Test-size Reduction Using Sparse Factor Analysis

Divyanshu Vats*, Christoph Studer, and Richard G. Baraniuk

Rice University, TX, USA; e-mail: $\{\text{dvats, studer, richb}\}$@sparfa.com

***Abstract*—Consider a large database of questions that test the knowledge of learners (e.g., students) about a range of different concepts. While the main goal of personalized learning is to obtain accurate estimates of each learner's concept understanding, it is additionally desirable to reduce the number of questions to minimize each learner's workload. In this paper, we propose a novel method to extract a small subset of questions (from a large question database) that still enables the accurate estimation of a learner's concept understanding. Our method builds upon the SPARse Factor Analysis (SPARFA) framework and chooses a subset of questions that minimizes the entropy of the error in estimating the level of concept understanding. We approximate the underlying combinatorial optimization problem using a mixture of convex and greedy methods and demonstrate the efficacy of our approach on real educational data.**

I. Introduction

There has been a recent surge in providing free and high-quality online education through ventures, such as Coursera, Udacity, and edX.[1] Among the key challenges of such systems is in the estimation of each learner's concept understanding. Such information is essential in order to automatically recommend remediation about concepts each learner has weak knowledge of (see, e.g., [6] for the details). In practice, accurate estimates for each learner's concept understanding can be extracted automatically by analyzing responses to large sets of questions about the concepts underlying the given class. To minimize each learner's workload, however, it is of paramount importance to reduce the test-size (compared to the size of the entire question database), while still enabling accurate estimates of each learner's concept understanding. We refer to this problem as *test-size reduction* (TeSR).

In this paper, we propose a novel algorithm for test-size reduction (TeSR), i.e., the problem of selecting a small number of questions from a large dataset, while enabling the accurate estimation of conceptual understanding of each learner. Our approach builds upon the *SPARse Factor Analysis* (SPARFA) framework proposed in [6] to automatically estimate the latent concepts associated with each question. Then, using theory of maximum likelihood (ML) estimators, we formulate the TeSR problem as a combinatorial optimization problem that minimizes the entropy of the asymptotic error in estimating the concept understanding of each learner. We show how the optimization problem can be solved approximately using a combination of convex and greedy methods. We then highlight the advantages of the proposed method by carrying out an experiment with real educational data.

Prior work on selecting a subset of questions mainly use statistical models that rely on a single parameter that captures the concept understanding of a learner [3]. In contrast, the SPARFA model used in this work assumes that there are multiple concepts involved in a database of questions. This scenario is more realistic in practice, since it is often the case that questions test knowledge from multiple concepts simultaneously. Several authors have considered the problem of selecting questions in an adaptive manner, see, e.g., [2], [7]. All these adaptive algorithms require a set of starting questions to gauge the adaptive process. Our proposed method can be used for this purpose and is designed to minimize the error of the initial concept understanding estimates, which eventually improves the performance of adaptive methods. We finally note that the problem of selecting questions is related to the problem of sensor selection [5]. The main difference is that the data in sensor network problems is typically real valued, whereas the SPARFA model focuses on binary-valued measurements (i.e., right and wrong answers to questions).

II. Problem Formulation

We begin by reviewing the SPARFA model [6] for extracting relationships between questions and concepts from graded question responses. We then detail the TeSR problem to select "good" subsets of questions for concept estimation.

A. The SPARFA Framework in a Nutshell

Suppose we have a total of Q questions that test knowledge from K concepts. For example, in a signal processing course, questions can test knowledge on concepts like convolution, the sampling theorem, or the Fourier transform. For each question $i = 1, \ldots, Q$, let $\mathbf{w}_i \in \mathbb{R}^{K \times 1}$ be a column vector that represents the association of question i to all concepts. Note that a question can test knowledge from multiple concepts. For example, a question on the convolution theorem (i.e., the Fourier transform of a convolution is the product of Fourier transforms of the two signals to be convoluted) in signal processing may test the learner's knowledge on both convolution and the Fourier transform.

