Joint Topic Modeling and Factor Analysis of Textual Information and Graded Response Data

Andrew S. Lan, Christoph Studer, Andrew E. Waters, Richard G. Baraniuk
Rice University, TX, USA
{mr.lan, studer, waters, richb}@sparfa.com

ABSTRACT

Modern machine learning methods are critical to the development of large-scale personalized learning systems that cater directly to the needs of individual learners. The recently developed SPARse Factor Analysis (SPARFA) framework provides a new statistical model and algorithms for machine learning-based learning analytics, which estimate a learner's knowledge of the latent concepts underlying a domain, and content analytics, which estimate the relationships among a collection of questions and the latent concepts. SPARFA estimates these quantities given only the binary-valued graded responses to a collection of questions. In order to better interpret the estimated latent concepts, SPARFA relies on a post-processing step that utilizes user-defined tags (e.g., topics or keywords) available for each question. In this paper, we relax the need for user-defined tags by extending SPARFA to jointly process both graded learner responses and the text of each question and its associated answer(s) or other feedback. Our purely data-driven approach (i) enhances the interpretability of the estimated latent concepts without the need of explicitly generating a set of tags or performing a post-processing step, (ii) improves the prediction performance of SPARFA, and (iii) scales to large test/assessments where human annotation would prove burdensome. We demonstrate the efficacy of the proposed approach on two real educational datasets.

Keywords

Factor analysis, topic model, personalized learning, machine learning, block coordinate descent

1. INTRODUCTION

Traditional education typically provides a "one-size-fits-all" learning experience, regardless of the potentially different backgrounds, abilities, and interests of individual learners. Recent advances in machine learning enable the design of computer-based systems that analyze learning data and provide feedback to the individual learner. Such an approach has great potential to revolutionize today's education by offering a high-quality, personalized learning experience to learners on a global scale.

1.1 Personalized learning systems

Several efforts have been devoted into building statistical models and algorithms for learner data analysis. In [5], we proposed a personalized learning system (PLS) architecture with two main ingredients: (i) *learning analytics* (analyzing learner interaction data with learning materials and questions to provide personalized feedback) and (ii) *content analytics* (analyzing and organizing learning materials including questions and text documents). We introduced the SPARse Factor Analysis (SPARFA) framework for learning and content analytics, which decomposes assessments into different knowledge components that we call *concepts*. SPARFA automatically extracts (i) a question–concept association graph, (ii) learner concept knowledge profiles, and (iii) the intrinsic difficulty of each question, solely from graded binary learner responses to a set of questions; see Fig. 2 for an example of a graph extracted by SPARFA. This framework enables a PLS to provide personalized feedback to learners on their concept knowledge, while also estimating the question–concept relationships that reveal the structure of the underlying knowledge base of a course.

The original SPARFA framework [5] extracts the concept structure of a course from binary-valued question–response data. The latent concepts are "abstract" in the sense that they are estimated from the data rather than dictated by a subject matter expert. To make the concepts interpretable by instructors and learners, SPARFA performs an ad-hoc post-processing step to fuse instructor-provided question tags to each estimated concept. Requiring domain experts to label the questions with tags is an obvious limitation to the approach, since such tags are often incomplete or inaccurate and thus provide insufficient or unreliable information.

Inspired by the recent success of modern text processing algorithms, such as latent Dirichlet allocation (LDA) [3], we posit that the text associated with each question can potentially reveal the meaning of the estimated latent concepts without the need of instructor-provided question tags. Such an data-driven approach would be advantageous as it would easily scale to domains with thousands of questions. Furthermore, directly incorporating textual information into the SPARFA statistical model could potentially improve the estimation performance of the approach.

