ASSESSING MODEL FIT WITH SYNTHETIC VS. REAL DATA

Behzad Beheshti
Polytechnique de Montréal
behzad.beheshti@polymtl.ca

Michel C. Desmarais
Polytechnique de Montréal
michel.desmarais@polymtl.ca

ABSTRACT
Assessing whether a model is a good fit to the data is non trivial. The standard practice is to compare a few machine learning techniques to learn a model from data, and pick the one with the highest predictive performance. The winner is considered the best fitting model. But each model may involve different machine learning algorithms that carry their own set of parameters and constraints imposed on the corresponding model. This results in a large space in which to explore model performance. The actual best fitting model may have been overlooked due to an unfortunate choice of the algorithm's parameters. We address this issue by complementing performance model comparison with a method that combines real data with synthetic data generated with the competing models. We naturally expect a model to perform best over its corresponding synthetic data, but the analysis across the other synthetic data sets provides some indication of the models generality and robustness under different assumptions about the data. Results of our investigation in the domain of educational data mining show that a model performance is, as expected, best when tested over synthetic data generated aligned with this model. But we observe much greater performance contrasts across synthetic data than across real data. The performance pattern of each model over a given synthetic data set results in a kind of "signature". We discuss the significance of this signature to assess model fit, and whether it can provide cues to the data's underlying ground truth.

KEYWORDS
Educational Data Mining, Learning Analytics, student skills assessment, IRT, synthetic data

1. INTRODUCTION

Data analysts that wish to build a classification or regression model over new and unknown data are faced with a very wide span of choices. Machine learning techniques nowadays offer the possibility to learn and train a large and an ever growing variety of models from data. Learning techniques such as the E-M algorithm MCMC methods eliminate many critical constraints on the models we can learn from data. They allow model parameters estimation that would otherwise represent an intractable problem using standard analytical or optimization techniques.

Along with this increased display of models that can be defined and trained from data, comes the question of deciding which are the most representative of the underlying ground truth. The consensus is that the model with the best predictive performance is the most likely to be the closest to the ground Truth. Then there are the issues of how sensitive is the model to sample size, noise, and biases that also need to be addressed before we can trust that this model is the best candidate. It can take numerous studies before a true consensus emerges as to which model is the best candidate for a given type of data.

Yet another approach to assess which model is closest to the ground Truth is to combine predictive performance analysis of real data with synthetic data. Comparing performance over synthetic and real data has been used extensively to validate models, but we further elaborate on the standard principle of comparison over both types of data by contrasting the predictive performance across types of synthetic data. The hypothesis we make is that the relative performance of different models will by stable and characteristic
of a given type of data, as defined by the underlying ground truth for real data, or by the model that generates the synthetic data. We explore this hypothesis in the domain of Educational Data Mining and the assessment of student skills, where a set of latent skills is mapped to question items and students skill mastery is inferred from item outcome results from test data.

2. SKILLS ASSESSMENT

There exists a large array of models to represent and assess student skills. We focus here on the assessment of static skills assessment, where we assume the test data represents a snapshot in time, as opposed to models that allow the representation of skills that changes in time (see Desmarais and Baker, 2012, for a review of both approaches), which is more typical of data from learning environments.

For modeling static test data, Item Response Theory (IRT) is likely the most established model. It dates back to the 1960’s and is still one of the prevailing approaches (Baker and Kim, 2004). But, akin to the trend in data mining in general, many other models have been introduced in recent years. Among them is the family of models that rely on slip and guess factors (Junker and Sijtsma, 2001), such as the DINA (Deterministic Input Noisy And-Gate), DINO (Deterministic Input Noisy Or-Gate), and other variants. Other approaches are based on the Knowledge Space theory of Doignon and Falmagne (Doignon and Falmagne, 1999; Desmarais et al., 1996), which does not directly attempt to model underlying skills but instead rely on observable items only. Finally, recent models based on matrix factorization have also emerged in the last decade (Winters et al., 2005; Thai-Nghe et al., 2011; Desmarais, 2012; Barnes, 2010). They factorize the student per item results matrix into the linear product of the so called Q-matrix (skills required per question item) and the skills mastery matrix.

