Non-compensatory Knowledge Tracing with Local Variational Approximation

Hiroshi Tamano[1,2] and Daichi Mochihashi[3]

[1] NEC Corporation
h-tamano@nec.com
[2] The Graduate University for Advanced Studies
[3] The Institute of Statistical Mathematics
daichi@ism.ac.jp

Abstract. Changes of latent skills of learners can be modeled by a state space model from sequential observations of their answers to questions. The model using continuous skill states and non-compensatory emissions has the potential to accurately predict whether learners can answer questions and explain which skill they are missing when they cannot answer the questions. To explore this potential, we propose a statistical model that combines a linear dynamical system with a non-compensatory model. Since this results in a complicated posterior of the skill states, we propose an approximation using a local variational method. We experimentally show that our variational posterior adequately approximates the true posterior using artificial data, and also our model outperforms two popular deep learning-based methods in prediction using open datasets.

Keywords: Personalized education · knowledge tracing · item response theory · Kalman filter · variational approximation.

1 Introduction

Changes of latent skills of learners can be modeled by a state space model from sequential observations of their answers to questions. This modeling technique has been studied in knowledge tracing (KT) [6, 8, 12] and cognitive diagnostic models (CDMs) [5, 7] to provide data-driven personalized education. To deal with a question that requires multiple skills, the models studied so far are organized into four categories: the variables for latent skill states are represented as binary or continuous vectors and the emission model is compensatory or non-compensatory [9], which means that each skill can complement other skills or not. Figure 1 shows the four categories.

The model using continuous skill states and non-compensatory emissions has the potential to accurately predict whether learners can answer questions and explain which skill they are missing when they cannot answer the questions. The assumptions of this model reflect situations in which correct or incorrect answers are emitted more realistically than other models. A person's skill states are considered to be continuous rather than binary; questions can require the same skill

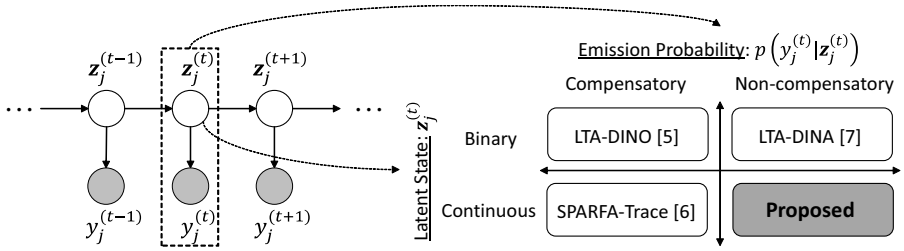

Fig. 1: Four categories of state space models for inferring changes of latent skills. The left panel shows a graphical model for these models. The right panel shows the four categories: the skill state variable $\boldsymbol{z}_j^{(t)}$ for learner j at time t is binary or continuous vector. The emission model is compensatory or non-compensatory. Our proposed model is shown in the lower right.

but different levels of mastery. The emission mechanism is non-compensatory rather than compensatory. For example, if we want to solve an equation such as $1/5\,x + 3/10 = 2x$, we need both fraction and equation skills. The fraction skill cannot complement the equation skill, and vice versa. The non-compensatory model can benefit learners by clearly explaining which skill they are missing when they cannot answer a question. However, the model in this category has not been explored yet.

To explore this potential, we propose a statistical model that combines a linear dynamical system with the non-compensatory model. Since this introduces non-conjugacy and makes it hard to obtain the exact posterior of the skill states, we propose an approximation using a local variational method [3]. In our experiments, we show that our variational posterior adequately approximates the true posterior using artificial data, and also our model outperforms two popular deep learning-based methods in prediction using open datasets.

2 Non-compensatory Knowledge Tracing

Non-compensatory KT is an extension of the linear dynamical system (LDS) [4] whose emission probability is the non-compensatory model in multidimensional item response theory (MIRT) [9]. Since this introduces the non-conjugacy of the non-compensatory emission and a Gaussian prior, it is hard to obtain the exact posterior analytically. We approximate non-compensatory emission as Gaussian so that we can simply treat our model same as the LDS.