1.2 Contributions

In this paper, we propose *SPARFA-Top*, which extends the SPARFA framework [5] to jointly analyze both graded learner responses to questions and the text of the question, response, or feedback. We augment the SPARFA model by statistically modeling the word occurrences associated with the questions as *Poisson* distributed. We develop a computationally efficient block-coordinate descent algorithm that,

given only binary-valued graded response data and associated text, estimates (i) the question–concept associations, (ii) learner concept knowledge profiles, (iii) the intrinsic difficulty of each question, and (iv) a list of most important keywords associated with each estimated concept.

SPARFA-Top is capable of automatically generating a human readable interpretation for each estimated concept in a purely data-driven fashion (i.e., no manual labeling of the questions is required), thus enabling a PLS to automatically recommend remedial or enrichment material to learners that have low/high knowledge level on a given concept. Our experiments on real-world educational datasets indicate that SPARFA-Top outperforms SPARFA in terms of both prediction performance and interpretability of the estimated concepts.

1.3 Related work

The joint analysis of binary-valued data and associated textual information has been studied in the context of congressional voting patterns using Bayesian inference methods [8, 11]. Our proposed approach uses a block-coordinate descent method [10] that is computationally more efficient; this aspect is crucial in practice as it enables to provide real-time feedback to each learner. In addition, the particular structure imposed by SPARFA (non-negativity and sparsity) distinguishes our framework from the methods in [8, 11]. Optimization-based topic model methods have been proposed in [6, 12]. None of these methods, however, consider the joint analysis of textual information and other forms of observed data (such as graded learner responses).

2. THE SPARFA-TOP MODEL

We start by summarizing the SPARFA framework [5], and then extend it by modeling word counts extracted from textual information available for each question. We then detail the SPARFA-Top algorithm, which jointly analyzes binary-valued graded learner responses to questions as well as question text to generate (i) a question–concept association graph and (ii) keywords for each estimated concept.

2.1 SPARse Factor Analysis (SPARFA)

SPARFA assumes that graded learner response data consist of N learners answering a subset of Q questions that involve $K \ll Q, N$ underlying (latent) concepts. Let the column vector $\mathbf{c}_j \in \mathbb{R}^K$, $j \in \{1, \dots, N\}$ represent the latent *concept knowledge* of the j^{th} learner, let $\mathbf{w}_i \in \mathbb{R}^K$, $i \in \{1, \dots, Q\}$ represent the *associations* of question i to each concept, and let the scalar $\mu_i \in \mathbb{R}$ represent the *intrinsic difficulty* of question i. The student–response relationship is modeled as

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

where $Y_{i,j} \in \{0, 1\}$ corresponds to the observed binary-valued graded response variable of the j^{th} learner to the i^{th} question, where 1 and 0 indicate correct and incorrect responses, respectively. $Ber(z)$ designates a Bernoulli distribution with success probability z, and $\Phi(x) = \frac{1}{1+e^{-x}}$ denotes the inverse logit link function, which maps a real value to the success probability $z \in [0, 1]$. The set Ω_{obs} contains the indices of the observed entries (i.e., the observed data may be incomplete).

The precision parameter $\tau_{i,j}$ models the *reliability* of the observed binary graded response $Y_{i,j}$. Larger values of $\tau_{i,j}$ indicate higher reliability on the observed graded learner responses, while smaller values indicate lower reliability. For the sake of simplicity of exposition, we will assume $\tau_{i,j} = \tau$, $\forall i, j$, throughout this paper.

To address the fundamental identifiability issue in factor analysis and to account for real-world educational scenarios, [5] imposed specific constraints on the model (1). Concretely, every row $\mathbf{w}_i$ of the question–concept association matrix $\mathbf{W}$ is assumed to be *sparse* and *non-negative*. The sparsity assumption dictates that one expects each question to be related to only a few concepts, which is typical for most education scenarios. The non-negativity assumption characterizes the fact that knowledge of a particular concept does not hurt one's ability of answering a question correctly.

2.2 SPARFA-TOP: Joint analysis of learner responses and textual information

SPARFA [5] utilizes a post-processing step to link predefined tags with the inferred latent concepts. We now introduce a novel approach to *jointly* consider graded learner response and associated textual information, in order to directly associate keywords with the estimated concepts.

