Learning Analytics via Sparse Factor Analysis

Andrew E. Waters, Andrew Lan, Christoph Studer, and Richard G. Baraniuk
Rice University; e-mail: {andrew.e.waters, sl29, studer, richb}@rice.edu

Introduction Textbooks, lectures, and homework assignments were the answer to the main educational challenges of the 19th century, but are now the main bottleneck of the 21st century. In particular, today's textbooks are typically static, linear in organization, time-consuming to develop, soon out-of-date, and expensive. Lectures remain a primarily passive experience of copying down what an instructor says. Homework assignments that are not graded for weeks provide poor feedback to students on their learning progress. Even more importantly, today's courses provide only a "one-size-fits-all" learning experience that does not cater to the background, interests, and goals of individual students. In contrast, we envision a statistically-minded, machine-learning based, cognitive tutor that is able to learn about the student as the student learns about the subject material being taught. This approach would allow the cognitive tutor to naturally assess which knowledge areas the student understands well, as well as the areas that remain problematic, enabling crucial tasks such as the automatic recommendation of remedial study material or additional practice problems [1].

Model and methods As a first step towards creating such a cognitive tutor, we propose a novel statistical framework for representing domain knowledge based on *sparse latent factor analysis* [2]. We assume that the underlying knowledge base is decomposable into a set of latent *knowledge concepts* that are to be learned by the student. As an example, an introductory calculus course might have latent concepts such as "integration by parts", "differentiation of polynomials", "l'Hôpital's rule", etc. For this model, we assume that there are N students answering a subset of P questions involving $K \ll P, N$ underlying (latent) concepts. Let the column vector $\mathbf{c}_j \in \mathbb{R}^K$, $j \in \{1, \ldots, N\}$, represent the latent *concept understanding* of the j^{th} student, $\mathbf{w}_i \in \mathbb{R}^K$ represent the *concept associations* of question i; and let the scalar $\mu_i \in \mathbb{R}$ model the *intrinsic difficulty* of question i. Then, we propose the following model for the student–response relationships:

$$Z_{i,j} = \mathbf{w}_i^T \mathbf{c}_j + \mu_i, \ \forall i, j, \quad \text{and} \quad Y_{i,j} \sim \mathit{Ber}(\Phi(Z_{i,j})), \ (i,j) \in \Omega_{\text{obs}}. \tag{1}$$

Here, $Y_{i,j} \in \{0, 1\}$ denotes the observed binary-valued response variable of the j^{th} student to the i^{th} question, where 0 and 1 indicates a wrong and correct response, respectively. $\mathit{Ber}(z)$ designates a Bernoulli distribution with success probability z and Φ denotes an inverse link function (e.g. logit or probit), which maps a real value to the success probability in $[0, 1]$ of a binary-valued random variable. The set Ω_{obs} contains the indices of the observed entries in $\mathbf{Y}$.

To improve the identifiability of our model, we impose additional constraints on $\mathbf{W}$, namely *sparsity* and *non-negativity.* The sparsity assumption implies that we expect each question to be related to only a small number of concepts, which is typical in most education scenarios. The non-negativity assumption implies that knowledge of a particular concept does not hurt one's chances of answering a question correctly. A particularly useful consequence of this assumption is that large, positive entries in $\mathbf{C}$ correspond to concepts that students have mastered well, while negative values indicate concepts with poor mastery.

We propose two novel SPARFA (short for SPARse Factor Analysis) algorithms for solving the inference problem in (1). The first method, SPARFA-M, is a low-complexity biconvex-optimization approach based on the fast iterative shrinkage-thresholding algorithm [3]. The second method, SPARFA-B, is a Markov chain Monte-Carlo (MCMC) algorithm that computes full posterior estimates for all parameters of interest (see [4] for a related algorithm). Both methods enable us to determine i) a model for how concepts intersect with questions and ii) each student's understanding of the concepts. In addition, both algorithms can further make use of side information, such as tags (or labels) on questions provided by content authors, which further enhances the intelligibility of the knowledge decomposition.

