Joint Sparse Factor Analysis and Topic Modeling for Learning Analytics (Poster)

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

Introduction Personalized education based on machine learning has the potential to revolutionize education and to improve learning for students of diverse backgrounds, abilities, and interests, at a large, global scale. We have previously developed a statistical framework for discovering and representing domain knowledge based on sparse latent factor analysis (SPARFA, for short) [1]. The framework assumes that N students answer 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, let $\mathbf{w}_i \in \mathbb{R}^K$, $i \in \{1, \ldots, P\}$, 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 model the student–response relationships as [1]:

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

Here, $Y_{i,j} \in \{0, 1\}$ corresponds to the observed binary-valued response variable of the j^{th} student to the i^{th} question, where 0 and 1 indicate a incorrect and correct response, respectively. $Ber(z)$ 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 $[0, 1]$. The set Ω_{obs} contains the indices of the observed entries. To address the inevitable identifiability issue in factor analysis, we impose additional constraints on the model (1), namely that $\mathbf{W}$ should be *sparse* and *non-negative.* Sparsity dictates that we expect each question to be related to only a few concepts, which is typical in most education scenarios; non-negativity dictates that knowledge of a particular concept does not hurt one's chances of answering a question correctly. In [1], We developed two algorithms, SPARFA-M and SPARFA-B to solve (1), which provide us a question–concept association graph, student concept mastery profile, and the intrinsic difficulty of the questions.

Incorporating topic models We have demonstrated the capabilities of the SPARFA framework (1) to provide a question–concept association graph and student masteries of concepts in [1] using real educational datasets. However, the concepts we learn are mathematical constructs and not necessarily interpretable by humans. Therefore, in [2] we developed a post-processing method that exploits pre-defined question tags to improve intelligibility of the extracted concepts [1]. We now consider the *joint analysis* of student–response information and textual information (e.g., available from the question or solution text) to further improve the identifiability and intelligibility of the decomposed factors. Text information provides a rich source of information and has been extensively studied in the topic model literature [3]. Specifically, assume that we additionally observe the matrix $\mathbf{B} \in \mathbb{N}^{P \times V}$, where V corresponds to the number of total words that have occurred among the P questions. Each entry $B_{i,v}$ represents how many times the v^{th} word occurs in the i^{th} question. To model the word frequencies contained in $\mathbf{B}$, we propose the following statistical topic model:

$$A_{i,v} = \mathbf{w}_i^T \mathbf{t}_v, \quad \text{and} \quad B_{i,v} \sim Pois(A_{i,v}), \quad \forall i, v, \tag{2}$$

where $\mathbf{t}_v \in \mathbb{R}_+^K$ is a column vector that characterizes how strongly the v^{th} word is expressed in every concept. Inspired by the topic model proposed in [4], we model the entries of the word-occurrence matrix $B_{i,v}$ in (2) as *Poisson* distributed, with the rate parameters determined by $A_{i,v}$.

In order to *jointly* estimate $\mathbf{W}$, $\mathbf{C}$, $\boldsymbol{\mu}$, and $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_V]$ from the observed student–response matrix $\mathbf{Y}$ and the word-frequency matrix $\mathbf{B}$, we solve the following optimization problem:

$$\underset{\mathbf{W},\mathbf{C},\mathbf{T}:\,\mathbf{W}\geq 0,\mathbf{T}>0}{\text{minimize}} \; \alpha \sum_{(i,j)\in\Omega_{\text{obs}}} - \log p(Y_{i,j}|\mathbf{w}_i, \mathbf{c}_j) + (1-\alpha) \sum_{i,v} - \log p(B_{i,j}|\mathbf{w}_i, \mathbf{t}_v) + \lambda \sum_i \|\mathbf{w}_i\|_1 + \frac{\gamma}{2} \sum_j \|\mathbf{c}_j\|_2^2 + \frac{\eta}{2} \sum_v \|\mathbf{t}_v\|_2^2, \tag{3}$$

