Test-size Reduction for Concept Estimation

Divyanshu Vats[1], Christoph Studer[1], Andrew S. Lan[1],
Lawrence Carin[2], Richard G. Baraniuk[1]
[1]Rice University, TX, USA; [2]Duke University, NC, USA
{dvats, studer, mr.lan, larry, richb}@sparfa.com

ABSTRACT

Consider a large database of questions that assess the knowledge of learners on a range of different concepts. In this paper, we study the problem of maximizing the estimation accuracy of each learner's knowledge about a concept while minimizing the number of questions each learner must answer. We refer to this problem as test-size reduction (TeSR). Using the SPARse Factor Analysis (SPARFA) framework, we propose two novel TeSR algorithms. The first algorithm is nonadaptive and uses graded responses from a prior set of learners. This algorithm is appropriate when the instructor has access to only the learners' responses after all questions have been solved. The second algorithm adaptively selects the "next best question" for each learner based on their graded responses to date. We demonstrate the efficacy of our TeSR methods using synthetic and educational data.

Keywords

Learning analytics, sparse factor analysis, maximum likelihood estimation, adaptive and non-adaptive testing

1. INTRODUCTION

A course instructor is naturally interested in estimating how well learners understand certain concepts (or topics) that are relevant to the course. Information about each learner's understanding is useful in (i) providing feedback to instructors to assess whether the material is suitable for the class and (ii) recommending remediation/enrichment for concepts a learner has weak/strong knowledge of. In practice, accurate estimates for each learner's concept knowledge can be extracted automatically by analyzing the responses to a (typically large) set of questions about the concepts underlying the given course (see, e.g., [8] for the details). In order to minimize each learner's workload, however, it is important to reduce the number of questions, or—more colloquially—the *test-size*, while still being able to retrieve accurate estimates of each learner's concept knowledge. In what follows, we refer to this problem as *test-size reduction* (TeSR).

Contributions: We propose two novel algorithms for test-size reduction (TeSR). Our algorithms build on the *SPARse Factor Analysis* (SPARFA) framework proposed in [8], which jointly estimates the question–concept relationships, question intrinsic difficulties, and the latent concept knowledge of each learner, based solely on binary-valued graded response data obtained in an homework, test, or exam. Given the SPARFA model, we leverage theory of maximum likelihood (ML) estimators to formulate TeSR as a combinatorial optimization problem of minimizing the uncertainty in estimating the concept knowledge of each learner. We then propose two algorithms, one nonadaptive and one adaptive, that approximates the combinatorial optimization problem at low computational complexity using a combination of convex optimization and greedy iterations. The nonadaptive TeSR algorithm, referred to as NA-TeSR, reduces the test size in a way that enables accurate concept estimates *for all learners* in a course. The adaptive TeSR algorithm, referred to as A-TeSR, adapts the test questions to *each individual learner*, based on their previous responses to questions. A range of experiments with synthetic data and two real educational datasets demonstrates the efficacy of both TeSR algorithms.

Prior Work: Prior results on test-size reduction build primarily on the Rasch model [1–3, 6, 9], which characterizes a learner using a single ability parameter [11]. In contrast, the SPARFA model used in this paper characterizes a learner using their concept knowledge on multiple latent concepts. In this way, SPARFA models educational scenarios of courses consisting of multiple concepts more accurately. Moreover, we show using experiments in Section 4 that the efficacy of the Rasch model for TeSR is inferior to SPARFA combined with our TeSR algorithms. The problem of selecting "good" questions is related to the sensor selection problem [5, 7], which finds use in environmental monitoring, for example. However, measurements from sensor-networks are typically real-valued, whereas, TeSR relies on discrete measurements.

2. PROBLEM FORMULATION

2.1 SPARFA in a nutshell

Suppose we have a total set of Q questions that test knowledge from K concepts. For example, in a high school mathematics course, questions can test knowledge from concepts like solving quadratic equations, evaluating trigonometric identities, or plotting functions on a graph. For each question $i = 1, \ldots, Q$, let $\mathbf{w}_i \in \mathbb{R}^K$ be a column vector that represents the association of question i to all K concepts. Note that each question can measure knowledge from multiple concepts[1]. 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,

[1] Solving $x^2 - x = \sin^2(x) + \cos^2(x)$ for $x \in \mathbb{R}$, for example, requires conceptual understanding of both solving quadratic equations as well as trigonometric identities.

non-negative $Q \times K$ matrix, assuming 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 each question. Finally, let $\mathbf{c}^* \in \mathbb{R}^K$ be a column vector that represents the concept knowledge of a particular learner. *It is this parameter vector that we are interested in estimating accurately.*

To model the interplay between $\mathbf{W}$, $\boldsymbol{\mu}$, and $\mathbf{c}^*$, we use the SPARFA framework proposed in [8]. 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, the SPARFA model assumes that $Y_i \in \{0, 1\}$ admits the following distribution:

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

where $\Phi(x) = 1/(1 + e^{-x})$ is the inverse logistic link function. In words, (1) says that the probability of answering a question correctly depends on a sparse linear combination of the entries in the concept understanding vector $\mathbf{c}^*$. This sparsity arises because of the assumption that $\mathbf{w}_i$ is sparse, i.e., it only contains a few non-zero entries. Given graded question responses from multiple learners, the factors $\mathbf{W}$ and $\boldsymbol{\mu}$ can be estimated using either the SPARFA-M or SPARFA-B algorithms introduced in [8].

