SPARSE PROBIT FACTOR ANALYSIS FOR LEARNING ANALYTICS

Andrew E. Waters, Andrew S. Lan, and Christoph Studer

Rice University; e-mail: {waters, mr.lan, studer@sparfa.com}

ABSTRACT

We develop a new model and algorithm for machine learning-based *learning analytics*, which estimate a learner's knowledge of the concepts underlying a domain. Our model represents the probability that a learner provides the correct response to a question in terms of three factors: their understanding of a set of underlying concepts, the concepts involved in each question, and each question's intrinsic difficulty. We estimate these factors given the graded responses to a set of questions. We develop a bi-convex algorithm to solve the resulting *SPARse Factor Analysis* (SPARFA) problem. We also incorporate user-defined tags on questions to facilitate the interpretability of the estimated factors. Experiments with synthetic and real-world data demonstrate the efficacy of our approach.

Index Terms— bi-convex optimization, content analytics, learning analytics, personalized learning, factor analysis

1. INTRODUCTION

Textbooks, lectures, and homework assignments were the answer to the main educational challenges of the 19th century, but they are the bottleneck of the 21st century. Today's textbooks are typically static, linear in organization, time-consuming to develop, soon out-of-date, and expensive.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 and writes on a board. Homework assignments that are not graded for weeks provide poor feedback to learners 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 learners. Modern machine-learning (ML) algorithms provide a golden opportunity to reinvent the way we teach and learn by making it more personalized and, hence, more efficient.

One attractive venue for ameliorating many of the shortcomings associated with the traditional education approach is intelligent tutoring systems (ITS) [1–3]. On one hand, classical ITS systems consist of pre-defined rules hard-coded by domain experts, which are used to provide automated feedback. While such systems have the potential to allow for an instantaneous, feedback to learners, they require enormous investment in both time and money. On the other hand, ML-based ITS provide an affordable solution by mining learner data to provide feedback, however they tend to use rather unsophisticated machine-learning techniques [4], which limits their efficacy. In our opinion, one way in which machine-learning can ameliorate this situation is by augmenting the process of learning analytics (LA), which monitors and analyzes the learners' interaction with the course's contents. LA then provides automatic, targeted feedback to learners, to their instructors, and to the content authors.

Developing a principled method for LA presents a number of challenges. For example, how should one model knowledge and learner response data? How can one reliably estimate a learner's understanding of the various subject material? Given a database of potential practice problems, how can one identify the problems that are relevant to certain knowledge components?

This paper develops a novel statistical framework for LA that enables the analysis of learner responses to a set of *questions*. We assume that the knowledge base is decomposable into a set of latent knowledge components, which we term *concepts*, that each learner should learn. For example, an introductory calculus course would include concepts such as "integration-by-parts", "l'Hôpital's rule", etc. Our goal is to develop efficient methods to i) discover the relationships between concepts and questions, ii) estimate each learner's concept mastery, and iii) estimate each question's difficulty.

Our LA framework is based on a principled statistical model relying on latent factor analysis [5] from binary response values. Inspired by our statistical framework, we propose SPARFA-M, a bi-convex optimization technique for probit factor analysis to compute point estimates of the parameters of interest at low computational complexity. Our proposed model and algorithms differ significantly from previous work in factor analysis [6–8] due to the additional structure (i.e., non-negativity combined with sparsity) arising from our target application. We demonstrate the efficacy of the developed LA framework on both synthetic and real educational data. In the case of real-world datasets, we show that SPARFA-M can i) visualize question–concept associations as a bi-partite graph, ii) analyze the concept knowledge of each learner and provide personalized feedback on what he/she

First author determined by coin toss. The authors would like to thank R. .G. Baraniuk for insightful discussions and his support. This work was supported by the National Science Foundation under Cyberlearning grant IIS-1124535, the Air Force Office of Scientific Research under grant FA9550-09-1-0432, the Google Faculty Research Award program, and in part by the Swiss National Science Foundation under grant PA00P2-134155.

should improve on, and iii) identifying potentially off-topic or ill-posed questions that could potentially be removed for conciseness of the course/assessment.

