University at Buffalo
The State University of New York
Department of Industrial and Systems Engineering
School of Engineering and Applied Sciences

Student Modeling for Learning Curve Estimation

Alexander Nikolaev, Alireza Farasat

CDSE Days, April, 2018

Outline:

- Motivation
- State of the Art in Student Modeling
- A Generative Model of Knowledge Acquisition Dynamics
- Tensor Based Modeling and Inference
- Constrained Optimization for Sparse Tensor Factorization (STF)
- Performance Evaluation
- Conclusion

Motivation

- The concerns for equity and student privacy protection present new challenges for *individual-focused learning* [1].

- The most commonly adopted teaching strategies are *NOT flexible* enough to account for the diversity of learning capabilities of students [2].

- Online tools and *Intelligent Tutoring* Algorithms to the rescue?..

[1] Visscher, Adrie J., and Robert Coe. "School performance feedback systems: Conceptualisation, analysis, and reflection." *School effectiveness and school improvement* 14.3 (2003): 321-349.

[2] Wu, Hsin-Kai, et al. "Current status, opportunities and challenges of augmented reality in education." *Computers & Education* 62 (2013): 41-49.

Motivation

- Assessing the level of *conceptual understanding* of the new material by the students is difficult.

- It is also hard to gauge *how challenging* a given problem may be for different students at different stages of the learning progress.

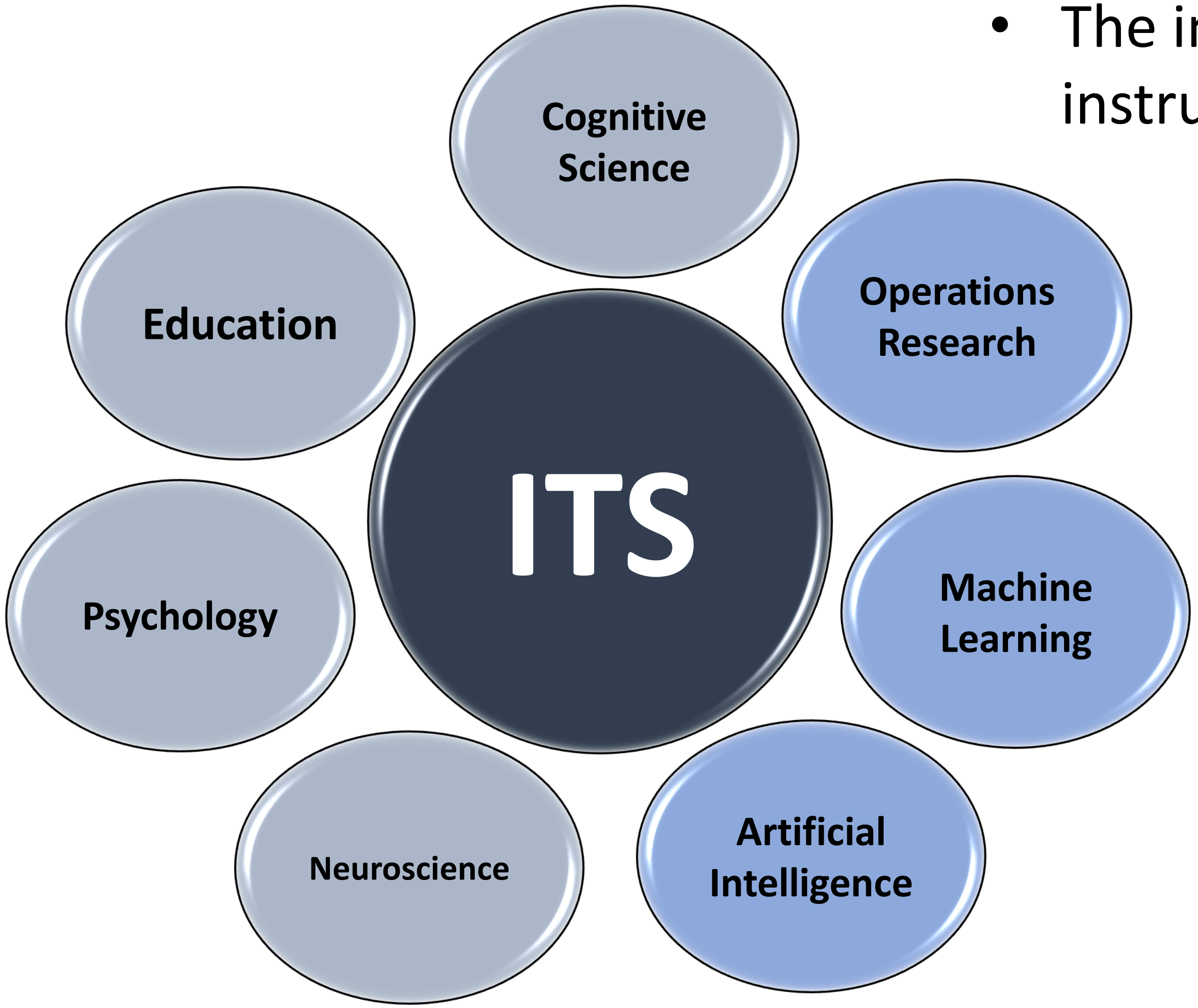
- The same quiz/exam problems tend *to be re-used* for instruction and assessment in many educational settings, so over time, large volumes of data can be compiled about their effectiveness. This is particularly true for courses that employ *Multiple Choice Questions*.

Contributions

This research:

- Advances the state-of-the-art in Sparse Factor Analysis by applying Probabilistic Sparse Tensor Factorization to analyze the *dynamics* of learning.
- Provides models with interpretable parameters describing students conceptual understanding.

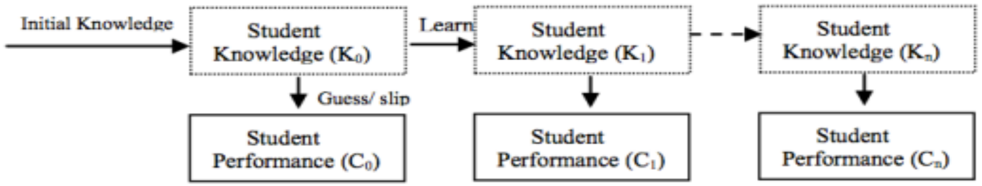
ITS Research Areas



- The introduction of ITS and computer-assisted instruction dates back to the 1960s [3].

[3] Polson, Martha C., and J. Jeffrey Richardson. *Foundations of intelligent tutoring systems*. Psychology Press, 2013.

ITS and Student Modeling



Student Modeling

Prescriptive Models

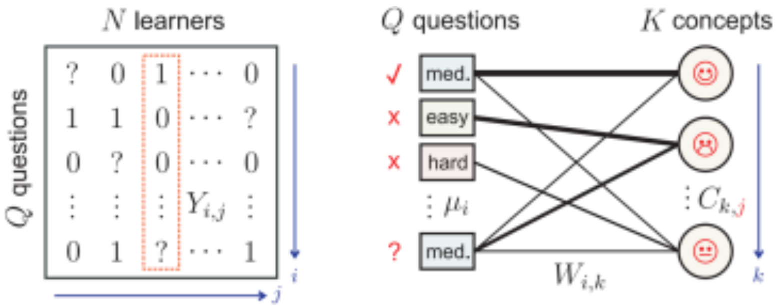
- Reinforcement Learning
- Markov Decision Process
- POMDP [6]

Predictive Models

- Performance Factors Analysis
- SPARse Factor Analysis (SPARFA) [5]
- Recurrent Neural Nets
- Time-varying SPARFA
- Deep knowledge Tracing
- …

Descriptive Models

- Bayesian Knowledge Tracing (BKT) [4]
- Extensions of BKT
- Hidden Markov Models
- …



[4] Yudelson, Michael V., Kenneth R. Koedinger, and Geoffrey J. Gordon. "Individualized bayesian knowledge tracing models." *Artificial Intelligence in Education*. Springer Berlin Heidelberg, 2013.

