An Empirical Research on Identifiability and Q-matrix Design for DINA model

Peng Xu
Polytechnique Montreal
peng.xu@polymtl.ca

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

ABSTRACT

In most contexts of student skills assessment, whether the test material is administered by the teacher or within a learning environment, there is a strong incentive to minimize the number of questions or exercises administered in order to get an accurate assessment. This minimization objective can be framed as a Q-matrix design problem: given a set of skills to assess and a fixed number of question items, determine the optimal set of items, out of a potentially large pool, that will yield the most accurate assessment. In recent years, the Q-matrix identifiability under DINA/DINO models has been proposed as a guiding principle for that purpose. We empirically investigate the extent to which identifiability can serve that purpose. Identifiability of Q-matrices is studied throughout a range of conditions in an effort to measure and understand its relation to student skills assessment. We compare identifiability to other Q-matrix design principles through simulation studies of skills assessment with both synthetic and real data.

1. INTRODUCTION

Q-matrix based models are widely researched during recent years in educational data mining. Among all of them, DINA model is the most used and discussed. Most of these research are focused on one of the 2 following problems. The first one is the Q-matrix validation problem, that is, to improve or validate an expert-given Q-matrix (de la Torre & Chiu, 2015; Chiu, 2013; Desmarais & Naceur, 2013). The other one is the Q-matrix derivation problem, that is, to directly derive a Q-matrix out of the test result matrix (Barnes, 2010; Liu, Xu, & Ying, 2012; Desmarais, Xu, & Beheshti, 2015; P. Xu & Desmarais, 2016). During the investigation of these two problems, a fundamental question has been proposed, the identifiability of model parameters, especially the Q-matrix. Detailed statistical analysis has been made under the DINA/DINO situation, which was first discussed under the situation that slip and guess is zero(Chiu, Douglas, & Li, 2009), and then the case that slip and guess exist but is known(Liu, Xu, & Ying, 2013), and finally the case that slip and guess is unknown(Chen, Liu, Xu, & Ying, 2015).

However, how do we use the conditions offered by the discussion to guide our educational test? First, the discussion is centered around the identifiability on Q-matrix, not on the identifiability of students, which is also a critical problem (Beck & Chang, 2007). Fortunately, for the case slip and guess are known, the identifiability of the parameters p are also given(Chen et al., 2015).

2. IDENTIFIABILITY

The general idea behind identifiability is that two or more configurations of model parameters can be considered as equivalent. Sets of parameters will be considered equivalent if, for example, their likelihood are equal given a data set. Or, conversely, if they are susceptible to generate data having the same characteristics of interest (see Doroudi & Brunskill, 2017, for more details).

The issue of identifiability for student skills assessment was first raised for the Bayesian Knowledge Tracing (BKT) model by Beck & Chang, 2007 and later discussed by Doroudi & Brunskill, 2017, and van De Sande, 2013. In this paper, we consider the identifiability of the Q-matrix which was studied by G. Xu & Zhang, 2015 Qin et al., 2015.

Identifiability of model parameters is a general concept for statistical models, its general definition is,

Definition (Casella & Berger, 2002) A parameter θ for a family of distribution $f(x|\theta : \theta \in \Theta$ is identifiable if distinct values of θ correspond to distinct pdfs or pmfs. That is, if $\theta \neq \theta'$, then $f(x|\theta)$ is not the same function of x as $f(x|\theta')$.

However, for Q-matrix related model, since exchanging columns does not yield an essentially different Q-matrix. During the research of G. Xu and Zhang (2015), the identifiability of Q-matrix has been redefined.

Definition (G. Xu & Zhang, 2015) We write $Q \sim Q'$ if and only if Q and Q' have identical column vectors that can be arranged in different orders. We say that Q is identifiable if there exists an estimator $\hat{Q}$ such that $\lim_{N\to\infty} P(\hat{Q} \sim Q) = 1$.

Definition (G. Xu & Zhang, 2015) The matrix Q is complete meaning that $e_i : i = 1, ...k \subset R_Q$, where R_Q is the set of row vectors of Q and e_i is a row vector such that the i-th

element is one and the rest are zero.

Proposition (G. Xu & Zhang, 2015) Under the DINA and DINO models, with Q, s and g being known, the population proportional parameter p is identifiable if and only if Q is complete.

3. EXPERIMENT

All experiments consider the situation that slip and guess are already known.

3.1 Experiment 1: Comparison of three strategies

In this experiment, we compare three different Q-matrix design strategies. They are all based on repetition of a pool of q-vectors.

Strategy 1: Using the identifiability condition. Only repeated using the vectors $e_i : i = 1, ...k$.Q-matrix used in this strategy is denoted as Q-matrix 1.

