent tutoring systems: Evolutions in design*. Psychology Press.

[6] Hao Cen, Kenneth Koedinger, and Brian Junker. 2006. Learning factors analysis–a general method for cognitive model evaluation



(a) Diagnosis (b) Analysis

Figure 8: Diagnosis results of a student from Mar to May, 2016 in Math2. Subfigure (a) shows her knowledge proficiency on six knowledge points in the 3 months where the above list gives the names of each knowledge point. Subfigure (b) gives the statistical analysis about knowledge point $K3$ and $K4$ which contains a graph about the numbers of exercises she does (Left) and a graph about the change of her knowledge proficiency (Right).

and improvement. In *Intelligent tutoring systems*. Springer, 164–175.

[7] Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.

[8] Jimmy De La Torre. 2009. DINA model and parameter estimation: A didactic. *Journal of educational and behavioral statistics* 34, 1 (2009), 115–130.

[9] Louis V DiBello, Louis A Roussos, and William Stout. 2006. 31a review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of statistics* 26 (2006), 979–1030.

[10] Hermann Ebbinghaus. 2013. Memory: a contribution to experimental psychology. *Annals of neurosciences* 20, 4 (2013), 155–156.

[11] Susan E Embretson and Steven P Reise. 2013. *Item response theory*. Psychology Press.

[12] Rebecca Grossman and Eduardo Salas. 2011. The transfer of training: what really matters. *International Journal of Training and Development* 15, 2 (2011), 103–120.

[13] Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question Difficulty Prediction for READING Problems in Standard Tests.. In *AAAI*. 1352–1359.

[14] Brian W Junker and Klaas Sijtsma. 2001. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement* 25, 3 (2001), 258–272.

[15] Mohammad Khajah, Rowan Wing, Robert Lindsey, and Michael Mozer. 2014. Integrating latent-factor and knowledge-tracing models to predict individual differences in learning. In *Educational Data Mining 2014*.

[16] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009).

[17] Chih-Jen Lin. 2007. Projected gradient methods for nonnegative matrix factorization. *Neural computation* 19, 10 (2007), 2756–2779.

[18] Jingchen Liu, Gongjun Xu, and Zhiliang Ying. 2012. Data-driven learning of Q-matrix. *Applied psychological measurement* 36, 7 (2012), 548–564.

[19] Qi Liu, Yong Ge, Zhongmou Li, Enhong Chen, and Hui Xiong. 2011. Personalized travel package recommendation. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 407–416.

[20] Zachary A Pardos and Neil T Heffernan. 2011. KT-IDEM: introducing item difficulty to the knowledge tracing model. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 243–254.

[21] Philip I Pavlik Jr, Hao Cen, and Kenneth R Koedinger. 2009. Performance Factors Analysis–A New Alternative to Knowledge Tracing. *Online Submission* (2009).

[22] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*. 505–513.

[23] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 452–461.

[24] Ruslan Salakhutdinov and Andriy Mnih. 2007. Probabilistic Matrix Factorization.. In *Nips*, Vol. 1. 2–1.

[25] Jiliang Tang, Huiji Gao, Xia Hu, and Huan Liu. 2013. Exploiting homophily effect for trust prediction. In *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 53–62.

[26] Nguyen Thai-Nghe, Lucas Drumond, Artus Krohn-Grimberghe, and Lars Schmidt-Thieme. 2010. Recommender system for predicting student performance. *Procedia Computer Science* 1, 2 (2010), 2811–2819.

[27] Nguyen Thai-Nghe, Tomáš Horváth, and Lars Schmidt-Thieme. 2010. Factorization models for forecasting student performance. In *Educational Data Mining 2011*.

[28] Heinz Von Foerster. 2007. *Understanding understanding: Essays on cybernetics and cognition*. Springer Science & Business Media.

[29] Xiaojing Wang, James O Berger, Donald S Burdick, et al. 2013. Bayesian analysis of dynamic item response models in educational testing. *The Annals of Applied Statistics* 7, 1 (2013), 126–153.

[30] Le Wu, Yong Ge, Qi Liu, Enhong Chen, Richang Hong, Junping Du, and Meng Wang. 2017. Modeling the Evolution of Users Preferences and Social Links in Social Networking Services. *IEEE Transactions on Knowledge and Data Engineering* 29, 6 (2017), 1240–1253.

[31] Runze Wu, Guandong Xu, Enhong Chen, Qi Liu, and Wan Ng. 2017. Knowledge or Gaming?: Cognitive Modelling Based on Multiple-Attempt Response. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 321–329.

[32] Run-ze Wu, Qi Liu, Yuping Liu, Enhong Chen, Yu Su, Zhigang Chen, and Guoping Hu. 2015. Cognitive Modelling for Predicting Examinee Performance.. In *IJCAI*. 1017–1024.

[33] Liang Xiong, Xi Chen, Tzu-Kuo Huang, Jeff Schneider, and Jaime G Carbonell. 2010. Temporal collaborative filtering with bayesian probabilistic tensor factorization. In *Proceedings of the 2010 SIAM International Conference on Data Mining*. SIAM, 211–222.

[34] Michael V Yudelson, Kenneth R Koedinger, and Geoffrey J Gordon. 2013. Individualized bayesian knowledge tracing models. In *International Conference on Artificial Intelligence in Education*. Springer, 171–180.

[35] Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. 2017. Dynamic Key-Value Memory Networks for Knowledge Tracing. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 765–774.