2.1 Generative Model

Our target data $\{(i_j^{(t)}, y_j^{(t)})\}_{j=1,\ldots,N,\ t=1,\ldots,T_j}$ are sequential observations of answers to questions by different learners over time. $i_j^{(t)} \in \{1, ..., M\}$ denotes the index of the question answered by learner $j \in \{1, ..., N\}$ at time step $t \in \{1, ..., T_j\}$, also denoted by $i(j, t)$, and we abbreviate it to i when it is clear from the context.

$y_j^{(t)} \in \{0, 1\}$ represents whether learner j answered question $i(j,t)$ correctly or not at time step t.

The generative model of non-compensatory KT is defined like the LDS, except that the emission probability is the non-compensatory item response model:

$$p(\boldsymbol{z}_j^{(1)}) = \mathcal{N}(\boldsymbol{z}_j^{(1)} | \boldsymbol{\mu}_0, P_0), \tag{1}$$

$$p(\boldsymbol{z}_j^{(t+1)} | \boldsymbol{z}_j^{(t)}) = \mathcal{N}\left(\boldsymbol{z}_j^{(t+1)} \middle| D_{i(j,t)} \boldsymbol{z}_j^{(t)} + \begin{bmatrix} \vdots \\ \boldsymbol{\beta}_k^T \boldsymbol{x}_{j,k}^{(t+1)} \\ \vdots \end{bmatrix}, \Gamma_{i(j,t+1)} \right), \tag{2}$$

$$p(y_j^{(t)} = 1 | \boldsymbol{z}_j^{(t)}) = \prod_k \sigma\left(a_{i,k}(z_{j,k}^{(t)} - b_{i,k}) \right)^{Q_{i,k}}. \tag{3}$$

$\boldsymbol{z}_j^{(t)}$ denotes the latent skill state of learner j at time step t. $\boldsymbol{z}_j^{(t)}$ is a K dimensional vector and $z_{j,k}^{(t)}$ denotes the state of k-th skill. Initially, the state of a learner is drawn from the Gaussian (1) with the mean $\boldsymbol{\mu}_0$ and covariance P_0. It then transits by a linear transformation (2) by D_i and $\boldsymbol{\beta}_k$ with Gaussian noise with zero mean and covariance $\Gamma_{i(j,t+1)}$; D_i is diagonal and $(D_i)_{k,k} = 1$ when k is not required from question i. $\boldsymbol{x}_{j,k} \in \mathbb{R}^{F_k}$ is any F_k dimensional covariate. Covariance Γ_i is also diagonal and $(\Gamma_i)_{k,k} = \gamma_k$ if question i requires skill k, otherwise $(\Gamma_i)_{k,k} = 0$. The response of a learner to question i is drawn from the item response model (3): $a_{i,k}$ and $b_{i,k}$ denote item discrimination and item difficulty, respectively. $Q_{i,k} \in \{0, 1\}$ denotes question-skill mapping, whether question j requires skill k or not, and we assume that the Q matrix is given in advance.

2.2 Posterior Inference and Parameter Estimation

The posterior inference and parameter estimation of the LDS can be processed through a forward-backward algorithm and an EM algorithm [2], respectively. By approximating emission probability as Gaussian, which is described in Section 2.3, we fully utilize this framework. A forward $\hat{\alpha}$ message can be obtained using the Bayesian update:

$$\hat{\alpha}(\boldsymbol{z}_j^{(t)}) \propto p(\boldsymbol{y}_j^{(t)} | \boldsymbol{z}_j^{(t)}) p(\boldsymbol{z}_j^{(t)} | \boldsymbol{y}_j^{(1)}, \ldots, \boldsymbol{y}_j^{(t-1)}). \tag{4}$$

Since the likelihood is approximated as Gaussian and the prior is also Gaussian, we can analytically calculate $\hat{\alpha}$ as Gaussian. Once we get the forward message as Gaussian, the backward message can be obtained as Gaussian in the same way as the LDS. In the parameter estimation, we employ Monte Carlo EM [10]. We omit the detail of EM algorithm and focus on the Gaussian approximation of the likelihood instead.

