A Matrix Factorization Method for Mapping Items to Skills and for Enhancing Expert-Based Q-matrices

Michel C. Desmarais and Rhouma Naceur

École Polytechnique de Montréal,
C.P. 6079, succ. Centre-ville
Montréal (Québec) H3C 3A7, Canada
{michel.desmarais,rhouma.naceur}@polymtl.ca

Abstract. Uncovering the right skills behind question items is a difficult task. It requires a thorough understanding of the subject matter and of the cognitive factors that determine student performance. The skills definition, and the mapping of item to skills, require the involvement of experts. We investigate means to assist experts for this task by using a data driven, matrix factorization approach. The two mappings of items to skills, the expert on one side and the matrix factorization on the other, are compared in terms of discrepancies, and in terms of their performance when used in a linear model of skills assessment and item outcome prediction. Visual analysis shows a relatively similar pattern between the expert and the factorized mappings, although differences arise. The prediction comparison shows the factorization approach performs slightly better than the original expert Q-matrix, giving supporting evidence to the belief that the factorization mapping is valid. Implications for the use of the factorization to design better item to skills mapping are discussed.

Keywords: student models, skills assessment, alternating least squares matrix factorization, latent skills, cognitive modeling

1 Introduction

Mapping items to latent skills is a notoriously difficult task and intelligent help to alleviate this difficulty would obviously be desirable. Although the complete automation of uncovering the skills behind question items for cognitive engineering purpose is beyond reach in the current state of research, means to help determine the number of skills and the common skills between items is a reasonable endeavour in the mid-term, and significant advances have been made recently. We review the state of the art towards this goal in recent years, and demonstrate how a matrix factorization technique can yield promising results to this end.

2 Skills Modeling, Q-matrices, and Matrix Factorization

Because of its importance, the problem of mapping items to underlying skills has been widely studied, and is still an ongoing topic of investigation in psychometrics and in educational data mining (see, for eg., [9, 12, 11, 7, 4, 3, 1] for recent contributions).

In the past ten years, a few groups of researchers have looked at *linear models* of item to skills mapping and of skills assessment, with promising results. We build upon this work which is briefly reviewed below.

2.1 Linear models

Linear models of skills are familiar to most teachers. An exam's weighted sums of individual score items, broken down by topic (skill), implicitly constitute a linear model model. Also highly familiar in the psychometric field is the Q-matrix formalism, investigated by Tatsuoka and her colleagues in the early 1980's, which maps skills to items [15, 16]. This formalism can also be considered a close parent of linear models.

Linear models were put to the task of assessing student skills mastery [2, 20, 19, 18]. In the 2010 KDD Cup, a tensor model was developed to model student skills and the mapping of items to skills. Thai-Nghe et al. used a multi-relational matrix and tensor-based factorization to model skills and learning curves to predict student success [19, 18]. A comparison with the widely recognized Bayesian Knowledge Tracing approach showed that it compares favorably [19].

The success of linear models and factorization methods raises the question of whether these methods could also be successful in deriving Q-matrices that maps items to skills. A few studies have shown that a mapping can, indeed, be derived from data [21, 6]. Winters et al. showed that item topic extraction can be obtained from different data mining techniques, one of which is matrix factorization [21]. However, only very distinct topics like French and mathematics can yield adequate mapping. This study was later corroborated by Desmarais [6] who also used simulated data to show that the low discrimination power of some topics might be explained by their lower weight in terms of skill factors, when compared to other factors such as item difficulty and student ability. Recent work by Lan et al. [10] combine a factor analysis framework, named SPARFA, with Bayesian techniques to uncover skills behind task and to label these skills from tags and from the question item texts.

The factorization methods in the studies mentioned above rely on the standard matrix operators ("dot product"), and therefore can be considered as *compensatory models* of skills: each skill required adds to the chances of success of an item. Barnes, Stampers, and other colleagues [1, 14] introduced a different algorithm to implement *conjunctive models* of skills, where any required skill missing will induce a failure to the item. We will borrow from this work and from [8] to implement both conjunctive and compensatory models in the current study. The foundations of these models is explained next.