[5] Lan, Andrew S., et al. "Sparse factor analysis for learning and content analytics." *The Journal of Machine Learning Research* 15.1 (2014): 1959-2008.

[6] Rafferty, Anna N., et al. "Faster teaching by pomdp planning." *Artificial intelligence in education*. Springer Berlin Heidelberg, 2011.

Challenges of Student Modeling

Expressing Knowledge
- Students' knowledge / skill levels are unobservable and hard to quantify

Noisy Data
- A student may forget / lose a skill; a correct answer may be a lucky guess...

Cold-start
- No data are available for new students or newly added activities

Missing Values
- Each student only answers a limited subset of questions (Sparsity)

Interpretability of Models
- Models should help generate insights

No unified framework yet exists that would address all these issues

A Promising Research Direction

Some Answers

The proposed modeling approach …

- Can infer students' levels of knowledge from their performance.
- Enables one to interpret the model parameters.
- Can be employed to find the best teaching policy for each student.
- Offers opportunities for us to learn more about how learning happens.

How: Constrained Sparse Tensor-Based Modeling

Static Student Model (SSM)

- SSM extends SPARFA by considering time as the 3rd dimension.
- It serves as a base model that accounts for the dynamics of knowledge acquisition.

Homogeneous Student Model (HSM)

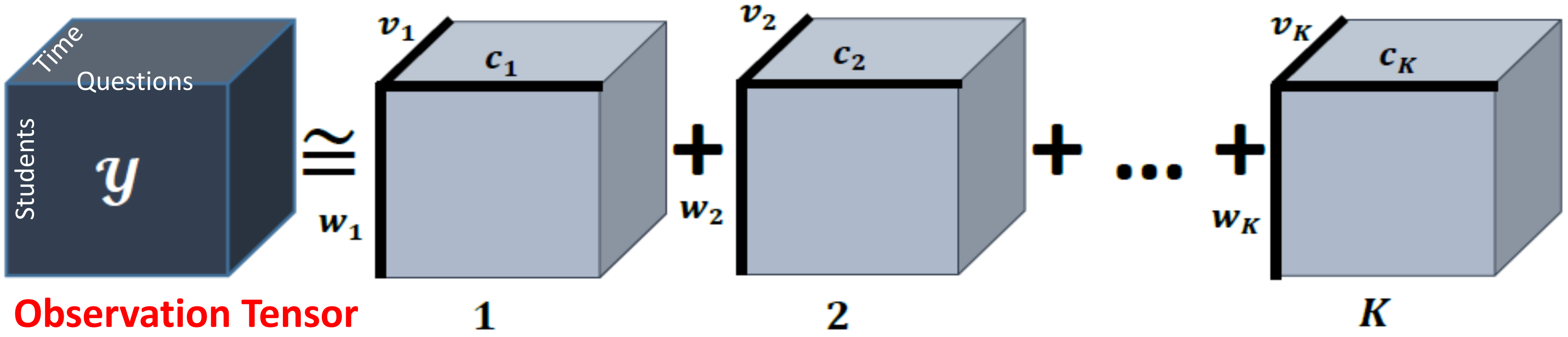
- HSM offers more interpretable parameters than SSM.
- It acknowledges the fact that knowledge and skills tend to improve in a fairly steady way.

Personalized Student Model (PSM)

- PSM employs a new tensor factorization model/technique
- It assumes learning by doing.
- Students have different learning rates in different areas.

Model 1 - Static Student Model (SSM)

- SSM views time as the 3rd dimension of the object under study and replaces MF with TF.



C: Student-knowledge area Matrix

$$\mathcal{Y} \approx W \otimes C \otimes V$$

W: Question-knowledge area Matrix

V: Time-knowledge area Matrix

- The probability that a student correctly solves a question comes from a Bernoulli PMF

$$\mathcal{Y}_{ijt} \sim Ber(\Phi(\mathcal{T}_{ijt})),$$

Logit Function

$$\boxed{P(\mathcal{Y}_{ijt}|\mathbf{w_i}, \mathbf{c_j}, \mathbf{v_t}, d_i, \theta_j)} = \Phi(\mathcal{T}_{ijt})^{\mathcal{Y}_{ijt}}[1 - \Phi(\mathcal{T}_{ijt})]^{1-\mathcal{Y}_{ijt}}$$

Probability of answering a question

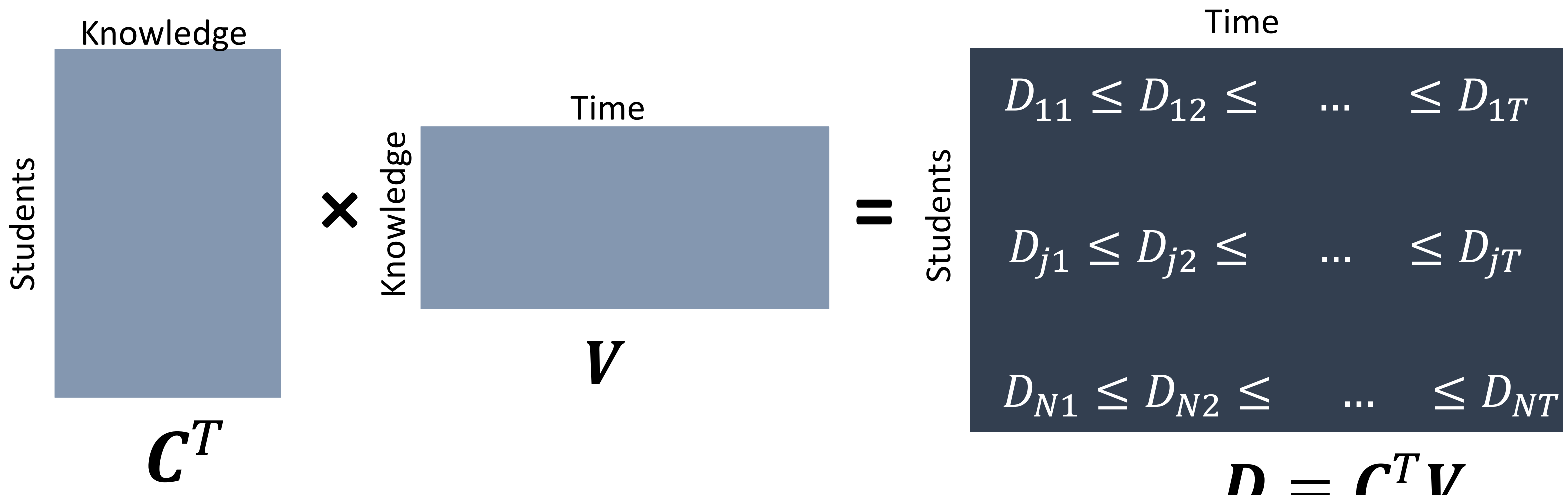
***θ*:** student parameter

$$\mathcal{T}_{ijt} = \sum_{k=1}^{K} W_{ki} C_{kj} V_{kt} - d_i + \theta_j \quad \forall \quad i = 1, \ldots, Q \;\; j = 1, \ldots, N, \; t = 1 \ldots T$$

***d*:** question difficulty

Model 2 - Homogenous Student Model (HSM)

- HSM builds upon the idea of SSM by controlling the way in which the knowledge is expected to be acquired.
  - $\boldsymbol{C}^T\boldsymbol{V}$ shows the students' learning trajectory over time.