2.3 Local Gaussian Approximation

This section describes a local Gaussian approximation to the likelihood function in (4). Our approximation is based on a method in [3]; therefore we approximate it by finding a variational lower bound of the likelihood in an exponential

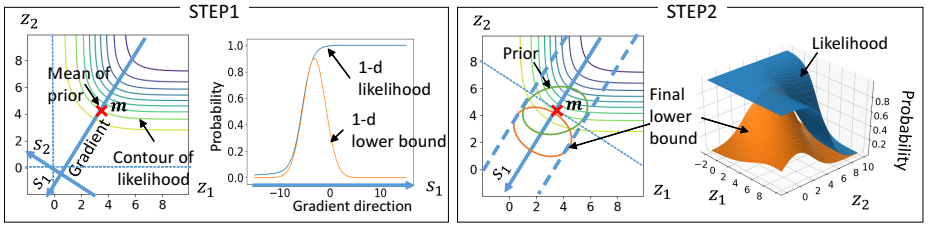

Fig. 2: Rough sketch of the Gaussian approximation.

quadratic form. Since the likelihood function for correct answers, $p(y_j^{(t)} = 1|\boldsymbol{z}_j^{(t)})$, is a product of sigmoid functions, applying the method in [3] for each sigmoid function yields a Gaussian approximation. Here, we focus on the approximation for incorrect answers. In this section, we consider the case where learner j answers question i at time step t. For simplicity, we omit j, t and slightly abuse the notation $\boldsymbol{z}, \boldsymbol{a}_i, \boldsymbol{b}_i$ to limit the dimensions to the skills required for question i.

The approximation of the likelihood for incorrect answers consists of two steps as shown in Figure 2. We just describe rough sketch of the derivation and the derived results due to the limitation of pages. In the first step, we take the gradient of the likelihood at the mean of the prior and convert the likelihood into a one-dimensional function along the gradient direction. Applying the complement rule, the lower bound of sigmoid function [3], and Jensen's inequality yields a variational lower bound in an exponential quadratic form. In the second step, we extend this lower bound to K_i dimensions, where K_i is the number of skill required to question i. The variance in the direction that is orthogonal to the s_1 axis is copied from the prior.

The derived Gaussian approximation is $\mathcal{N}(\boldsymbol{z}|\boldsymbol{\eta}, \Psi)$:

$$\boldsymbol{\eta} = W \begin{bmatrix} \frac{\sum_{l=0}^{2^{K_i}-2} q_l B_l}{2\sum_{l=0}^{2^{K_i}-2} q_l A_l} \\ \underline{W}^T \boldsymbol{m} \end{bmatrix}, \quad \Psi^{-1} = W \begin{bmatrix} 2\left(\sum_{l=0}^{2^{K_i}-2} q_l A_l\right) & \mathbf{0} \\ \mathbf{0} & \Lambda_{\setminus s_1} \end{bmatrix} W^T.$$

The prior in (4), $p(\boldsymbol{z}_j^{(t)}|\boldsymbol{y}_j^{(1)}, \ldots, \boldsymbol{y}_j^{(t-1)})$, is assumed to be $\mathcal{N}(\boldsymbol{z}|\boldsymbol{m}, G)$. Each column of matrix W forms an orthonormal basis, the direction of the first column vector is the gradient $\nabla p(y = 0|\boldsymbol{z})|_{\boldsymbol{m}}$, the directions of other column vectors can be any directions as far as the orthonormal property holds, and $\underline{W}^T$ is a matrix where the first row of W^T is removed. Other definitions are the followings:

$$\Lambda = W^T G^{-1} W, \quad \Lambda_{\setminus s_1} = \Lambda_{2:K_i,2:K_i} - \Lambda_{2:K_i,1}\Lambda_{1,1}^{-1}\Lambda_{1,2:K_i},$$

