The refinement of a Q-matrix: Assessing methods to validate tasks to skills mapping

Michel C. Desmarais
Polytechnique Montreal
michel.desmarais@polymtl.ca

Behzad Beheshti
Polytechnique Montreal
behzad.beheshti@polymtl.ca

Peng Xu
Polytechnique Montreal
peng.xu@polymtl.ca

ABSTRACT

The objective of specifying which skills are required in a given task is fundamental for the accurate assessment of a student's knowledge and for personalizing tutor interaction towards more relevant and effective assessment and learning. We compare three data driven techniques for the validation of skills-to-tasks mappings. All methods start from a given mapping, typically obtained from domain experts, and use optimization techniques to suggest a refined version of the skills-to-task mapping. To validate the different techniques, we inject perturbations in the Q-matrix and verify whether the original Q-matrix can be recovered. Tests are run over both simulated and real data. The analysis of the Q-matrix refinements of each technique over ten data sets shows that, in general, around 1/2 to 2/3 of the perturbations can be restored to their original values, but a number of potentially wrong perturbations are also introduced. The number of correctly restored and falsely switched values vary across the three techniques and between synthetic and real data. For 1 to 10 perturbations injected, simulated data recovery rate is around 2/3, and invalid alterations introduced vary around 2 to 3. For real data, the two best techniques generally recover about half the perturbations injected, but introduce between 5 and 7 alterations inconsistent with the original, expert defined Q-matrix, although some of them may be real improvements.

Keywords

Student model, Skills modeling, Psychometrics, Q-matrix, Matrix Factorization, Alternate Least-Squares, DINA

1. INTRODUCTION

Detailed assessment of skills rely on a fine grained mapping of tasks to skills. Student success and failures over these tasks provide evidence of which skills are mastered. Many intelligent tutors use such information to tailor their behaviour (for eg. [9]).

However, defining the mapping of tasks to skills is non trivial and error prone. The validation of such mapping from student test results has been the focus of recent developments in the field of psychometrics and educational data mining in recent years [3, 1, 11, 2, 6]. The ever growing abundance of student assessment traces from e-learning environments further enhances our capacity to validate expert defined mappings through data mining techniques.

This paper compares three families of techniques to refine a given mapping of skills to tasks, which we will refer to as *items*. All methods compared start with a given skill to item mapping, and typically suggest a few changes. We define a methodology to validate whether the proposed changes are appropriate. This validation rests on a number of experiments with artificial and real data to compare the quality of the changes recommended by each technique. The background work of item to skills mapping is first reviewed, followed by the description of the methodology and results of the experiments.

2. Q-MATRICES, THEIR INTERPRETATION AND VALIDATION

A mapping of item to skills is termed a Q-matrix [14]. If all specified skills are required to succeed the item, the Q-matrix is labelled **conjunctive**. If a any of the required skill is sufficient to the item's success, then it is labelled **disjunctive**. The conjunctive/disjunctive distinction is also referred to as AND/OR gates. A well known model, DINA for "Deterministic Input Noisy AND", corresponds to the conjunctive version. A variant of DINA, the DINO model (Deterministic Input Noisy OR) corresponds to a disjunctive Q-matrix [8].

Two techniques for Q-matrix validation surveyed here rely on the DINA model. A third one relies on a matrix factorization technique called ALS (Alternative Least Squares). We refer to them as (1) de la Torre (2008), (2) Chiu (2013), and (3) ALS:

(1) **de la Torre (2008)**. The method defined by de la Torre [3] searches for a Q-matrix that maximizes the difference in the probabilities of a correct response to an item between examinees who possess all the skills required for a correct response to that item and examinees who do not.

(2) **Chiu (2013)**. Chiu defines a method that minimizes the residual sum of square (RSS) between the real responses and the ideal responses that follow from a given Q-matrix [2]. The algorithm adjusts the Q-matrix by choosing the item with the worst RSS over to the data, and replaces it with the one has the lowest RSS, and iterates until convergence.

(3) **Alternate Least-Square Factorization (ALS)**. The (ALS) method is defined in [6]. Contrary to the other two methods, it does not rely on the DINA model. Instead, it decomposes the results matrix $\mathbf{R}_{m\times n}$ of m items by n students as the inner product two smaller matrices: $\mathbf{R} = \mathbf{Q}\,\mathbf{S}$,

where **R** is the results matrix, **Q** is the m items by k skills Q-matrix, and **S** is the mastery matrix of k skills by n students. The factorization consists of alternating between estimates of **S** and **Q** until convergence.

