VarFA: A Variational Factor Analysis Framework For Efficient Bayesian Learning Analytics

Zichao Wang[1], Yi Gu[2], Andrew S. Lan[3], Richard G. Baraniuk[1,4]
[1]Rice University, [2]Northwestern University, [3]University of Massachusetts Amherst, [4]OpenStax
jzwang@rice.edu, Yi.Gu@u.northwestern.edu, andrewlan@cs.umass.edu, richb@rice.edu

ABSTRACT

We propose VarFA, a variational inference factor analysis framework that extends existing factor analysis models for educational data mining to efficiently output uncertainty estimation in the model's estimated factors. Such uncertainty information is useful, for example, for an adaptive testing scenario, where additional tests can be administered if the model is not quite certain about a students' skill level estimation. Traditional Bayesian inference methods that produce such uncertainty information are computationally expensive and do not scale to large data sets. VarFA utilizes variational inference which makes it possible to efficiently perform Bayesian inference even on very large data sets. We use the sparse factor analysis model as a case study and demonstrate the efficacy of VarFA on both synthetic and real data sets. VarFA is also very general and can be applied to a wide array of factor analysis models. Code and instructions to reproduce results in this paper are available at https://tinyurl.com/tvm4332. An extended version of this paper is available at https://arxiv.org/abs/2005.13107.

1. INTRODUCTION

A core task for many practical educational systems is *student modeling*, i.e., estimating students' mastery level on a set of skills or knowledge components (KC) [14, 6]. Such estimates allow in-depth understanding of students' learning status and form the foundation for automatic, intelligent learning interventions. A fruitful line of research for student modeling follows the *factor analysis (FA)* approach. FA models usually assume that an unknown, potentially multidimensional student parameter, in which each dimension is associated with a certain skill, explains how a student answers questions and is to be estimated.

Most of the aforementioned FA models compute a single point estimate of skill levels for each student [13, 1, 3, 9, 5, 15]. Often, however, it is not enough to obtain mere point estimates of students' skill levels; knowing the model's uncertainty in its estimation is crucial because it potentially helps improve the model's performance and improve both students' and instructors' experience with educational systems. For example, in adaptive testing systems [4, 16], knowing the uncertainty in model's estimation could help the model intelligently pick the next test items to most effectively reduce its uncertainty about estimated students' skill levels. This will help to potentially reduce the number of items needed to have a confident, accurate estimation of the students' skill mastery level, saving time for both students to take the test and instructors to have a good assessment of the student's skills.

In this work, we propose VarFA, a novel framework based on *variational inference* (VI) to perform efficient, scalable Bayesian inference for FA models. The key idea is to approximate the true posterior distribution, whose costly computation slows down Bayesian inference, with a *variational distribution*. In addition, this variational distribution is very flexible and we have full control specifying it, allowing us to freely use the latest development in machine learning, e.g., deep neural networks (DNNs), to design the variational distribution that closely approximates the true posterior. Thus, we also regard our work as a first step in applying DNNs to FA models for student modeling, achieving efficient Bayesian inference (enabled by DNNs) without losing interpretability (brough by FA models). We demonstrate the efficacy of our framework on three real data sets, showcasing that VarFA substantially accelerates classic Bayesian inference for FA models with no compromise on performance.

2. BACKGROUND

We first set up the problem and review related work. Assume we have a data set $\boldsymbol{Y} \in \mathbb{R}^{N\times Q}$ organized in matrix format where N is the total number of students and Q is the number of questions. This is a binary students' answer record matrix where each entry y_{ij} represents whether student i correctly answered question j. Usually, not all students answer all questions. Thus, $\boldsymbol{Y}$ contains missing values. We use $\{i,j\} \in \Omega_{\text{obs}}$ to denote entries in $\boldsymbol{Y}$, i.e., the i-th student's answer record to the j-th question, that are observed.