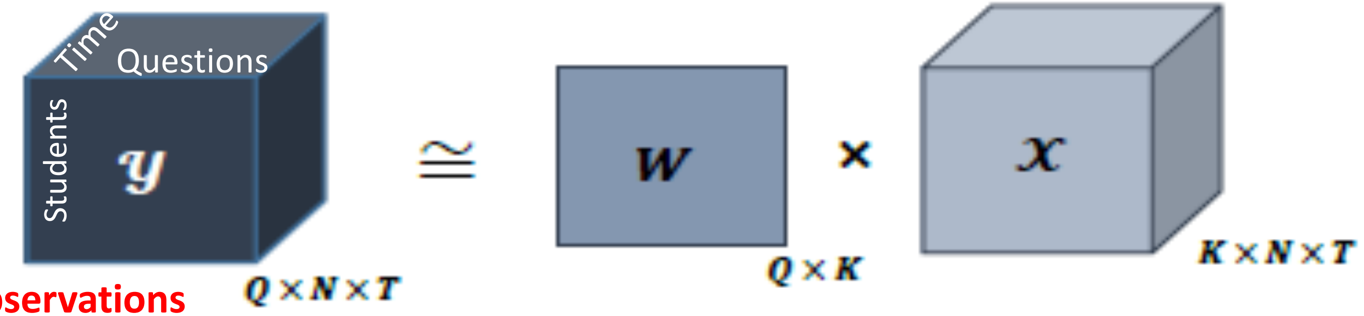
$$\boldsymbol{D} = \boldsymbol{C}^T\boldsymbol{V}$$

- $D_{kt} \leq D_{kt+1}$ means that:

$$\sum_{k=1}^{K} C_{kj}V_{kt+1} - \sum_{k=1}^{K} C_{kj}V_{kt} \geq -\epsilon \quad \forall\ j = 1\dots,N, \quad t = 1\dots,T,$$

Model 3 - Personalized Student Model (PSM)

- PSM assumes that learning occurs in a personalized way, i.e., different students may have different learning curves.



Observations

W: Question-knowledge are Matrix

$$\mathcal{Y} \approx \boldsymbol{W} \otimes \mathcal{X}$$

X: Student-knowledge area-time Tensor

- The probability that a student correctly solves a problem follows a Bernoulli distribution

$$\mathcal{Y}_{ijt} \sim Ber(\Phi(\mathcal{T}_{ijt})),$$

Logit Function

$$\boxed{P(\mathcal{Y}_{ijt}|\mathbf{w}_i,\mathcal{X}_{:jt},d_i,\theta_j)} = \Phi(\mathcal{T}_{ijt})^{\mathcal{Y}_{ijt}}[1-\Phi(\mathcal{T}_{ijt})]^{1-\mathcal{Y}_{ijt}}$$

Probability of answering a question

$\boldsymbol{\theta}$: student parameter

$$\mathcal{T}_{ijt} = \sum_{k=1}^{K} W_{ik}\mathcal{X}_{kjt} - d_i + \theta_j \quad \forall \quad i = 1,\ldots,Q \quad j = 1,\ldots,N, \; t = 1\ldots T$$

Personalized Student Model – Optimization Problem Formulation

$$(P3): \max_{\mathbf{w},\mathcal{X},v,\theta} \sum_{(i,j,t)\in\Omega_{obs}} log(P(\mathcal{Y}_{ijt}|\mathbf{w_i},\mathcal{X}_{:jt},d_i,\theta_j))$$

Maximize Log-Likelihood of Observation

$$P(\mathcal{Y}_{ijt}|\mathbf{w_i},\mathcal{X}_{:jt},d_i,\theta_j) = \Phi(\mathcal{T}_{ijt})^{\mathcal{Y}_{ijt}}[1-\Phi(\mathcal{T}_{ijt})]^{1-\mathcal{Y}_{ijt}}$$

s.t.

$\|\mathbf{w_i}\|_1 \leq \delta \ \forall \ i = 1,\ldots,Q$ — Norm 1 controls sparsity

$\|\mathbf{w_i}\|_2 \leq \beta \ \forall \ i = 1,\ldots,Q$ — Norm 2 helps convergence

$\|\mathcal{X}_{::t}\|_F = \xi \ \forall \ t = 1,\ldots,T$ — Prevent unbound growth in tensor X

$\mathcal{X}_{kjt} \leq \mathcal{X}_{kjt+1} \ \forall \ k = 1,\ldots,K, \ j = 1,\ldots,N, \ t = 0,\ldots,T-1$

Knowledge accumulation/acquisition

$W_{ik} \geq 0 \ \forall \ i = 1,\ldots,Q, \ K = 1,\ldots,K,$

$\mathcal{X}_{kj0} \geq 0 \ \forall \ k = 1,\ldots,K, \ j = 1,\ldots,N.$

Non-negativity constraints

Gradient:

$$\nabla\Gamma_{W_{ki}} = -\sum_{j=1}^{N}\sum_{t=1}^{T} C_{kj}V_{kt}(\mathcal{Y}_{ijt} - \frac{1}{1+e^{-\mathcal{T}_{ijt}}}) + \frac{\lambda_2 W_{ki}}{\|W\|_F} \quad (24)$$

$$\nabla\Gamma_{C_{kj}} = -\sum_{i=1}^{Q}\sum_{t=1}^{T} W_{ki}V_{kt}(\mathcal{Y}_{ijt} - \frac{1}{1+e^{-\mathcal{T}_{ijt}}}) + \frac{\lambda_3 C_{kj}}{\|C\|_F} \quad (25)$$

$$\nabla\Gamma_{V_{kt}} = -\sum_{j=1}^{N}\sum_{i=1}^{Q} C_{kj}W_{ki}(\mathcal{Y}_{ijt} - \frac{1}{1+e^{-\mathcal{T}_{ijt}}}) + \frac{\lambda_4 V_{ki}}{\|V\|_F} \quad (26)$$

$$\nabla\Gamma_{d_i} = \sum_{j=1}^{N}\sum_{t=1}^{T} (\mathcal{Y}_{ijt} - \frac{1}{1+e^{-\mathcal{T}_{ijt}}}) \quad (27)$$

$$\nabla\Gamma_{\theta_j} = -\sum_{i=1}^{Q}\sum_{t=1}^{T} (\mathcal{Y}_{ijt} - \frac{1}{1+e^{-\mathcal{T}_{ijt}}}) \quad (28)$$

Optimization Algorithm - BCD

Algorithm 1 Block Coordinate Descent Algorithm-STF

Input: Observed Tensor $\mathcal{Y}$, $N, Q, T, K, \lambda, \mu, \gamma, \beta$.
Output: Completed Tensor $\mathcal{Y}$, Matrices $\mathbf{W}$, $\mathbf{C}$ and $\mathbf{V}$.
BlockCoordinateDescent()