The j^{th} entry in $\mathbf{w}_i$, which we denote by w_{ij}, measures the association of question i to concept j. In other words, if question i does not test any knowledge from concept j, then $w_{ij} = 0$. Let $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_Q]^T$ be a sparse $Q \times K$ matrix with non-negative entries so that each question only tests a subset of all concepts. Let $\mu_i \in \mathbb{R}$ be a scalar that represents the intrinsic difficulty of a question. A larger (smaller) μ_i corresponds to an easier (harder) question. Let $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_Q]^T$ be a $Q \times 1$ column vector that represents the difficulty of

*Also affiliated with the Institute for Mathematics and its Applications, University of Minnesota - Twin Cities, USA.

[1] https://www.coursera.org ; https://www.udacity.com ; https://www.edx.org

Notation	Description
$\mathbf{W}$	A *sparse* non-negative matrix that characterizes the relationship between questions and knowledge concepts
$\boldsymbol{\mu}$	A vector that specifies the intrinsic difficulty of each question
$\mathbf{c}^*$	A vector that represents a learner's concept knowledge

TABLE I
MAIN PARAMETERS OF THE SPARFA MODEL.

each question. Finally, let $\mathbf{c}^* \in \mathbb{R}^K$ be a column vector that represents the concept understanding of a particular learner. It is this parameter vector that personalized learning systems are naturally interested in estimating accurately.

To model the interaction between $\mathbf{W}$, $\boldsymbol{\mu}$, and $\mathbf{c}^*$, we use the SPARFA framework proposed in [6]. Let Y_i be a binary random variable that indicates whether question i has been answered correctly or not, indicated by 1 and 0, respectively. More specifically, we assume that $Y_i \in \{0, 1\}$ admits the following distribution:

$$\Pr(Y_i = 1 \,|\, \mathbf{w}_i, \mu_i, \mathbf{c}^*) = \Phi(\mathbf{w}_i^T \mathbf{c}^* + \mu_i)\,, \tag{1}$$

where $\Phi(\cdot)$ is an appropriate link function. In this paper, we consider the logistic link function, i.e., $\Phi(x) = 1/(1 + e^{-x})$. Assuming that all the random variables $Y_1, \ldots, Y_Q$ are independent of each other, the joint probability distribution of the random vector $\mathcal{Y} = [Y_1, \ldots, Y_Q]^T$ can be written as

$$\Pr(\mathcal{Y} = \mathbf{y} \,|\, \mathbf{W}, \boldsymbol{\mu}, \mathbf{c}^*) = \prod_{i=1}^{Q} \frac{\exp(y_i(\mathbf{w}_i^T \mathbf{c}^* + \mu_i))}{1 + \exp(\mathbf{w}_i^T \mathbf{c}^* + \mu_i)}\,, \tag{2}$$

where $\mathbf{y} = [y_1, \ldots, y_Q]^T \in \{0, 1\}^Q$ is the response of a learner to all the questions. Given graded question responses from multiple learners, the problem of computing the factors $\mathbf{W}$, $\boldsymbol{\mu}$, and the concept understanding vector for each learner can be solved using either the SPARFA-M or SPARFA-B algorithm proposed in [6].

B. Problem Statement: Test-size Reduction (TeSR)

The problem we consider here is to select an appropriate subset of $q < Q$ questions so that $\mathbf{c}^*$, the unknown concept understanding vector of a learner, can be estimated accurately. We assume that prior data, a binary-valued matrix $\widetilde{\mathbf{Y}}$, is known such that an entry $\widetilde{\mathbf{Y}}_{i,j}$ refers to whether a learner j answered question i correct or incorrect. This data matrix can be easily obtained in real educational settings by looking at past offerings of a course, for example. As mentioned in Section II-A, we can compute $\mathbf{W}$ for all the Q questions in the database using $\widetilde{\mathbf{Y}}$.