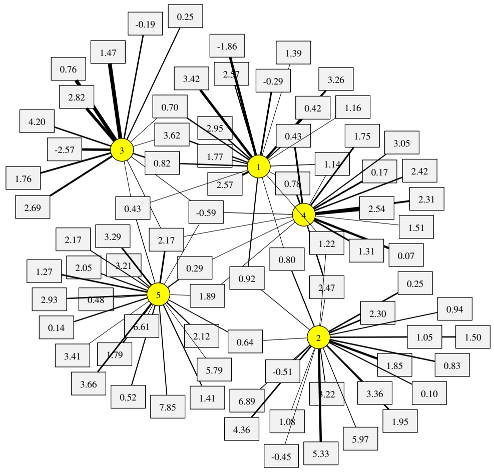

Figure 1: Question–concept association graph recovered by SPARFA-TOP. Circles and rectangles represent concepts and questions, respectively; the values in the rectangles represent question difficulties.

Concept 1	Concept 2	Concept 3
Water	Water	Energy
Soil	Rock	Light
Container	Sand	Thermal
Sample	Moon	Temperature
Cans	Form	Bulb
Plants	Heat	Grams
Substances	Canyon	Noise

Concept 4	Concept 5
Water	Water
Objects	Sand
Dense	Earth
Resources	Percentage
Energy	Wagon
Glass	Buffalo
River	High

Table 1: Seven most important words for the five concepts recovered by SPARFA-TOP for an 8$^{\text{th}}$ grade Earth-science curriculum.

where the probabilities $p(Y_{i,j}|\mathbf{w}_i, \mathbf{c}_j)$ and $p(B_{i,j}|\mathbf{w}_i, \mathbf{t}_v)$ follow the statistical models in (1) and (2), respectively. The ℓ_1-norm penalty term $\lambda \sum_i \|\mathbf{w}_i\|_1$ induces sparsity on $\mathbf{W}$, while the ℓ_2-norm penalty terms $\frac{\gamma}{2} \sum_j \|\mathbf{c}_j\|_2^2$ and $\frac{\eta}{2} \sum_v \|\mathbf{t}_v\|_2^2$ gauge the norms of $\mathbf{C}$ and $\mathbf{T}$. The parameter $0 \leq \alpha \leq 1$ controls the relative importance of the question–answer model (1) vs. the Poisson topic model (2); smaller values of α favor the topic model, while larger values favor the question–answer model. We solve (3) using an efficient block-coordinate-descent algorithm relying on the fast iterative shrinkage-thresholding algorithm [5], which we dub SPARFA-TOP (SPARse Factor Analysis and TOPic Modeling).

Results We demonstrate the validity of SPARFA-TOP on a real educational dataset consisting of an 8$^{\text{th}}$ grade Earth-science curriculum maintained by STEMscopes [6]. The dataset consists of 145 students answering 80 questions, with only 13.5% of the total question/answer pairs being observed. Excluding common stop-words, the question and answer text vocabulary consists of 326 words. Figure 1 and Table 1 show the question–concept associations along with the recovered intrinsic difficulties and the top 7 words characterizing each concept extracted by SPARFA-TOP, respectively. Compared to the approach in [1], we see that SPARFA-TOP is able to relate all questions to concepts, including those that were found in [1] to be ill-posed or off-topic, by taking advantage of topic models. Furthermore, Table 1 demonstrates that SPARFA-TOP can automatically provide an interpretable summary of the true meaning of each concept.

Conclusions The SPARFA-TOP method proposed here extends our SPARFA framework to automatically decompose an educational domain into its constituent knowledge concepts by jointly considering binary-valued student response data to a set of questions as well as the actual question and answer text. The framework enables the easy interpretation of the concepts, which enables SPARFA-TOP to automate a number of vital tasks for personalized learning, including automating personalized feedback to students, recommending new questions for remediation or enrichment, and refining the course content.

References

[1] A. E. Waters, A. S. Lan, C. Studer, and R. G. Baraniuk, "Learning analytics via sparse factor analysis," in *submitted to Personalizing Education With Machine Learning, NIPS 2012 Workshop*, (Tahoe, NV), Sep. 2012.

[2] A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk, "Learning analytics: A sparse factor analysis approach," *Journal of Machine Learning Research (JMLR)*, 2012, submitted.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Jan. 2003.

[4] J. Zhu and E. P. Xing, "Sparse topical coding," in *Proc. of 27th Conference on Uncertainty in Artificial Intelligence (UAI)*, (Barcelona, Spain), Mar. 2011.

[5] 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.

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