We are interested in models capable of inferring each i-th student's skill mastery level that can accurately predict the student's answers given the above data. These models are often evaluated on the prediction accuracy and whether the inferred student skill mastery levels are easily interpretable and educationally meaningful. We now review factor anal-

Zichao Wang, Yi Gu, Andrew Lan and Richard Baraniuk "VarFA: A Variational Factor Analysis Framework For Efficient Bayesian Learning Analytics" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 696 - 699

ysis models (FA), one of the most widely adopted and successful methodologies for the student modeling task.

Many FA models, despite differences in their respective mathematical formulae, modeling assumptions and the available auxiliary data used, can be unified into a canonical formulation below

$$\mathbb{P}(y_{ij} = 1) = \sigma(\mathbf{c}_i^\top \mathbf{m}_j + \mu_j)\,, \tag{1}$$

where $\mathbf{c}_i \in \mathbb{R}^K$, $\mathbf{m}_j \in \mathbb{R}^K$ and $\mu_j \in \mathbb{R}$ are factors whose dimension, interpretations and subscript indices depend on the specific instantiations of the FA model. We will use this general formulation in the rest of this paper. Usually, FA models obtains a point estimate of $\mathbf{c}_i$, $\mathbf{m}_j$ and μ_j. We will show next how to obtain uncertainty estimation of these variables of interest.

3. VarFA: A VARIATIONAL INFERENCE FACTOR ANALYSIS FRAMEWORK

The core idea of VarFA follows the variational principle, i.e., we use a parametric variational distribution to approximate the true posterior distribution. VarFA is highly flexible and efficient, making it suitable for large scale Bayesian inference for FA models in the context of educational data mining. In this current work, we focus on obtaining credible interval for the student skill mastery factor $\mathbf{c}_i$'s as a first step of VarFA. Extension to VarFA to full Bayesian inference for all unknown factors is part of an ongoing research; see 5 for more discussions.

Now, we explain in detail how to apply variational inference for FA models for efficient Bayesian inference. Because the posterior distribution is intractable to compute, we approximate the true posterior distribution for $\mathbf{c}_i$'s with a parametric variational distribution

$$p(\mathbf{C}|\boldsymbol{Y}, \mathbf{M}, \boldsymbol{\mu}) \approx q_\phi(\mathbf{C}|\boldsymbol{Y}) = \prod_{i=1}^{N} q_\phi(\mathbf{c}_i|\boldsymbol{y}_i)\,, \tag{2}$$

where ϕ is a collection of learnable parameters that parametrize the variational distribution and $\boldsymbol{y}_i$ is all the answer records by student i. Notably, we have removed the dependency of the variational distribution on ψ and θ so that the variational distribution is solely controlled by the variational parameter ϕ. Thus, the design of the variational distribution is highly flexible. All we need to do is to specify a class of distributions and design a function parametrized by ϕ to output the parameters of q_ϕ. Common in prior literature is to use a Gaussian with diagonal covariance for q_ϕ:

$$q_\phi(\mathbf{c}_i|\boldsymbol{y}_i) = \mathcal{N}(\boldsymbol{u}_i, \text{diag}(\boldsymbol{v}_i))\,, \tag{3}$$

where its mean and variance $[\boldsymbol{u}_j^\top, \boldsymbol{v}_j^\top]^\top = f_\phi(\boldsymbol{y}_i)$. We can use arbitrarily complex functions such as a deep neural network for f_ϕ as long as they are differentiable. With the above approximation, Bayesian inference turns into an optimization problem under the variational principle, where we now optimize a lower bound, known as the evidence lower bound (ELBO) [2], of the marginal data log likelihood.

Table 1: Student answer prediction erformance comapring VarFA to SPARFA-M on Assistment, Algebra and Bridge data sets. ↑ and ↓ denote higher and lower is better, respectively.