2.2 Results matrix, Q-Matrix, and skills matrix

Student test data can be represented in the form of a results matrix, $\mathbf{R}$, with m row items by n column students. We use the term *item* to represent exercises, questions, or any task where the student has to apply a skilled performance to accomplish it correctly. If a student successfully answers an item, the corresponding value in the results matrix is 1, otherwise it is 0. Intermediate values could also be used to indicate partial success.

A results matrix $\mathbf{R}$ can be decomposed into two smaller matrices:

$$\mathbf{R} \approx \mathbf{Q}\,\mathbf{S} \tag{1}$$

The process of matrix factorization is to determine the matrices $\mathbf{Q}$ and $\mathbf{S}$ from $\mathbf{R}$. The $\mathbf{Q}$ matrix is equivalent in form to the Q-matrix developed in the cognitive modeling field [16, 15], although various semantics apply to each formalism, such as the *conjunctive* or *compensatory* versions explained below. This matrix is an m items by k skills matrix that defines which skills are necessary to correctly answer an item. It allows a "compressed" representation of the data that assumes the item outcome results are determined by the skills involved in each item and the skills mastered by each student. The k skills by n student matrix $\mathbf{S}$ represents the student skills mastery profiles. The product of $\mathbf{Q}$ and $\mathbf{S}$ yields an estimated results matrix $\hat{\mathbf{R}}$. The goal of factorization algorithms is to minimize $||\hat{\mathbf{R}} - \mathbf{R}||$.

As mentioned above, the Q-matrix ($\mathbf{Q}$) can take different interpretations. A *conjunctive* Q-matrix assumes *all* skills in an item row are necessary for success, whereas a *disjunctive* Q-matrix assumes *any* skill is sufficient, and finally a *compensatory* Q-matrix assumes each skill *adds* to item success, which can be interpreted as increasing the chances of success if each item is either succeeded or failed. Equation (1) corresponds to the *compensatory* version of the Q-matrix, but it can be transformed into a *conjunctive* version through negation of the $\mathbf{R}$ and $\mathbf{S}$ matrices [8].

3 Comparing a Q-matrix Induced from Data with an Expert Defined Matrix

Given the factorization obtained from equation (1), the question we address here is how to compare the matrix $\mathbf{Q}$ obtained from item outcome data, with an expert defined Q-matrix, in the hope that this comparison can help validate and improve the expert matrix.

3.1 Comparison Issues and Principle of the Proposed Method

One issue with the factorization of equation (1) is the interpretation of the $\mathbf{Q}$ matrix obtained. Although factorization techniques allow, or require, the specification of the number of skills, k, the skills appear in matrix $\mathbf{Q}$ in some unpredictable order. Moreover, the matrix can contain numerical values of various signs and amplitude that may not lend themselves to a sharp interpretation.

Another issue has to do with the factorization technique used. Some techniques, such as non-negative matrix factorization (NMF), lead to non unique and to local minima solutions. Experience shows that these solutions can be widely different, worsening the problem of interpretation and comparison with the expert Q-matrix.

To alleviate these issues, we rely on the principle of starting the factorization process with an initial matrix $\mathbf{Q}$ set to the expert Q-matrix. Many factorization algorithms could be used, as long as this condition can be met. The initial condition ensures that the matrix $\mathbf{Q}$ obtained after minimizing $||\hat{\mathbf{R}} - \mathbf{R}||$ will minimally diverge from the initial one, thereby rendering the comparison with, and enhancement of the expert's Q-matrix more feasible.