Strategy 2: Using the vectors $e_i : i = 1, ...k$ plus a all-one vector $(1, 1, 1)$ or $(1, 1, 1, 1)$. In this way it forms an orthogonal array, which is a commonly seen design of experiments. Q-matrix used in this strategy is denoted as Q-matrix 2.

Strategy 3: Repeated using all q-vectors.Q-matrix used in this strategy is denoted as Q-matrix 3.

For the three skill case, all these three Q-matrices have been shown below.

Q-matrix 1

	k_1	k_2	k_3
q_1	1	0	0
q_2	0	1	0
q_3	0	0	1
...	...	...	...
q_{19}	1	0	0
q_{20}	0	1	0
q_{21}	0	0	1

Q-matrix 2

	k_1	k_2	k_3
q_1	1	0	0
q_2	0	1	0
q_3	0	0	1
q_4	1	1	1
...	...	...	...
q_{17}	1	0	0
q_{18}	0	1	0
q_{19}	0	0	1
q_{20}	1	1	1
q_{21}	1	0	0

Q-matrix 3

	k_1	k_2	k_3
q_1	1	0	0
q_2	0	1	0
q_3	0	0	1
q_4	1	1	0
q_5	1	0	1
q_6	0	1	1
q_7	1	1	1
...	...	...	...
q_{15}	1	0	0
q_{16}	0	1	0
q_{17}	0	0	1
q_{18}	1	1	0
q_{19}	1	0	1
q_{20}	0	1	1
q_{21}	1	1	1

3.2 Experiment 2: Find best configuration

For a given pool of q-vectors to choose from and an integer indicating the number of questions, we need to know the number of possible configurations of Q-matrices we have. This is equivalent to a classical combinatorial problem, that is, to allocate distinguished balls(q-vectors) to indistinguished cells(questions). It can be easily computed by combinatorial coefficients and interpreted by using stars and bars methods. For example, in 3-skill case, we have 7 q-vectors, and if we have 4 questions to allocate them, then we have $\binom{4+7-1}{7-1} = 210$ possible configurations. To compare, in 4-skill case, and if we have 5 questions to allocate them, then we have $\binom{5+15-1}{15-1} = 11628$ possible configurations. For each configuration, we will calculate the MAP estimation for all categories of each student, and compare with the one-hot encoding for their true categories. The total loss is reported as the performance index.

For the 3-skill case, all the 8 profile patterns are,

	k_1	k_2	k_3
p_1	0	0	0
p_2	1	0	0
p_3	0	1	0
p_4	0	0	1
p_5	1	1	0
p_6	1	0	1
p_7	0	1	1
p_8	1	1	1

4. RESULT

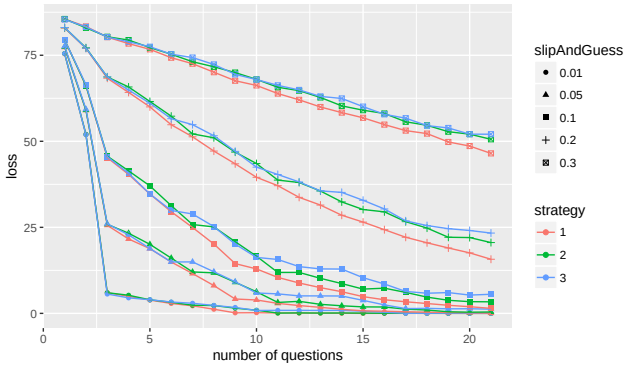

Figure 1: Three Strategy Comparison on 3-skill case

5. DISCUSSION

From the result in experiment 1 we can see that strategy 1 always works better than the other two strategies.

From the result in experiment 2, when slip and guess parameters are as low as 0.01, we can see obvious graded patterns among different configurations. Those configuration that satisfies the identifiability requirement always work better than other configurations. There is no clear distinction between the configuration using only e_j (the pure strategy) and those using at least one other q-vectors(mixed strategy). However, when slip and guess becomes higher, the pure strategy shows better than the mixed strategy.