(a) Assistment

Metric	Algorithm	
	SPARFA-M	VarFA
ACC ↑	0.7074±0.0044	**0.7101**±0.0048
AUC ↑	0.756±0.048	**0.7635**±0.0036
F1 ↑	0.7746±0.0029	**0.7765**±0.0014
Run time (s) ↓	**5.3319**±0.2774	6.9167±0.1074

(b) Algebra

Metric	Algorithm	
	SPARFA-M	VarFA
ACC ↑	0.7735±0.0037	**0.7774**±0.0031
AUC ↑	0.8137±0.003	**0.8245**±0.002
F1 ↑	0.8465±0.0021	**0.8486**±0.001
Run time (s) ↓	**8.464**±0.4568	10.3335±0.4435

(c) Bridge

Metric	Algorithm	
	SPARFA-M	VarFA
ACC ↑	**0.8492**±0.0016	0.8468±0.0016
AUC ↑	0.837±0.0024	**0.8419**±0.0028
F1 ↑	**0.9121**±0.0005	0.912±0.0009
Run time (s) ↓	**15.6048**±0.7314	15.8558±1.046

We form the following optimization objective to estimate ϕ and θ:

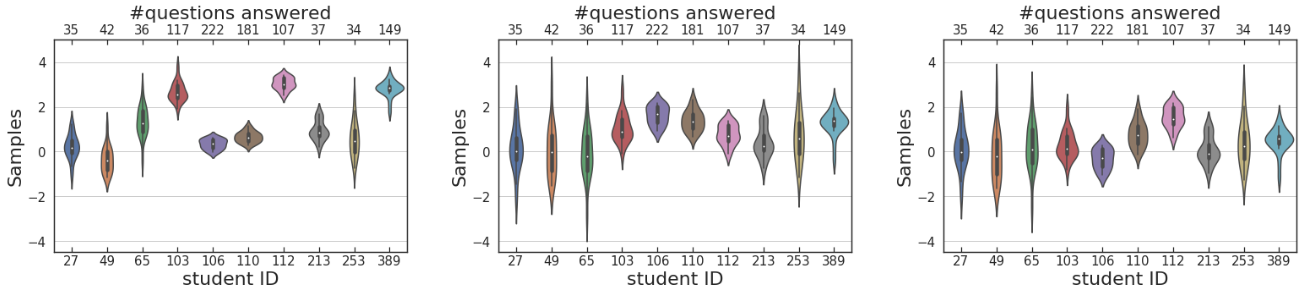
$$\widehat{\theta}, \widehat{\phi} = \underset{\theta,\phi}{\text{argmin}} \; -\mathcal{L}_{\text{ELBO}}(\phi, \theta) + \lambda\mathcal{R}(\theta)\,, \tag{4}$$

where $\theta = \{\mathbf{m}_1, ..., \mathbf{m}_Q, \mu_1, ..., \mu_Q\}$ and $\mathcal{R}(\theta)$ is a regularization term. That is, we perform VI on the student factor $\mathbf{c}_i$'s and MLE inference on the remaining factors denoted as θ.

4. EXPERIMENTS

We demonstrate the efficacy of VarFA variational inference framework using the sparse factor analysis model (SPARFA-M) as the underlying FA model. On three real-world data sets, we demonstrate that 1) VarFA predicts students' answers more accurately than SPARFA-M; 2) VarFA can output the same insights as SPARFA-M, including point estimate of students' skill levels and questions' associations with skill tags; 3) VarFA can additionally output meaningful uncertainty quantification for student skill levels, which SPARFA-M is incapable of, without sacrifice to computational efficiency. Note that SPARFA-B can also compute uncertainty for small data sets but fails for large data sets due to scalability issues and thus we do not compare to SPARFA-B for real data sets. The code along with instructions to reproduce our experiments can be downloaded from https://tinyurl.com/tvm4332.

Data sets. We perform experiments on three large-scale, publicly available, real educational data sets including AS-



(a) 3rd latent concept

(b) 4th latent concept

(c) 7th latent concept

Figure 1: Violin plot showing the mean and standard deviation of the estimated skill mastery levels on 10 selected students on the 3rd, 4th and 7th latent skills that VarFA computes. In each sub-figure, bottom and top axises respectively shows student IDs and top axis shows the number of questions each student answered.

SISTments 2009-2010 (Assistment) [7], Algebra I 2006-2007 (algebra) [10] and Bridge to Algebra 2006-2007 (bridge) [11, 12]. The details of the data sets, including data format and data collection procedure can be found in the preceding references.