3.2 Alternate Least-square Factorization (ALS)

The factorization method we use is the Alternate Least-square (ALS). Starting with the results matrix $\mathbf{R}$ and an initial Q-matrix, $\mathbf{Q}_0$, a least-squares estimate of the skills matrix $\hat{\mathbf{S}}_0$ can be obtained by:

$$\hat{\mathbf{S}}_0 = (\mathbf{Q}_0^{\mathrm{T}}\,\mathbf{Q}_0)^{-1}\,\mathbf{Q}_0^{\mathrm{T}}\,\mathbf{R} \tag{2}$$

The initial matrix $\mathbf{Q}_0$ will be the expert defined Q-matrix. Then, a new estimate of the Q-matrix, $\hat{\mathbf{Q}}_1$, is again obtained by the least-squares estimate:

$$\hat{\mathbf{Q}}_1 = \mathbf{R}\,\hat{\mathbf{S}}_0^{\mathrm{T}}\,(\hat{\mathbf{S}}_0\,\hat{\mathbf{S}}_0^{\mathrm{T}})^{-1} \tag{3}$$

And so on for estimating $\hat{\mathbf{S}}_1$, $\hat{\mathbf{Q}}_2$, etc. Alternating between equations (2) and (3) yields progressive refinements of the matrices $\hat{\mathbf{Q}}_i$ and $\hat{\mathbf{S}}_i$ that more closely approximate $\mathbf{R}$ in equation (1). In our experiments, the convergence at a delta of 0.001 occurs after 7–8 iterations and in a fraction of a second for factorizing a matrix of dimension 20×536. This performance makes the technique many times more efficient than factorizations that rely on gradient descent, for example.

It is worth mentioning that, by starting with non negative matrices $\mathbf{Q}_0$ and $\mathbf{R}$, the convergence process will generally end with positive values for both matrices $\mathbf{Q}_i$ and $\mathbf{S}_i$. The vast majority of values obtained are between -0.5 and 1.5 if both the results matrix and the initial Q-matrix have {0,1} values. No regularization terms are used in the current implementation of the algorithm to force non-negative or integer values.

4 Experiments and Data

We use the ALS method described above to compare an expert defined Q-matrix and a factor Q-matrix. Unless otherwise mentioned, factorization is based on the conjunctive model of skills (see [8]), which essentially consists in using the negation of the $\mathbf{R}$ matrix instead of the raw values.

The data comes from Tatsuoka's fraction algebra problems [17] which is available through the R package CDM [13]. It is composed of 20 question items

and 536 respondents (see table 1 in [4] for a description of the problems and of the skills).

When cross-validation experiments are performed, they consists in breaking down the data into 8 sets of 67 students each. Training is done on 7 of the 8 sets and testing on the remaining set.

4.1 Visual comparison of Q-matrices

Figure 1 shows three versions of the Q-matrices. The left matrix is the one defined by the expert, as provided in [13]. Dark cells represent the required skills. The middle matrix is derived from the full data set. The gradients of colors represent the values that range between -0.5 and 1.5, where the darker color indicate higher values. The right matrix is the rounded version of the middle matrix: Real values of the middle matrix are rounded to 0 or 1.

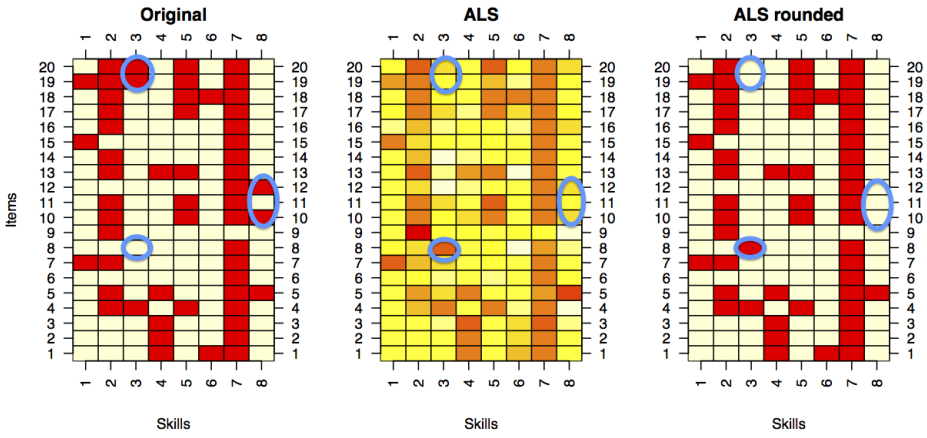

Fig. 1. Three Q-matrices of Tatsuoka's fraction algebra items.