These models rest on fundamentally different theoretical assumptions about skills representation and skilled performance. But they all share the same goal of providing a framework to assess individual skills from observed performance on a set of tasks, which we loosely refer here as “items”.

The current work aims to develop a new method to estimate model fit, and to characterize assessment data and yield insights about this data's ground truth structure. It extends previous work in this direction (Desmarais and Pelczer, 2010). It also relates to the question of whether the maximum prediction performance has been reached (Beck and Xiong, 2013). The general intuition is that if data is best modeled by a linear or a Bayesian model, for example, this will be reflected by higher predictive performance for models that rely on the respective linear or Bayesian framework. If real data perfectly fits an approach underlying model, then the expected maximum predictive accuracy will be reached by this specific model. Furthermore, we could also expect that the relative performance of each assessment approach will be the same over real and synthetic data if they both share the same underlying model. Furthermore, when analyzing performance across models and data sets, we can refer to the relative performance of all models over a data set as a performance “signature” pattern. Data generated from an underlying ground truth model may have a unique signature, whether it is real or synthetic data. We return to this conjecture in the results analysis and conclusion sections.

The following section describes the specific skills models that are compared. It is followed by the methodology and the results of the experiments.

3. SKILLS ASSESSMENT MODELS COMPARED

The skills assessment model we compare can be grouped into four categories: (1) the single skill Item Response Theory (IRT) approach, (2) the Knowledge Space frameworks which models a knowledge state as a set of observable items without explicit reference to skills, (3) the matrix factorization approach which decomposes the student results matrix into a Q-matrix that maps items to skills, and a skills matrix that maps skill to students, and which relies on standard matrix algebra for parameter estimation and item outcome prediction, and finally (4) the multi-skills family of DINA/DINO approaches which also refers to a Q-matrix, but incorporates slip and guess factors and relies on different parameter estimation techniques than the matrix factorization method.

The details of the specific are described below.
3.1 Item Response Theory (IRT)

The IRT family is based on a logistic regression framework. It models a single latent skill (although variants exist for modeling multiple skills) (Baker and Kim, 2004). Each item has a difficulty and a discrimination parameter.

IRT assumes the probability of success to an item \( X_j \) is a function of a single ability factor \( \theta \):

\[
P(X_j = 1 \mid \theta) = \frac{1}{1 + e^{-a_j(\theta-b_j)}}
\]

In the two parameter form above, referred to as IRT-2pl, parameter \( a \) represents the item discrimination and parameter \( b \) represents the item difficulty.

The ability of a single student, \( \theta_i \), is estimated by maximizing the likelihood of the observed response outcomes probabilities:

\[
P(X_1, X_2, ..., X_J, \theta_i) = \prod_j P(X_j \mid \theta_i)
\]

This corresponds to the usual logistic regression procedure.

The specific IRT skills assessment version retained for this study is the Rash model, for which the discrimination parameter \( a \) is fixed to 1. Fixing this parameter reduces over fitting, as the discrimination can sometimes take unrealistically large values. Note however that we do generate synthetic data from the more general IRT-2pl model to make this data more realistic.

3.2 Knowledge Spaces and Partial Order Knowledge Structure (POKS)

The knowledge spaces approach models a knowledge domain as a set of observable items. An individual’s knowledge state is represented as a subset of these items. The “knowledge space” determines the valid subsets of items. There are no explicit latent skills, but the skills can be obtained from the estimated item outcomes, akin to how every teacher breaks down an exam’s results by topic as a weighted sum.

The model adopted to assess knowledge in the Knowledge Spaces family is POKS (Partial Order Knowledge Structures), which is a more constrained version of Knowledge Spaces theory (Desmarais et al., 1996). POKS assumes that items are learned in a strict partial order. It uses this order to infer that the success to hard items increases the probability of success to easier ones, or conversely, that the failure to easy items decreases the chances of success to harder ones.

The structure of the partial order of items is obtained from a statistical hypothesis test that reckons the existence of a link between two items, say \( A \rightarrow B \), on the basis of two Binomial statistical tests \( P(B | A) > p_1 \) and \( P(\neg A | \neg B) > p_2 \) and under a predetermined alpha error \( (a < p_2) \). The values of \( p_1 = 0.85 \) and \( p_2 = 0.10 \) are chosen in this study across all experiments.

