Goodness of fit of skills assessment approaches: Insights from patterns of real vs. synthetic data sets

Behzad Beheshti
Polytechnique Montreal
behzad.beheshti@polymtl.ca

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

ABSTRACT

This study investigates the issue of the goodness of fit of different skills assessment models using both synthetic and real data. Synthetic data is generated from the different skills assessment models. The results show wide differences of performances between the skills assessment models over synthetic data sets. The set of relative performances for the different models create a kind of "signature" for each specific data. We conjecture that if this signature is unique, it is a good indicator that the corresponding model is a good fit to the data.

1. INTRODUCTION

There exists a large array of models to represent and assess student skills. Item Response Theory (IRT) is probably the most established method. It dates back to the 1960's and is still one of the prevailing approaches (see [1]). But many other methods have been introduced in recent years. Among them is the family of models that rely on slip and guess factors [12, 11], such as the DINA (Deterministic Input Noisy And-Gate), DINO (Deterministic Input Noisy Or-Gate), and other variants (see [7]). Other approaches are based on the Knowledge Space theory of Doignon and Falmagne [10, 8], which does not directly attempt to model underlying skills but instead rely on observable items only. Finally, recent methods based on matrix factorization have also emerged in the last decade [16, 15, 5, 2]. They factorize the student per item results matrix into the linear product of the so called Q-matrix (skills required per item) and the skills mastery matrix.

We undertook the effort of comparing prevailing and widely different methods to assess skills. The comparison is based on each method's ability to predict item/task outcome. However, in addition to providing a comprehensive comparison of skills assessment approaches, this research also aims to develop a method that uses synthetic data to characterize item outcome data and yield insights about this data's ground truth structure. Beyond the obvious expectation that the model behind the generation of synthetic data will outperform all others on this data set, we conjecture that the relative performance of all other methods will be unique and can represent a kind of "performance signature" that characterizes this type of data. Therefore, if a data set from a real setting reflects that signature, it would constitute a good indicator that the corresponding model is a good fit.

This work is an extension of [3], and is similar in its general principles to the approach of Rosenberg-Kima and Pardos [13], who take the likelihood of a model's parameter space as a signature instead of the performance of different techniques as we do here. Their idea is that the likelihood function of two parameters of Bayesian Knowledge tracing is a unique characterization of a data set. If the likelihood function of synthetic data generated with estimates of these parameters from real data has the same "signature" as the likelihood function of that real data, then the model is a good fit.

2. SKILLS ASSESSMENT METHODS

We compare a total of seven different skills assessment methods. We briefly describe them here and refer the reader to [7] and [6] for details. They can be grouped into four categories:

(1) The single skill Item Response Theory (IRT) approach. IRT is a well known framework based on logistic regression and represents student proficiency by a single skill (although we also find multiple skills version of IRT, MIRT).

(2) The POKS (Partial Order Knowledge Structrures) represents the order in which items are learned and uses a Naive Bayes framework to make inferences based on this order. It does not represent latent skills, but a Q-matrix can be used aposteriori on the estimated item outcome to assess skills.

(3) The matrix factorization approach decomposes the matrix of m students by n items into the product of m students by k skills representing the latent skills assessment, and an k by n Q-matrix.

(4) The multi-skills family of DINA/DINO approaches are equivalent to a binary matrix factorization framework, where the skill outcome is a boolean product of binary vectors, but they also contain *guess* and *slip* parameters. In the DINA version, the boolean product is based on the AND operator, whereas DINO is based on the OR operator.

Finally, as a baseline for comparison we also consider the *Expected value* as the simplest model. It takes into account the mean item difficulty and student ability to compute the expected score of the corresponding item. The mean difficulty is the average success rate of an item obtained from the training data, while the student ability is the mean success rate obtained from the observed data. The Expected value is the geometric mean of the product of these two means.

3. METHODOLOGY

The performance of each method is assessed on the basis of 10-folds cross-validation, and on observing all items from a student except the one that is to be predicted. For each fold, each item in the set is taken as a target prediction once.

For the IRT and POKS models, the parameters of each models are trained and the testing is based on feeding the models with all but one question. A probability of mastery is obtained and rounded, resulting in a 0/1 error loss function. We report the mean accuracy as the performance measure. The R package `ltm` is used for parameter and skills estimation.

For the other models, they rely on a Q-matrix to estimate the remaining item outcome. For the linear conjunctive and compensatory models, the Q-matrix needs to be normalized such that if all skills for an item are mastered, the inner product of the skills mastered vector and the skills required will be 1. Here too, results are rounded for obtaining a 0/1 loss function. Normalization of the Q-matrix is not necessary for the DINA and DINO models.