Assume that we observe the word–question occurrence matrix $\mathbf{B} \in \mathbb{N}^{Q \times V}$, where V corresponds to the size of the vocabulary, i.e., the number of *unique* words that have occurred among the Q questions. Each entry $B_{i,v}$ represents how many times the v^{th} word occurs in the associated text of the i^{th} question; as is typical in the topic model literature, common stop words ("the", "and", "in" etc.) are excluded from the vocabulary. The word occurrences in $\mathbf{B}$ are modeled as follows:

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

where $\mathbf{t}_v \in \mathbb{R}_+^K$ is a non-negative[1] column vector that characterizes the expression of the v^{th} word in every concept. Inspired by the topic model proposed in [12], the entries of the word-occurrence matrix $B_{i,v}$ in (2) are assumed to be *Poisson* distributed, with rate parameters $A_{i,v}$.

We emphasize that the models (1) and (2) share the same question–concept association vector, which implies that the relationships between questions and concepts manifested in the learner responses are assumed to be exactly the same as the question–topic relationships expressed as word co-occurrences. Consequently, the question–concept associations generating the associated question text are also sparse and non-negative, coinciding with the standard assumptions made in the topic model literature [3, 9].

3. SPARFA-TOP ALGORITHM

We now develop the SPARFA-Top algorithm by using block multi-convex optimization, to *jointly* estimate $\mathbf{W}$, $\mathbf{C}$, $\boldsymbol{\mu}$, and $\mathbf{T} = [\mathbf{t}_1, \dots, \mathbf{t}_V]$ from the observed student–response matrix $\mathbf{Y}$ and the word-frequency matrix $\mathbf{B}$. Specifically, we

[1]Since the Poisson rate $A_{i,v}$ must be strictly positive, we assume that $A_{i,v} \geq \varepsilon$ with $\varepsilon = 10^{-6}$ in all experiments.

seek to solve the following optimization problem:

$$\underset{\mathbf{W},\mathbf{C},\mathbf{T}:\ W_{i,k}\geq 0\,\forall i,k,T_{k,v}\geq 0\,\forall k,v}{\text{minimize}} \sum_{(i,j)\in\Omega_{\text{obs}}} -\log p(Y_{i,j}|\mathbf{w}_i^T\mathbf{c}_j+\mu_i,\tau)+\sum_{i,v}-\log p(B_{i,v}|\mathbf{w}_i^T\mathbf{t}_v) + \lambda\sum_i\|\mathbf{w}_i\|_1+\tfrac{\gamma}{2}\sum_j\|\mathbf{c}_j\|_2^2+\tfrac{\eta}{2}\sum_v\|\mathbf{t}_v\|_2^2. \quad (3)$$

Here, the probabilities $p(Y_{i,j}|\mathbf{w}_i^T\mathbf{c}_j+\mu_i,\tau)$ and $p(B_{i,v}|\mathbf{w}_i^T\mathbf{t}_v)$ 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 the question–concept matrix $\mathbf{W}$. 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 the matrices $\mathbf{C}$ and $\mathbf{T}$. To simplify the notation, the intrinsic difficulty vector $\boldsymbol{\mu}$ is added as an additional column of $\mathbf{W}$ and with $\mathbf{C}$ augmented with an additional all-ones row. The precision parameter τ serves as a balance between the observed learner responses and question text. For $\tau\to\infty$, SPARFA-Top corresponds to SPARFA, while for $\tau\to 0$, SPARFA-Top corresponds to topic models.

The optimization problem (3) is block multi-convex, i.e., the subproblem obtained by holding two of the three factors $\mathbf{W}$, $\mathbf{C}$, and $\mathbf{T}$ fixed and optimizing for the other is convex. This property inspires us to deploy a block coordinate descent approach to compute an approximate to (3). The SPARFA-Top algorithm starts by initializing $\mathbf{W}$, $\mathbf{C}$, and $\mathbf{T}$ with random matrices and then optimizes each of these three factors iteratively until convergence. The subproblems of optimizing over $\mathbf{W}$ and $\mathbf{C}$ are solved iteratively using algorithms relying on the FISTA framework (see [2] for the details).