2.2 Test-size reduction (TeSR)

The problem we consider in this paper is the selection of an appropriate subset of $q < Q$ questions so that $\mathbf{c}^*$, a learner's unknown concept understanding vector, can be estimated accurately. We assume that a set of responses from N learners, i.e., a binary-valued matrix $\widetilde{\mathbf{Y}}$, is known *a-priori*; an entry $\widetilde{Y}_{i,j}$ of $\widetilde{\mathbf{Y}}$ refers to whether a learner j answered question i correctly or incorrectly. In many educational settings, such a data matrix can be obtained by looking at past offerings of the same course. As mentioned in Section 2.1, the matrix $\widetilde{\mathbf{Y}}$ can be used to estimate the question to concept matrix $\mathbf{W}$ and the intrinsic difficulty vector $\boldsymbol{\mu}$ using the algorithms proposed in [8].

Suppose, hypothetically, that we choose a subset $\mathcal{I}$ of $q < Q$ questions, and we are given a response vector $\mathbf{y}_{\mathcal{I}}$. Let $\widehat{\mathbf{c}}$ be an estimate of the unknown concept knowledge vector $\mathbf{c}^*$ that can be computed using standard maximum likelihood (ML) estimators. The *test-size reduction* (TeSR) problem is to choose an appropriate set of questions $\mathcal{I}$ so that the error $\widehat{\mathbf{c}} - \mathbf{c}^*$ is as small possible. Although this problem seems impossible since we do not have access to the response vector $\mathbf{y}_{\mathcal{I}}$, it turns out that the covariance of the error $\sqrt{q}(\widehat{\mathbf{c}} - \mathbf{c}^*)$ can be approximated by the inverse of the Fisher information matrix [4], which is defined as follows:

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

The notation $\mathbf{W}_{\mathcal{I}}$ refers to the rows of $\mathbf{W}$ indexed by $\mathcal{I}$. Similarly, $\boldsymbol{\mu}_{\mathcal{I}}$ refers to the entries in $\boldsymbol{\mu}$ indexed by $\mathcal{I}$. Thus, a natural strategy for choosing a "good" subset of questions $\mathcal{I}$, is to minimize the uncertainty (formally, the differential entropy) of a multivariate normal random vector with mean zero and covariance $\mathbf{F}(\mathbf{W}_{\mathcal{I}}, \boldsymbol{\mu}_{\mathcal{I}}, \mathbf{c}^*))^{-1}$. Consequently, the optimization problem considered in the remainder of the paper, referred to as the *test-size reduction* (TeSR) problem, corresponds to

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

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

Step 1) First choose K questions by solving

$$\widehat{\mathcal{I}}_{[K]} = \underset{\mathcal{I} \subset \{1,\ldots,Q\}, |\mathcal{I}| = K}{\arg\max} \; \log\det\left(\mathbf{W}_{\mathcal{I}}^T \widehat{\mathbf{V}} \mathbf{W}_{\mathcal{I}}\right) \qquad (3)$$

using the convex optimization, see [7]. The entries of the diagonal matrix $\overline{\mathbf{V}}$ are defined as $\widehat{V}_{kk} = \exp(\widehat{v}_k)$, where $\widehat{v}_i = \frac{1}{N}\sum_{j=1}^{N} \log\left(\widetilde{Y}_{ij} - \frac{1}{N}\sum_{j=1}^{N} \widetilde{Y}_{ij}\right)^2$

Step 2) Select questions $K + 1, \ldots, q$ in a greedy manner:

$$\widehat{\mathcal{I}}_{j+1} = \underset{i \in \{1\ldots,Q\} \setminus \widehat{\mathcal{I}}_{[j]}}{\arg\max} \; \widehat{v}_i \mathbf{w}_i^T \left(\mathbf{W}_{\mathcal{I}_{[j]}}^T \widehat{\mathbf{V}}_{\mathcal{I}_{[j]}} \mathbf{W}_{\mathcal{I}_{[j]}}\right)^{-1} \mathbf{w}_i.$$

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.

3. TESR ALGORITHMS

Our proposed algorithms, that are data driven and computationally efficient, for solving TeSR are summarized in Algorithms 1 and 2. Due to space constraints, in what follows, we only present a high level summary of the methods.