Five cells differ between the expert (left) and the ALS factorization (right) matrices. They are highlited by three ellipses. Except for cell (8,3), all of the differences are cells missing on the ALS matrix. We could bring back the missing values by tweaking the threshold to the 0/1 function, but that would come at the cost of creating false positives. Their absence in the ALS factorized matrix suggests that the corresponding skills may not contribute to the response outcome as much as the other skills. Equally interesting are the different color brightnesses in the true positive, suggesting that some skills may be more important than others. Finally, we note that the differences all come from only 2 skills (see [4] for a description of all skills):

(3): simplify before subtracting (eg. $3\frac{1}{2} - 2\frac{3}{2}$, $4 - 1\frac{4}{3}$, $4\frac{1}{3} - 1\frac{5}{3}$),

(8): reduce answers to simplest form (eg. $4\frac{3}{5} - 3\frac{4}{10}$, $4\frac{1}{2} - 2\frac{7}{12}$, $1\frac{1}{8} - \frac{1}{8}$).

The high level of discrepancies between the matrices for these two skills may hint at some issues with these particular skills. This observation is congruent with different analyis by DeCarlo of Tatsuoka's Q-matrix and based on the DINA latent factor model which identifies Skill 3 as a source of error: "Together, these results suggest that the Q-matrix might be misspecified, in that Skill 3 should not be included" (in [5], p. 20).

4.2 Validity of the ALS Q-matrix

The ALS method clearly meets the criteria of interpretation and ease of comparison with an expert defined Q-matrix. However, does the ALS Q-matrix represent a "better" mapping of skills to items than the expert's, or even a "good" mapping at all?

Let us define the goodness of a Q-matrix by its ability to make accurate predictions. Accordingly, we compare the expert Q-matrix and the ALS Q-matrix over their performance for predicting response outcomes.

A cross-validation simulation is conducted and consists in predicting, in turn, each of the 20 items given the answers of the other 19 items. The individual respondents are split into 8 bins. The data from 7 bins serve for the training (deriving the ALS Q-matrices) and the remaining bin serves for testing. 8-folds simulations are conducted, one for each bin. The skills of each examinee are assessed from 19 items based on the Q-matrix of the training phase. The item that remains is used for testing prediction.

Skills are assessed according to equation (2). However, for skills assessment, the Q-matrix requires response vectors of 20 items, whereas only 19 are given. Therefore, the expected value is used in place of the missing item outcome to predict: the geometric mean of the average item difficulty and the examinee ability over the 19 given items is used (the value is not rounded to 0/1). Then, the predicted item outcome is computed according to equation (1).

To assess the performance of the original, expert Q-matrix, the same process as described above is used, except that there is no training phase. The expert Q-matrix is used in place of the ALS Q-matrix derived from training.

The results of the simulation are reported in figure 2. Results from both conjunctive and compensatory models are reported. The predictions based on expected values are also reported. Expected values are computed according to the method described above when assigning the value of the item to predict for the ALS Q-matrix predictions. Note that the average success rate is 53%, and therefore a *0-rule* type of prediction (predicting all 1s) would yield 47% MAE.