2. RELATED WORK

A significant body of work in ML-based ITS uses Bayesian belief networks to probabilistically model and analyze learner response data [9, 10]. The deployed models, however, rely on pre-defined question–concept dependencies. In contrast, the LA framework proposed next discovers question–concept dependencies relying only on learner responses to questions. Matrix and tensor factorization methods, as, e.g., proposed in [11, 12], treat learner responses as real-values and do not consider a probabilistic model in their analysis, which differs substantially from our framework. We finally note that recent results presented in [13–15] address the problem of *predicting* missing entries in a binary-valued learner–question response matrix, an emerging field in educational data mining [16]. However, both methods retrieve data in a way that *inhibits the interpretation* of the underlying knowledge, which makes them unsuitable for LA. Item response theory (IRT) uses statistical models to analyze graded question response data [17, 18]. Although the SPARFA model shares similarity to the Rasch model [19], corresponding algorithms do not provide disciplined algorithms to estimate the model parameters.

3. STATISTICAL MODEL FOR LEARNING ANALYTICS

We view a given knowledge domain/course as consisting of both content and assessments. Each assessment consists of a number of *questions* that test the learner's understanding of various portions of the course's content. The latent factors governing the learners' answers to questions are referred to as *concepts*. For the sake of simplicity of exposition, we model the answers to questions as binary-valued entries, with 0 and 1 denoting incorrect and correct responses, respectively.

3.1. Sparse probit model for learner response data

Assume that there are N learners, Q questions, and K underlying concepts. Let the column vector $\mathbf{c}_j \in \mathbb{R}^K$, $j \in \{1, \dots, N\}$, represent the latent concept mastery of the j^{th} learner, with its k^{th} component representing the j^{th} learner's understanding of the k^{th} concept. Then, for the i^{th} question, with $i \in \{1, \dots, Q\}$, we propose the following model for the learner–response relationships:

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

Here, $Y_{i,j} \in \{0, 1\}$ denotes the binary-valued response variable of the j^{th} learner to the i^{th} question. The set $\Omega_{\text{obs}} \subseteq \{1, \dots, Q\} \times \{1, \dots, N\}$ in (1) contains the indices associated to the observed learner–response data, in case $\mathbf{Y}$ is not fully observed. The slack variable $\Phi(Z_{i,j}) \in [0, 1]$ governs the probability of learner j answering question i correctly, $\textit{Ber}(z)$ designates a Bernoulli distribution with mean z, and $\Phi(x) = \int_{-\infty}^{x} \mathcal{N}(g|0, 1) \mathrm{d}g$ denotes the inverse *probit* function with $\mathcal{N}(g|0, 1)$ representing the value of the probability density function of a standard normal distribution evaluated at g. The column vector $\mathbf{w}_i \in \mathbb{R}^K$ models the *concept associations*; that is, it encodes how question i is related to each concept. The scalar $\mu_i \in \mathbb{R}$ models the *intrinsic difficulty* of question i. In the remainder of the paper, we will often write (1) in matrix form as

$$\begin{aligned} \mathbf{Z} &= \mathbf{W}\mathbf{C} + \mathbf{M}, \\ Y_{i,j} &\sim \textit{Ber}\big(\Phi(Z_{i,j})\big), \ (i,j) \in \Omega_{\text{obs}}, \end{aligned} \tag{2}$$

where $\mathbf{Y}$, $\mathbf{M}$, and $\mathbf{Z}$ are $Q \times N$ matrices, $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_Q]^T$ and $\mathbf{C} = [\mathbf{c}_1, \dots, \mathbf{c}_N]$ are $Q \times K$ and $K \times N$ matrices, respectively. Matrix $\mathbf{M} = \boldsymbol{\mu}\mathbf{1}_{1\times N}$ is formed by $\boldsymbol{\mu} = [\mu_1, \dots, \mu_Q]^T$ and the N-dimensional all-ones row vector $\mathbf{1}_{1\times N}$.

3.2. Fundamental assumptions

Estimating latent factors $\mathbf{W}$ and $\mathbf{C}$ from observation matrix $\mathbf{Y}$ usually is in an ill-posed inverse problem. To reduce the number of parameters and to improve identifiability, we build our framework on the following three assumptions, which we argue are reasonable for typical exam and homework questions across all levels of education:

(A1) *Low-dimensionality*: K is small relative to N and Q, implying question redundancy and low-dimensionality of the learner–question response space.

(A2) *Sparsity*: The ability to answer a question correctly depends only on a few concepts (relative to all concepts covered by a given course). Therefore, $\mathbf{W}$ is sparse.

(A3) *Non-negativity*: The learners' mastery of concepts should not reduce the chance of correctly answering questions, i.e., mastery should never be "harmful." The entries of $\mathbf{W}$ are therefore non-negative; this provides additional interpretability to the element of $\mathbf{C}$, where large, positive values denote strong concept mastery while large negative values denote poor mastery.

