NEURAL INFORMATION
PROCESSING SYSTEMS

An Autoencoder-Like Nonnegative Matrix Co-Factorization for Improved Student Cognitive Modeling

Shenbao Yu, Yinghui Pan*, Yifeng Zeng, Prashant Doshi, Guoquan Liu, Kim-Leng Poh, Mingwei Lin

November 2024

Introduction

Two Learning Tasks of Student Cognitive Modeling

T1: estimate students' cognitive levels on knowledge concepts (cognitive diagnosis).
T2: predict the exercise performance they have never done before.

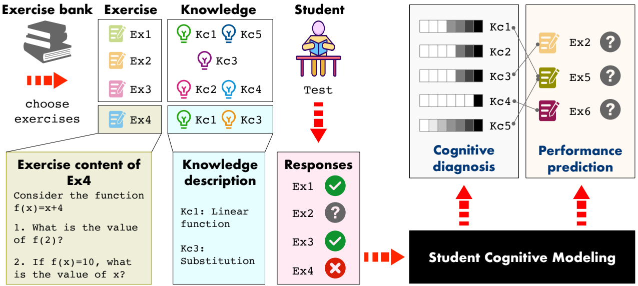


Figure: A schematic illustration of the student cognitive modeling problem.

- **Left:** A set of exercises with the expert-labeled knowledge concepts.
- **Middle:** A student's binary-value responses with missing values that are input to the modeling.
- **Right:** Two learning tasks as the output of the modeling.

Related Work and Challenges

Previous approaches have differed along two dimensions:

Methods	**Pros & Cons**
Cognitive diagnosis models	Good in T1, but may trigger cascading errors in T2.
Data mining methods	Can perform well in T2, but the students' knowledge levels are unknown.

Main Challenges

- The ground truth of students' knowledge proficiency is unknown (could not observe).
- How to frame the two learning tasks as the building blocks of an optimization framework to reduce the cascading errors?

Key Ideas

- We leverage the **monotonicity** to sidestep the issue of unknown knowledge proficiency.
- We use the **autoencoder** mechanism to meet the requirement of the monotonic constraint.

Motivations and Framework Overview

We propose the **autoencoder-like nonnegative matrix co-factorization** (AE-NMCF) framework.

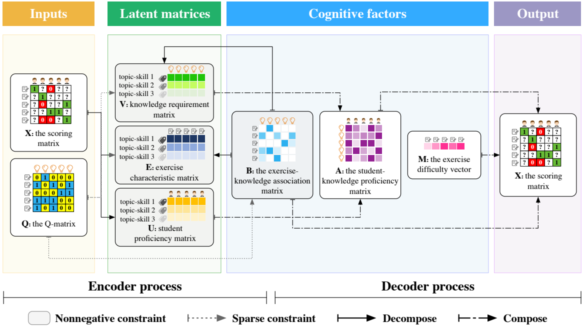


Figure: The end-to-end pipeline of AE-NMCF.

- The inputs are the students' partial responses on exercises ($\mathbf{X}$) and the knowledge-exercise relations ($\mathbf{Q}$).
- The encoder targets the specification of students' knowledge proficiency vectors.
- The decoder reconstructs the students' response to the exercises via the specified knowledge proficiency.

Model Formulation - Problem Statement

Given students **St** $= \{\text{St}_m\}_{m=1}^{\mathrm{M}}$ and exercises **Ex** $= \{\text{Ex}_n\}_{n=1}^{\mathrm{N}}$, all the students' responses are recorded in a binary scoring matrix $\mathbf{X} \in \mathbb{R}^{\mathrm{N}\times\mathrm{M}}$.

Also, suppose that there is a set of related knowledge concepts **Kc** $= \{\text{Kc}_k\}_{k=1}^{\mathrm{K}}$, we have a question matrix (Q-matrix) $\mathbf{Q} \in \mathbb{R}^{\mathrm{N}\times\mathrm{K}}$ that maps each exercise Ex_n to a set of knowledge concepts.

Table: An example of a scoring matrix (the left half) and a Q-matrix (the right side)

Exercises	Students			Knowledge concepts		
	St_1	St_2	St_3	Kc_1	Kc_2	Kc_3
Ex_1	1	?	1	0	1	1
Ex_2	?	0	?	1	0	1
Ex_3	?	1	0	1	0	0