Nonadaptive TeSR: Algorithm 1 summarizes a nonadaptive method (NA-TeSR) for solving the TeSR problem. To deal with the problem of the unknown $\mathbf{c}^*$ in (2), we notice that the coefficient of the term $\mathbf{w}_i \mathbf{w}_i^T$ in (2) is simply the variance of a learner in answering a question i. This variance can easily be estimated using the prior student response data $\widetilde{\mathbf{Y}}$. The first step in NA-TeSR is to estimate K questions, where K is the number of concepts involved in the question database. We are able to make use of properties of the determinant to formulate TeSR as a convex optimization problem, which we solve using low complexity methods in [7]. The second step is to select the remaining $q - K$ questions using a greedy algorithm that selects the "best" question iteratively until all q questions have been selected.

Remark 1: Note that when $\mathbf{W}$ is a $Q \times 1$ vector of all ones, the SPARFA model reduces to the Rasch model [11]. 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 among questions, the questions can no longer be chosen independently as it is evident from Algorithm 1.

Adaptive TeSR: Our second algorithm, A-TeSR, is designed for the situation where one can iteratively and individually ask questions to a learner and then use the responses to *adaptively* select the next "best" question based on the previous responses. Such an approach is often referred to as *computerized adaptive testing* [12].

Algorithm 2: Adaptive test-size reduction (A-TeSR)

```
Choose $K$ questions $\mathcal{I}_{[K]}$ as in Step 1 of Algorithm 1.
Acquire graded learner responses $\mathbf{y}_{\mathcal{I}_{[K]}}$.
for $j = K+1, \ldots, q$ do
    Compute the ML estimate $\widehat{\mathbf{c}}$ using $\mathbf{y}_{\mathcal{I}_{[j-1]}}$
    if $\widehat{\mathbf{c}}$ exists then
        Find $\mathcal{I}_j$ using Step 2 of Algorithm 2 by replacing $\widehat{v}_k$
        with $\mathbb{V}\mathrm{ar}[Y_i|\widehat{\mathbf{c}}]$.
    else
        Find $\mathcal{I}_j$ using Step 2 of Algorithm 2 by searching
        only amongst questions so that $\widehat{\mathbf{c}}$ will most likely
        exist in subsequent iterations.
    Acquire graded learner responses $\mathbf{y}_{\mathcal{I}_j}$.
```

The main idea behind A-TeSR is to use NA-TeSR until a maximum likelihood estimate (MLE) $\widehat{\mathbf{c}}$ of $\mathbf{c}^*$ can be computed. Then, we use $\widehat{\mathbf{c}}$ to evaluate the objective function of the TeSR problem and keep updating $\widehat{\mathbf{c}}$ as the learner responds to adaptively chosen questions. The main challenge of such an adaptive algorithm is the fact that a solution may not exist for certain patterns of the graded response of a given learner when computing the MLE. Thus, we would like our proposed adaptive algorithm to select questions such that the MLE can be computed using less number of questions than the nonadaptive algorithm. To this end, whenever the MLE does not exist, we choose the next question (using a simple modification of Step 2 of NA-TeSR) in such a way that the MLE may exist with higher probability in each subsequent iteration.

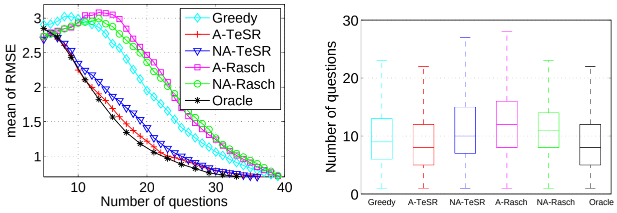
Remark 2: Just as in the case of NA-TeSR, A-TeSR reduces to an adaptive Rasch model-based method when $\mathbf{W}$ is a $Q \times 1$ vector of ones; see [3] for examples of such algorithms. The main differences when using the SPARFA model for selecting questions, as opposed to using the Rasch model, are that the condition for the MLE to exist changes and in each iteration we estimate a multidimensional concept vector as opposed a scalar parameter.

4. EXPERIMENTAL RESULTS

Baseline algorithms: We compare NA-TeSR and A-TeSR to four baseline algorithms.

- NA-Rasch and A-Rasch: Nonadaptive and adaptive methods that use the Rasch model to select questions. See Remark 1 and 2 for more details.
- 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 intrinsic difficulty of a question when performing TeSR.
- Oracle: 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 $\mathsf{RMSE} = \|\widehat{\mathbf{c}} - \mathbf{c}^*\|_2$. 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.



(a) (b)

Figure 1: TeSR methods for synthetic data.

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 [8] to estimate $\mathbf{W}$, $\boldsymbol{\mu}$, 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 of the RMSE evaluated over all M learners.