4. DATA SETS AND SYNTHETIC DATA GENERATION

The performance of the methods is assessed over a total of 14 data sets, 7 of which are synthetic, and 7 are real data. They are listed in table 1), along with the number of skills of their Q-matrix, their number of items, the number of the student respondents, and the average score. Table 1 also reports the Q-matrix used. To make these data sets more comparable to their real counter part we used Q-matrices and other parameters from real data sets to generate synthetic datasets.

Of the 7 real data sets, only three are independent. The other 4 are variations of a well known data set in fraction Algebra from Tatsuoka's work [14]. The real data sets were obtained from different sources and are freely available from the CDM and NPCD R packages. The Q-matrices of the real data sets were made by experts.

The synthetic data sets are generated from their underlying respective skills assessment model.

For POKS, the structure was obtained from the Fraction data set and the conditional probabilities were generated stochastically, but in accordance with the semantic constraints of these structures and to obtain an average success rate of 0.5.

For IRT, the student ability distributions was obtained from the Fraction data set, and the item difficulty was set to reasonable values: averaging to 1 and following a Poisson distribution that kept most values between 0.5 and 2[1].

Data set	Number of			Mean Score	Q-matrix
	Skills	Items	Students		
Synthetic					
1. Random	7	30	700	0.75	$\mathbf{Q}_{01}$
2. POKS	7	20	500	0.50	$\mathbf{Q}_{02}$
3. IRT-Rasch	5	20	600	0.44	$\mathbf{Q}_{04}$
4. DINA	7	28	500	0.31	$\mathbf{Q}_{5}$
5. DINO	7	28	500	0.69	$\mathbf{Q}_{6}$
6. Linear Conj.	8	20	500	0.24	$\mathbf{Q}_{1}$
7. Linear Comp.	8	20	500	0.57	$\mathbf{Q}_{1}$
Real					
8. Fraction	8	20	536	0.53	$\mathbf{Q}_{1}$
9. Vomlel	6	20	149	0.61	$\mathbf{Q}_{4}$
10. ECPE	3	28	2922	0.71	$\mathbf{Q}_{3}$
Fraction subsets and variants of $\mathbf{Q}_1$					
11. 1	5	15	536	0.53	$\mathbf{Q}_{10}$
12. 2/1	3	11	536	0.51	$\mathbf{Q}_{11}$
13. 2/2	5	11	536	0.51	$\mathbf{Q}_{12}$
14. 2/3	3	11	536	0.51	$\mathbf{Q}_{13}$

Table 1: Datasets

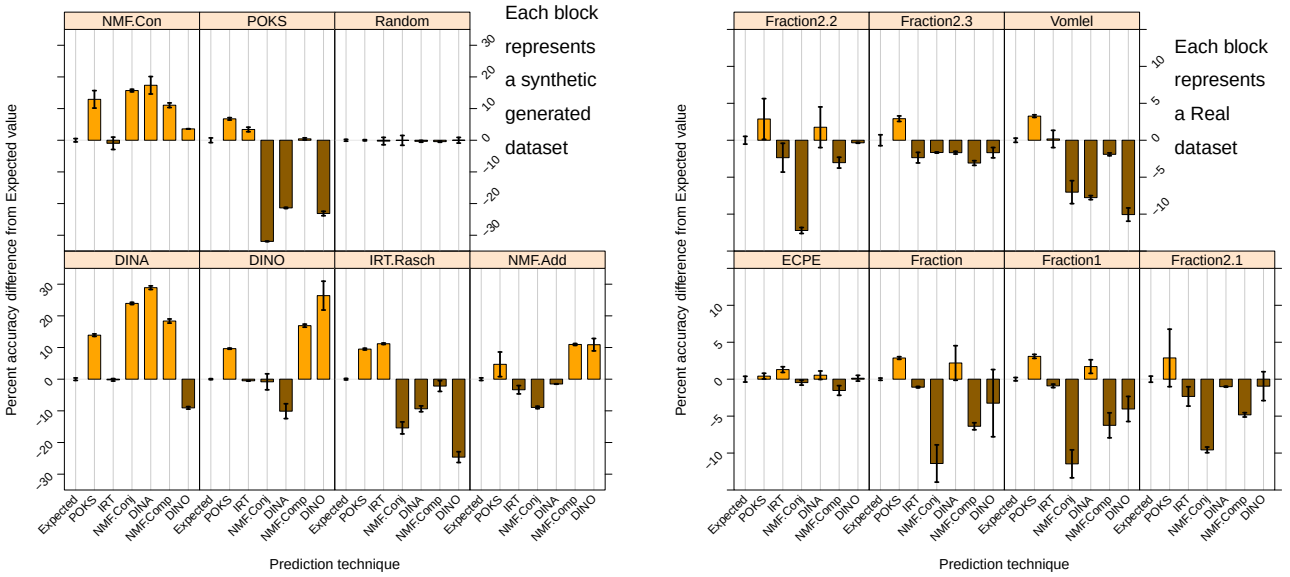
The matrix factorization synthetic data sets of DINO and DINA were generated by taking a Q-matrix of 7 skills that contains all possible combinations of 1 and 2 skills, which gives a total of 28 combinations and therefore the same number of items. Random binary skills matrix were generated and the same process was used for both the DINO and DINA data sets. Item outcome is then generated with a slip and guess factor of 0.1.