Suppose, hypothetically, that we choose a subset $\mathcal{I}$ of $q < Q$ questions and we are given a response vector $\mathbf{y}_{\mathcal{I}}$. Using the model in (2), the maximum likelihood (ML) estimate $\widehat{\mathbf{c}}$ can be computed as follows:

$$\begin{aligned}\widehat{\mathbf{c}} &= \arg\max_{\mathbf{c} \in \mathbb{R}^K} \log \Pr(\mathcal{Y}_{\mathcal{I}} = \mathbf{y}_{\mathcal{I}} \,|\, \mathbf{W}, \boldsymbol{\mu}, \mathbf{c}) \\ &= \arg\max_{\mathbf{c} \in \mathbb{R}^K} \sum_{i \in \mathcal{I}} \Big[y_i(\mathbf{w}_i^T \mathbf{c} + \mu_i) - \log(1 + e^{\mathbf{w}_i^T \mathbf{c} + \mu_i}) \Big]\,. \end{aligned} \tag{3}$$

Given $\mathbf{y}_{\mathcal{I}}$, the result of (3) can be solved via standard convex optimization algorithms [1]. Our main objective is to find a subset $\mathcal{I}$ so that $|\mathcal{I}| = q$ and the ML estimate $\widehat{\mathbf{c}}$ is as close to the ground truth $\mathbf{c}^*$ as possible. To do this, we make use of the following asymptotic normality property of ML estimators (see, e.g., [4] for the details). First, define the Fisher information matrix as follows:

$$\mathbf{F}(\mathbf{W}_{\mathcal{I}}, \boldsymbol{\mu}_{\mathcal{I}}, \mathbf{c}^*)) = \sum_{i \in \mathcal{I}} \frac{\exp(\mathbf{w}_i^T \mathbf{c}^* + \mu_i)}{(1 + \exp(\mathbf{w}^T \mathbf{c}^* + \mu_i))^2} \mathbf{w}_i \mathbf{w}_i^T\,, \tag{4}$$

where the notation $\mathbf{W}_{\mathcal{I}}$ refers to the rows of $\mathbf{W}$ indexed by $\mathcal{I}$ and $\boldsymbol{\mu}_{\mathcal{I}}$ refers to the entries in $\boldsymbol{\mu}$ indexed by $\mathcal{I}$.

Theorem II.1. *Let $\mathcal{I}_r$ for $r = 1, \ldots, q$ be a fixed sequence of q subsets of size r. Assume that there exists a $q_0 < q$ such that $\mathbf{F}(\mathbf{W}_{\mathcal{I}_q}, \boldsymbol{\mu}_{\mathcal{I}_q}, \mathbf{c}^*))$ is invertible for all $r > q_0$. Then, the random vector $\sqrt{q}(\widehat{\mathbf{c}} - \mathbf{c}^*)$ converges in distribution to a multivariate normal vector with mean zero and covariance $\mathbf{F}(\mathbf{W}_{\mathcal{I}_q}, \boldsymbol{\mu}_{\mathcal{I}_q}, \mathbf{c}^*))^{-1}$, i.e., $\sqrt{q}(\widehat{\mathbf{c}} - \mathbf{c}^*) \xrightarrow{d} \mathcal{N}(0, \mathbf{F}(\mathbf{W}_{\mathcal{I}_q}, \boldsymbol{\mu}_{\mathcal{I}_q}, \mathbf{c}^*))^{-1})$.*

Theorem II.1 states that as the number of questions q gets large, the covariance of the error $\sqrt{q}(\widehat{\mathbf{c}} - \mathbf{c}^*)$ can be approximated by the inverse of the Fisher information matrix. This motivates a natural strategy to choose a subset of questions $\mathcal{I}$ that minimizes the differential entropy[2] of a multivariate normal random vector with mean zero and covariance $\mathbf{F}(\mathbf{W}_{\mathcal{I}}, \boldsymbol{\mu}_{\mathcal{I}}, \mathbf{c}^*))^{-1}$, which intuitively minimizes the uncertainty in the error $\sqrt{q}(\widehat{\mathbf{c}} - \mathbf{c}^*)$. Consequently, the optimization problem considered in the remainder of the paper, referred to as the *test-size reduction* (TeSR) problem, corresponds to

$$\text{(TeSR)} \quad \widehat{\mathcal{I}} = \underset{\mathcal{I} \subset \{1, \ldots, Q\}, |\mathcal{I}| = q}{\arg\max} \log\det(\mathbf{F}(\mathbf{W}_{\mathcal{I}}, \boldsymbol{\mu}_{\mathcal{I}}, \mathbf{c}^*))\,.$$