$$q_l \propto \left[\prod_{k=1}^{K_i} \sigma(\xi_{l,k})\right] \exp(-A_l\xi^2 + B_l\xi + C_l), \quad \sum_{l=0}^{2^{K_i}-2} q_l = 1,$$

$$A_l = \sum_{k=1}^{K_i} \lambda(\xi_{l,k})\tilde{a}_{l,i,k}^2, \quad B_l = \sum_{k=1}^{K_i} 2\lambda(\xi_{l,k})\tilde{a}_{l,i,k}\tilde{b}_{l,i,k} + \frac{1}{2}\tilde{a}_{l,i,k},$$

$$C_l = \sum_{k=1}^{K_i} \lambda(\xi_{l,k})(\xi_{l,k}^2 - \tilde{b}_{l,i,k}^2) - \frac{1}{2}(\xi_{l,k} + \tilde{b}_{l,i,k}),$$

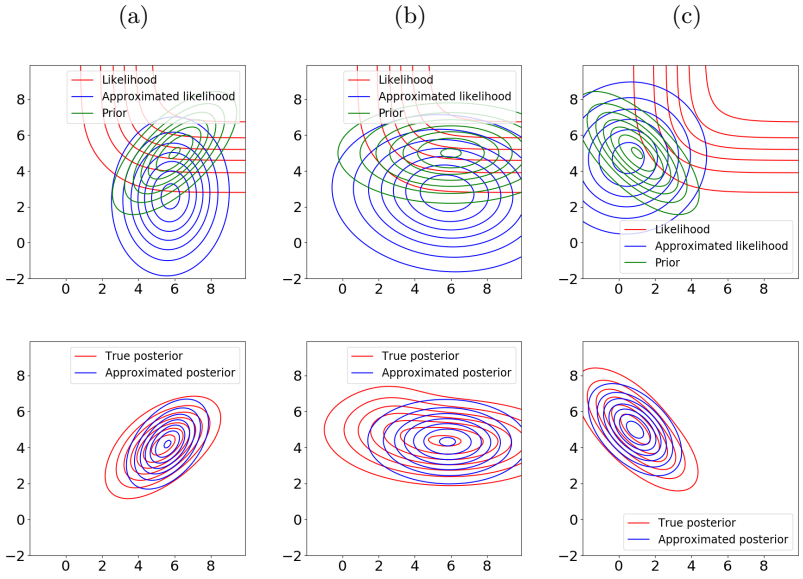

Fig. 3: Comparison between the true posterior and the approximated posterior.

$$\tilde{a}_{l,i,k} = \mathrm{sgn}(l,k)a_{i,k}w_{k,1}, \quad \tilde{b}_{l,i,k} = -\mathrm{sgn}(l,k)a_{i,k}(\boldsymbol{w}_{k,2:}\underline{W}^T\boldsymbol{m} - b_{i,k}),$$
$$\xi_{l,k} = \tilde{a}_{l,i,k}\xi - \tilde{b}_{l,i,k}, \quad \mathrm{sgn}(l,k) = \begin{cases} 1 & \text{if } l \bmod 2^k \neq l \bmod 2^{k-1} \\ -1 & \text{otherwise} \end{cases}.$$
$$\lambda(\xi) = \frac{1}{2\xi}(\sigma(\xi) - \frac{1}{2}),$$

ξ is a variational parameter. It can be optimized by maximizing lower bound of marginal likelihood.

3 Experiments

We show some visualizations of our approximation of $\hat{\alpha}$ messages to see how it works, and results of prediction performance on the Assistments datasets [1].