The lower MAEs of the ALS Q-matrix, compared to the Original expert Q-matrix in both the conjunctive and compensatory models, provide support for the validity of the ALS Q-matrix. Not surprisingly, the expert Q-matrix performs better under the conjunctive model than the compensatory. This is expected to the extent that it was designed by experts as a conjunctive rather than a

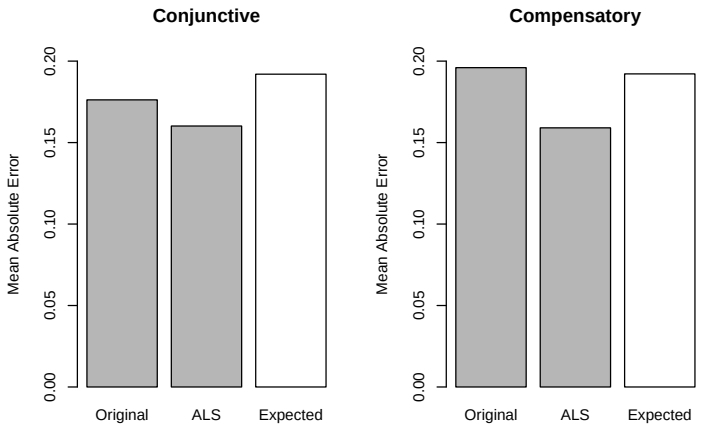

Fig. 2. Mean Absolute Error (MAE) predictive accuracy of Q-matrices with conjunctive and compensatory models. Predictions based on expected values is also provided for comparison. Standard deviation for ALS is 0.012 for the 8-folds simulation and 0.013 for the Original. A paired t-test of the conjunctive model shows the difference is statistically significant ($p = 0.001$, pairing is per fold).

compensatory model. However, the ALS Q-matrix predictions have practically the same accuracy for both models.

Turning to the question of whether the expert and ALS Q-matrices are "good" at all, we compare the predictive performance of ALS Q-matrices derived from the expert Q-matrix with ALS Q-matrices derived from random starting points.

We computed the MAE of ALS Q-matrices for randomly generated initial Q-matrices. The MAE for this experiment based on 10-folds is 0.159 (sd. 0.001), which is practically the same as the ALS Q-matrices obtained when the starting Q-matrix is the expert one. However, convergence is slower, requiring between 8 and 14 iterations.

The fact that the MAE is relatively similar regardless of the initial Q-matrix further supports the belief that the ALS Q-matrix obtained from starting with the expert one is valid and could be regarded as a legitimate improvement.

4.3 Convergence/Divergence from Original Matrices

It is comforting to believe that there is one "true" Q-matrix for a given set of items and skills, and that, given a close approximation of this matrix, there exists a means to converge towards this true matrix and avoid divergence away from it. If the ALS factorization method allowed such outcome, it would truly offer useful guidance for the design of Q-matrices by reliably indicating the "faulty" cells regardless of which they are in the matrix.

Table 1. Discrepancies as a function of the number of perturbations.

Number of	**Number of discrepancies**																
perturbations	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	(1) Discrepancies with original ALS Q-matrix																
1	42	50	28	16	9	10	2	2	1	0	0	0	0	0	0	0	0
2	11	27	31	30	26	10	8	6	4	3	2	1	1	0	0	0	0
3	3	9	14	25	24	20	20	11	12	10	6	2	1	1	1	1	0
4	1	2	8	15	12	19	25	20	19	11	10	9	5	3	0	0	1
	(2) Discrepancies with Original Q-matrix																
1	0	0	0	0	0	47	55	33	17	8	0	0	0	0	0	0	0
2	0	0	0	0	0	18	20	36	45	22	12	5	1	0	1	0	0
3	0	0	0	0	0	5	21	29	35	27	21	15	5	1	1	0	0
4	0	0	0	0	1	3	7	15	34	27	28	17	18	4	4	1	1

To explore this conjecture, we design an experiment where perturbations are introduced in the Original expert Q-matrix, and perform ALS factorization on this corrupted matrix. If we had started with a perfect Q-matrix, we would like the method to detect the perturbations and return the original Q-matrix. If we had started with a close approximation, we would expect the ALS Q-matrix derived from the corrupted matrix to still converge towards the same, hopefully "best" Q-matrix as the one obtained without perturbations.

