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 recent years, the identifiability problem of Q-matrix under DINA/DINO model has been proposed and researched. Strict statistical discussion and formulation have been given. One useful result is that under the DINA and DINO models, with Q-matrix, slip and guess being known, the population proportion parameter p is identifiable if and only if Q is complete(meaning it contains all unit vectors). In order to diagnose students, this condition should be used as a basic requirement for Q-matrix design. However, it does not suffice by just using a small Q-matrix which merely satisfies identifiability condition to offer a correct diagnosis for each individual student, particularly because the condition is about the parameters for the whole population. Fortunately, by increasing the number of questions being asked, i.e the number of q-vectors, we can improve the accuracy of estimation on student profiles. The ways to achieve a balance between conciseness and diagnosability of Q-matrix are discussed empirically in this paper.

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; 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 there is no slip or guess, and then the case that slip and guess exist but is known, and finally the case that slip and guess is unknown.

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 parameters p are also given.

2. IDENTIFIABILITY

3. EXPERIMENT

3.1 Simulations

3.1.1 Experiment 1: Comparison of three strategies

The Q-matrix used are,

Strategy 1: Repetition of e_j, or in other words, the identifiability conditions.

Strategy 2: Repetition of orthogonal arrays.

Strategy 3: Repetition of full combinations.

Q-matrix 1
$$\begin{array}{c} \\ q_1\\ q_2\\ q_3\\ q_4\\ q_5\\ q_6\\ q_7\\ q_8\\ q_9\\ q_{10}\\ q_{11}\\ q_{12}\\ q_{13}\\ q_{14}\\ q_{15}\\ q_{16}\\ q_{17}\\ q_{18}\\ q_{19}\\ q_{20}\\ q_{21} \end{array}
\begin{array}{c} \begin{array}{ccc} k_1 & k_2 & k_3 \end{array} \\
\begin{bmatrix} 1&0&0\\0&1&0\\0&0&1\\1&0&0\\0&1&0\\0&0&1\\1&0&0\\0&1&0\\0&0&1\\1&0&0\\0&1&0\\0&0&1\\1&0&0\\0&1&0\\0&0&1\\1&0&0\\0&1&0\\0&0&1\\1&0&0\\0&1&0\\0&0&1 \end{bmatrix} \end{array}$$

Q-matrix 2

$$\begin{array}{cc} & \begin{array}{ccc} k_1 & k_2 & k_3 \end{array} \\ \begin{array}{c} q_1 \\ q_2 \\ q_3 \\ q_4 \\ q_5 \\ q_6 \\ q_7 \\ q_8 \\ q_9 \\ q_{10} \\ q_{11} \\ q_{12} \\ q_{13} \\ q_{14} \\ q_{15} \\ q_{16} \\ q_{17} \\ q_{18} \\ q_{19} \\ q_{20} \\ q_{21} \end{array} & \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} \end{array}$$

Q-matrix 3(full)

$$\begin{array}{cc} & \begin{array}{ccc} k_1 & k_2 & k_3 \end{array} \\ \begin{array}{c} q_1 \\ q_2 \\ q_3 \\ q_4 \\ q_5 \\ q_6 \\ q_7 \\ q_8 \\ q_9 \\ q_{10} \\ q_{11} \\ q_{12} \\ q_{13} \\ q_{14} \\ q_{15} \\ q_{16} \\ q_{17} \\ q_{18} \\ q_{19} \\ q_{20} \\ q_{21} \end{array} & \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \end{array}$$

3.1.2 *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, if we have 7 q-vectors, and 4 questions to allocate them, then we have $\binom{4+7-1}{7-1} = 210$ possible configurations. For each configuration, we will calculate the MAP estimation for the category of each student, and evaluate the total accuracy.

In order to have a more refined analysis, we go deeper to check the accuracy for each profile pattern. For a 3 skill model, we have $2^3 = 8$ profile patterns. For these patterns, we want to check the precision and recall for each of them under every possible configuration. The result is reported by figures. And in order to show the usefulness of the identifiability condition, the configurations containing all the e_j and only use e_j will be denoted by red colour, while configurations containing all the e_j but used at least one other q-vectors are denoted by blue colour. J denotes the number of questions.