The subproblem of optimizing over $\mathbf{C}$ with $\mathbf{W}$ and $\mathbf{T}$ fixed was detailed in [5]. The subproblem of optimizing over $\mathbf{T}$ with $\mathbf{W}$ and $\mathbf{C}$ fixed is separable in each column of $\mathbf{T}$, with the problem for $\mathbf{t}_v$ being:

$$\underset{\mathbf{t}_v:T_{k,v}\geq 0,\forall k}{\text{minimize}} \quad \textstyle\sum_i -\log p(B_{i,v}|\mathbf{w}_i^T\mathbf{t}_v)+\frac{\eta}{2}\|\mathbf{t}_v\|_2^2. \quad (4)$$

This subproblem can be efficiently solved using FISTA, where the gradient of the objective function with respect to $\mathbf{t}_v$ being:

$$\nabla_{\mathbf{t}_v}\textstyle\sum_i -\log p(B_{i,v}|\mathbf{w}_i^T\mathbf{t}_v)+\frac{\eta}{2}\|\mathbf{t}_v\|_2^2=\mathbf{W}^T\mathbf{r}+\eta\,\mathbf{t}_v, \quad (5)$$

where $\mathbf{r}$ is a $Q\times 1$ vector with its i^{th} element being $r_i=1-\frac{B_{i,v}}{\mathbf{w}_i^T\mathbf{t}_v}$. The projection step corresponds to the simple operation of setting negative entries in $\mathbf{t}_v$ to zero.

The subproblem of optimizing over $\mathbf{W}$ with $\mathbf{C}$ and $\mathbf{T}$ fixed is also separable in each row of $\mathbf{W}$. The problem for each $\mathbf{w}_i$ is:

$$\underset{\mathbf{w}_i:\ W_{i,j}\geq 0\forall j}{\text{minimize}} \quad \textstyle\sum_{j:(i,j)\in\Omega_{\text{obs}}} -\log p(Y_{i,j}|\mathbf{w}_i^T\mathbf{c}_j+\mu_i,\tau) + \sum_v -\log p(B_{i,v}|\mathbf{w}_i^T\mathbf{t}_v)+\lambda\|\mathbf{w}_i\|_1, \quad (6)$$

which can be efficiently solved using FISTA. Specifically, analogous to [5, Eq. 5], the gradient of the smooth part of the objective function with respect to $\mathbf{w}_i$ corresponds to:

$$\nabla_{\mathbf{w}_i}\Big(\textstyle\sum_{j:(i,j)\in\Omega_{\text{obs}}} -\log p(Y_{i,j}|\mathbf{w}_i^T\mathbf{c}_j+\mu_i,\tau) + \sum_v -\log p(B_{i,v}|\mathbf{w}_i^T\mathbf{t}_v)\Big)=-\mathbf{C}^T(\mathbf{y}_i-\mathbf{p})+\mathbf{T}^T\mathbf{s}, \quad (7)$$

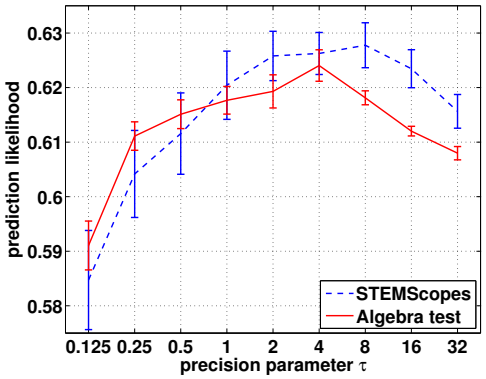


Figure 1: Average predicted likelihood on 20% holdout data in Y using SPARFA-Top with different precision parameters τ. For $\tau\to\infty$ SPARFA-Top corresponds to SPARFA as proposed in [5].