The main challenges in solving (TeSR) are (i) the TeSR problem is a combinatorial optimization problem and (ii) the concept knowledge vector $\mathbf{c}^*$ is *unknown*, so the objective function cannot be evaluated exactly. In the next section, we outline a data-driven approach for approximating the (TeSR) objective function. We then develop a computationally efficient algorithm that delivers good approximations to the combinatorial TeSR problem.

III. TEST-SIZE REDUCTION ALGORITHM

We start by noting that the scalar term in the summation in (4) is equivalent to the variance of the random variable Y_i

[2] Note that the differential entropy of $X = (X_1, \ldots, X_q) \sim \mathcal{N}(0, \Sigma)$ is given by $\log\big((2\pi e)^q \det(\Sigma)\big)$.

Algorithm 1: Nonadaptive test-size reduction (NA-TeSR)

```
Step 1) First choose $K$ questions by solving
    $\widehat{\mathcal{I}}_{[K]} = \underset{\mathcal{I}\subset\{1,\ldots,Q\},|\mathcal{I}|=K}{\arg\max} \log\det\left(\mathbf{W}_{\mathcal{I}}^T\widehat{\mathbf{V}}\mathbf{W}_{\mathcal{I}}\right)$    (5)
    using the convex optimization method in (8). The entries
    of the diagonal matrix $\widehat{\mathbf{V}}$ are defined as $\widehat{\mathbf{V}}_{kk} = \widehat{v}_k$,
    where $\bar{v}_k$ specified in (6).
Step 2) Select questions $K+1,\ldots,q$ in a greedy manner:
    $\widehat{\mathcal{I}}_{j+1} = \underset{i\in\{1\ldots,Q\}\setminus\widehat{\mathcal{I}}_{[j]}}{\arg\max} \widehat{v}_i\mathbf{w}_i^T\left(\mathbf{W}_{\mathcal{I}_{[j]}}^T\widehat{\mathbf{V}}_{\mathcal{I}_{[j]}}\mathbf{W}_{\mathcal{I}_{[j]}}\right)^{-1}\mathbf{w}_i\,.$
```

conditioned on $\mathbf{c}^*$, i.e.,

$$\mathbb{V}\mathrm{ar}[Y_i|\mathbf{c}^*] = \frac{\exp(\mathbf{w}_i^T\mathbf{c}^* + \mu_i)}{(1+\exp(\mathbf{w}^T\mathbf{c}^* + \mu_i))^2}. \tag{6}$$

The variance $\mathbb{V}\mathrm{ar}[Y_i|\mathbf{c}^*]$ captures the variability of a learner in answering the i^{th} question. By defining $\mathbf{V}$ as a $Q\times Q$ diagonal matrix with entries $\mathbf{V}_{ii} = \mathbb{V}\mathrm{ar}[Y_i|\mathbf{c}^*]$, the TeSR problem can be rewritten in matrix form as

$$(\text{TeSR}) \quad \widehat{\mathcal{I}} = \underset{\mathcal{I}\subset\{1,\ldots,Q\},|\mathcal{I}|=q}{\arg\max} \log\det(\mathbf{W}_{\mathcal{I}}^T\mathbf{V}_{\mathcal{I}}\mathbf{W}_{\mathcal{I}})\,.$$