3. METHODOLOGY AND DATA SETS

The two first methods, de la Torre (2008) [3] and Chiu (2013) [2], have been shown to perform well on artificial data. On real data, their performance is more blurry. The ALS factorization method [6] has only been tested on one real data set. But the methodologies used to validate all three techniques in each respective study vary considerably and do not allow for a proper comparison.

To validate and compare the effectiveness of each technique for refining a given Q-matrix, we follow a methodology based on recovering the Q-matrix from a number perturbations: the binary value of a number of cells of the Q-matrix is inverted, and this "corrupted" matrix is given as input to each technique. If the technique recovers the original value of each altered cell, then we consider that it successfully "refined" the Q-matrix. This approach is similar to the studies mentioned [3, 2, 6].

Ten levels of perturbations are defined, from 1 to 10. For each level, we conduct up to 30 experiments that consists in choosing Q-matrix cells to be altered. If the Q-matrix contains 30 or less cells, all of them are altered in turn. If it is larger, combinations of cells are chosen at random. We refer to this procedure as a single perturbation run. The runs are repeated for each of the 10 levels of perturbation, and over the different data sets.

The measures of performance are the number of *true positives* and *false positives*:

- **Mean true positives**: a *true positive* corresponds to an alteration that was injected in the input, and was correctly switched back to its original value by the method. The measure reported is the number of correctly recovered alterations averaged over the 8 runs and by level of perturbation.
- **Mean false positive ratio**: a *false positive* corresponds to a changed Q-matrix entry returned by the method, but that was not injected in the input. Hereto, averages by perturbation runs are reported.

For real data, the definition of true/false positives rests on the assumption that the original matrix is better than the corrupted one, which is not necessarily the case with an expert generated Q-matrix. The expert may be wrong. However, we have no other means to inform us of the "real" Q-matrix and it is reasonable to assume that most of the cells in the Q-matrix are correct. Of course, for synthetic data, this assumption is correct as the Q-matrix is at the source of the generation of the data.

A total of 10 data sets are used for the validation. They are freely available from two R packages: CDM (`http://cran.r-project.org/web/packages/CDM/index.html`) [12] and Chiu (2013) (`http://cran.r-project.org/web/packages/NPCD/NPCD.pdf`). Table 1 contains a short description of each data set. Note that for the last five data sets, the source data is the same, but different Q-matrices are defined over them and subsets of items are used in the last four: the fraction data set data is used to create four variations through subsets of questions and alternative Q-matrices (Fraction 1, Fraction 2/1, Fraction 2/2, and Fraction 2/3). The artificial data sets are generated from the well known DINA and DINO models.

For obtaining the results of the de la Torre (2008) method, we used the R implementation found in the CDM package [12]. A DINA model and parameter estimation is first built with the default arguments to the `din` function, and fed to the `din.validate.qmatrix` function to obtain a refined version of the Q-matrix. For the results of the Chiu (2013) method, the R NPCD packaged is used (function `Qrefine`).

4. RESULTS

The three methods are evaluated over the 10 data sets and for 8 runs. Each run is conducted over a set of 30 different random combinations of perturbations, from 1 up to 10 perturbations. For the 1-perturbation condition, the total number of possible combinations is the size of the Q-matrix itself.

Figures 1 and 2 show the results broken down by real and synthetic data sets respectively, as space does not allow to report individual data set results.

Performance of each method is reported as a function of the number of perturbations. Recoveries are labelled "True Positives" (TP) whereas changes introduced by a method, but which do not correspond to perturbations introduced, are labelled "False Positives" (FP). The two graphs of figure 1 show the averages of the 6 real data sets, whereas the graphs of 2 show the averages for the 4 synthetic data sets. The "Total" line is shown to visually indicate the maximum that can be reached by a TP curve.

The ALS method shows the greatest ability to recover alterations, but at the cost of a higher rate of FP: changes that do not correspond to perturbations. It is followed closely by the Chiu (2013) method. The de la Torre (2008) method has a very low rate of recovery (TP) that makes it impractical. In general, the ALS and Chiu (2013) methods recover about 2/3 of the perturbations for synthetic data, and this rate falls to 1/2 for real data with ALS, and about 1/3 for Chiu (2013). For real data, the number of FP is around 5 for Chiu (2013) and around 6 for ALS, whereas it is respectively 2 and 3 for synthetic data. The relative performance of Chiu (2013) with respect to ALS is better for synthetic data and this might be explained by the fact that the data generation process is directly based on the DINA model.