Table 1 reports the results of this experiment. From 1 to 4 random perturbations are introduced in the Original expert matrix, $\mathbf{Q}_0$. With 1 perturbation, all 160 values of the 20 items $\times$ 8 skills Q-matrix are changed. With 2 and more perturbations, 160 random samples of combinations of values in the Q-matrix are changed. The table reports the number of discrepancies of the derived ALS Q-matrix between (1) the original ALS Q-matrix and (2) between the Original Q-matrix. Each row contains 160 values and the frequency of discrepancies from 0 to the observed maximum of 16 are reported. A line is drawn at the value of 5 discrepancies as a reminder of the original number of discrepancies.

We observe that with 1 perturbation, 42 of the 160 ALS Q-matrix derived are identical to the one derived with the unperturbed expert Q-matrix, and 50 show 1 discrepancy. This leaves 68 ALS Q-matrices that have 2 or more discrepancies, i.e. more discrepancies than perturbations introduced.

This trend increases with the number of perturbations: with 4 perturbations induced, only 28 ALS Q-matrices show 4 or less discrepancies, which leaves 132 with more discrepancies than the number of perturbations introduced.

Comparing the ALS Q-matrices with the Original expert Q-matrix, we see that for 1 perturbation, 47 of the 160 ALS Q-matrices derived correspond to the Original Q-matrix. For these matrices, the perturbation was removed. For the remaining 113, they show 6 to 9 discrepancies. However, given that for 1 perturbation, only 5 ALS Q-matrices diverge from the original ALS Q-matrix, this means that the overwhelming bulk of the discrepancies introduced are in fact

changes towards the original ALS Q-matrix. The same argument can be made for 2 and for 3 perturbations, albeit to a lesser extent. Therefore, small perturbations still result in inducing ALS Q-matrices that converge towards the original ALS Q-matrix induced with the expert Q-matrix as a starting point.

However, starting at 4 perturbations, we see more divergences that are not aligned with the original ALS Q-matrix. Nevertheless, even at this number of perturbations, the large majority of the 160 cells remain intact, and so does the majority of the 56 cells which have a value of 1.

5 Discussion

The ALS factorization method offers a promising means of deriving Q-matrices from data given an expert defined Q-matrix to start with. One important advantage of this method is that it lends itself to an unambiguous comparison with the initial expert Q-matrix, and consequently to a clear interpretation.

The fact that the ALS Q-matrix derived generates slightly better predictive item outcome performance supports the hypothesis that the discrepancies between the this matrix and expert matrix are potentially valuable hints towards improving the expert Q-matrix.

The exploration of the space of Q-matrices through the experiment with perturbations showed that, up to 2 or 3 changes in an initial Q-matrix of the ALS factorization, the changes induced converge towards the original ALS factorization. This result suggests that a small number of errors will not affect the method's capacity to derive "better" Q-matrices (as defined by their predictive power) and make useful hints for enhancements.

In spite of these encouraging results, this study is limited to a single expert Q-matrix. Generalization to different dimensions of Q-matrices and different domains remain unknown and further studies are called for. Furthermore, a more in-depth, qualitative, and domain expert analysis of the discrepancies would be highly useful to better understand the results and assess the value of the method.

References

1. Barnes, T.: Novel derivation and application of skill matrices: The Q-matrix method. Handbook on Educational Data Mining (2010)
2. Cetintas, S., Si, L., Xin, Y.P., Hord, C.: Predicting correctness of problem solving in ITS with a temporal collaborative filtering approach. In: Aleven, V., Kay, J., Mostow, J. (eds.) Intelligent Tutoring Systems, 10th International Conference, ITS 2010, Pittsburgh, PA, USA, June 14-18, 2010, Proceedings, Part I. Lecture Notes in Computer Science, vol. 6094, pp. 15–24. Springer (2010)
3. De La Torre, J.: 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. DeCarlo, L.T.: 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)