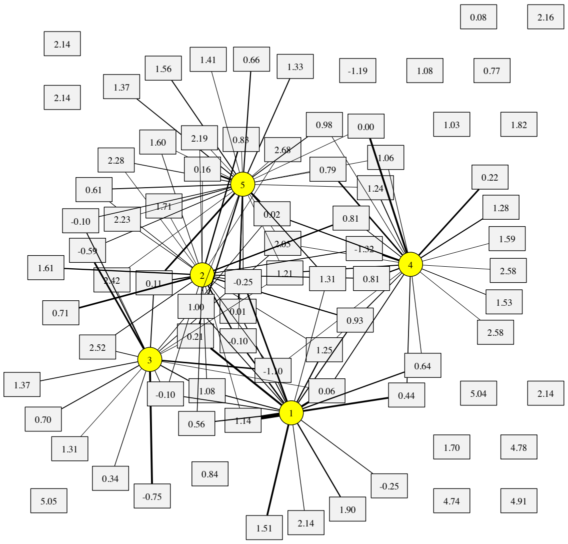

Figure 1: Question–concept association graph recovered by SPARFA-B. Circles and rectangles designate concepts and questions, respectively; the values in the rectangles indicate intrinsic question difficulties.

Concept 1	Concept 2
Changes to land (45%) Properties of soil (28%) Uses of energy (27%)	Evidences of the past (74%) Mixtures and solutions (14%) Environmental changes (12%)
Concept 3	Concept 4
Alternative energy (76%) Environmental changes (19%) Changes from heat (5%)	Properties of soil (77%) Environmental changes (17%) Classifying matter (6%)
Concept 5	
Formation of fossil fuels (54%) Mixtures and solutions (28%) Uses of energy (18%)	

Table 1: Three most important tags and the associated relative weights for the five concepts recovered in an 8th grade Earth-science curriculum.

Results We have demonstrated the validity of the SPARFA approach on several real-world educational datasets. An example of its capabilities is provided in Figure 1, which shows the recovered knowledge base for an 8th grade Earth-science curriculum maintained by the STEMscopes organization [5]. The data consists of 145 students answering 80 questions that have been tagged by the content authors. The observed data is highly incomplete, with only 13.5% of the total question/answer pairs being observed. Our proposed SPARFA algorithms are able to recover the underlying question–concept associations, interpret the meaning of each concept using tag information (shown in Table 1), and measure the intrinsic difficulty of each question. In addition, SPARFA can determine the concept mastery for individual students and provide human-readable feedback to students, i.e., which tags they have mastered well and which they have not.

Conclusions Our proposed statistical framework and both SPARFA algorithms automatically decompose an educational domain into its constituent knowledge concepts by only evaluating binary-valued student response data to a set of questions and user-provided tags on each question. Moreover, the output of the SPARFA algorithms allow for the convenient visualization of course content via a bipartite graph consisting of question and concept nodes. Our algorithms further provide estimates of the concept mastery for each student in the course, which enables a number of vital tasks for cognitive tutoring, including automating personalized feedback to students, recommending new questions, and refining course-content.

References

[1] B. P. Woolf, *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning.* Morgan Kaufman Publishers, 2008.

[2] M. West, "Bayesian factor regression models in the "large p, small n" paradigm," *Bayesian statistics*, vol. 7, pp. 723–732, Sep. 2003.

[3] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, pp. 183–202, Mar. 2009.

[4] P. R. Hahn, C. M. Carvalho, and J. G. Scott, "A sparse factor-analytic probit model for congressional voting patterns," *J. Royal Stat. Soc.*, vol. 61, no. 2, 2012, to appear.

[5] STEMscopes K-12 Science Curriculum, *http://stemscopes.com/*. Sep. 2012.