The goal of the algorithm we develop next is to estimate $\mathbf{W}$, $\mathbf{C}$, and $\boldsymbol{\mu}$ from the binary-valued observation matrix $\mathbf{Y}$ while enforcing (A1), (A2), and (A3).

4. SPARFA-M: SPARSE PROBIT BINOMIAL FACTOR ANALYSIS

Our algorithm to estimate the factors $\mathbf{W}$, $\mathbf{C}$, and $\boldsymbol{\mu}$ from response data $\mathbf{Y}$ solves a factor-analysis problem using *probit regression*, in contrast to previous work that perform principal component analysis on binary matrix data [8].

4.1. Problem formulation

To estimate $\mathbf{W}$, $\mathbf{C}$, and $\boldsymbol{\mu}$, we maximize the log-likelihood of $\mathbf{Y}$ subject to the assumptions (A1), (A2), and (A3), where the likelihood of each response variable is given by

$$p(Y_{i,j}|\mathbf{w}_i, \mathbf{c}_j) = \Phi\big(\mathbf{w}_i^T \mathbf{c}_j\big)^{Y_{i,j}} \big(1 - \Phi(\mathbf{w}_i^T \mathbf{c}_j)\big)^{1-Y_{i,j}}.$$

To this end, we seek to solve the following ℓ_1/ℓ_2-norm regularized optimization problem:

$$\text{(P)} \quad \underset{\mathbf{W},\mathbf{C}\colon \mathbf{W}\geq 0}{\text{minimize}} \; -\textstyle\sum_{i,j:(i,j)\in\Omega_{\text{obs}}} \log p(Y_{i,j}|\mathbf{w}_i, \mathbf{c}_j) + \lambda \textstyle\sum_i \|\mathbf{w}_i\|_1 + \frac{\gamma}{2} \sum_j \|\mathbf{c}_j\|_2^2.$$

Here, the first regularization term $\lambda \sum_i \|\mathbf{w}_i\|_1$ induces sparsity on each vector $\mathbf{w}_i$ as required by (A2); the parameter $\lambda > 0$ controls the sparsity level. The constraint in (P) imposes non-negativity in $\mathbf{W}$. Since one can arbitrarily increase the scale of either $\mathbf{W}$ or $\mathbf{C}$ while decreasing the scale of the other accordingly, we gauge the vectors $\mathbf{c}_j$ using the second regularization term $\frac{\gamma}{2} \sum_j \|\mathbf{c}_j\|_2^2$ with regularization parameter $\gamma > 0$. We incorporate the intrinsic difficulty vector $\boldsymbol{\mu}$ in (2) by adding it as a column to $\mathbf{W}$, and by accordingly augmenting $\mathbf{C}$ with a fixed all-ones row vector.

4.2. The SPARFA-M algorithm

Since the probit log-likelihood function is concave in the product $\mathbf{WC}$ [20], the problem (P) is *bi-convex* with respect to the individual factors $\mathbf{W}$ and $\mathbf{C}$. To arrive at a practicable way of finding an approximate solution to (P), we propose SPARFA-M, an algorithm that bases on the following alternating optimization approach.

After initializing $\mathbf{C}$ and $\mathbf{W}$ with random matrices, we iteratively optimize the objective of (P) for both factors in an alternating fashion. Thus, each (outer) iteration consists of two phases. In the first phase, we hold $\mathbf{W}$ constant and separately optimize each vector $\mathbf{c}_j$; in the second phase, we hold $\mathbf{C}$ constant and separately optimize each vector $\mathbf{w}_i$. Individual subproblems in each phase are carried out by iterative methods, which form the inner iterations of SPARFA-M, discussed below. The outer loop is terminated either if a maximum number of iterations I is reached, or if the decrease in the objective function of (P) is smaller than a predefined threshold.