We first address the problem of approximating the objective function using a graded question response matrix $\widetilde{\mathbf{Y}}$ acquired in, e.g., a previous offering of a course. Since the vector $\mathbf{c}^*$ is not known, we need to make some assumptions on $\widetilde{\mathbf{Y}}$ so that the objective function can be estimated. As it turns out, a natural, and convenient, assumption is for the prior data to be chosen in such a way that the concept understanding of the learners in the response matrix $\widehat{\mathbf{Y}}$ is roughly equal to $\mathbf{c}^*$. Using this assumption, we can easily estimate $\mathbb{V}\mathrm{ar}[Y_i|\mathbf{c}^*]$ to be the sample variance of the data $\widetilde{\mathbf{Y}}$:

$$\widehat{v}_i = \mathbb{V}\mathrm{ar}[Y_i|\mathbf{c}^*] = \frac{1}{N}\sum_{j=1}^{N}\left(\widetilde{Y}_{ij} - \frac{1}{N}\sum_{j=1}^{N}\widetilde{Y}_{ij}\right), \tag{7}$$

where $\widetilde{Y}_{ij}$ is the $(i,j)^{\text{th}}$ entry of $\widetilde{\mathbf{Y}}$. Using the sample variance, (TeSR) can be rewritten as

$$(\text{TeSR}) \quad \widehat{\mathcal{I}} = \underset{\mathcal{I}\subset\{1,\ldots,Q\},|\mathcal{I}|=q}{\arg\max} \log\det(\mathbf{W}_{\mathcal{I}}^T\widehat{\mathbf{V}}_{\mathcal{I}}\mathbf{W}_{\mathcal{I}})\,,$$

where $\widehat{\mathbf{V}}$ is a diagonal matrix with entries $\widehat{\mathbf{V}}_{kk} = \widehat{v}_k$. In the above formulation, there is no longer any dependence on $\mathbf{c}^*$.

Algorithm 1 summarizes a nonadaptive method for solving the TeSR problem. The first step is to find the "best" K questions, where K is the number of concepts in the Q questions. Next, we select the remaining questions $K+1,\ldots,q$ in an iterative manner. Note that selecting less than K questions would inhibit estimating the K-dimensional concept knowledge vector.

For any subset $\mathcal{I}$, let $\mathcal{I}_{[K]}$ denote the first K elements. To select the initial K questions $\widehat{\mathcal{I}}_{[K]}$, we use methods in [5] to formulate the combinatorial optimization problem in (5) as a convex optimization problem. More specifically, we can obtain an approximate solution to (5) by solving the following convex optimization problem:

$$\begin{aligned} &\text{maximize} && \log\det\left(\mathbf{W}_{\mathcal{I}}^T\widehat{\mathbf{V}}\mathbf{Z}\mathbf{W}_{\mathcal{I}}\right) \\ &\text{subject to} && \text{diagonal matrix } \mathbf{Z} \text{ with } Z_{kk} = z_k \\ &&& \textstyle\sum z_k = K \text{ and } 0 \le z_k \le 1 \end{aligned} \tag{8}$$

Once (8) has been computed, $\widehat{\mathcal{I}}_{[K]}$ can be approximated as the indices corresponding to the top K largest values of the diagonal elements $Z_{kk} = z_k$ of the matrix $\mathbf{Z}$.

The second step in Algorithm 1 chooses the remaining $q-K$ questions in a greedy manner. Using the identity

$$\det(\mathbf{X} + \mathbf{b}\mathbf{b}^T) = \det(\mathbf{X})(1 + \mathbf{b}^T\mathbf{X}^{-1}\mathbf{b}),$$

where $\mathbf{X}$ is a square matrix and $\mathbf{b}$ is a column vector, the quantity $\log\det(\mathbf{W}_{\mathcal{I}_{[j+1]}}^T\widehat{\mathbf{V}}_{\mathcal{I}_{[j+1]}}\mathbf{W}_{\mathcal{I}_{[j+1]}})$ can be rewritten as

$$\log\det(\mathbf{W}_{\mathcal{I}_{[j]}}^T\widehat{\mathbf{V}}_{\mathcal{I}_{[j]}}\mathbf{W}_{\mathcal{I}_{[j]}}) + \log(1+F) \tag{9}$$

with the definition

$$F = \widehat{\mathbf{V}}_{\mathcal{I}_{j+1},\mathcal{I}_{j+1}}\mathbf{w}_{\mathcal{I}_{j+1}}^T(\mathbf{W}_{\mathcal{I}_{[j]}}^T\widehat{\mathbf{V}}_{\mathcal{I}_{[j]}}\mathbf{W}_{\mathcal{I}_{[j]}})^{-1}\mathbf{w}_{\mathcal{I}_{j+1}}. \tag{10}$$