Therefore, we argue that the best Q-matrix design is to use only the boolean unit vectors e_j since it offers quicker con-

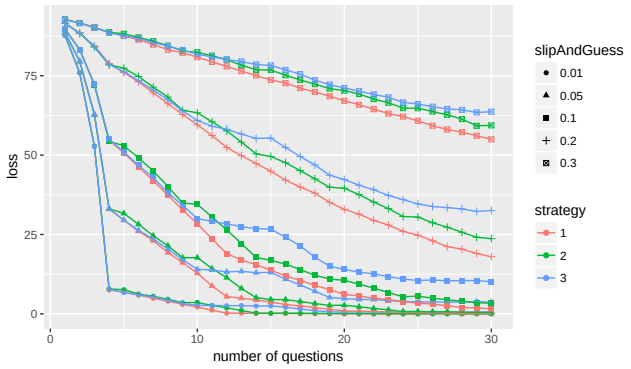

Figure 2: Three Strategy Comparison on 4-skill case

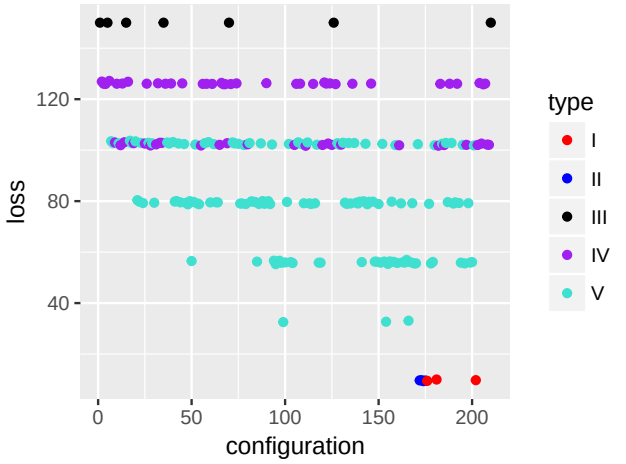

Figure 3: 3-skill case, slip=guess=0.01, J=4

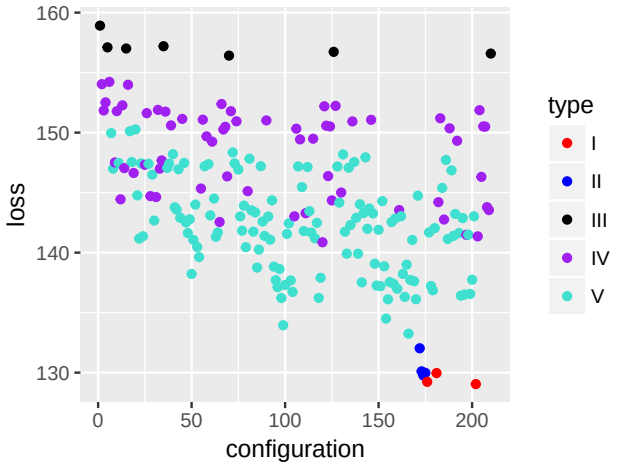

Figure 4: 3-skill case, slip=guess=0.2, J=4

vergence speed (as shown in experiment 1) and better robustness (as shown in experiment 2).

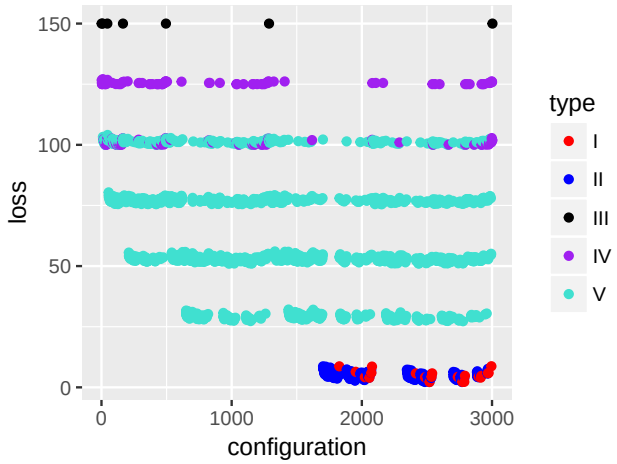

Figure 5: 3-skill case, slip=guess=0.01, J=8

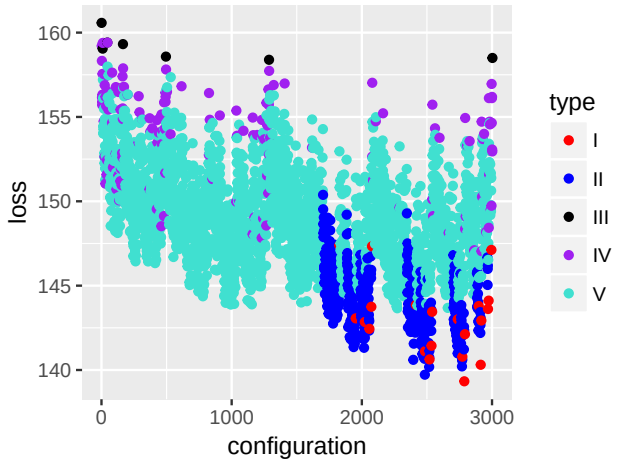

Figure 6: 3-skill case, slip=guess=0.3, J=8

6. FUTURE WORK

More different experiments settings can be considered with different choice on student profiles distribution, and different number of skills involved. Besides, the case that slip and guess are unknown should also be considered, which involves a different identifiability requirement. Moreover, besides the empirical exploration, rigorous mathematical discussion can be done for the best Q-matrix design under DINA model.