The individual sub-problems to be solved in each phase correspond to the following convex ℓ_1-norm and ℓ_2-norm regularized probit regression problems:

$$\underset{\mathbf{w}_i\,:\,W_{i,k}\geq 0\;\forall k}{\arg\min} \; -\textstyle\sum_{j\colon (i,j)\in\Omega_{\text{obs}}} \log p(Y_{i,j}|\mathbf{w}_i, \mathbf{c}_j) + \lambda \|\mathbf{w}_i\|_1,$$

$$\underset{\mathbf{c}_j}{\arg\min} \; -\textstyle\sum_{i\colon (i,j)\in\Omega_{\text{obs}}} \log p(Y_{i,j}|\mathbf{w}_i, \mathbf{c}_j) + \frac{\gamma}{2} \|\mathbf{c}_j\|_2^2 .$$

In contrast to logistic regression, the second derivative of the *probit* log-likelihood function inhibits efficient evaluation[1]. Consequently, we develop two novel first-order methods that efficiently solve these regularized probit regression subproblems, by building on the fast iterative soft-thresholding algorithm (FISTA) [21]. Both algorithms iteratively perform a gradient step on the log-likelihood terms followed by a proximal mapping step with respect to the regularization terms, which form the inner loop of SPARFA-M.

While SPARFA-M reduces the objective function over the iterations, it does not necessarily converge to a global optimum due to its bi-convex nature. Nevertheless, using recent results in [22], we have established the following global convergence guarantee of SPARFA-M from any arbitrary starting point to a local optimum. Details can be found in [23].

Theorem 1 *(Global convergence of SPARFA-M) From any starting point, SPARFA-M converges to a critical point of (P). Moreover, if the starting point is within a close neighborhood of a global optimum of (P), then SPARFA-M converges to this global optimum.*

[1]This fact inhibits the use of iteratively reweighted second-order algorithms [20]. Note that our probit-based LA model can easily be adjusted to a logit-based model, which we do not discuss here due to space limitations.

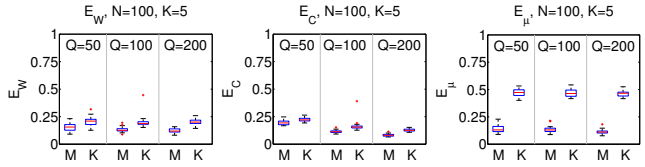

Fig. 1. Performance comparison (SPARFA-M and K-SVD$_+$).

5. EXPERIMENTS

5.1. Synthetic data

We start by evaluating the performance of SPARFA-M using synthetic test data. To arrive at a fair comparison, we use a non-negative variant of the K-SVD algorithm [24], referred to as K-SVD$_+$, as a base-line. For K-SVD$_+$, we use a non-negative variant of orthogonal matching pursuit [25] in the sparse-coding step; in the dictionary update stage we impose non-negativity in $\mathbf{W}$ as in [26, Fig. 4]. We provide K-SVD$_+$ with the true number of non-zero elements for each $\mathbf{w}_i$, which clearly favors this algorithm over SPARFA-M.

In all synthetic experiments, we estimate $\widehat{\mathbf{W}}$, $\widehat{\mathbf{C}}$, and $\hat{\boldsymbol{\mu}}$ of the true parameters $\mathbf{W}$, $\mathbf{C}$, and $\boldsymbol{\mu}$. Since factor analysis is susceptible to permutations and scaling in $\mathbf{W}$ and $\mathbf{C}$, we normalize each row of $\mathbf{C}$, $\widehat{\mathbf{C}}$ and each column of $\mathbf{W}$ and $\widehat{\mathbf{W}}$ to unit ℓ_2-norm. We then permute the columns of $\widehat{\mathbf{W}}$ and $\widehat{\mathbf{C}}$ to match best with the ground truth and use the following performance metrics:

$$E_{\mathbf{W}} = \frac{\|\mathbf{W} - \widehat{\mathbf{W}}\|_F^2}{\|\mathbf{W}\|_F^2}, \; E_{\mathbf{C}} = \frac{\|\mathbf{C} - \widehat{\mathbf{C}}\|_F^2}{\|\mathbf{C}\|_F^2}, \; E_{\boldsymbol{\mu}} = \frac{\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_2^2}{\|\boldsymbol{\mu}\|_2^2},$$

We generate synthetic data as follows: Set $K = 5$ for all trials. For each trial we generate $\mathbf{C}_{i,j}, \boldsymbol{\mu}_i \sim \mathcal{N}(0, 1)$. For $\mathbf{W}$ we choose the number of active concepts per row in $\{1, 2, 3\}$

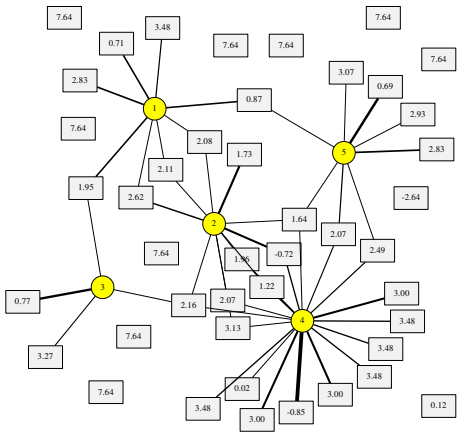

Fig. 2. Question–concept association graph extracted by SPARFA-M on an undergraduate DSP course. Circles correspond to latent concepts. Squares correspond to questions with the numerical label denoting the intrinsic difficulty.