A similar process was followed to generate the Q-matrices and the skills matrices $\mathbf{S}$ of the linear matrix factorization data sets

Note that the first 3 models do not rely on any Q-matrix for the data generation process, but the DINO/DINA and matrix factorization assessment methods still require one. To define these Q-matrices (denoted $\mathbf{Q}_{0x}$ in table 1, a wrapper method was used to first determine the number of skills according to [4], then a Q-matrix was derived with the ALS method (see [9]).

All data sets are considered *static* in the sense that they represent a snapshot of student test performance data. This corresponds to the assumption that the student has not mastered new skills during the process of assessment, as we would expect from data from learning environments. This assumption is common to all models considered for this study.

[1]Done by generating random numbers from a Poisson distribution with lambda parameter set to 10 and dividing by 10.



(a) Synthetic datasets

(b) Real dataset

Figure 1: Item outcome prediction accuracy results. Each plot reports the prediction accuracy of the different techniques, whereas each bar shows the percentage difference in accuracy from the Expected value baseline (square root of item × student average success rates).

5. RESULTS AND DISCUSSION

Figure 1 shows the difference between the performance of each technique and the Expected value accuracy as computed by the geometric mean: square root of item × student average success rates. An error bar of 1 standard deviation is reported and computed over the 10 random sampling simulation runs and provides an idea of the variability of the results. Also reported is the performance of random data with a 0.75 average success rate.

As expected, when the generative model behind the synthetic data set is the same as the skills assessment technique, the corresponding technique's performance is generally the best. Exceptions are found for the linear conjunctive case, where the corresponding technique performance comes second. For real data, the performance of many techniques is often lower than the Expected value baseline. This is likely due to the fact that all but one item is observed, the target, and therefore the Expected value is a reliable predictor.

The most consistent performance across the synthetic data sets are those of POKS and IRT, with POKS showing a greater accuracy on average. This consistency also transfers to the real data sets, although the differences are smaller and the Expected value method performance is sometimes better than the IRT one. But as mentioned the good performance of the Expected value may well depend on the relatively high number of observations for each data sets (1 less than the total number of questions per data set).

Also worth noticing is that the random data set has a flat performance across techniques which corresponds to the dominant class prediction. This is not necessarily surprising, but it is reassuring in a sense to know that they all perform the same in the face of random data and this performance is indeed the best that could be obtained.

For the independent real data sets, the differences between techniques are less divergent and closer to the Expected value technique, although the best performers are still significantly better than the Expected value for the Fraction (POKS and DINA) and Vomlel (POKS) data sets. However, for the ECPE data set, the pattern corresponds closely to that of random data: The Expected value performance is close to the dominant class performance, and all techniques are aligned towards this performance. One possibility is that all student perform more or less the same and therefore no technique is good at discriminating high/low performers.

The results from the subsets of the Fraction data shows that the pattern of the Fraction performance data set repeats over Fraction-1, Fraction-2/1 and Fraction-2/2, in spite of the different number of skills and different subsets of questions. However, it differs substantially from Fraction-2/3 for the NMF conjunctive performance which reaches that of the NMF compensatory one. This is readily explained by the fact that the Q-matrix of this data set has the property of assigning a single skill to each item, in which case the two matrix factorization techniques become equivalent.

As mentioned, the performance of the Expected value technique is high for real data, and systematically close to the best performers, POKS and DINA, which only have 2–4% better performance than the Expected value. Note that this is still substantial because we have to look at this difference relative to the remaining error (about 20%), but it is far less than for the synthetic data sets, especially on a relative difference basis.

6. CONCLUSION

This study relies on the assumption that better skills models result in better item outcome prediction. The results do show wide differences in the performance of the techniques for different synthetic data sets. For real data sets, the differences are smaller, though still significant, especially in terms of relative residual errors. Based on the results, we could conclude that POKS and DINA would provide more accurate estimates of skills.