Three examples of approximating $\hat{\alpha}$ messages are shown in Figure 3. Given a Gaussian prior (green on the upper panel) and a non-compensatory emission likelihood (red on the upper panel), we display the approximated likelihood (blue on the upper panel), the approximated posterior (blue on the lower panel), and true posterior (red on the lower panel) in three cases (a)-(c). The likelihood functions are $p(y=0|\boldsymbol{z}) = 1 - \sigma(z_1 - 3.0)\sigma(z_2 - 5.0)$ in all cases and the priors are set at different locations and with different rotations. From the upper panels, we can see that the likelihood is differently and locally approximated depending on the location and shape of prior. From the lower panels, we can see that our posteriors adequately approximate the true posteriors.

The results of prediction performance on the Assistments 2009-2010 "skill builder" and "non-skill builder" datasets [1] are shown in Table 1. We compared our method with the state-of-the-art deep learning based methods: DKT [8]

Table 1: Prediction scores on Assistments dataset.

	Skill Builder		Non Skill Builder	
Method	F1	AUC	F1	AUC
Proposed	$\mathbf{0.565} \pm 0.009$	$\mathbf{0.764} \pm 0.005$	$\mathbf{0.612} \pm 0.004$	$\mathbf{0.798} \pm 0.005$
DKT	0.525 ± 0.009	0.756 ± 0.006	0.585 ± 0.005	0.779 ± 0.005
DKVMN	0.521 ± 0.024	0.755 ± 0.007	0.604 ± 0.018	0.785 ± 0.007

and DKVMN [12]. In DKT and DKVMN, we created joint skills for questions requiring multiple skills to prevent a leakage problem reported in [11]. We used 5-fold cross-validation and evaluated AUC and F1 score for predicting incorrect answers. We can see our method outperforms DKT and DKVMN.

4 Conclusion

In this paper, we proposed a state space model for knowledge tracing which combines the LDS with the non-compensatory model in MIRT. Introducing the non-compensatory emission precludes from inferring the exact posterior of the latent skill states. We derived Gaussian approximation of the posterior using local variational methods. In the experiments, we showed that our variational posterior adequately approximates the true posterior using artificial data, and also our model achieves better prediction accuracy compared to two popular deep learning-based methods using open datasets.

References

1. Feng, M., Heffernan, N., Koedinger, K.: Addressing the assessment challenge with an online system that tutors as it assesses. User Modeling and User-Adapted Interaction **19**(3), 243–266 (2009)
2. Ghahramani, Z., Hinton, G.E.: Parameter estimation for linear dynamical systems. Tech. Rep. CRG-TR-96-2, University of Toronto (1996)
3. Jaakkola, T.S., Jordan, M.I.: Bayesian parameter estimation via variational methods. Statistics and Computing **10**(1), 25–37 (2000)
4. Kalman, R.E.: A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering **82**(1), 35–45 (1960)
5. Kaya, Y., Leite, W.L.: Assessing Change in Latent Skills Across Time With Longitudinal Cognitive Diagnosis Modeling: An Evaluation of Model Performance. Educational and Psychological Measurement **77**(3), 369–388 (2017)
6. Lan, A.S., Studer, C., Baraniuk, R.G.: Time-varying learning and content analytics via sparse factor analysis. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 452–461. ACM (2014)
7. Li, F., Cohen, A., Bottge, B., Templin, J.: A Latent Transition Analysis Model for Assessing Change in Cognitive Skills. Educational and Psychological Measurement **76**(2), 181–204 (2016)
8. Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L.J., Sohl-Dickstein, J.: Deep knowledge tracing. In: Advances in neural information processing systems. pp. 505–513 (2015)
9. Reckase, M.D.: Multidimensional Item Response Theory. Springer Publishing Company, Incorporated, 1st edn. (2009)
10. Wei, G.C., Tanner, M.A.: A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American statistical Association **85**(411), 699–704 (1990)
11. Xiong, X., Zhao, S., Van Inwegen, E.G., Beck, J.E.: Going deeper with deep knowledge tracing. International Educational Data Mining Society (2016)
12. Zhang, J., Shi, X., King, I., Yeung, D.Y.: Dynamic key-value memory networks for knowledge tracing. In: Proceedings of the 26th international conference on World Wide Web. pp. 765–774 (2017)