Concept 1	Concept 2
Frequency response (46%)	CTFT (40%)
Sampling rate (23%)	Laplace transform (36%)
Aliasing (21%)	DTFT (24%)
Concept 3	**Concept 4**
Z-transform (66%)	CTFT (43%)
Pole/zero plot (22%)	Systems/circuits (31%)
Laplace transform (12%)	Transfer function (26%)
Concept 5	
Impulse response (74%)	
Transfer function (15%)	
DTFT (11%)	

Table 1. Three most important tags and relative weights for the recovered concepts.

with equal probability. Non-zero entries in $\mathbf{W}$ are generated from the exponential distribution with rate parameter $2/3$. The sparsity regularization parameter λ for SPARFA-M is selected according to Bayesian information criterion (BIC) [20].

We first consider the impact of problem size on estimation error. To this end, we fix $N = 100$ and sweep $Q \in \{50, 100, 200\}$ for $K = 5$ concepts. We retrieve parameter estimates for both algorithms and compute performance using metrics. The results we show are averaged over 25 Monte-Carlo trials. Box-and-Whisker plots of the estimation error for each algorithm are shown in Fig. 1. We observe that the performance error metrics decrease as the problem size increases. SPARFA-M has superior performance to K-SVD$_+$ in all cases, despite the fact that K-SVD$_+$ is provided with the oracle sparsity level of each question.

5.2. Real-world data

We analyze a small database consisting of 15 learners answering 44 questions taken from the final exam of an introductory course on digital signal processing (DSP).

We estimate $\mathbf{W}$, $\mathbf{C}$, and $\boldsymbol{\mu}$ from the fully-populated 44×15 binary-valued matrix $\mathbf{Y}$ using SPARFA-M assuming $K = 5$ concepts to prevent over-fitting. In practice, it is crucial to identify the meaning of each concept. The proposed LA framework enables us to infer this information. To this end, we make use of 12 user-defined tags that were assigned manually to each question by the instructor. We note that this incorporation of tags is not necessary for SPARFA-M but is done to provide additional interpretability to the estimates provided by SPARFA-M. We form a sparse 44×12 matrix $\mathbf{T}$, where the columns correspond to each of the 12 pre-defined tags, and we set a 1 at the locations where a tag is present in a given question (and set 0 otherwise). We postulate that the learned question concept association matrix $\mathbf{W}$ can be further decomposed as $\mathbf{W} = \mathbf{TA}$, where $\mathbf{A}$ is a 12×5 sparse non-negative matrix representing the tags-to-concept mapping. This assumption enables us to extract $\mathbf{A}$ using ℓ_1-norm regularized least-squares [27]. From $\mathbf{A}$, we can now associate tags with concepts. In the bipartite graph shown in Fig. 2, circles represent concepts, while squares represent questions (along with their intrinsic difficulty μ_i). Connecting lines indicate whether a concept is present in a question with thick, dark lines representing stronger question–concept associations. Tbl. 1 lists the three most important tags and the proportions they contribute to each concept, by examining the magnitude of non-zero entries in $\mathbf{A}$.

Note that there are some questions that are not linked to any concept. Here, all 15 learners answer these questions correctly, and so nothing can be inferred about the underlying concept structure of these questions.

6. CONCLUSIONS

We have proposed a sparse probit factor analysis framework for learning analytics (LA). Our approach enables joint recovery of question–concept associations, question intrinsic difficulties as well as learner concept knowledge profile from binary graded learner response data. The estimated question–concept association can be used to visualize the knowledge base structure in a given course or assessment. Together with the estimated question intrinsic difficulties, they enable an ITS to provide feedback to course instructors to identify off-topic or ill-posed questions. The estimated learner concept knowledge also enables an ITS to provide feedback to students about their learning progress. Leveraging this information, an ITS is able to make personalized recommendations to learners, e.g., recommending remedial material to learners who have not demonstrated sufficient concept knowledge, or providing new material to learners who have. This personalized approach to education has the potential to greatly enhance learning efficiency and reduce instructor workload.