where $\mathbf{y}_i$ represents the transpose of the i^{th} row of $\mathbf{Y}$, $\mathbf{p}$ represents a $N\times 1$ vector with $p_j=1/(1+e^{-\mathbf{w}_i^T\mathbf{c}_j})$ as its j^{th} element, and $\mathbf{s}$ is a $V\times 1$ vector with $s_v=1-\frac{B_{i,v}}{\mathbf{w}_i^T\mathbf{t}_v}$ as its v^{th} element. The projection step is a soft-thresholding operation, as detailed in [5, Eq. 7]. The step-sizes are chosen via back-tracking line search as described in [4].

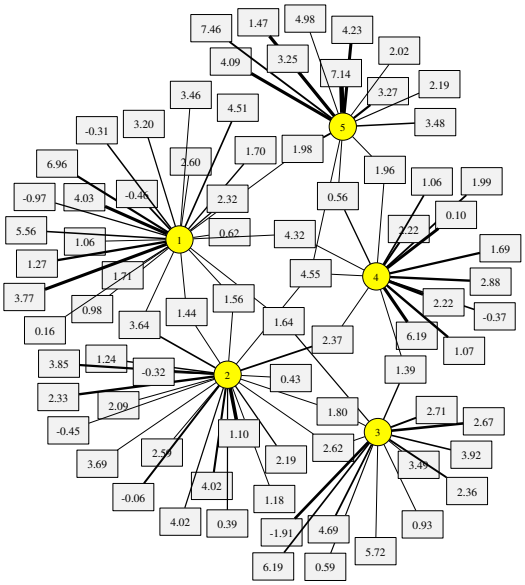
Note that we treat τ as a fixed parameter. Alternatively, one could estimate this parameter *within* the algorithm by introducing an additional step that optimizes over τ. A throughout analysis of this approach is left for future work.

4. EXPERIMENTS

We now demonstrate the efficacy of SPARFA-Top on two real-world educational datasets: an 8^{th} grade Earth science course dataset provided by STEMscopes [7] and a high-school algebra test dataset administered on Amazon's Mechanical Turk [1], a crowdsourcing marketplace. The STEMscopes dataset consists of 145 learners answering 80 questions, with only 13.5% of the total question/answer pairs being observed. The question–associated text vocabulary consists of 326 words, excluding common stop-words. The algebra test dataset consists of 99 users answering 34 questions, with the question–answer pairs fully observed. As no informative question text is available, we use the tags on each question to from a vocabulary of 13 words.

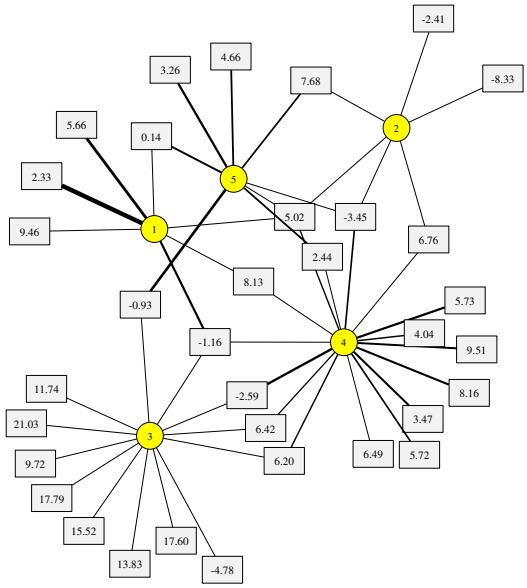
The regularization parameters λ, γ and η, together with the precision parameter τ of SPARFA-Top, are selected via cross-validation. In Figure 1, we show the prediction likelihood defined by $p(Y_{i,j}|\mathbf{w}_i^T\mathbf{c}_j+\mu_i,\tau),(i,j)\in\bar{\Omega}_{\text{obs}}$ for SPARFA-Top on 20% holdout entries in $\mathbf{Y}$ and for varying precision values τ. We see that textual information can slightly improve the prediction performance of SPARFA-Top over SPARFA (which corresponds to $\tau\to\infty$), for both the STEMscopes dataset and the algebra test dataset. The reason for (albeit slight) improvement in prediction performance is the fact that textual information reveals additional structure underlying a given test/assessment.

