Conditions for effectively deriving a Q-Matrix from data with Non-negative Matrix Factorization

MICHEL C. DESMARAIS, Polytechnique Montréal

The process of deciding which skills are involved in a given task is tedious and challenging. Means to automate it are highly desirable, even if only partial automation that provides supportive tools can be achieved. A recent technique based on Non-negative Matrix Factorization (NMF) was shown to offer valuable results, especially due to the fact that the resulting factorization allows a straightforward interpretation in terms of a Q-matrix. We investigate the factors and assumptions under which NMF can effectively derive the underlying high level skills behind assessment results. We demonstrate the use of different techniques to analyse and interpret the output of NMF. We propose a simple model to generate simulated data and to provide lower and upper bounds for quantifying skill effect. Using the simulated data, we show that, under the assumption of independent skills, the NMF technique is highly effective in deriving the Q-matrix. However, the NMF performance degrades under different ratios of variance between subject performance, item difficulty, and skill mastery. The results corroborates conclusions from previous work in that high level skills, corresponding to general topics like World History and Biology, seem to have no substantial effect on test performance, whereas other topics like Mathematics and French do. The analysis and visualization techniques of the NMF output, along with the simulation approach presented in this paper, should be useful for future investigations using NMF for Q-matrix induction from data.

1. INTRODUCTION

The construction of a Q-matrix from data is a highly desirable goal for tutoring systems. Not only would it waive the expertise and labour intensive task of assigning which skills are involved in which task, but it would also offer a more objective and replicable means of getting the correct skill-to-task mapping. Furthermore, it might also allow a more effective means to build Q-matrices, as machine learning methods often outperform humans over a range of complex tasks.

However, the success in achieving this goal remains limited. Nowadays, we find no reliable method to automate the mapping of skills to tasks from data, but some progress has been made.

Working with log data from tutoring systems, data which is characterized by the fact that the knowledge state of the student dynamically changes in the data as the student learns, Cen et al. [2006; 2005] have used a technique known as Learning Factor Analysis (LFA) in order to bring improvements over an initially hand built Q-matrix (also termed a *transfer model*). This technique was shown useful for bringing improvements to the Q-matrix composed of fine-grained skills which are deemed necessary to complete certain exercises.

Inspired from the work of Tatsuoka [1983], Barnes [2006] developed a method of mapping skills to items based on a measure of the fit of a potential Q-matrix to the data. This method and the other methods described below rely on static student knowledge states, as opposed to the dynamically changing knowledge states of the LFA technique. Barnes method is fully automated and it was shown to perform at least as well as Principal Component Analysis for skill clustering analysis. However, it involves an algorithm that does not scale well to a Q-matrix that comprises 20 or more items.

Winters et al. [2005] investigated how a number of standard clustering techniques can effectively match skills to test items. They applied these techniques to a wide array of test outcomes, from SAT topics such as Mathematics, Biology and French, to computer science exams, and to different trivia topics. Their findings show that for skills associated to topics within a single course, for example, the techniques were essentially no better at classifying test items than random clustering. The same conclusion applies for topics like World history and Biology. However, the techniques were relatively successful at separating items that belongs to totally different topics, such as Mathematics and French.

In this paper, we replicate parts of the study by Winters et al. [2005] and focus on one of the cluster algorithms they used, Non-negative Matrix Factorization (NMF). We use visualization techniques to analyze in greater details the results of the factorization. We propose a model to simulate student data and show that the NMF technique is indeed effective under certain assumptions. We use the simulation data model parameters as a means to quantify and estimate the effect of skills over the observed examinee performance in some of the real data of Winters et al. original study. First, let us give some details about NMF.