All the 8 patterns are,

$$\begin{array}{cc} & \begin{array}{ccc} k_1 & k_2 & k_3 \end{array} \\ \begin{array}{c} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \\ p_7 \\ p_8 \end{array} & \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \end{array}$$

4. RESULT

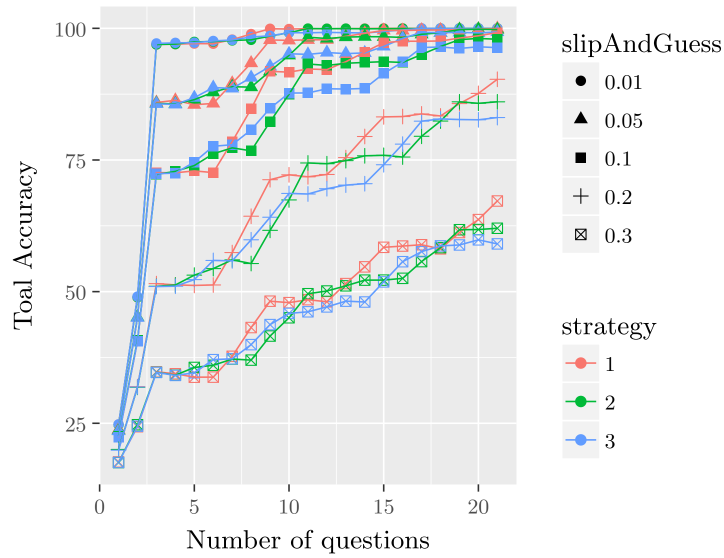

Figure 1: Strategy Comparison

5. DISCUSSION

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).

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

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).

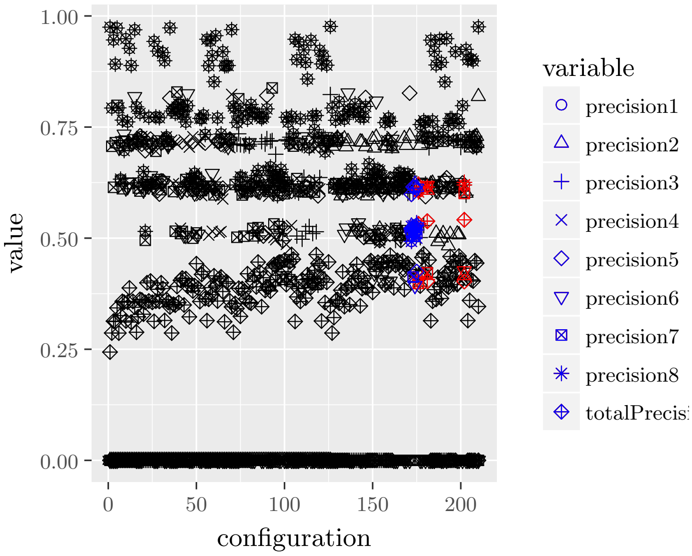

Figure 2: Precision on slip=guess=0.2, J=4

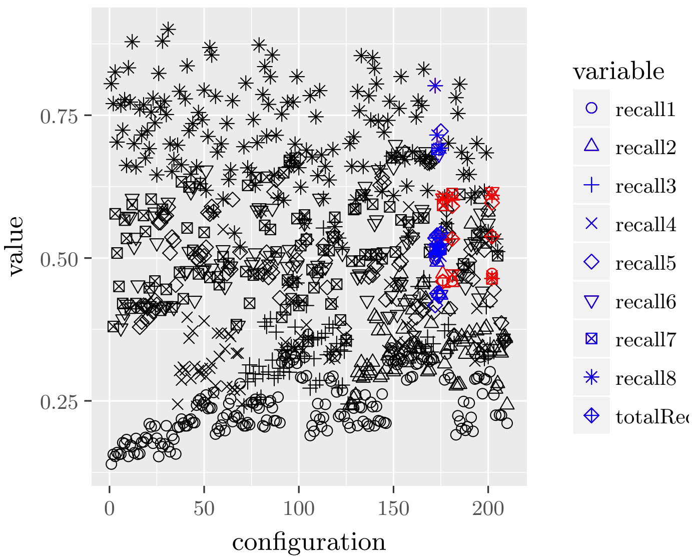

Figure 3: Recall on slip=guess=0.2, J=4

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*.