Figures 2 and 3 show the question–concept association graphs along with the recovered intrinsic difficulties, as well as the top three words characterizing each concept, for both



Concept 1	Concept 2	Concept 3
Energy	Water	Plants
Water	Percentage	Buffalo
Earth	Sand	Eat
Concept 4	**Concept 5**	
Water	Water	
Soil	Heat	
Sample	Objects	

Figure 2: Question–concept association graph and most important keywords recovered by SPARFA-Top for the STEMscopes dataset; boxes represent questions, circles represent concepts, and thick lines represent strong question–concept associations.



Concept 1	Concept 2	Concept 3
Solving equations	Slope	Arithmetic
Quadratic function	Fractions	Trigonometry
Fractions	Simplifying expressions	System of equations
Concept 4	**Concept 5**	
Simplifying expressions	Inequality	
Factoring polynomials	Plotting functions	
Fractions	Geometry	

Figure 3: Question–concept association graph and 3 most important keyword recovered by SPARFA-Top for the algebra test dataset; boxes represent questions, circles represent concepts, and thick lines represent strong question–concept associations.

datasets. Compared to SPARFA (see [5]), we observe that SPARFA-Top is able to relate all questions to concepts, including those questions that were found in [5, Figs. 2 and 9] to be unrelated to any concept. Furthermore, Figures 2 and 3 demonstrate that SPARFA-Top is capable of automatically generating an interpretable summary of the meaning of each concept.

5. CONCLUSIONS

We have introduced the SPARFA-Top framework, which extends SPARFA [5] by jointly analyzing both the binary-valued graded learner responses to a set of questions and the text associated with each question via a topic model. As our experiments have shown, our purely data-driven approach avoids the manual assignment of tags to each question and significantly improves the interpretability of the estimated concepts by automatically associating keywords extracted from question text to each estimated concept.

6. REFERENCES

[1] Amazon Mechanical Turk. http://www.mturk.com/mturk/welcome, Sep. 2012.

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. on Imaging Science*, 2(1):183–202, Mar. 2009.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, Jan. 2003.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk. Sparse factor analysis for learning and content analytics. Oct. 2012, submitted.

[6] H. Lee, R. Raina, A. Teichman, and A. Ng. Exponential family sparse coding with applications to self-taught learning. In *Proc. 21st Intl. Joint Conf. on Artificial Intelligence*, pages 1113–1119, July 2009.

[7] STEMscopes Science Education. http://stemscopes.com, Sep. 2012.

[8] E. Wang, D. Liu, J. Silva, D. Dunson, and L. Carin. Joint analysis of time-evolving binary matrices and associated documents. *Advances in neural information processing systems (NIPS)*, Dec. 2010.

[9] S. Williamson, C. Wang, K. Heller, and D. Blei. The IBP compound Dirichlet process and its application to focused topic modeling process and its application to focused topic modeling. In *Proc. 27th Intl. Conf. on Machine Learning*, pages 1151–1158, June 2010.

[10] Y. Xu and W. Yin. A block coordinate descent method for multi-convex optimization with applications to nonnegative tensor factorization and completion. Technical report, Rice University CAAM, Sep. 2012.

[11] X. X. Zhang and L. Carin. Joint modeling of a matrix with associated text via latent binary features. *Advances in neural information processing systems (NIPS)*, Dec. 2012.

[12] J. Zhu and E. P. Xing. Sparse topical coding. In *Proc. 27th Conf. on Uncertainty in Artificial Intelligence*, Mar. 2011.