Initialize randomly $\mathbf{W}$, $\mathbf{C}$ and $\mathbf{V}$.
while (stopping criteria) **do**
  randomly select $i \in \{1, \ldots, Q\}$, $j \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T\}$;
  calculate $\nabla\Gamma_{\mathbf{w}_i}$, $\nabla\Gamma_{\mathbf{c}_j}$, $\nabla\Gamma_{\mathbf{v}_t}$, $\nabla\Gamma_{d_i}$ and $\nabla\Gamma_{\theta_j}$ using (24)-(28)
  update $\mathbf{w}_i^{new} \leftarrow \max\{0, \mathbf{w}_i^{old} - \beta\nabla\Gamma_{\mathbf{w}_i}\}$
  update $\mathbf{c}_j^{new} \leftarrow \frac{1}{1+\beta\gamma}\left(\max\{0, \mathbf{c}_j^{old} - \beta\nabla\Gamma_{\mathbf{c}_j}\}\right)$
  update $\mathbf{v}_t^{new} \leftarrow \frac{1}{1+\beta\gamma}\left(\max\{0, \mathbf{v}_t^{old} - \beta\nabla\Gamma_{\mathbf{v}_t}\}\right)$
  update $d_i^{new} \leftarrow d_i^{old} - \beta\nabla\Gamma_{d_i}$
  update $\theta_j^{new} \leftarrow \theta_j^{old} - \beta\nabla\Gamma_{\theta_j}$
  calculate objective function using (23)
end
report AIC, AIC_c, BIC, MSE

- The idea of BCD is to randomly select a set of coordinate to update in each iteration [10].
- Very fast and scalable.
- Convergence is an issue.

[10] Xu, Yangyang, and Wotao Yin. "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion." *SIAM Journal on imaging sciences* 6.3 (2013): 1758-1789.

Optimization Algorithm - ADMM

Alternating Direction Method of Multipliers (ADMM)

Algorithm 2 Alternating Direction Method of Multipliers-STF

Input: Observed Tensor $\mathcal{Y}$, N, Q, T, K, $\lambda, \mu, \gamma, \beta$.
Output: Completed Tensor $\mathcal{Y}$, Matrices $\mathbf{W}$, $\mathbf{C}$ and $\mathbf{V}$.
AlternatingDirectionMethodMultipliers()

Initialize randomly $\mathbf{W}$, $\mathbf{C}$ and $\mathbf{V}$.
while (stopping criteria) **do**
  choose step sizes β_W, β_d, β_C, β_θ and β_V
  while (stopping criteria) **do**
    calculate $\nabla\Gamma_{\mathbf{w}}$ and $\Gamma_{\mathbf{d}}$ using (24) and (27)
    update $\mathbf{w}^{new} \leftarrow \max\{0, \mathbf{w}^{old} - \beta_W \nabla\Gamma_{\mathbf{w}}\}$
    update $\mathbf{d}^{new} \leftarrow \mathbf{d}^{old} - \beta_d \nabla\Gamma_{\mathbf{d}}$
  while (stopping criteria) **do**
    calculate calculate $\nabla\Gamma_{\mathbf{c}}$ and $\nabla\Gamma_{\boldsymbol{\theta}}$ using (25) and (28)
    update $\mathbf{C}^{new} \leftarrow \frac{1}{1+\beta_C\gamma}\left(\max\{0, \mathbf{C}^{old} - \beta_C \nabla\Gamma_{\mathbf{C}}\}\right)$
    update $\boldsymbol{\theta}^{new} \leftarrow \boldsymbol{\theta}^{old} - \beta_\theta \nabla\Gamma_{\boldsymbol{\theta}}$
  while (stopping criteria) **do**
    calculate $\nabla\Gamma_{\mathbf{V}}$ using (26) and (28)
    update $\mathbf{V}^{new} \leftarrow \frac{1}{1+\beta_V\gamma}\left(\max\{0, \mathbf{V}_t^{old} - \beta\nabla\Gamma_{\mathbf{V}}\}\right)$
  calculate objective function using (23)
end
report AIC, AIC_c, BIC, MSE

- ADMM is an extension of BCD that selects convex sub-problems and finds the optimum of each sub-problem [11].
- Not as fast as BCD but still scalable.
- Convergence is good in practice.

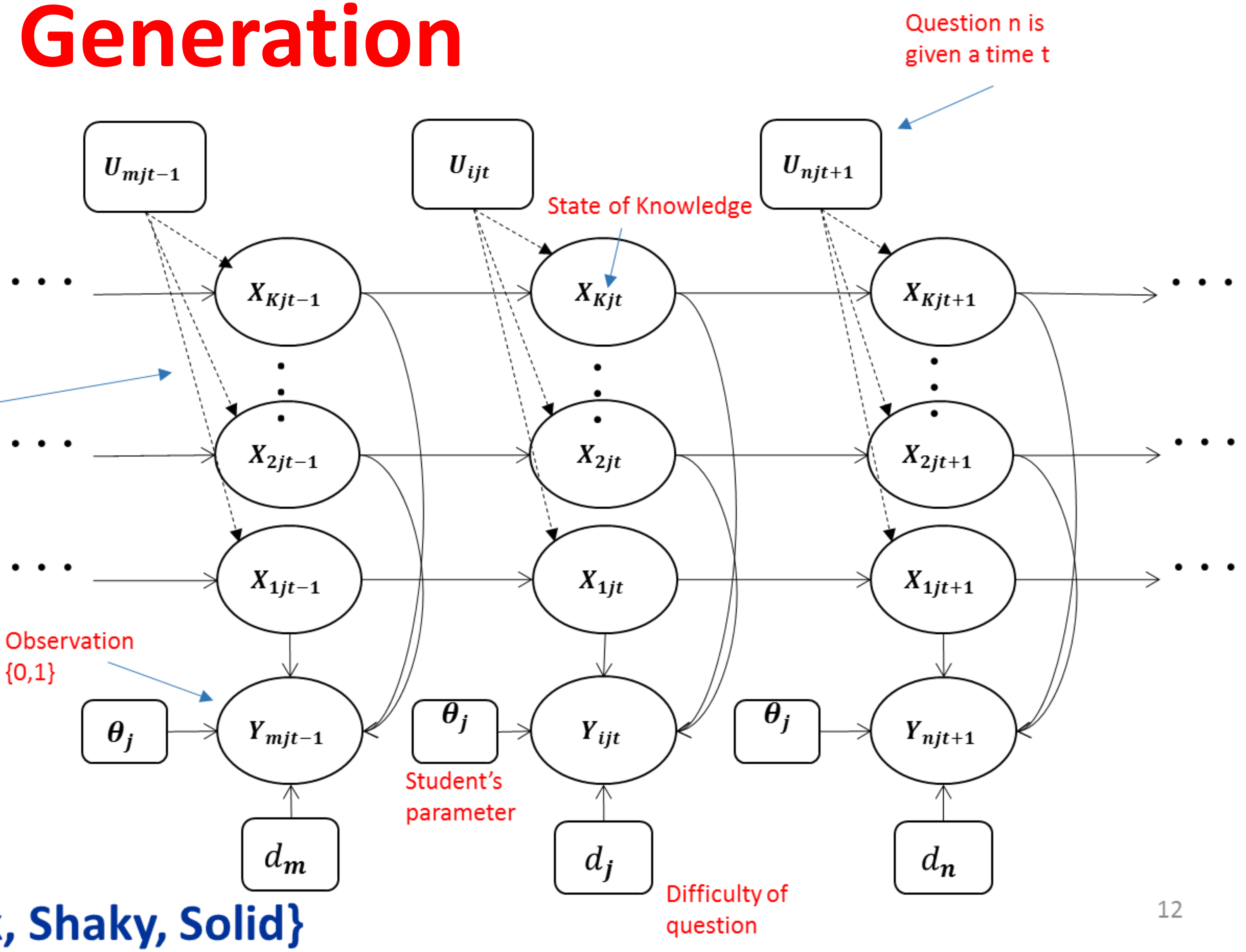
[11] Ouyang, Hua, et al. "Stochastic Alternating Direction Method of Multipliers."*ICML (1)* 28 (2013): 80-88.