2. NON-NEGATIVE MATRIX FACTORIZATION AND Q-MATRIX INTERPRETATION

Non-negative matrix factorization (NMF) decomposes a matrix into two smaller matrices. It is used for dimensionality reduction, akin to Principal Component Analysis and Factor analysis. NMF decomposes a matrix of $n \times m$ positive numbers, $\mathbf{V}$, as the product of two matrices:

$$\mathbf{V} \approx \mathbf{WH} \tag{1}$$

The matrices $\mathbf{W}$ and $\mathbf{H}$ are respectively $n \times r$ and $r \times m$, where r is called the rank of the factorization. For our purpose, matrix $\mathbf{V}$ represents the observed test outcome data for n question items and m respondents. Therefore, the product of $\mathbf{W}$ and $\mathbf{H}$ reproduces the observed patterns of success/failures of the m examinee to the n items. The matrix $\mathbf{W}$ can be considered as a Q-matrix, whereas $\mathbf{H}$ can be considered as the skills mastery for each m examinee. In the case of a Q-matrix, r represents the number of skills, which can take any value but should normally conform to: $r < nm/(n+m)$ [Lee and Seung 1999].

Let us take an example to better explain NMF in our context. Assume the following Q-matrix, $\mathbf{W}$, composed of 3 skills and 4 items, and the following skills mastery matrix, $\mathbf{H}$, for 5 examinees:

$$\mathbf{W} = \begin{array}{cc} & \text{skills} \\ \text{items} & \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array} \qquad \mathbf{H} = \begin{array}{cc} & \text{examinees} \\ \text{skills} & \begin{pmatrix} 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix} \end{array}$$

Given this Q-matrix and the skills mastered by the 5 examinees, the expected results are:

$$\mathbf{V} = \mathbf{WH} = \begin{array}{cc} & \text{examinees} \\ \text{items} & \begin{pmatrix} 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix} \end{array}$$

For example, taking the first item and the first examinee, we have, from $\mathbf{W}$, that item 1 requires skill 2, but, from $\mathbf{H}$, we see that examinee 1 only masters skill 1, therefore item 1 is failed by examinee 1. In fact, examinee 1's only success is over item 3 since all other items require either skills 2 or 3.

It is important to emphasize that there are many solutions to $\mathbf{V} = \mathbf{WH}$. For example, the same results as those above can be obtained with different Q-matrix and skills matrix:

$$\begin{array}{cc} & \text{examinees} \\ \text{items} & \begin{pmatrix} 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix} \end{array} = \begin{array}{cc} & \text{skills} \\ \text{items} & \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 1 \\ \frac{1}{2} & 0 & 0 \end{pmatrix} \end{array} \begin{array}{cc} & \text{examinees} \\ \text{skills} & \begin{pmatrix} 0 & 2 & 0 & 2 & 2 \\ 0 & 0 & 2 & 0 & 2 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Notice that the weights are changed as well as the ordering of rows and columns compared to the first solution. Nevertheless, it remains a valid factorization of $\mathbf{V}$ that could be derived by some NMF algorithm.

Indeed, there are many NMF algorithms that were developed since its introduction by Lee and Seung [1999] and they can yield different solutions. We refer the reader to Berry et al. [2007] for a more thorough and recent review of this technique which has gained strong adoption in many different fields.

Whereas the other matrix factorization techniques often impose constraints of orthogonality among factors, NMF imposes the constraint that the two matrices, $\mathbf{W}$ and $\mathbf{H}$, be non-negative. This constraint makes the interpretation much more intuitive in the context of using this technique for building a Q-matrix. It implies that the skills (latent factors) are additive "causes" that contribute to the success of items, and that they can only increase the probability of success and not decrease it, which makes good sense for skill factors. Note that negative values in $\mathbf{W}$ can be interpreted as misconceptions and would lower the expected score to items, but allowing negative values in the factorization also opens up the space of possible solutions and raises the issue of convergence and of the multiplicity of solutions, making the interpretation of $\mathbf{W}$ much more speculative.

The non-negative constraint and the additive property of the skills bring a specific interpretation of the Q-matrix. For example, if an item requires skills a and b with the same weight each, then each skill will contribute equally to the success of the item. This corresponds to the notion of a *compensatory* or *additive* model of skills.