Results: Performance Comparison. Table 1 shows the average performance on the test set of each data set comparing VarFA and SPARFA-M for all three data sets and additionally run time. We can see that VarFA achieves slightly better student answer prediction on most data sets and on most metrics. Table 1 also shows the run time comparison between VarFA and SPARFA-M; see the last row in each sub-table. We see that both inference algorithms have very similar run time, showing that VarFA is applicable for very large data sets. Notably, VarFA achieves this efficiency while also performing Bayesian inference on the student knowledge level factor.

Results: Bayesian Inference With VarFA. We now illustrate VarFA's capability of outputting credible intervals using the Assistment data set. Fig. 1 presents violin plots that show the sampled student latent skill levels for a random subset of 10 students. Plots 1a, 1b and 1c shows the inferred students ability for the 3rd, 4th and 7th latent skill dimension. In each plot, the bottom axis shows the student ID and the top axis shows the total number of questions answered by the corresponding student. For each student, the horizontal width of the violin represents the density of the samples; the skinnier the violin, the more widespread the samples are, implying the model's less certainty on its estimations.

Results in Fig. 1 confirms our intuition that the more questions a student answers, the more certain the model is about its estimation. For example, students with ID 106, 110 and 389 answered 222, 181 and 149 questions, respectively, and the credible intervals of their ability estimation is quite small. In contrast, students with ID 27, 49 and 65 answered far less questions and the credible intervals of their ability estimation is quite large. This result implies that VarFA outputs sensible and interpretable credible intervals.

Results: Post-Processing for Improved Interpretability. SPARFA assumes that each student factor $\mathbf{c}_i$ identifies a multi-dimensional skill level on a number of "latent" skills (recall that we use 8 latent skills in our experiments). As mentioned earlier, these latent skills are not interpretable without the aid of additional information. To improve interpretability, [8] proposed that, when the skill tags for each question is available in the data set, we can associate each latent skill with skill tags via a simple matrix factorization. Then, we can compute each students' mastery levels on the actual skill tags.

We again use the Assistment data set for illustration. We compute the association of skill tags in the data set with each of the latent skills and show 4 of the latent skills with their top 3 most strongly associated skill tags. We can see that each latent skill roughly identify the same group of skill tags. For example, latent skill 4 clusters skill tags on statistics and probability while latent skill 7 clusters skill tags on geometry. Thus, by simple post-processing, we obtain an interpretation of the latent skills by associating them with known skill tags in the data.

We can similarly obtain VarFA's estimations of the students' mastery levels on each skill tags through the above process. In Fig. 2, we compare the predicted mastery level for each skill tag (only for the questions this student answered) with the percent of correct answers for that skill tag. Blue curve shows the empirical student's mastery level on a skill tag by computing the percentage of correctly answered questions belonging to a particular skill tag. Orange curve shows VarFA's estimated student mastery level on a skill tag, normalized to range $[0, 1]$. Even though the two curves show different numeric values, they nevertheless demonstrate similar trends, showing that the predictions reasonably match our intuition about student's skill mastery levels.

5. CONCLUSIONS AND FUTURE WORK

We have presented VarFA, a variational inference factor analysis framework to perform efficient Bayesian inference for learning analytics. VarFA is general and can be applied to a wide array of FA models. We have demonstrated the effectiveness of our VarFA using the sparse factor analysis (SPARFA) model as a case study. We have shown that VarFA can very efficiently output interpretable, education-

Table 2: Illustration of the estimated latent skills with the their top 3 most strongly associated skill tags in the Assistment data set. The percentage in the parenthesis shows the association probability (summed to 1 for each latent skill). We see that the tagged skills associated with each estimated latent skill form intuitive and interpretable groups.