Synthetic Data Generation

Why synthetic data?

- There is no ground truth of students’ states of knowledge.
- There is limited publicly available data (due to privacy policies).
- It makes sense to work under the assumption that learning takes place *over time*.
- The flexibility of conducting scalability testing of inference algorithms is valuable.

Synthetic Data Generation



Factorial Hidden Markov Model

X_{kjt}: {Know nothing, Weak, Shaky, Solid}

Synthetic Data Generation

- The objective is to model the underlying dynamics of the learning process.

- Transition Probability Matrix:

State of Knowledge

Question

Probability of transition

Probability of forgetting

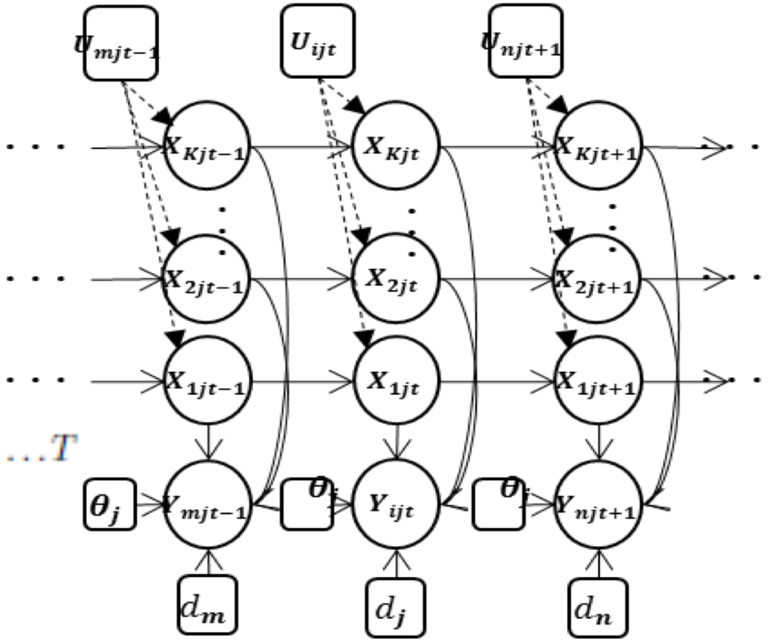
$$
P(\mathcal{X}_{kjt}|\mathcal{X}_{kjt-1},\mathcal{U}_{ijt}) =
\begin{array}{r c}
 & \begin{array}{cccc} \text{Know nothing} & \text{Weak} & \text{Shaky} & \text{Solid} \end{array} \\
\begin{array}{r} \text{Know nothing} \\ \\ \\ \text{Solid} \end{array} &
\begin{bmatrix}
1-\rho_{kj} & \rho_{kj} & 0 & 0 \\
\rho_F & 1-(\rho_F+\rho_{kj}) & \rho_{kj} & 0 \\
0 & \rho_F & 1-(\rho_F+\rho_{kj}) & \rho_{kj} \\
0 & 0 & \rho_F & 1-\rho_F
\end{bmatrix}
\end{array}
\qquad (35)
$$

$$\rho_{kj} = A_{kj} + \mathcal{B}_{ijt} \times \mathcal{U}_{ijt},$$

Learning rate of student

Learning by questions

$$\mathcal{T}_{ijt} = \sum_{k=1}^{K} W_{ik}\mathcal{X}_{kjt} - d_i + \theta_j \quad \forall,\ i = 1,\dots,Q \ \ j = 1,\dots,N,\ t = 1\dots T$$



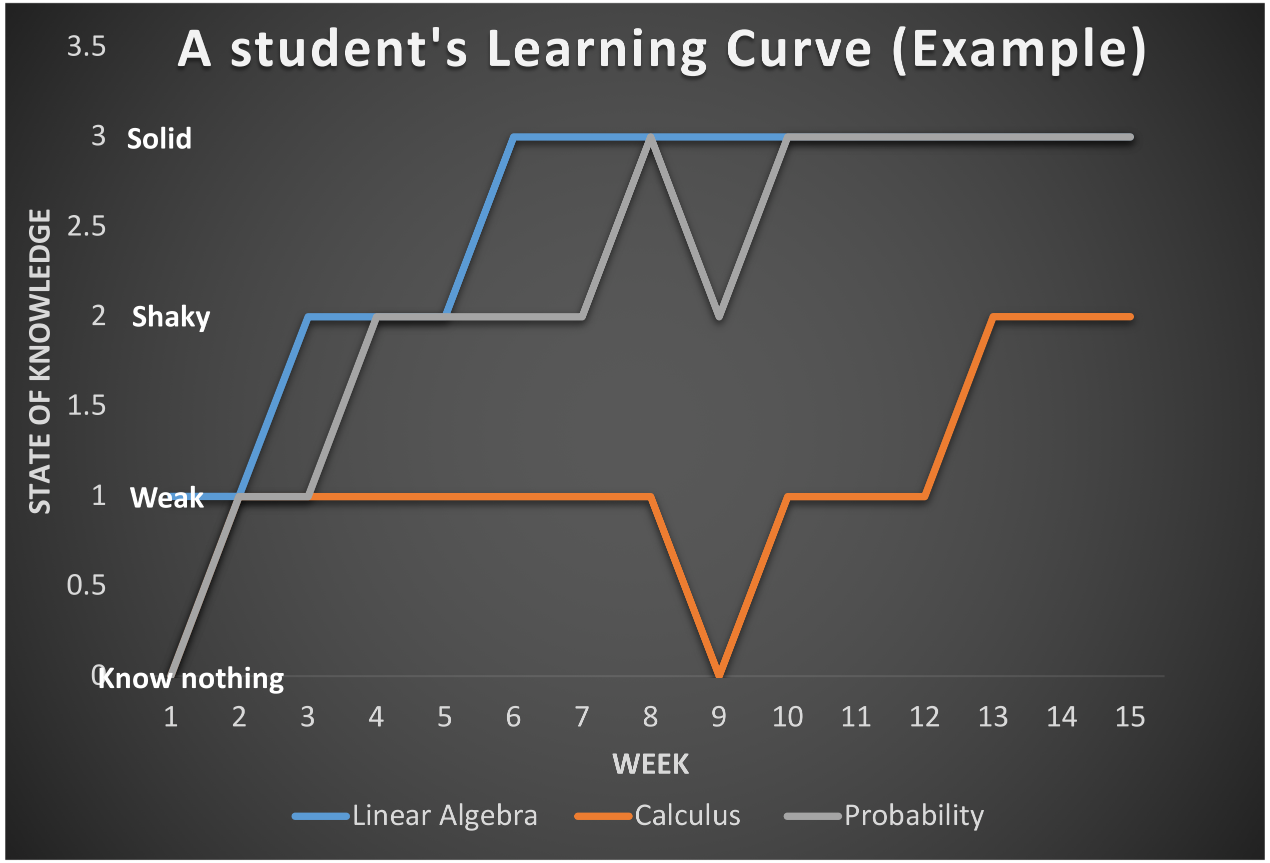
$$P(\mathcal{Y}_{ijt}|\mathbf{w}_i, \mathcal{X}_{:jt}, d_i, \theta_j) = \Phi(\mathcal{T}_{ijt})^{\mathcal{Y}_{ijt}}[1-\Phi(\mathcal{T}_{ijt})]^{1-\mathcal{Y}_{ijt}}$$