In our study, we focus on high level skills, which we term *topic skills*. However, if an item requires two specific lower level skills, such as mastery of the rules $a/b + c/b = (a + b)/c$ and $a/b \cdot b = a$, a *conjunctive* model would be necessary, indicating that a failure is expected if any skill is not mastered. The standard interpretation of the Q-matrix corresponds to the conjunctive model, and the $\mathbf{W}$ matrix of NMF does not correspond to this interpretation, unless and as mentioned, we assume that each item belongs to a single skill and for which case the two interpretations are indiscernible.

A last remark on NMF: as mentioned above, the factorization can produce multiple solutions, even with a sigle algorithm, which raises the issue of stability of the results. However, Schachtner et al. [2010] discuss this issue and suggest that for binary data the problem may not appear at all. Nevertheless, we will assess the extent to which the multiple solution issue impacts the validity and usefullness of the approach by running multiple folds simulations.

3. Q-MATRIX EXTRACTION FROM SIMULATED DATA

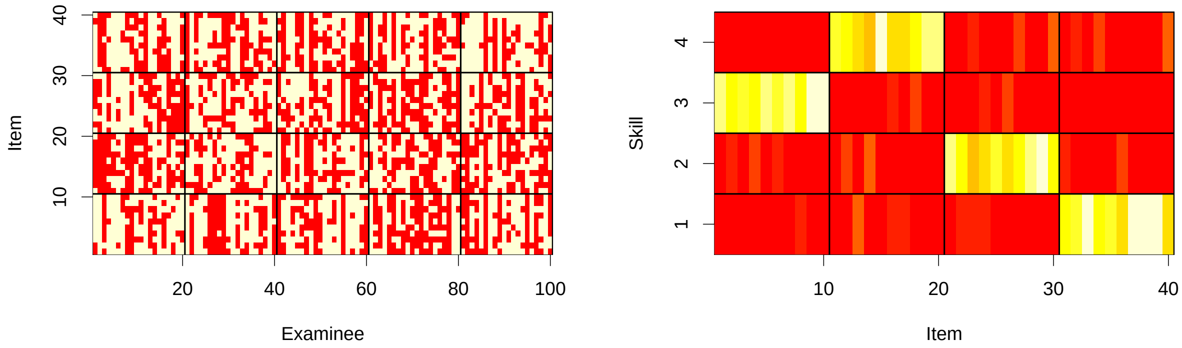
Let us start with an assessment of the validity of the NMF technique to extract the Q-matrix from simulated data and ascertain under which assumptions its effectiveness can be shown.

For the sake maintaining the similarity with real data analyzed later in this paper, let us use a 4 skill Q-matrix. Under the assumption that the topic (skill) is the only factor that affects performance and that each item depends on a single topic, the simulated data for 40 items and 100 examinees can be generated from a matrix 40×100, $\mathbf{P}$, where each column contains 40 probabilities, one probability per item, structured as a sequence of 10×4 probabilities:

$$(p_{1,1}, p_{1,2}, ..., p_{1,10}, p_{2,1}, ..., p_{2,10}, p_{3,1}, ..., p_{3,10}, p_{4,1}, ..., p_{4,10})$$

where $p_{1,1}$ to $p_{1,10}$ are all equal, $p_{2,1}$ to $p_{2,10}$ are all equal, and so on. Each column contains therefore only 4 distinct and independent probabilities, one for each skill. These probabilities are generated from a random variable, z, taken from a normal standard distribution and transformed into a probability by computing the *cumulative distribution function* (the area $[-\infty, z]$).

Given the probability matrix $\mathbf{P}$, a data matrix having the same dimensions as $\mathbf{P}$ is generated, $\mathbf{D}$, by sampling a success or failure, $\{0, 1\}$ using $P_{i,j}$ as the probability of success and $1 - P_{i,j}$ for failure. The matrix $\mathbf{D}$ corresponds to $\mathbf{V}$ in equation (1). A sample of this data is provided in figure 1(a). By grouping



(a) Simulated item outcome data of 40 questions and 100 examinees.