Model Formulation - Encoder

Using $\mathbf{X}$ and $\mathbf{Q}$, the encoder gets the feature matrices for exercise, students, and knowledge concepts, respectively: $\mathbf{E} \in \mathbb{R}^{\mathrm{N}\times\mathrm{T}}$, $\mathbf{U} \in \mathbb{R}^{\mathrm{T}\times\mathrm{M}}$, and $\mathbf{V} \in \mathbb{R}^{\mathrm{T}\times\mathrm{K}}$, as

$$\begin{aligned} \min_{\mathbf{B},\mathbf{E},\mathbf{U},\mathbf{V}} \quad & \|\mathbf{W} \odot (\mathbf{X} - \mathbf{EU})\|_{\mathsf{F}}^2 + \|\mathbf{Q} \odot (\mathbf{B} - \mathbf{EV})\|_{\mathsf{F}}^2 \\ \text{s.t.} \quad & \mathbf{B} \geq \mathbf{0}, \mathbf{E} \geq \mathbf{0}, \mathbf{U} \geq \mathbf{0}, \mathbf{V} \geq \mathbf{0}. \end{aligned} \tag{1}$$

We introduce $\mathbf{B} \in \mathbb{R}^{\mathrm{N}\times\mathrm{K}}$ to quantify the knowledge-exercise linkage strength.

In problem 1, $\mathbf{X}$ and $\mathbf{Q}$ share the matrix $\mathbf{E}$, which project the two nonnegative vectors $\mathbf{U}_{:m}$ and $\mathbf{V}_{:k}$ into the new basis $\mathbf{E}$. Hence, we specify the students' knowledge proficiency via the matrix $\mathbf{A} \in \mathbb{R}^{\mathrm{K}\times\mathrm{M}}$.

Model Formulation - Decoder

We introduce the exercise Ex$_n$'s difficulty level μ_n, and $\mathbf{M} = [\mu_1, \mu_2, \cdots, \mu_\mathrm{N}]^\top$. Using $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{M}$, the decoder is

$$\min_{\mathbf{B}_{n:}, \mathbf{A}_{:m}, \mu_n \forall n,m} -\ell + \frac{\gamma}{2} \sum_{n=1}^{\mathrm{N}} \|\mathbf{B}_{n:}\|_2^2. \tag{2}$$

ℓ is a likelihood function:

$$\ell = \sum_{(n,m)\in\Omega_\text{obs}} \log \left\{ \Phi(\mathbf{B}_{n:}\mathbf{A}_{:m} + \mu_n)^{\mathbf{X}_{nm}} \left[1 - \Phi(\mathbf{B}_{n:}\mathbf{A}_{:m} + \mu_n)\right]^{(1-\mathbf{X}_{nm})} \right\}$$

$\Phi(x)$ is a standard *inverse link* function, we choose the probit function as

$$\Phi(x) = \int_{-\infty}^{x} \mathcal{N}(t)\mathrm{d}t = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-t^2/2}\mathrm{d}t.$$

Model Formulation - Objective Function

By combining the encoder and the decoder components, we use $\mathcal{O}_{\mathsf{AF}}$ to denote the objective function of AE-NMCF. The optimization problem is given as

$$\begin{aligned} \min_{\mathbf{B},\mathbf{E},\mathbf{U},\mathbf{V},\mathbf{M}} \quad & \mathcal{O}_{\mathsf{AF}} = -\ell + \|\mathbf{W} \odot (\mathbf{X} - \mathbf{E}\mathbf{U})\|_{\mathsf{F}}^2 + \|\mathbf{Q} \odot (\mathbf{B} - \mathbf{E}\mathbf{V})\|_{\mathsf{F}}^2 + \frac{\gamma}{2} \sum_{n=1}^{\mathrm{N}} \|\mathbf{B}_{n:}\|_2^2, \\ \text{s.t.} \quad & \mathbf{B} \geq \mathbf{0}, \mathbf{E} \geq \mathbf{0}, \mathbf{U} \geq \mathbf{0}, \mathbf{V} \geq \mathbf{0}. \end{aligned}$$

In this problem, the negative log-likelihood term ℓ is convex; while the second and third terms are convex in either $\mathbf{B}$ only, $\mathbf{E}$ only, $\mathbf{U}$ only, or $\mathbf{V}$ only.

Since the nonnegative constraints on $\mathbf{B}$, $\mathbf{E}$, $\mathbf{U}$, $\mathbf{V}$, and the blocks of variables $\{\mathbf{B}_{n:}\}_{n=1}^{\mathrm{N}}$ and $\{\mathbf{A}_{:m}\}_{m=1}^{\mathrm{M}}$, we apply the *projected gradient method via a block coordinate descent* (PG-BCD) approach.