Synthetic Data Generation

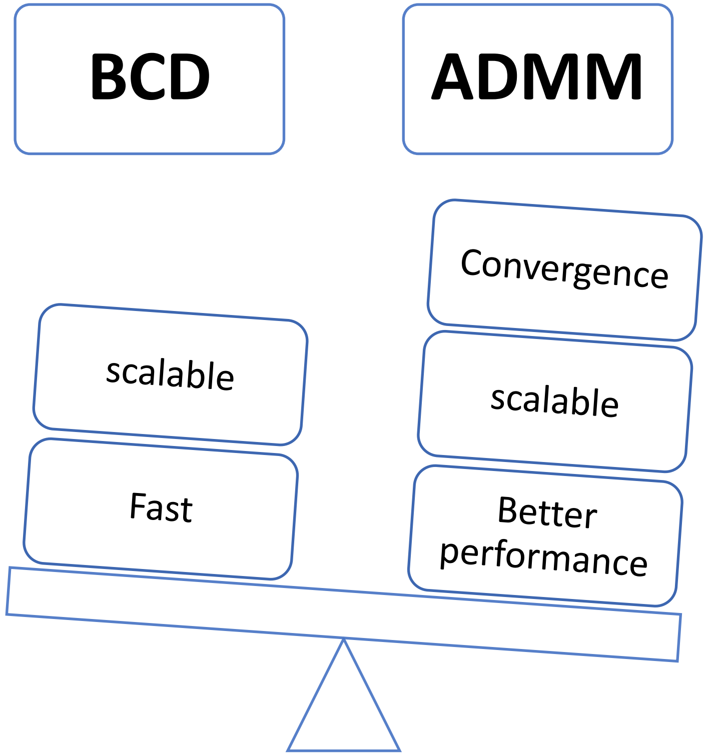
Algorithm 3 Synthetic Data Generator based on Factorial Hidden Markov Model

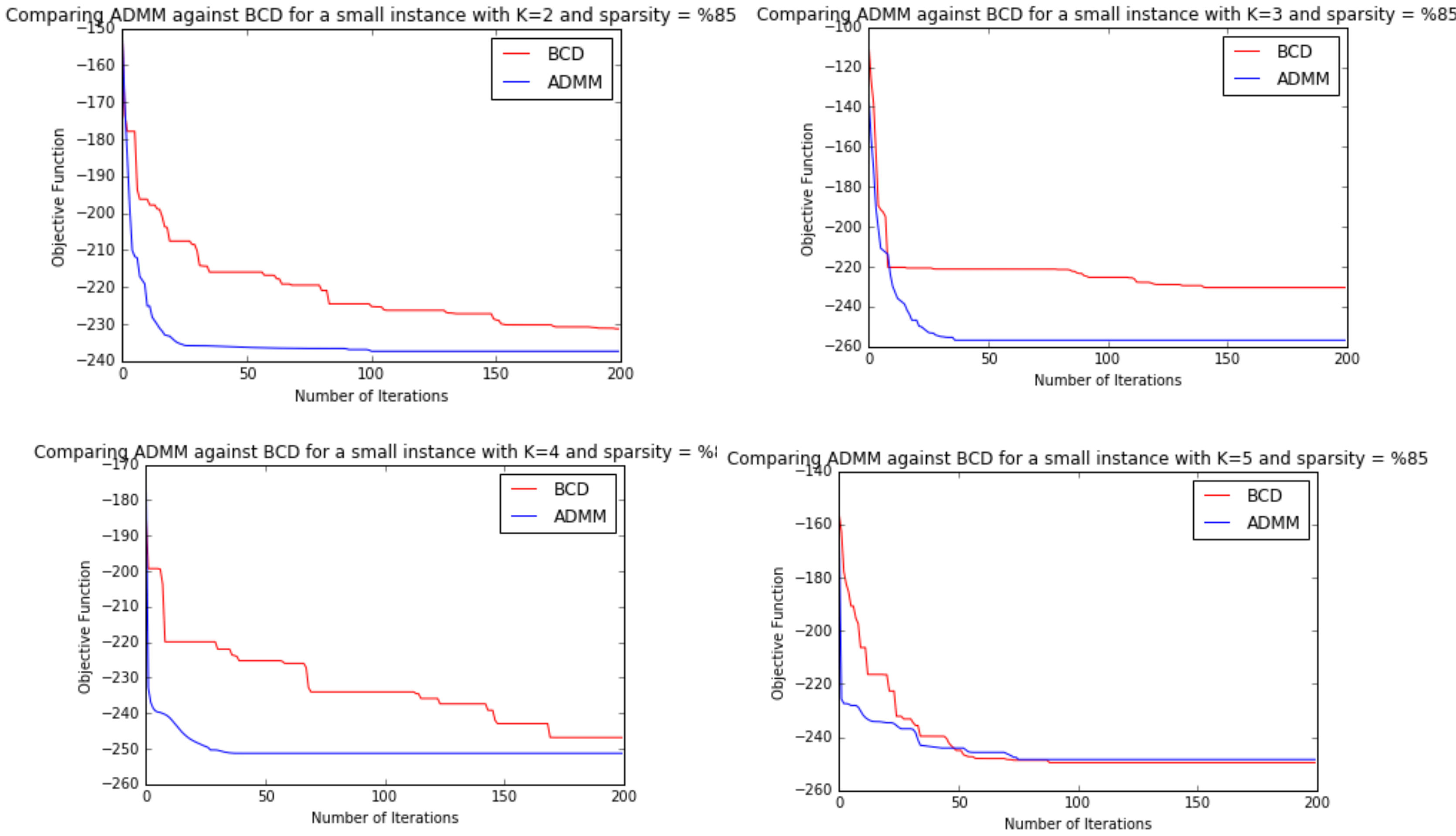
Input: Tensors $\mathcal{B}, \mathcal{U}$, matrices $\mathbf{W}$, $\mathbf{A}$, vectors $\mathbf{d}$, $\boldsymbol{\theta}$ and parameters N, Q, T, K.
Output: Tensors $\mathcal{Y}$ and $\mathcal{X}$.
SyntheticDataGenerator()

for student $j = 1 : N$
    while $(t \leq T - 1)$ **do**
        $\mathcal{U}_{ijt}$, $\mathcal{B}_{ijt} \leftarrow$ propose a question
        for knowledge component $k = 1 : K$
            calculate $P(\mathcal{X}_{kjt}|\mathcal{X}_{kjt-1}, \mathcal{U}_{ijt})$ using (35)
            given $\mathcal{X}_{kjt-1}$, $\mathcal{X}_{kjt} \leftarrow 0, 1, 2$ or 3 with probability $P(\mathcal{X}_{kjt}|\mathcal{X}_{kjt-1}, \mathcal{U}_{ijt})$
            calculate $\mathcal{T}_{ijt} \leftarrow \sum_{k=1}^{K} W_{ik}\mathcal{X}_{kjt} - d_i + \theta_j$
            calculate $P(\mathcal{Y}_{ijt}|\mathbf{w_i}, \mathcal{X}_{:jt}, d_i, \theta_j) \leftarrow \Phi(\mathcal{T}_{ijt})^{\mathcal{Y}_{ijt}}[1 - \Phi(\mathcal{T}_{ijt})]^{1-\mathcal{Y}_{ijt}}$
            choose $\mathcal{Y}_{ijt} = 1$ with $P(\mathcal{Y}_{ijt} = 1|\mathbf{w_i}, \mathcal{X}_{:jt}, d_i, \theta_j)$ and 0 otherwise
        end
        $t \leftarrow t + 1$
    end
end