7. REFERENCES

[1] J. A. Kulik, *Meta-analytic studies of findings on computer-based instruction*, Routledge, in "Technology Assessment in Education and Training," Eds. E. Baker, H. F. O'Neil Jr., and H. F. O'Neil, 1994.

[2] P. Brusilovsky and C. Peylo, "Adaptive and intelligent web-based educational systems," *International Journal of Artificial Intelligence in Education*, vol. 13, no. 2-4, pp. 159–172, Apr. 2003.

[3] J. A. Dijksman and S. Khan, "Khan Academy: the world's free virtual school," in *Bulletin APS*, Mar. 2011, vol. 56.

[4] D. Hu, "How Khan academy is using machine learning to assess student mastery," *online: http://david-hu.com/*, Nov. 2011.

[5] D. Child, *The essentials of factor analysis*, Continuum Intl Pub Group, New York, NY, 3rd edition, 2006.

[6] M. N. Schmidt, O. Winther, and L. K. Hansen, "Bayesian non-negative matrix factorization," in *Independent Component Analysis and Signal Separation*, 2009, vol. 5441, pp. 540–547.

[7] S. Chib and E. Greenberg, "Analysis of multivariate probit models," *Biometrika*, vol. 85, no. 2, pp. 347–361, June 1998.

[8] S. Lee, J. Z. Huang, and J. Hu, "Sparse logistic principal components analysis for binary data," *Annals of Applied Statistics*, vol. 4, no. 3, pp. 1579–1601, 2010.

[9] B. P. Woolf, *Building Intelligent Interactive Tutors: Student-centered Strategies for Revolutionizing E-learning*, Morgan Kaufman Publishers, 2008.

[10] G. A. Krudysz and J. H. McClellan, "Collaborative system for signal processing education," in *2011 IEEE International Conference on on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 2904 – 2907.

[11] T. Barnes, "The Q-matrix method: Mining student response data for knowledge," in *Proceedings of the AAAI Workshop Educational Data Mining*, Pittsburg, PA, July 2005.

[12] N. Thai-Nghe, L. Drumond, T. Horvath, and L. Schmidt-Thieme, "Multi-relational factorization models for predicting student performance," *KDD 2011 Workshop on Knowledge Discovery in Educational Data (KDDinED)*, Aug. 2011.

[13] M. Desmarais, "Conditions for effectively deriving a Q-matrix from data with non-negative matrix factorization," in *Proceedings of the 4th International Conference on Educational Data Mining*, Eindhoven, the Netherlands, Jul. 2011, pp. 41–50.

[14] B. Beheshti, M. Desmarais, and R. Naceur, "Methods to find the number of latent skills," in *Proceedings of the 5th International Conference on Educational Data Mining*, Chania, Greece, Jun. 2012, pp. 81–86.

[15] Y. Bergner, S. Droschler, G. Kortemeyer, S. Rayyan, D. Seaton, and D. Pritchard, "Model-based collaborative filtering analysis of student response data: Machine-learning item response theory," in *Proceedings of the 5th International Conference on Educational Data Mining*, Chania, Greece, Jun. 2012, pp. 95–102.

[16] KDD-2010, "Educational data mining challenge," 2010, http://pslcdatashop.web.cmu.edu/KDDCup/.

[17] F. M. Lord, *Applications of Item Response Theory to Practical Testing Problems*, Erlbaum Associates, 1980.

[18] M. D. Reckase, *Multidimensional Item Response Theory*, Springer Publishing Company, Incorporated, 1st edition, 2009.

[19] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests*, MESA Press, 1993.

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2010.

[21] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Science*, vol. 2, no. 1, pp. 183–202, Mar. 2009.

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

[23] A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk, "Sparse factor analysis for learning and content analytics," Oct. 2012, submitted to Journal of Machine Learning Research (JMLR).

[24] M. Aharon, M. Elad, and A. M. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transaction on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Dec. 2006.

[25] A. M. Bruckstein, M. Elad, and M. Zibulevsky, "On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations," *IEEE Transactions on Info Theory*, vol. 54, no. 11, pp. 4813–4820, Nov. 2008.

[26] M. Aharon, M. Elad, and A. M. Bruckstein, "K-SVD and its non-negative variant for dictionary design," in *Proceedings of the SPIE Conference on on Wavelets*, Jul. 2005, vol. 5914, pp. 327–339.

[27] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, Mar. 1998.