Model Solution - First-order Method

Considering second-order methods do not scale well to high-dimensional problems, we build our algorithm on first-order methods. For example, for $\mathbf{B}_{n:}$:

$$\mathbf{B}_{n:}^{(l+1)} \leftarrow \left[\mathbf{B}_{n:}^{(l)} - \eta_{\mathbf{B}_{n:}}^{(l)} \nabla \mathcal{O}_{\mathsf{AF}}^{(l)}(\mathbf{B}_{n:})\right]_{+}, \tag{3}$$

where $[x]_+ = max(\epsilon, x)$ ensures the nonnegativity. The gradient $\nabla \mathcal{O}_{\mathsf{AF}}(\mathbf{B}_{n:})$ is

$$-\sum_{m} \left\{\mathbf{D}_{nm}[\mathbf{X}_{nm} - \Phi(\Delta_{nm})]\mathbf{U}_{:m}^{\top}\mathbf{V}\right\} + 2[\mathbf{Q}_{n:} \odot \mathbf{B}_{n:} - \mathbf{Q}_{n:} \odot (\mathbf{E}_{n:}\mathbf{V})] + \gamma \mathbf{B}_{n:};$$

where $\Delta_{nm} = \mathbf{B}_{n:}\mathbf{V}^{\top}\mathbf{U}_{:m} + \mu_n$, and $\mathbf{D}_{nm}$ is given by

$$\mathbf{D}_{nm} = \frac{\mathcal{N}(\Delta_{nm})}{\Phi(\Delta_{nm})[1 - \Phi(\Delta_{nm})]}.$$

Model Solution - Lipschitz Constant Step Size

In Eq. (3), a key issue is choosing the appropriate $\eta_{\mathbf{B}_{n:}}^{(l)}$, we determine $\eta_{\mathbf{B}_{n:}}^{(l)}$ by the *Lipschitz* constant to enable efficient solution, and set $\eta^{(l)} = 1/L$, where L is the *Lipschitz* constant of ∇f.

For $\mathbf{B}_{n:}$, we establish an upper bound of the l_2-norm of the difference between the gradients at two arbitrary points $\mathbf{y}$ and $\mathbf{z}$ for $\nabla \mathcal{O}_{\mathsf{AF}}(\mathbf{B}_{n:})$ as

$$\|\nabla \mathcal{O}_{\mathsf{AF}}(\mathbf{y}) - \nabla \mathcal{O}_{\mathsf{AF}}(\mathbf{z})\|_2 \leq \underbrace{\left(L_p \sigma_1^2(\mathbf{U}^\top \mathbf{V}) + 2\sqrt{\sum_{k=1}^{\mathrm{K}} \mathbf{Q}_{nk}^2} + \gamma \right)}_{\text{Lipschitz constant}} \|\mathbf{y} - \mathbf{z}\|_2,$$

where $L_p = 1$ is the scalar Lipschitz constant of the probit function, and $\sigma_1(\mathbf{X})$ denotes the largest singular value of the matrix $\mathbf{X}$.

Experiments - Data sets and Baselines

We use real-world students' response data with different sparsities and knowledge-exercise relations, which cover diversified academic subjects, as shown in Table 2.

Table: The statistics of data sets

Statistics	Data Set					
	FrcSub	Junyi-s	Quanlang-s	SLP-Bio-s	SLP-His-s	SLP-Eng
# Student	536	1,091	50	100	1057	360
# Exercise	20	9	107	129	326	362
# Knowledge concept	8	9	14	7	14	19
Subject	Math	Math	Math	Biology	History	English
Relations	many-to-many	one-to-one	one-to-many	one-to-many	one-to-many	one-to-many
Sparsity	0%	75.03%	68.67%	54.92%	84.28%	96.92%

The baselines include data mining approaches and cognitive diagnosis models.

- **Data mining approaches:** NMF [1], MCF [2], GNMF [3], NMMF [4], and SNMCF [5].
- **Cognitive diagnosis models:** DINA [6], DIRT [7], DeepCDF [8], and QRCDM [9]

The evaluation metrics are (*a*) ACC and RMSE for student performance prediction; and (*b*) KRC (r_c) for knowledge proficiency estimation.

Experiments - Overall Performance

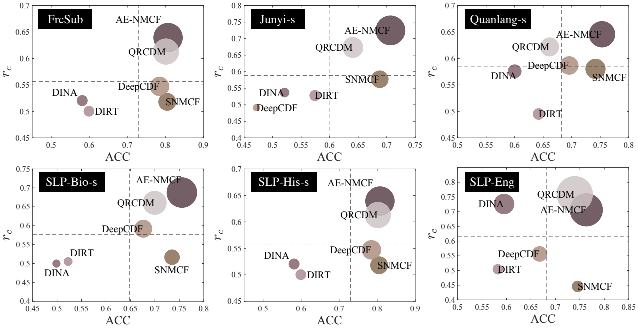


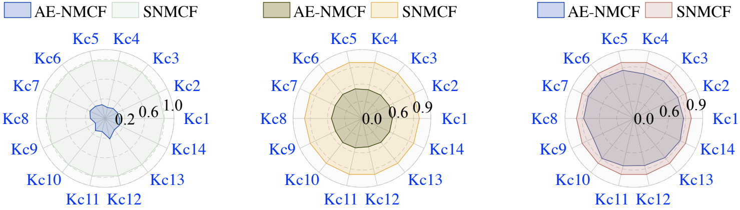
Figure: Model comparison in balancing the two learning tasks.