(b) Image output of Q-matrix from NMF for 4 skills and 40 question items.

Fig. 1. Simulated data (a) and the corresponding Q-matrix (b) under the assumption that topic is the only factor that affects success. Skill mastery follows a standard normal distribution. A perfect match from items to skills is obtained with this Q-matrix.

items in 4 contiguous groups of 10, the effect of the different levels of skill is apparent: a high probability of mastery will appear as a vertical pattern (single examinee) consisting mostly of pale square dots between vertical stretches of 10 contiguous items, whereas a low probability appears as a pattern composed of mostly dark red dots. No horizontal pattern is apparent since we do not define an item difficulty factor in this data. Similarly, no vertical pattern is apparent *across* the groups of 10 contiguous items because no general ability factor is attributed to examinee (however, vertical patters are apparent *within* the 10 contiguous item arrangement).

When NMF is applied to $\mathbf{D}$ the resulting $\mathbf{W}$ matrix can be considered as the Q-matrix. For simulated data generated according to the procedure described above, the NMF algorithm is perfectly accurate in assigning the contiguous items in the same group, as can be seen in figure 1(b) where we find 4 bright squares representing the clusters. The figure's image represent the values of the 40×4 $\mathbf{W}$ matrix in NMF (transposed in this image) that directly represents what can be considered as a Q-matrix. Values are mapped to color gradients ranging from pale yellow to dark red.

Items 1 to 10 can readily be assigned to skill 3, items 10 to 20 to skill 4, and so on. The pattern is very obvious to the eye. A simple algorithm, that takes the maximum value for each item in the Q-matrix of figure 1(b) as the main skill, can systematically and correctly classify all question items in the correct skill cluster. These results are, for all practical purposes, deterministic, even though some variance could theoretically occur (we report variance when it becomes substantial later).

The visual results of the Q-matrix leaves little doubt that, under the assumption that topic skill is the only factor that affects performance, the NMF technique is highly effective. We now turn to real data and replicate some experiments by Winters et al. [2005] to verify how the results come out under realistic conditions.

4. Q-MATRIX EXTRACTION FROM REAL DATA

Winters et al. [2005] experimented with NMF over SAT Subject Test data (see CollegeBoard [2011])[1]. The data is broken down in 4 topics: (1) Mathematics, (2) Biology, (3) World History, and (4) French. These topics are sufficiently far apart that we can expect that they have strong intra-topic correlation and are therefore discernible for clustering. The data is composed of a total of 297 respondents who completed the 40 question items tests over the Internet. The profile of the respondents is unknown but they are probably from the university student community.

This data has the same structure as the simulated data of section 3: 40 question items broken down into 4 topics of 10 items each. The results of the NMF algorithm over this data is reported in figure 2. Variation in the difficulty of each topic is apparent in figure 2(a), where items 1 to 10 show a higher success rate than items 10 to 20. Individual item difficulty is also apparent by the horizontal patterns, as can be expected. Although we can discern some vertical patterns across item groups, it is far less apparent (except intra-topic vertical patterns), suggesting that examinee ability does not span very much across topics.

Figure 2(b) shows the Q-matrix obtained from the SAT data. It is consistent with the results from Winters et al. [2005]. Clustering of the Mathematics (items 1 to 10) and the French items (31 to 40) is relatively well defined, but not so with the Biology (21 to 30) and World History (31 to 40).