References

Barnes, T. (2010). Novel derivation and application of skill matrices: The q-matrix method. *Handbook on educational data mining*, 159–172.

Beck, J. E., & Chang, K.-m. (2007). Identifiability: A fundamental problem of student modeling. In *International conference on user modeling* (pp. 137–146).

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). Duxbury Pacific Grove, CA.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of q-matrix based diagnostic classification

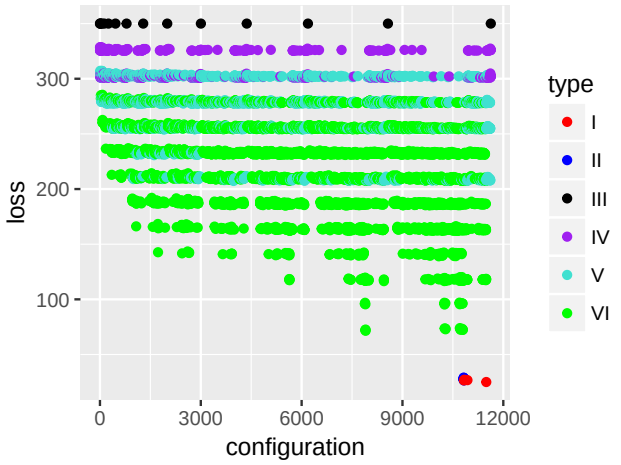

Figure 7: 4-skill case, slip=guess=0.01, J=5

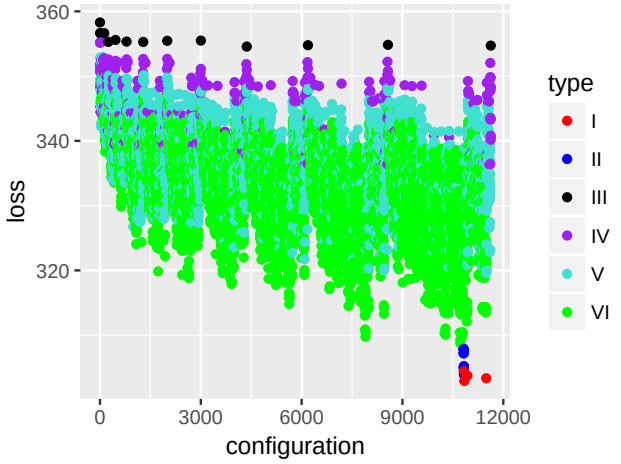

Figure 8: 4-skill case, slip=guess=0.2, J=5

models. *Journal of the American Statistical Association*, *110*(510), 850–866.

Chiu, C.-Y. (2013). Statistical refinement of the q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*(8), 598–618.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*(4), 633.

de la Torre, J., & Chiu, C.-Y. (2015). A general method of empirical q-matrix validation. *Psychometrika*, 1–21.

Desmarais, M. C., & Naceur, R. (2013). A matrix factorization method for mapping items to skills and for enhancing expert-based q-matrices. In *Artificial intelligence in education* (pp. 441–450).

Desmarais, M. C., Xu, P., & Beheshti, B. (2015). Combining techniques to refine item to skills q-matrices with a partition tree. In *Educational data mining 2015*.

Doroudi, S., & Brunskill, E. (2017). *The misidentified identifiability problem of bayesian knowledge tracing*. International Conference on Educational Data Mining, EDM2017.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of q-matrix. *Applied psychological measurement*, *36*(7), 548–564.

Liu, J., Xu, G., & Ying, Z. (2013). Theory of the self-learning q-matrix. *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability*, *19*(5A), 1790.

Qin, C., Zhang, L., Qiu, D., Huang, L., Geng, T., Jiang, H., ... Zhou, J. (2015). Model identification and q-matrix incremental inference in cognitive diagnosis. *Knowledge-Based Systems*, *86*, 66–76.

van De Sande, B. (2013). Properties of the bayesian knowledge tracing model. *Journal of Educational Data Mining*, *5*(2), 1–10.

Xu, G., & Zhang, S. (2015). Identifiability of diagnostic classification models. *Psychometrika*, 1–25.

Xu, P., & Desmarais, M. (2016). Boosted decision tree for q-matrix refinement. In *Edm* (pp. 551–555).