Results

- The closer to the upper right corner with a larger bubble size, the better the balance achieved.
- AE-NMCF is well above the model average (indicated by dash lines) on all data sets, which achieves the best balance between prediction accuracy and diagnostic ability and works with multiple relation cases.

Experiments - Case Study

We present the diagnostic results for case students on Quanlang-s, compared with the advanced baseline SNMCF.



(*a*) St50 (46.67%) (*b*) St33 (66.36%) (*c*) St42 (84.62%)

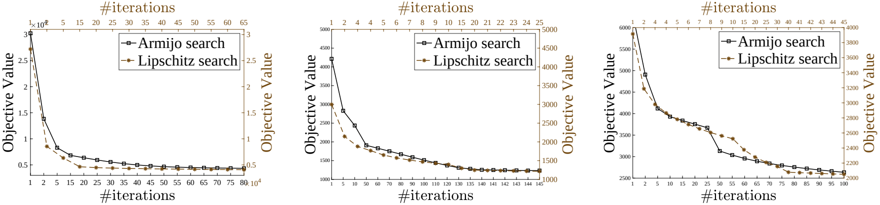
Figure: Diagnosis results of three case students between AE-NMCF and SNMCF.

Results

- The radar charts measure the three students' knowledge proficiency, their answer accuracy rates (the ratio of correctly answering all exercises) are in the parentheses.
- Intuitively, the proficiency levels of St_{42} should be the highest, and St_{50} is at the lowest level.
- SNMCF gives an extreme estimation, instead, AE-NMCF gives reasonable results.

Lipschitz Search v.s. Armijo Search

We compare the step-size search based on *Lipschitz* constant with the one based on the “Armijo” rule.



(a) FrcSub (b) Junyi-s (c) Quanlang-s

Figure: #iterations vs. objective values for the *Lipschitz* search and *Armijo* search.

Results

- The Armijo search at first quickly decreases the objective function value but slows down in sequence, which takes more time to converge.
- The Lipschitz search achieves the fastest convergence while maintaining a relatively small objective function value.

Conclusion and Future Work

Conclusion

- We study student cognitive modeling from a data mining perspective, in which students' knowledge proficiency estimation is our primary concern.
- We introduce AE-NMCF for improved student cognitive modeling, which provides an end-to-end and data-driven way of specifying and assessing students' understanding of a set of knowledge concepts.
- To learn the model, we present a novel projected gradient method based on block coordinate descent with Lipschitz constants, for which theoretical convergence is guaranteed.
- AE-NMCF provides a good fit to the students' knowledge proficiency while maintaining student performance prediction that is comparable to other student cognitive models.

Future Work

- Considering the learning dependency of knowledge concepts.
- Investigating other efficient parameter learning methods and exploring their scalability.

[1] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *International Conference on Neural Information Processing Systems*, pages 556–562, 2000.

[2] Ajit P Singh and Geoffrey J Gordon. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 650–658, 2008.

[3] Hyekyoung Lee and Seungjin Choi. Group nonnegative matrix factorization for eeg classification. In *Artificial Intelligence and Statistics*, pages 320–327. PMLR, 2009.

[4] Koh Takeuchi, Katsuhiko Ishiguro, Akisato Kimura, and Hiroshi Sawada. Non-negative multiple matrix factorization. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 1713–1720, 2013.

[5] Shenbao Yu, Yifeng Zeng, Yinghui Pan, and Fan Yang. Snmcf: a scalable non-negative matrix co-factorization for student cognitive modeling. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–14, 2023.

[6] Jimmy De La Torre. Dina model and parameter estimation: a didactic. *Journal of Educational and Behavioral Statistics*, 34(1):115–130, 2009.

[7] Song Cheng, Qi Liu, Enhong Chen, Zai Huang, Zhenya Huang, Yiying Chen, Haiping Ma, and Guoping Hu. Dirt: deep learning enhanced item response

theory for cognitive diagnosis. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2397–2400, 2019.

[8] Lina Gao, Zhongying Zhao, Chao Li, Jianli Zhao, and Qingtian Zeng. Deep cognitive diagnosis model for predicting students' performance. *Future Generation Computer Systems*, 126:252–262, 2022.

[9] Haowen Yang, Tianlong Qi, Jin Li, Longjiang Guo, Meirui Ren, Lichen Zhang, and Xiaoming Wang. A novel quantitative relationship neural network for explainable cognitive diagnosis model. *Knowledge-Based Systems*, 250:109156, 2022.