Thus, once j questions $\widehat{\mathcal{I}}_{[j]}$ have been selected, the next question, $\widehat{\mathcal{I}}_{j+1}$, can be selected so that the quantity F defined above is maximized.

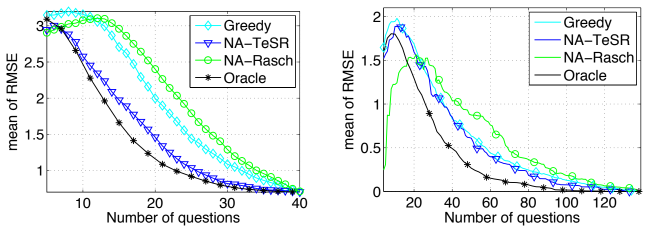
Remark 1: The computational complexity of Step 1 of Algorithm 1 is rather low when using the convex optimization relaxation approach outlined in (TeSR). We refer to [5] for iterative methods that solve (8). We note that although Step 2 requires computing an inverse of a $K\times K$ matrix multiple times, this inverse can be computed recursively once $(\mathbf{W}_{\mathcal{I}_{[K]}}^T\widehat{\mathbf{V}}\mathbf{W}_{\mathcal{I}_{[K]}})^{-1}$ has been computed. Finally, we can directly solve (TeSR) using the convex relaxation in (8). However, the computational complexity of this approach can be large, especially when q is large.

Remark 2: Note that when $\mathbf{W}$ is a $Q\times 1$ vector of all ones, the SPARFA model reduces to the Rasch model [9]. In this case, (TeSR) reduces to a problem of maximizing the sum of the variance terms over the selected questions. Thus, all the questions can be selected independently of the others when using the Rasch model. On the other hand, when using SPARFA, since we account for the statistical dependencies amongst questions, the questions can no longer be chosen independently as it is evident from Algorithm 1.

IV. Experimental Results

In this section, we assess the performance of our algorithms for test-size reduction (TeSR) using synthetic and real educational datasets.

Baseline algorithms: We compare NA-TeSR to three baseline algorithms. The first, referred to as NA-Rasch, uses the Rasch model [9] and selects questions in a non-adaptive manner



(a) Synthetic Data

(b) Real Data

Fig. 1. TeSR and baseline methods for synthetic data and real data.

(see Remark 2). The second, referred to as Greedy, iteratively selects a question from each concept until the required number of q questions has been selected. If all questions from a given concept have been exhausted, then Greedy skips to the next concept to select a question. Note that this approach completely ignores the variability of a learner in answering various questions. Finally, we also compare to an oracle algorithm, referred to as Oracle, that uses the true underlying (but in practice unknown) vector $\mathbf{c}^*$ to solve the TeSR problem. Note that the oracle algorithm is not practical and is only used to characterize the performance limits of TeSR.

Performance measure: We assess the performance of the algorithms using the root mean-square error (RMSE), defined as $\mathsf{RMSE} = \|\hat{\mathbf{c}} - \mathbf{c}^*\|_2$, where $\hat{\mathbf{c}}$ is the estimate delivered by each method and $\mathbf{c}^*$ is the ground truth. Although $\mathbf{c}^*$ is known for synthetic experiments, for real data, we assume that the ground truth is the concept vector estimated when asking all Q available questions.