A student knowledge state is represented as a vector of probabilities, one per item. Probabilities are updated under a Naive Bayes assumption as simple posterior probabilities given observed items.

3.3 Matrix Factorization

The matrix factorization approach decomposes a set of student results, represented as a matrix \( R \) of \( m \) student by \( n \) items, into the inner product of two other matrices:

\[
R \approx QS
\]

where \( Q \) is a Boolean matrix that represents the skills required by the \( n \) items (normalized for the skills to sum to 1), whereas the matrix \( S \) represents the skills mastered by the \( m \) students. Larger values in matrix \( S \) represent greater mastery level of the corresponding skill row by the corresponding student column. The matrix \( Q \) is also used by the family DINO/DINA of models reviewed below, and termed the “Q-matrix”.

Equation (1) represents the compensatory version of a Q-matrix, where each skill increases the chances of success to an item. However, it is often the case that a Q-matrix is defined to require that all skills be mastered, which does not correspond to the usual inner product in equation (1). This version of the Q-matrix is termed conjunctive and is obtained by the negation of the student results matrix \( R \) and skill matrix \( S \):

\[
\neg R \approx Q \neg S
\]

The operator \( \neg \) is the Boolean negation that maps 0 values to 1 and other values to 0. Following this logic, an
examinee who mastered all required skills for an item will get 1 in the result matrix, and otherwise he will get a 0 value, even if the required skills are partially mastered.

We use the Alternating Least Squares algorithm which starts with a matrix of observed test results, \( R \), to get a first skills estimate \( S_1 \) through least squares: \( S_1 = (Q^TQ)^{-1}Q^TR \), and then a new estimate of \( Q_1 \) through least squares again, and alternates between estimates of the Q-matrix and the skill matrix until convergence. The predictions are based on the result of the product of the matrices at convergence: \( Q_nS_n \).

### 3.4 Deterministic Input Noisy And/Or (DINA/DINO)

The last skills assessment model we consider are based on what is referred to as Deterministic Input Noisy And/Or (DINO/DINA) (Junker and Sijtsma, 2001). They also rely on a Q-matrix. The DINA model (Deterministic Input Noisy And) corresponds to the conjunctive model whereas the DINO (Deterministic Input Noisy Or) corresponds to the disjunctive one, where the mastery of a single skill is sufficient to succeed an item. The acronyms makes reference to the AND/OR gates terminology.

These models predict item outcome based on three parameters: the slip and guess factors of items, and the different “gate” function between the student's ability and the required skills. The gate functions are equivalent to the conjunctive and disjunctive vector product logic described for the matrix factorization above. In the DINA case, if all required skills are mastered, the result is 1, and 0 otherwise. In the DINO case, mastery of any skills is sufficient to output 1. Assuming \( \xi_i \) is the output of the corresponding DINA or DINO model and \( s_j \) and \( g_j \) are the slip and guess factors, the probability of a successful outcome to item \( X_{ij} \) is:

\[
p(X_{ij} = 1 | \xi_{ij}) = (1 - s_j)^{\xi_{ij}}g_j^{1-\xi_{ij}}
\]  

(3)

A few methods have been developed to estimate the slip and guess parameters from data and we use the one implemented in the R CDM package (Robitzsch et al., 2012).

### 3.5 Expected value

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.

### 4. METHODOLOGY

The performance of each model is assessed on the basis of 10-folds cross-validation, and on feeding the model with the outcome of all items from a student, except the item that is to be predicted from these “observed items”. The folds consists in splitting the data into 10 bins, taking 9 of them for the training and 1 for the testing.

For the IRT and POKS models, the process is straightforward. The parameters of each model 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.
5. DATA SETS AND SYNTHETIC DATA GENERATION

The performance of the models is assessed over a total of 15 data sets, 8 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. As can be seen, some synthetic data sets share their Q-matrix with real data sets. This sharing allows greater similarity between the synthetic data and a real data counterpart that shares a Q-matrix. Other parameters used to create the synthetic data sets were also obtained from real data sets with the same intent of allowing better comparison.