5. DeCarlo, L.T.: On the analysis of fraction subtraction data: The dina model, classification, latent class sizes, and the q-matrix. Applied Psychological Measurement 35(1), 8–26 (2011)
6. Desmarais, M.C.: Conditions for effectively deriving a q-matrix from data with non-negative matrix factorization. In: Conati, C., Ventura, S., Calders, T., Pechenizkiy, M. (eds.) 4th International Conference on Educational Data Mining, EDM 2011. pp. 41–50 (Eindhoven, Netherlands, June 6–8 2011)
7. Desmarais, M.C.: Mapping question items to skills with non-negative matrix factorization. ACM KDD-Explorations 13(2), 30–36 (2011)
8. Desmarais, M.C., Beheshti, B., Naceur, R.: Item to skills mapping: Deriving a conjunctive q-matrix from data. In: 11th Conference on Intelligent Tutoring Systems, ITS 2012. pp. 454–463 (Chania, Greece, 14–18 June 2012 2012)
9. Koedinger, K.R., McLaughlin, E.A., Stamper, J.C.: Automated student model improvement. In: Proceedings of the 5th International Conference on Educational Data Mining. pp. 17–24 (2012)
10. Lan, A.S., Waters, A.E., Studer, C., Baraniuk, R.G.: Sparse factor analysis for learning and content analytics. arXiv preprint arXiv:1303.5685 (2013)
11. Li, N., Cohen, W.W., Matsuda, N., Koedinger, K.R.: A machine learning approach for automatic student model discovery. In: Proceedings of the 4th International Conference on Educational Data Mining. pp. 31–40 (2011)
12. Liu, J., Xu, G., Ying, Z.: Data-driven learning of q-matrix. Applied Psychological Measurement 36(7), 548–564 (2012), `http://apm.sagepub.com/content/36/7/548.abstract`
13. Robitzsch, A., Kiefer, T., George, A., Uenlue, A., Robitzsch, M.: Package CDM (2012), `http://cran.r-project.org/web/packages/CDM/index.html`
14. Stamper, J.C., Barnes, T., Croy, M.J.: Extracting student models for intelligent tutoring systems. In: AAAI 2007. pp. 1900–1901. AAAI Press (2007)
15. Tatsuoka, K.K.: Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement 20, 345–354 (1983)
16. Tatsuoka, K.: Cognitive Assessment: An Introduction to the Rule Space Method. Routledge Academic (2009)
17. Tatsuoka, K., of Illinois at Urbana-Champaign. Computer-based Education Research Laboratory, U., of Education (US), N.I.: Analysis of errors in fraction addition and subtraction problems. Computer-based Education Research Laboratory, University of Illinois (1984)
18. Thai-Nghe, N., Drumond, L., Horváth, T., Nanopoulos, A., Schmidt-Thieme, L.: Matrix and tensor factorization for predicting student performance. In: Verbraeck, A., Helfert, M., Cordeiro, J., Shishkov, B. (eds.) CSEDU 2011 - Proceedings of the 3rd International Conference on Computer Supported Education, Volume 1, Noordwijkerhout, Netherlands, 6-8 May, 2011. pp. 69–78. SciTePress (2011)
19. Thai-Nghe, N., Horváth, T., Schmidt-Thieme, L.: Factorization models for forecasting student performance. In: Conati, C., Ventura, S., Pechenizkiy, M., Calders, T. (eds.) Proceedings of EDM 2011, The 4th International Conference on Educational Data Mining. pp. 11–20. www.educationaldatamining.org (Eindhoven, Netherlands, July 6–8 2011)
20. Toscher, A., Jahrer, M.: Collaborative filtering applied to educational data mining. Tech. rep., KDD Cup 2010: Improving Cognitive Models with Educational Data Mining. (2010)
21. Winters, T.: Educational Data Mining: Collection and Analysis of Score Matrices for Outcomes-Based Assessment. Ph.D. thesis, University of California Riverside (2006)