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

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

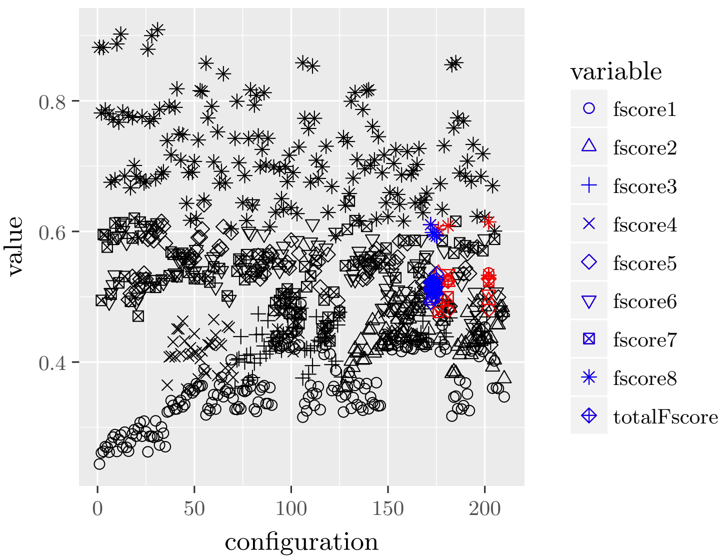

Figure 4: Fscore on slip=guess=0.2, J=4

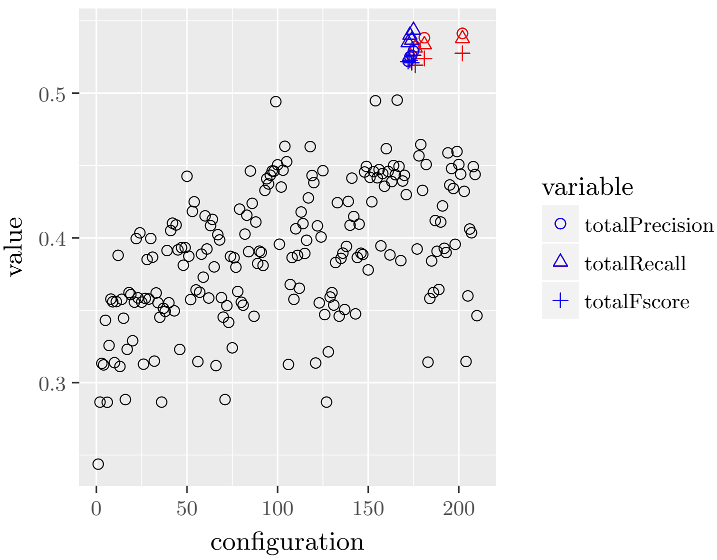

Figure 5: Total measures on slip=guess=0.2, J=4

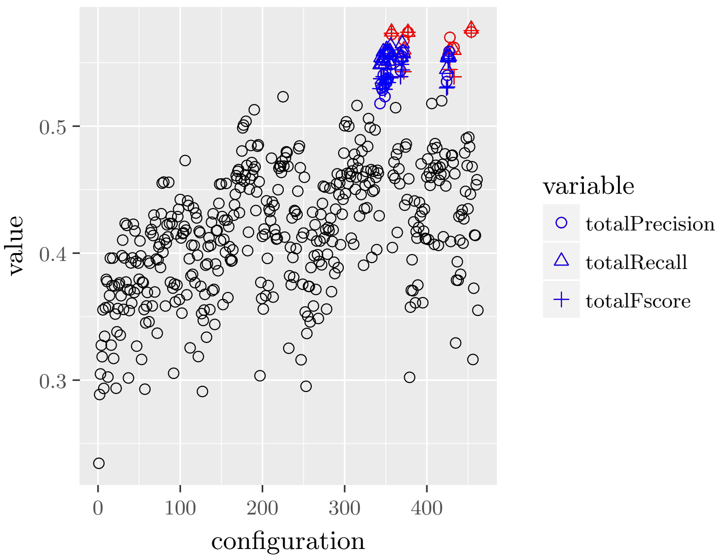

Figure 6: Total measures on slip=guess=0.2, J=5