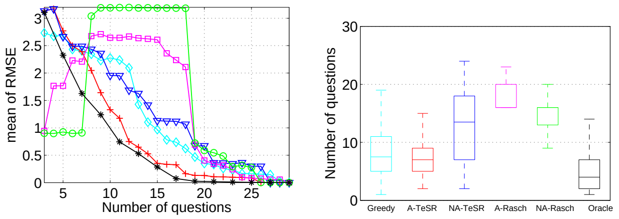
MLE convergence: As mentioned in Section 3, the MLE may not exist for certain patterns of the response vectors. In the case of inexistent ML estimates, we make use of the sign of the ML estimates (since each value in $\widehat{\mathbf{c}}$ will either be ∞ or $-\infty$) to compute the RMSE. We then assign each entry in $\widehat{\mathbf{c}}$ 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 $\widetilde{\mathbf{Y}}$.

Synthetic Data: 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 $q = 5, 6, \ldots, 44$.

Figure 1(a) 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 and A-TeSR are superior to the baseline algorithms A-Rasch, NA-Rasch, and Greedy. 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.

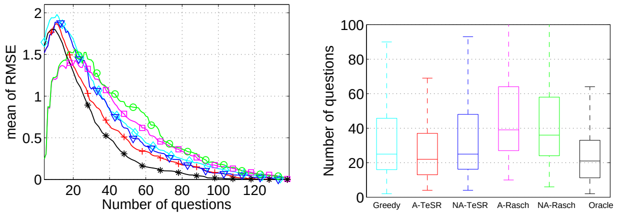
Algebra test dataset: The first dataset was obtained by a high school algebra test administered on Amazon's Mechanical Turk (see [8] for more details). This dataset contains no missing data and consists of responses from 99 learners on 34 questions.

We used SPARFA-M assuming that there are $K = 3$ latent concepts. The estimated concept–question matrix $\mathbf{W}$ contains roughly 40% non-zero values. Figure 2(a) shows the



(a) (b)

Figure 2: Mechanical Turk algebra test with 3 concepts; see Figure 1(a) for the legend.



(a) (b)

Figure 3: AssisMENT system data with 4 concepts; see Figure 1(a) for the legend.

mean RMSE over 78 learners[2]. We see similar trends as in the synthetic experiments. The main difference is that the performance of Rasch-based algorithms is much worse when compared to the synthetic data case. As we will explain later, this behavior is mainly because the ML estimates for the Rasch model did not converge for $q < 17$. Interestingly, for $q < 7$, the mean RMSE of NA-TeSR is much lower when compared to other algorithms. This behavior can be addressed to the fact that we deal with convergence failures of the ML estimates. Furthermore, we note that the mean performance of Greedy is better, in some regimes, than NA-TeSR. However, this gain in performance, for some questions, comes at the cost of slightly higher variability in the estimation of concept knowledge.

ASSISTment system dataset: The second real educational dataset corresponds to response data obtained from the ASSISTment system that was studied in [10]. The original data contained responses from 4354 learners on 240 questions. There are a large number of missing responses in this dataset. 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. Figures 3(a) shows the associated results and we observe trends that are similar to the algebra test dataset.

How many questions are needed? Another interesting measure to evaluate the performance of the TeSR algorithms is the number of questions needed for the ML to converge. Intuitively, this measure signifies the number of questions needed to get accurate estimates of each learner's concept knowledge. Figures 1(b)–3(b) show box plots of the number of questions needed for the ML estimates to converge for each algorithm and for each dataset considered here. Each box corresponds to the 25th and 75th percentiles over all learners in a class. We see that the A-TeSR algorithm is the fastest to converge amongst all the practical algorithms (the oracle algorithm is not practical since it utilizes information about the unknown concept vector of interest $\mathbf{c}^*$).

Acknowledgements

Thanks to Andrew Waters for inspiring discussions and Daniel Calderon for administering the Algebra Test on Amazon's Mechanical Turk. DV thanks the Institute for Mathematics and its Applications (IMA) for financial 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 the Swiss National Science Foundation under grant PA00P2-134155.

5. REFERENCES

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

[2] H. Chang and Z. Ying. A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20(3):213–229, 1996.

[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] D. Golovin, M. Faulkner, and A. Krause. Online distributed sensor selection. In *Proc. ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, 2010.

[6] U. Graßhoff, H. Holling, and R. Schwabe. Optimal designs for the Rasch model. *Psychometrika*, pages 1–14.

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

[8] 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.

[9] 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.

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

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

[12] W. J. van der Linden and C. A. W. Glas. *Computerized adaptive testing: Theory and practice*. Springer, 2000.

[2] For some of the questions, the ML estimate did not exist when using all the 34 questions; hence, the ground truth could not be computed.