BCD and ADMM Comparison (SSM)





Tensor-based Student Models-Results

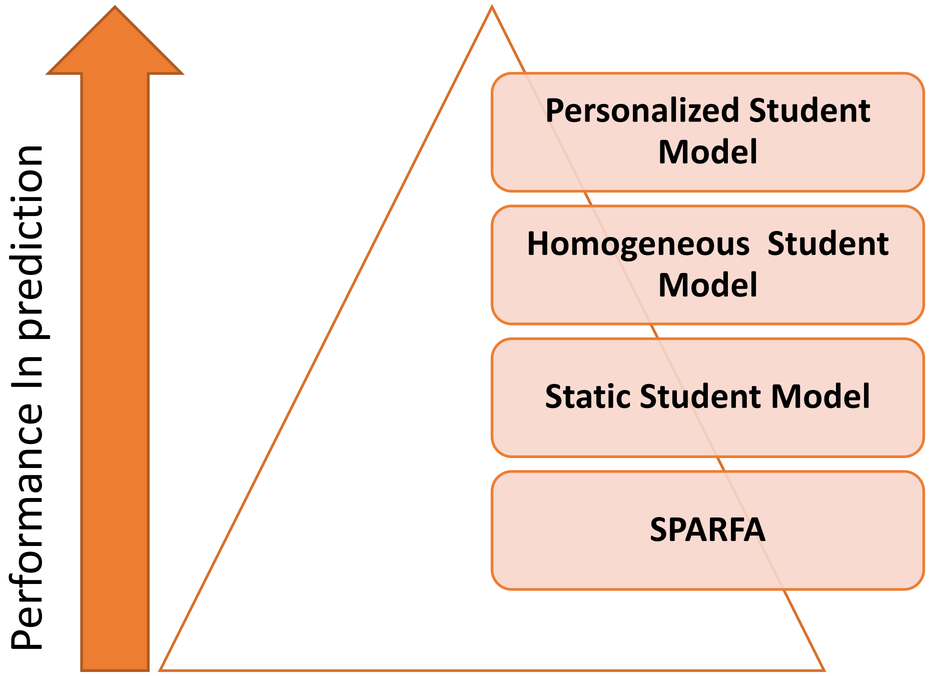
Small-sized Instances

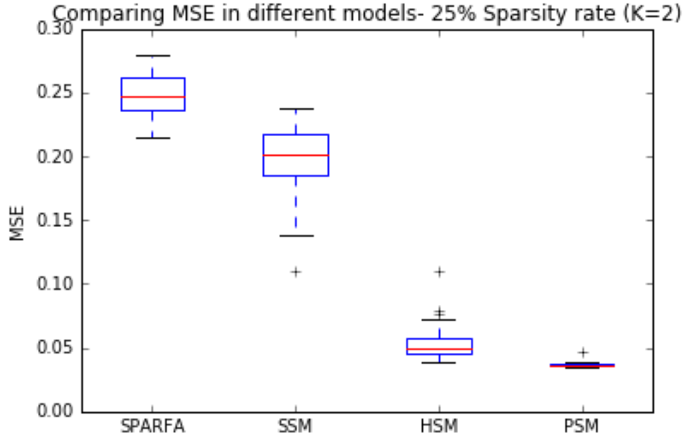
100 questions, 20 students in 6 weeks

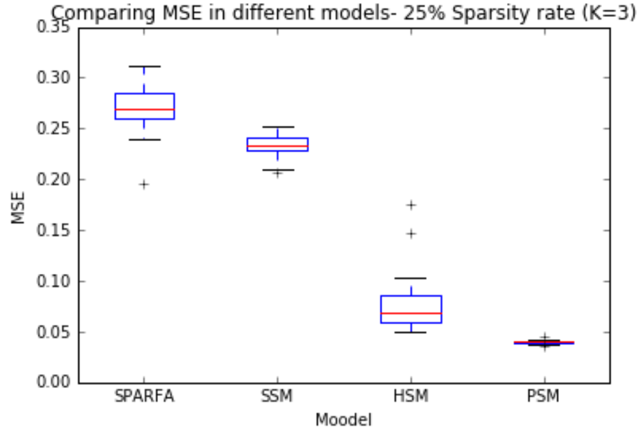
$$\text{MSE}=\frac{1}{QNT}\left\|P-\widehat{P}\right\|_2$$

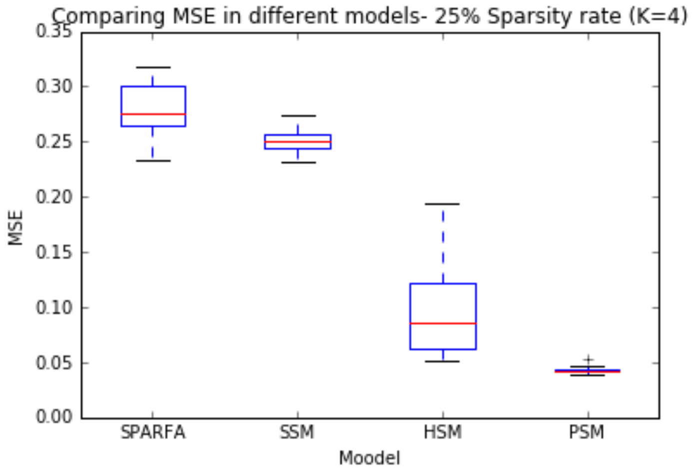
$\boldsymbol{P}$: True probability Matrix

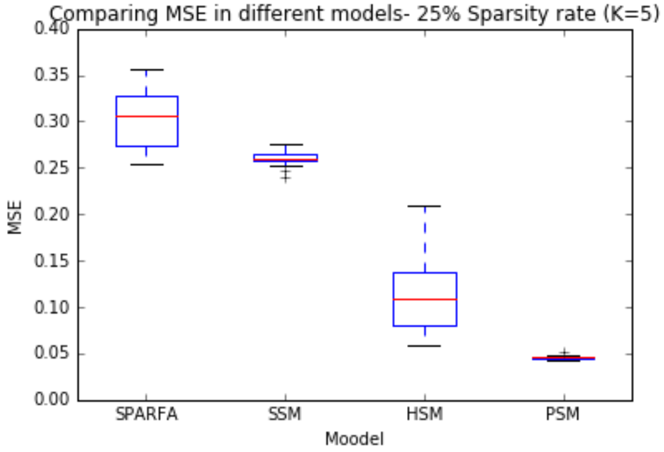
$\widehat{\boldsymbol{P}}$: Estimated probability Matrix