As mentioned, clustering is based on the simple algorithm which assigns each item to one of the 4 clusters based on the maximum column value in matrix $\mathbf{W}$. Given that we know the actual category of each item, the accuracy of the clustering can be computed. This is obtained by a two step process. First, a contingency table is compiled from the clustering algorithm. Next, the lines are reordered so that the sum of the diagonal is maximized. The ratio of this sum over the total represents the accuracy of the assignment. An example of the contingency table obtained is given below for the SAT data along with its reordering:

	Cluster			
Category	1	2	3	4
1	5	5	0	0
2	0	0	10	0
3	1	0	1	8
4	10	0	0	0

Reordering
$\Longrightarrow$

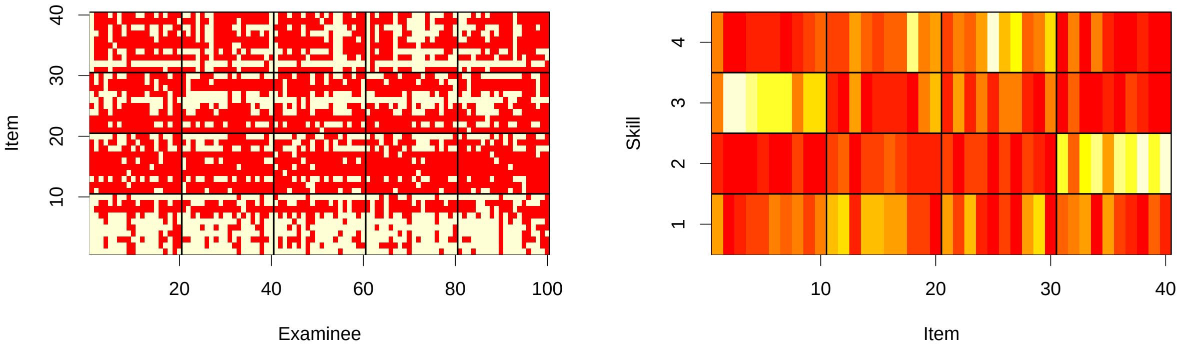
	Cluster			
Category	1	2	3	4
4	10	0	0	0
1	5	5	0	0
2	0	0	10	0
3	1	0	1	8

Note that the category and the cluster labels are irrelevant for measuring accuracy, but it it interesting to note that in this example the values of 10 are the Mathematics and French categories/clusters. As mentioned, the sum of the diagonal over the sum of all values represents the accuracy of this assignment: $33/40 = 0.825$.

Let us now turn to another data set from Winters et al. [2005] for which the task of deriving a Q-matrix from data was shown very challenging. They used used questions published from the Trivial Pursuit game and assembled a test that mimics the 4 topic structure of the SAT with 10 questions on each of: (1) Arts and entertainment, (2) Sports and leisure, (3) Science and nature, and (4) Geography. The results of our replication of this experiment are reported in figure 3.

Winters et al. [2005] results over the Trivia data concurs with our experiment and show that the NMF is no better than chance at correctly clustering items and building a Q-matrix. The most troubling findings from their experiments is that the Trivia results are similar to the results they obtain over a number of test outcome from different computer science courses: “Nearly every course behaves the same as the trivia data. Only our smallest data set, the Algorithms course data, showed any significant hint of topic structure.” This

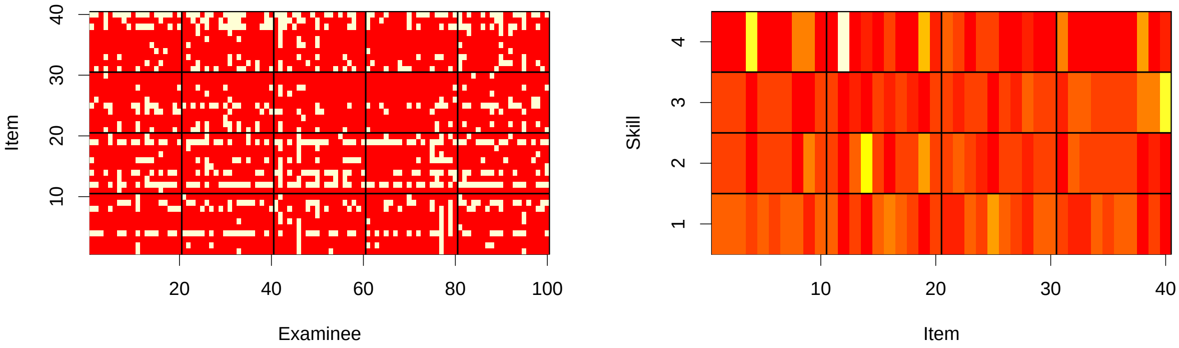
[1] The data sets from [Winters et al. 2005] were made available from `http://alumni.cs.ucr.edu/~titus/`. The simulation scripts of this paper are available from `http://www.professeurs.polymtl.ca/michel.desmarais/Papers/EDM2011/scripts.html`. They are based on the NMF package from the statistical software R.