A common pattern across methods is the relatively stable number of FP as a function of the number of perturbations. ALS does show an increase of close to 1 FP between 1 to 10 perturbations, whereas the increase for the Chiu (2013) and de la Torre (2008) methods is closer to 1/2 for real data, and even less for synthetic data (in fact it decreases for de la Torre (2008)). As a result, the rate of TP over FP increases with the number of perturbations.

Table 1: Data sets

Name	Number of			Description
	Skills	Items	Cases	
Sim. DINA	3	9	400	Artificial data available from the (`sim.dina`) data set of the CDM package.
Sim. DINO	3	9	400	Same parameters as No. 1 but using the DINO model (`sim.dino` data set).
Sim. CDM DINA	3	12	4000	Artificial data generated through the CDM function `sim.din`.
Sim. DCM	3	7	10000	Artificial data from chapter 9 of the book *Diagnostic Measurement* [13]
ECPE	3	28	2922	Dataset from [15] in [4]
Fraction	8	20	536	Tatsuoka's fraction algebra problems [14] (see table 1 in [5] for a description of the problems and of the skills).
Fraction 1	5	15	536	15 questions subset of Fraction with Q-matrix defined in [4].
Fraction 2/1	3	11	536	11 questions subset of Fraction with Q-matrix from [7].
Fraction 2/2	5	11	536	11 questions subset of Fraction with Q-matrix from [4].
Fraction 2/3	3	11	536	3 skills version of Fraction 1.

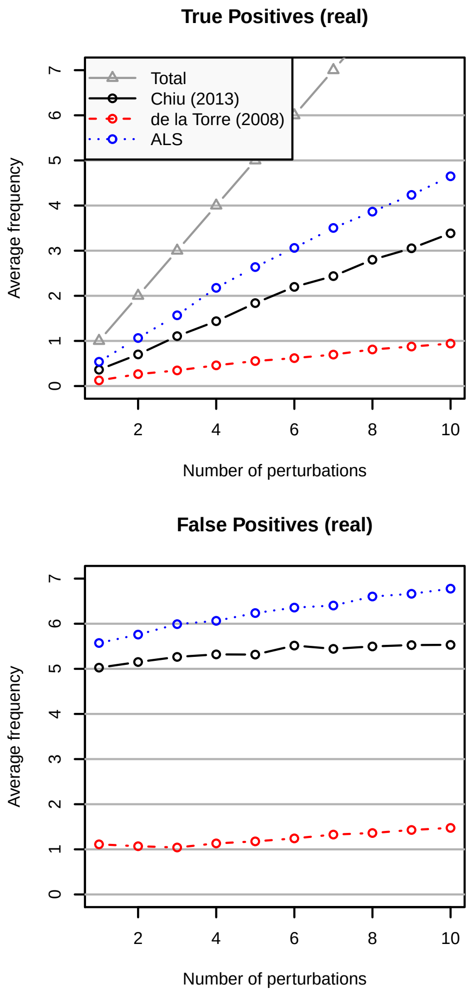

Figure 1: Average recovery rate by number of perturbations (real data)

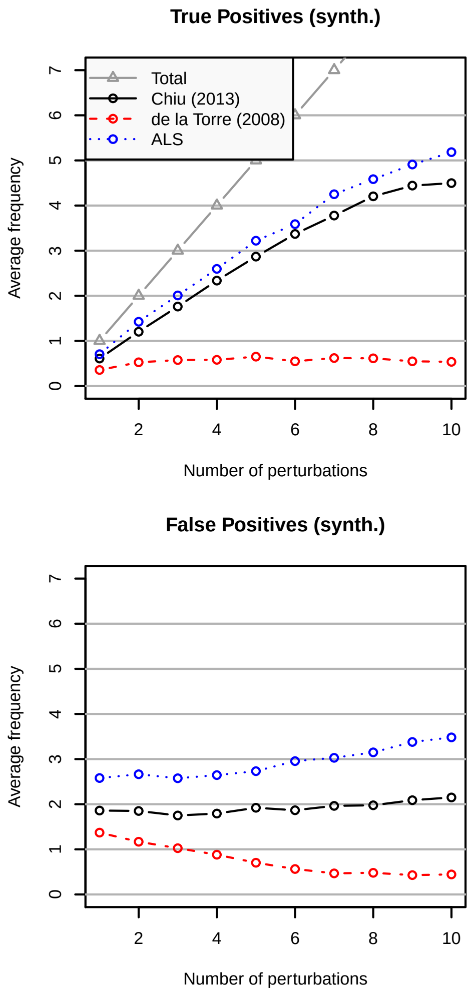

Figure 2: Average recovery rate by number of perturbations (synthetic data)

5. DISCUSSION

The contribution of this work is to provide performance assessments and compare existing methods of Q-matrix refinements based on a methodology and on metrics that allow meaningful comparisons. Previous work was limited to showing their ability to make Q-matrix refinements on an individual basis and in a restricted context.