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 (Tatsuoka, 1984). They consist in subsets of questions and variations of the Q-matrix. These variants allow us to explore the effect of different models (Q-matrices) over the same data source.

The Vomlel data was obtained from (Vomlel 2004) and is also on the topic of fraction algebra. The Q-matrix for this data is derived from the Bayesian Network defined over the 20 item test by experts.

The ECPE data (Examination for the Certificate of Proficiency in English) is an English as a foreign language examination. It is recognized in several countries as a test of advanced proficiency in English and used by a number of universities.

These real data sets were obtained from different sources and are freely available from the CDM (Robitzsch et al., 2012) and NPCD (http://cran.r-project.org/web/packages/NPCD/) R packages. The Q-matrices of the real data sets were made by experts.

Table 1. Datasets

<table>
<thead>
<tr>
<th>Data set</th>
<th>Number of Skills</th>
<th>Number of Items</th>
<th>Number of Students</th>
<th>Mean score</th>
<th>Q-matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthetic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Random</td>
<td>7</td>
<td>30</td>
<td>700</td>
<td>0.75</td>
<td>Q_{01}</td>
</tr>
<tr>
<td>2. POKS</td>
<td>7</td>
<td>20</td>
<td>500</td>
<td>0.50</td>
<td>Q_{02}</td>
</tr>
<tr>
<td>3. IRT-2pl</td>
<td>5</td>
<td>20</td>
<td>600</td>
<td>0.50</td>
<td>Q_{03}</td>
</tr>
<tr>
<td>4. IRT-Rasch</td>
<td>5</td>
<td>20</td>
<td>600</td>
<td>0.44</td>
<td>Q_{04}</td>
</tr>
<tr>
<td>5. DINA</td>
<td>7</td>
<td>28</td>
<td>500</td>
<td>0.31</td>
<td>Q_{05}</td>
</tr>
<tr>
<td>6. DINO</td>
<td>7</td>
<td>28</td>
<td>500</td>
<td>0.69</td>
<td>Q_{06}</td>
</tr>
<tr>
<td>Linear (Matrix factorization)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Conj.</td>
<td>8</td>
<td>20</td>
<td>500</td>
<td>0.24</td>
<td>Q_{01}</td>
</tr>
<tr>
<td>8. Comp.</td>
<td>8</td>
<td>20</td>
<td>500</td>
<td>0.57</td>
<td>Q_{01}</td>
</tr>
<tr>
<td>Real</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Fraction</td>
<td>8</td>
<td>20</td>
<td>536</td>
<td>0.53</td>
<td>Q_{1}</td>
</tr>
<tr>
<td>10. Vomlel</td>
<td>6</td>
<td>20</td>
<td>149</td>
<td>0.61</td>
<td>Q_{4}</td>
</tr>
<tr>
<td>11. ECPE</td>
<td>3</td>
<td>28</td>
<td>2922</td>
<td>0.71</td>
<td>Q_{3}</td>
</tr>
<tr>
<td>Fraction subsets and Variants of Q_{1}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. 1</td>
<td>5</td>
<td>15</td>
<td>536</td>
<td>0.53</td>
<td>Q_{10}</td>
</tr>
<tr>
<td>13. 2/1</td>
<td>3</td>
<td>11</td>
<td>536</td>
<td>0.51</td>
<td>Q_{11}</td>
</tr>
<tr>
<td>14. 2/2</td>
<td>5</td>
<td>11</td>
<td>536</td>
<td>0.51</td>
<td>Q_{12}</td>
</tr>
<tr>
<td>15. 2/3</td>
<td>3</td>
<td>11</td>
<td>536</td>
<td>0.51</td>
<td>Q_{13}</td>
</tr>
</tbody>
</table>

The synthetic data sets are generated from each skills assessment model, with an effort to fit the parameters as closely as possible to a real data counterpart that shares the same Q-matrix.