Methods: In the experiments shown next, we assume that a matrix $\mathbf{Y}$ is given that contains graded responses of Q questions from M students. As mentioned in Section 2, for real data, we use SPARFA-M [6] to estimate $\mathbf{W}$ and the ground truth concept values of each learner. For each learner, we apply the baseline and our proposed TeSR algorithms using $\mathbf{W}$ and a training data $\widetilde{\mathbf{Y}}$ obtained after removing the responses of the learner from the matrix $\mathbf{Y}$. To show the performance of our TeSR algorithms, we report the mean and standard deviation of the RMSE evaluated over all M learners.

MLE convergence: It turns out that the maximum likelihood estimate (MLE) may not converge for certain patterns of the response vectors. In the case of inexistent ML estimates, we make use of the sign of the ML estimates to compute the RMSE. We then assign each entry in $\hat{\mathbf{c}}$ to the worst (for $-\infty$) or best (for $+\infty$) value obtained from a prior set of learners who have taken the course. In our simulations, these worst and best concept values are computed using the training data $\widetilde{\mathbf{Y}}$.

Results: We generated a sparse 50×5 matrix $\mathbf{W}$ that maps 50 questions to 5 concepts. There were roughly 30% non-zero entries in $\mathbf{W}$ with the non-zero entries chosen from an exponential random variable with parameter $\lambda = 2/3$. Each entry in the intrinsic difficulty vector $\boldsymbol{\mu}$ was generated from a standard normal distribution. We assumed 25 learners whose concept understanding vectors were again generated from a standard normal distribution. For each $\mathbf{Y}$, we computed the reduced test-size with $q = 5, 6, \ldots, 44$. Fig. 1 shows the mean value of the RMSE over 100 randomly generated response vectors $\mathbf{Y}$. Note that the mean RMSE is taken over all 25 learners. We observe that NA-TeSR is superior to all practical baseline algorithms. This observation suggests that the Rasch model is not an appropriate model for selecting questions for the purpose of test-size reduction in courses having more than one underlying concept.

Fig. 1(b) shows results on real educational dataset corresponding to graded response data obtained from the ASSISTment system [8]. The original data contained responses from 4354 learners on 240 questions. There is a large number of missing responses in this dataset, i.e., not every learner answered all problems. In order to get a dataset with a sufficient number of observed entries, we focused on a subset of 219 questions answered by 403 learners. The resulting trimmed $\mathbf{Y}$ matrix has roughly 75% missing values. Fig. **??**(b) shows the associated results and we observe similar trends as for synthetic data set. The main difference is that the performance of the Greedy algorithm is almost as good as the NA-TeSR algorithm in certain domains. This may be a result of the several missing values present in the dataset that does not allow for accurate computations of the variability in answering each question. We note that we have extended the NA-TeSR algorithm in [10] to an adaptive algorithm where each question selected by the greedy step in NA-TeSR uses prior responses to form an estimate of $\widehat{\mathbf{c}}$. This method leads to results that are closer to the Oracle algorithm. We refer to [10] for more details.

References

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

[2] S. Buyske. *Applied optimal designs*, chapter Optimal design in educational testing, pages 1–16. John Wiley & Sons Inc, 2005.

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

[4] L. Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1):342–368, 1985.

[5] S. Joshi and S. Boyd. Sensor selection via convex optimization. *IEEE Transactions on Signal Processing*, 57(2):451–462, 2009.

[6] 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*, Nov. 2012, submitted.

[7] W. J. Linden and P. J. Pashley. *Elements of adaptive testing*, chapter Item selection and ability estimation in adaptive testing, pages 3–30. Springer, 2010.

[8] Z. Pardos and N. Heffernan. Modeling individualization in a Bayesian networks implementation of knowledge tracing. *User Modeling, Adaptation, and Personalization*, pages 255–266, 2010.

[9] G. Rasch. *Probabilistic Models for Some Intelligence and Attainnment Tests*. Studies in mathematical psychology. Danmarks paedagogiske Institut, 1960.

[10] D. Vats, C. Studer, A. Lan, L. Carin, and R. Baraniuk. Test-size reduction for concept estimation. In *International Conference on Educational Data Mining (EDM)*, 2013.