The experiments conducted confirm that two methods, ALS and Chiu (2013) can recover the original Q-matrix from an altered one, as shown in previous work [2, 6], but the performance of the de la Torre (2008) method is considerably lower than the other two. The comparison of their performance over a number of data sets, and based on a common measure of performance, reveals wide differences across data sets (not reported here). As expected, all methods fare better on synthetic data sets, for which a close to perfect performance is reached with large samples. For real and synthetic data sets alike, the ALS and Chiu (2013) methods overall performances are comparable, but the advantage is spread between the two across the different data sets.

Can the methods be useful for refining Q-matrices in practice? Some issues clearly arise in the results. One issue is the size of the data sets required. For example, the Sim. DINA set has 400 cases and yet the best method only finds a single perturbation 1 time over 2. This result suggest that small samples of 100 cases or less are likely to be too small for being useful. In the days of big data from web deployment, for example, this is not such a major issue, but it does rule out some context of validation of a Q-matrix.

Another potential issue is that the results generally show more False Negatives than False Positives with real data. Note that, for real data, we cannot assume that all False Positives are wrong corrections. Some of them may represent potential improvements. Empirical evidence will be required to verify whether the suggested corrections do lead to real improvements when experts are presented with these corrections. Further work will also be required to validate if we can use the cross-evidence to filter out weak suggestions. For example, the recurrence of the same False Positives across perturbations and across techniques may yield stronger support to a suggestion.

Future work should also extend the comparison to more techniques such as [10, 11]. Finally, we stress the need for open access to the data and the code used in such studies. This particular study was highly facilitated by the CDM [12] and NPCD packages which provided both the code and the data.

Acknowledgements: This research was funded under the NSERC Discovery grant of the first author.

6. REFERENCES

[1] T. Barnes. Novel derivation and application of skill matrices: The Q-matrix method. *Handbook on Educational Data Mining*, 2010.

[2] C.-Y. Chiu. Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, 2013.

[3] J. De La Torre. An empirically based method of Q-Matrix validation for the DINA model: Development and applications. *Journal of educational measurement*, 45(4):343–362, 2008.

[4] J. de la Torre. DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34(1):115–130, 2009.

[5] L. T. DeCarlo. On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-Matrix. *Applied Psychological Measurement*, 35:8–26, 2011.

[6] M. C. Desmarais and R. Naceur. A matrix factorization method for mapping items to skills and for enhancing expert-based Q-Matrices. In *6th International Conference, AIED 2013, Memphis, TN, USA*, pages 441–450, 2013.

[7] R. A. Henson, J. L. Templin, and J. T. Willse. Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74(2):191–210, 2009.

[8] B. Junker and K. Sijtsma. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3):258–272, 2001.

[9] K. R. Koedinger, J. R. Anderson, W. H. Hadley, and M. A. Mark. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8:30–43, 1997.

[10] J. Liu, G. Xu, and Z. Ying. Data-driven learning of Q-Matrix. *Applied Psychological Measurement*, 36(7):548–564, 2012.

[11] N. Loye, F. Caron, J. Pineault, M. Tessier-Baillargeaon, C. Burney-Vincent, and M. Gagnon. La validité du diagnostic issu d'un mariage entre didactique et mesure sur un test existant. *Des mécanismes pour assurer la validité de l'interprétation de la mesure en éducation*, 2:11–30, 2011.

[12] A. Robitzsch, T. Kiefer, A. George, A. Uenlue, and M. Robitzsch. Package CDM. 2012.

[13] A. A. Rupp, J. Templin, and R. A. Henson. *Diagnostic measurement: Theory, methods, and applications*. Guilford Press, 2010.

[14] K. Tatsuoka, U. of Illinois at Urbana-Champaign. Computer-based Education Research Laboratory, and N. I. of Education (US). *Analysis of errors in fraction addition and subtraction problems*. Computer-based Education Research Laboratory, University of Illinois, 1984.

[15] J. Templin and L. Hoffman. Obtaining diagnostic classification model estimates using mplus. *Educational Measurement: Issues and Practice*, 32(2):37–50, 2013.