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 (done by generating random numbers from a Poisson distribution with lambda parameter set to 10 and dividing by 10).
Figure 1: Item outcome prediction accuracy results of synthetic data sets

Figure 2: Item outcome prediction accuracy results of real data sets
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 (which corresponds to matrix $S$ in equation (1)) were generated and the same process was used for both the DINO and DINA data sets. Item outcome is then generated according to equation (3) with a slip and guess factor of 0.1.

A similar process was followed to generate the Q-matrices and the skills matrices $S$ of the linear matrix factorization data sets, except that item outcome follows equation (1) and is discretized.

Note that the first 4 models do not rely on any Q-matrix for the data generation process, but the DINO/DINA and matrix factorization assessment models still require one. To define these Q-matrices (denoted $Q_{0x}$ in table 1, a wrapper method was used to first determine the number of skills according to (Beheshti et al., 2012), then a Q-matrix was derived with the deterministic ALS algorithm as described in section 3.3, starting with an initial random Q-matrix.

6. RESULTS AND DISCUSSION

Figure 1 and 2 show the performance of each technique over the synthetic and real data sets. For each data set, the predictive accuracy of each model is shown, in addition to the average success rate for the data set (rightmost column of each technique), which represents the majority class prediction (or (1-perf) when the value is below 0.5). An error bar of 1 standard deviation is reported, computed over the 10–random sampling simulation runs, provides an idea of the variability of the results. A data set of random data is also reported for 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 the best, or close to the best.

For synthetic data, we note that the pattern of relative differences of performances across techniques varies considerably and is unique to each data set: no two data sets have the same pattern of relative performance across models. In other words, the synthetic data sets seem to have a performance “signature” over these techniques. A conjecture from this observation is that if a real data set signature was perfectly aligned with one of these synthetic data sets, we could conclude that the underlying structure of the data is consistent with the generative model, and therefore that the generative model may represent the ground truth. We return to this conjecture below.

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 confirms that all models perform the same in the face of random data, and this performance is indeed the best that could be obtained.

For the real data sets, the relative performance among the techniques shows smaller discrepancies and is closer to the simple expected values technique, although the best performers are still significantly better than the expected values for the majority of the data sets. However, for the ECPE data set, the “signature” is close to that of random data, although we do observe significant but very small differences.

The results from the subsets of the Fraction data show 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. But aside from the Fraction-2/3 case, this similarity among Fraction data set and its derivative suggests that in spite of the model differences (different Q-matrices and item subsets), the performance “signature” remains constant across these data sets.

Finally, we note that none of the real data sets show the large the discrepancies found in the synthetic data sets models. One exception is the linear compensatory synthetic data set which displays smaller variance across models and which “signature” resembles the Vomlel data, although the performance difference with the majority class is substantially higher for the synthetic data than the Vomlel data, suggesting that the real data is yet not a perfect fit to this model.
7. CONCLUSION

Let us assume that the comparison of real vs. synthetic data can help determine whether a specific skill model corresponds to the ground truth of some data set. This is a complex question but results give some hints. A clear finding is that the synthetic data sets have very distinct performance patterns, showing sharp differences across models. In that respect, synthetic data do have a distinct signatures. None of the real data sets display the sharp differences found in all but the data generated from the Linear compensatory model.

The Fraction data sets do display a similar pattern of performances across different subsets of items, different number of skills (latent factors), and different variants of the models as expressed by variations in the Q-matrices. Only when the Q-matrix has the property of a single skill per item do we observe a very different performance signature for the models that depend on the Q-matrix (NMF conjunctive, DINA, NMF Compensatory, and DINO). The other models are not affected (expected, POKS and IRT). Therefore, in the domain of skills modeling, we find evidence that data from a common source does have signature as long as the models do not have large formal differences.

Even though the Vomlel data is within the same domain as the Fraction data, namely arithmetic, the performance signature of the Q-matrix dependent models display two very different patterns. This could be attributed to the Q-matrices involved (although both are multiple skills and therefore the difference is not due to the formal property of single skill per item), but other factors (like sample size) could also be involved.

Therefore, we do find evidence to support the claim that the relative performance of the different skills modeling approaches create signatures over data sets. And although the current findings are still exploratory and require further investigations, the evidence so far yields some confidence that these signatures carry the potential to provide evidence about the ground truth.

REFERENCES