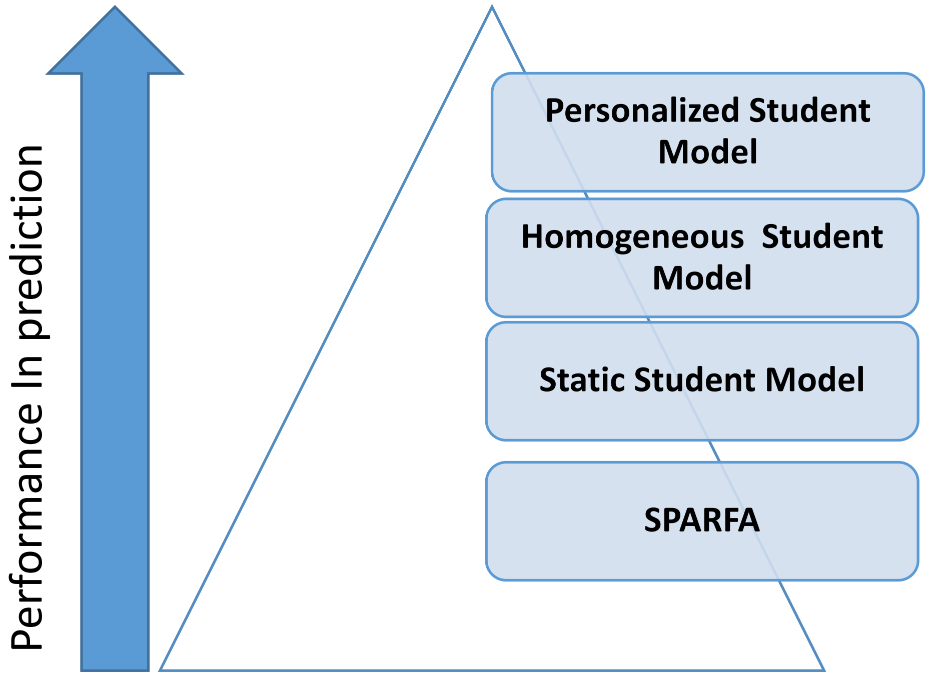




Tensor-based Student Models-Results

Large-sized Instances

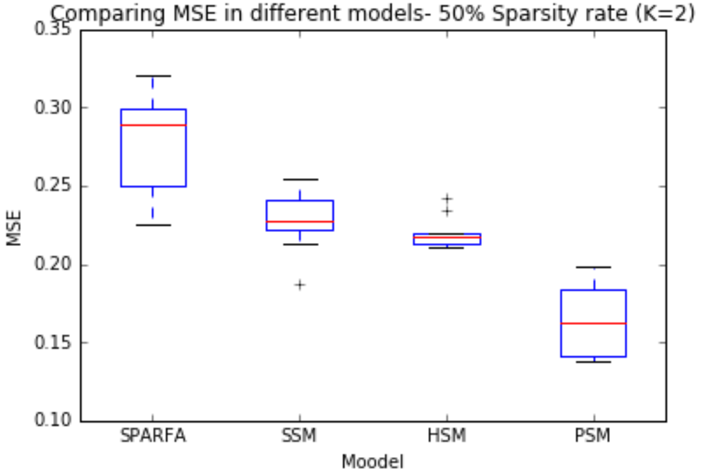
500 questions, 100 students in 15 weeks

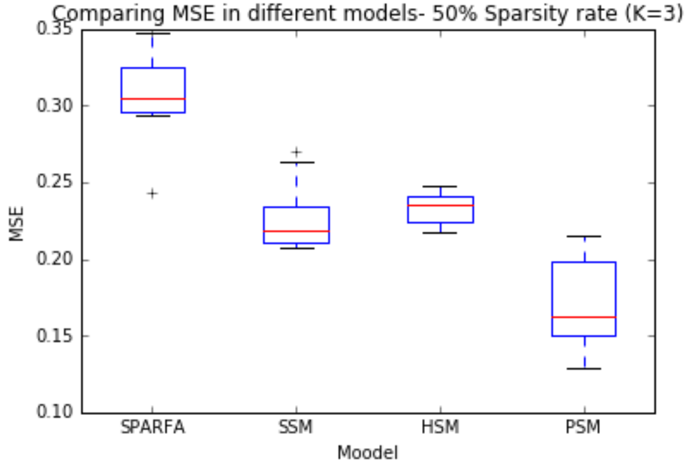


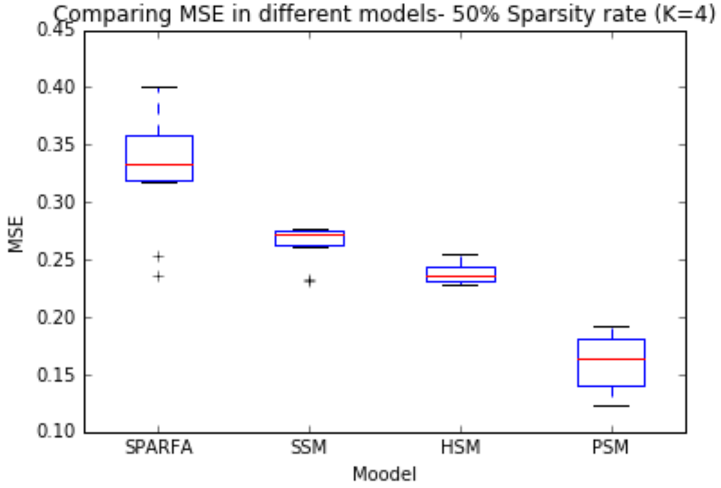
$$\text{MSE}=\frac{1}{QNT}\left\|P-\widehat{P}\right\|_2$$

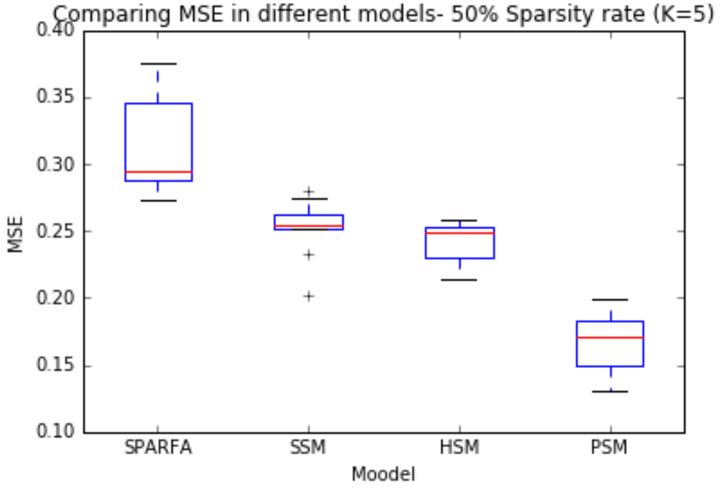
$\boldsymbol{P}$: True probability Matrix

$\widehat{\boldsymbol{P}}$: Estimated probability Matrix









Student Modeling Advances - Summary

Optimization

- All three models are parameterized via Maximum Likelihood Estimation – a non-convex optimization problem

Constraints

- Constraints to control Knowledge accumulation, Sparsity, achieve Convergence, and prevent Unbounded growth and Non-negativity are considered

Algorithms

- BCD and ADMM algorithms are employed to deal with the optimization problem (costumed code in Python).

Experiments

- Models are evaluated using small-, medium- and large-sized instances including 12 experiments with different sparsity levels and numbers of latent variables.

Conclusions

This Research …

- Incorporates time as an important component of dynamics of learning.
- Provides models with interpretable parameters describing students conceptual understanding.
- Proposes new Probabilistic Sparse Tensor Factorization methods for modeling students’ learning.

The Future …

- A model that explicitly takes the learning gain as a result of the interaction of the learner with a learning material.
- Enabling learning curve optimization.

thank you!