Let us return to the comparison of real vs. synthetic data and to the conjecture that this comparison can help determine whether a specific skill model corresponds to the ground truth of some data set. This is a complex question but some clear hints are given in the results. There is a clear evidence in the DINA vs. DINO performance of figure 1 data that, if a Q-matrix is conjunctive vs. disjunctive, the results show a much better the fit to the corresponding model. Evidence is also some evidence to the claim that unidimensional data sets, i.e. a domain for which a single skill best characterizes the performance data, are best modelled by the IRT single skill IRT or the skill-less POKS models, and the multi-skills NMF conjunctive and DINA approaches do rather poorly. Conversely, multiple skills data sets of the DINO/DINA and linear family of models are better characterized by multi-skills approaches, and the IRT single skill performance is much lower in relative terms.

Another interesting finding is that random data does have a signature of its own: all methods converge towards the score of the majority class. Now, this result could stem from a set of highly similar response patterns from students, but it is clearly different from, for example, the Fraction-2/3 data set, for which all methods have relatively similar performance but they are all well above the majority class condition (AVG Success rate).

Therefore, we do conclude that there is evidence to support the claim that the relative performance of the different skills modelling approaches do create signatures over data sets and can yield some evidence about the ground truth. And if we accept this perspective, then we can also conclude that the real data sets we studied do not correspond to any of the prototypical synthetic data sets. The ground truth may involve correlations between skills, which we did not take into account. Or, the Q-matrices we have studied are not faithful to the reality and, for example, may involve combinations of conjunctive and disjunctive skills. In fact, many explanations can be evoked, but the hope is that by looking at the relative performances of each method we can gain some insights of the best explanations.

References

[1] F. B. Baker and S.-H. Kim. *Item Response Theory, Parameter Estimation Techniques (2nd ed.)*. Marcel Dekker Inc., New York, NY, 2004.

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

[3] B. Beheshti and M. C. Desmarais. Predictive performance of prevailing approaches to skills assessment techniques: Insights from real vs. synthetic data sets. In *5th International conference on Educational Data Mining, EDM 2014*, pages 409–410, London, UK 2014.

[4] B. Beheshti, M. C. Desmarais, and R. Naceur. Methods to find the number of latent skills. In *5th International conference on Educational Data Mining, EDM 2012, Chania, Greece, 19–21 June 2012*, pages 81–86. Springer, 2012.

[5] M. Desmarais. Conditions for effectively deriving a Q-matrix from data with non-negative matrix factorization. In *4th International Conference on Educational Data Mining, EDM*, pages 41–50, 2011.

[6] M. C. Desmarais, B. Beheshti, and R. Naceur. Item to skills mapping: Deriving a conjunctive Q-matrix from data. In *11th Conference on Intelligent Tutoring Systems, ITS 2012*, pages 454–463, Chania, Greece, 14–18 June 2012 2012.

[7] M. C. Desmarais and R. S. d Baker. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22(1-2):9–38, 2012.

[8] M. C. Desmarais, P. Meshkinfam, and M. Gagnon. Learned student models with item to item knowledge structures. *User Modeling and User-Adapted Interaction*, 16(5):403–434, 2006.

[9] 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.

[10] J.-P. Doignon and J.-C. Falmagne. *Knowledge Spaces*. Springer-Verlag, Berlin, 1999.

[11] B. Junker and K. Sijtsma. Cognitive assessment models with few assumptions, and connections with nonparametric IRT, Dec. 27 2000.

[12] B. W. 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.

[13] R. B. Rosenberg-Kima and Z. Pardos. Is this data for real? In *Twenty Years of Knowledge Tracing Workshop (colocated with EDM 2014*, pages 141–145, London, UK.

[14] K. K. Tatsuoka. *Analysis of errors in fraction addition and subtraction problems*. Computer-based Education Research Laboratory, University of Illinois, 1984.

[15] N. Thai-Nghe, L. Drumond, T. Horváth, A. Nanopoulos, and L. Schmidt-Thieme. Matrix and tensor factorization for predicting student performance. In A. Verbraeck, M. Helfert, J. Cordeiro, and B. Shishkov, editors, *CSEDU 2011 - Proceedings of the 3rd International Conference on Computer Supported Education, Volume 1, Noordwijkerhout, Netherlands, 6-8 May, 2011*, pages 69–78. SciTePress, 2011.

[16] T. Winters, C. Shelton, T. Payne, and G. Mei. Topic extraction from item level grades. In *American Association for Artificial Intelligence 2005 Workshop on Educational Datamining*, 2005.