Latent Skill 1	**Latent Skill 3**
Division Fractions (29.1%) Least Common Multiple (18.1%) Write Linear Equation from Ordered Pairs (17.8%)	Conversion of Fraction Decimals Percents (7.3%) Addition and Subtraction Positive Decimals (6.8%) Probability of a Single Event (5.7%)
Latent Skill 4	**Latent Skill 7**
Pattern Finding (17.4%) Histogram as Table or Graph (11.3%) Percent Of (10.5%)	Volume Sphere (13.4%) Volume Cylinder (10.4%) Surface Area Rectangular Prism (10.2%)

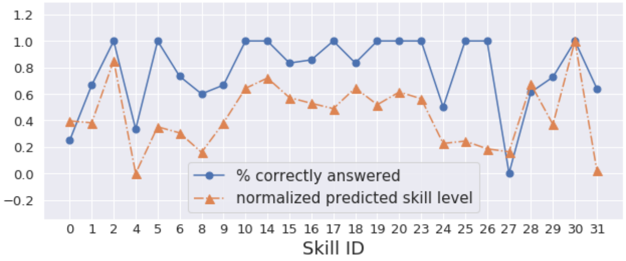


Figure 2: Comparison between the estimated skill mastery levels using VarFA's predictions and using empirical observations for student with ID 110.

ally meaningful information, in particular credible intervals, much faster than classic Bayesian inference methods. Thus, VarFA has potential application in many educational data mining scenarios where efficient credible interval computation is desired, i.e., in adaptive testing and adaptive learning systems. We have also provided open-source code to reproduce our results and facilitate further research efforts.

Acknowledgement

This work was supported by NSF grants CCF-1911094, IIS-1838177, IIS-1730574, DRL-1631556, IUSE-1842378; ONR grants N00014-18-12571 and N00014-17-1-2551; AFOSR grant FA9550-18-1-0478; and a Vannevar Bush Faculty Fellowship, ONR grant N00014-18-1-2047.

6. REFERENCES

[1] T. A. Ackerman. Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, 7(4):255–278, 1994.

[2] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

[3] H. Cen, K. Koedinger, and B. Junker. Learning factors analysis – a general method for cognitive model evaluation and improvement. In M. Ikeda, K. D. Ashley, and T.-W. Chan, editors, *Proc. Intelligent Tutoring Systems*, pages 164–175, 2006.

[4] H.-H. Chang, Z. Ying, et al. Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics*, 37(3):1466–1488, 2009.

[5] M. Chi, K. R. Koedinger, G. J. Gordon, P. Jordon, and K. VanLahn. Instructional factors analysis: A cognitive model for multiple instructional interventions. 2011.

[6] K. Chrysafiadi and M. Virvou. Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications*, 40(11):4715–4729, 2013.

[7] N. T. Heffernan and C. Heffernan. The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24:470–497, 2014.

[8] A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk. Sparse factor analysis for learning and content analytics. *Journal of Machine Learning Research*, 15:1959–2008, 2014.

[9] P. I. Pavlik, H. Cen, and K. R. Koedinger. Performance factors analysis –a new alternative to knowledge tracing. In *Proc. Artificial Intelligence in Education*, page 531–538, NLD, 2009. IOS Press.

[10] J. Stamper, A. Niculescu-Mizil, S. Ritter, G. Gordon, and K. Koedinger. Algebra i 2006-2007. *Development data set from KDD Cup 2010 Educational Data Mining Challenge*, 2010.

[11] J. Stamper, A. Niculescu-Mizil, S. Ritter, G. Gordon, and K. Koedinger. Bridge to algebra 2006-2007. *Development data set from KDD Cup 2010 Educational Data Mining Challenge*, 2010.

[12] J. C. Stamper and Z. A. Pardos. The 2010 kdd cup competition dataset: Engaging the machine learning community in predictive learning analytics. 2016.

[13] W. J. van der Linden and R. K. Hambleton. *Handbook of modern item response theory*. Springer Science & Business Media, 2013.

[14] K. VanLehn. Student modeling. *Foundations of Intelligent Tutoring Systems*, 55:78, 1988.

[15] J.-J. Vie and H. Kashima. Knowledge tracing machines: Factorization machines for knowledge tracing. In *Proc. AAAI Conference on Artificial Intelligence*, volume 33, pages 750–757, 2019.

[16] D. Yan, A. A. Von Davier, and C. Lewis. *Computerized multistage testing: Theory and applications*. CRC Press, 2016.