(a) Item outcome from SAT scores of 4 topics and a sample of 100 examinees.

(b) Image output of Q-matrix from NMF for 4 skills and 40 question items.

Fig. 2. NMF results over SAT data.



(a) Item outcome from Trivia scores of 4 topics and a sample of 100 examinees.

(b) Image output of Q-matrix from NMF for 4 skills and 40 question items.

Fig. 3. NMF results over Trivia data.

conclusion casts a gloomy picture for high level transfer models, where we aim to assess the mastery of topic specific skills from similar topic skills.

However, statistical characteristics of the test data may also influence what can be extracted from this data. For example, skewness of the scores towards 0% or 100% will result in sparsity of success/failure that can can negatively affect the ability to extract a valid Q-matrix from the data. The Trivia data shows such skewness towards low success rate and we can question whether this is not the source of the low accuracy.

In the next section, we investigate the influence of the success rates and item and examinee variance over the Q-matrix validity.

5. INVESTIGATING THE PARAMETER SPACE OF SIMULATED STUDENT DATA

We turn back to simulated data to assess how the validity of the Q-matrix degrades under relaxed assumptions and under different ratios variance ratios between the skill, item difficulty, and examinee ability factors. This will allow us to better quantify the effect of the skill factor on examinee performance with respect to item difficulty and examinee ability.

First, we verify if NMF can extract the Q-matrix if we drop the unrealistic assumption set in section 3 and assume that item difficulty and person ability each contribute to the probability of success of an item by the same amount that topic skill can influence the probability of success.

Recall that the matrix **P**, as defined in section 3, contains independent normally distributed probabilities, each probability representing the chances of success to items of a single topic and for a single examinee. To account for the fact that item difficulty also affects item success, the probability of each item is modulated by a random quantity that is normally distributed with the same mean and variance (0,1) as the topic skill probability. Akin to item difficulty, examinee ability is accounted for by a similar quantity added on an examinee basis. Therefore, the probability of success by an examinee, m, to an individual item, n, belonging to topic, q, is defined as:

$$P(X_{mnq}) = \Phi(\beta_m + \beta_n + \beta_q) \tag{2}$$

where $\Phi(x)$ is the *cumulative distribution function* of the standard normal distribution, and where β_m, β_n and β_q are random Gaussian variables where the mean and standard deviation of β_m and β_n are:

$$\begin{aligned} \beta_m &\sim \mathcal{N}(\overline{X}, s_m) \\ \beta_n &\sim \mathcal{N}(\overline{X}, s_n) \end{aligned}$$

The variable $\overline{X}$ is constrained to be the mean of the whole data (matrix **D**). Variables s_m and s_n are respectively the individual examinee and item specific standard deviations. In the case of β_q, the mean can vary across each skill and is therefore defined as:

$$\beta_q \sim \mathcal{N}(\overline{X}_q, s_q)$$

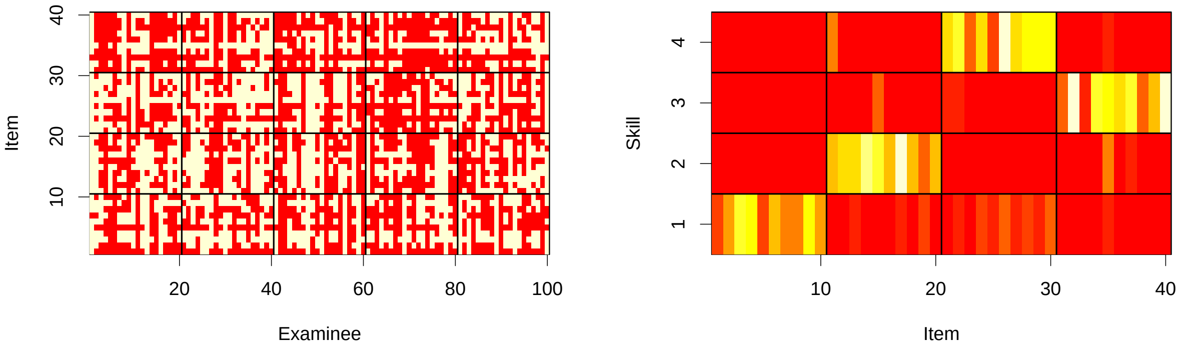
The parameter $\overline{X}_q$ is the specific mean of a skill and the different values must be congruent with $\overline{X}$ (the weighted sum of the mean for each skill times the number of items belonging to that skill must be equal to $\overline{X}$). s_q is the inter-skill standard deviation, measured by averaging the standard deviations of cluster means on an examinee basis.

Equation (2) can be considered as a simple model of examinee performance as a function of topic skill mastery, item difficulty, and examinee general ability (which spans across topics). In spite of its simplicity compared to other means of generating simulated data (for eg., see [Desmarais and Pelczer 2010]), it remains realistic for our context where we assume that topic skills are relatively independent, or at least this is an assumption we want to investigate and therefore it makes sense that our model follows that same assumption.

Figure 4(b) displays the Q-matrix (**W**) obtained from applying NMF over the data generated according to equation (2) with values of 0 for the mean and of 1 for the standard deviation for all β parameters. The raw data is displayed in figure 4(a).

Although we can visually appreciate that the clustering in the Q-matrix is not quite as sharp as in figure 1, these results still yield a perfect match of item to skills using the simple algorithm outlined in section 4.

Figure 4 shows that, when the mean and variance of the different β parameters in equation (2) are all equal (standard normal), the Q-matrix from NMF perfectly matches the underlying Q-matrix. Of course, as the effect of the topic skill parameter, β_q, becomes weaker compared to the other two parameters, the



(a) Item outcome.

(b) Image output of Q-matrix from NMF.

Fig. 4. NMF results over random data from randomly distributed data according to equation (2), reflecting equal effect of topic, item difficulty, and examinee ability over the probability of success.

accuracy of the item-skill match will become lower. This can be observed in table I where the link between accuracy and parameter ratios is quantified.

Table I reports the accuracy results of 14 N-folds simulation experiments conducted with different parameters. For simplicity, we consider a single mean of 0 for β_q. We also restrict the standard deviations to 1 for β_m and β_n given that they have the same effect according to equation (2) and and that we are interested in the values of the parameters respective to one another, therefore we can keep them fixed and vary s_q only. Note also that positive and negative values for the means β_n and β_m have symmetric effect such that only positive values are reported.

The first experiment reports an accuracy of 0.36 when no topic skill is defined[2]. As the variance increases ("S.d.": standard deviation column in the table), the accuracy over a 20-fold simulation gradually reaches 1 as its variance approaches that of the other two parameters. This trend is expected, but it quantifies, in terms of relative variance, the relation between the effect of the topic skill and the item and examinee effect. When the variance of the topic factor is comparable to that of item and examinee factors, the method yields very high accuracy.

Experiments 6 to 9 show the results of variations over the means of β_n and β_n. Experiment 7 shows that when both means of β_n and β_m are increased to 1 (in z score of the standard normal distribution), the accuracy starts to drop slightly to 0.98. Only for means of 1.5 and 2.0 does the performance decrease noticeably to 0.90 and 0.81 respectively.

In experiment 10, the simulation parameters replicate those of the Trivia data set, whereas experiment 12 is done with parameters from the SAT data set. Experiments 11 and 13 report the accuracies of NMF over the real data, corresponding respectively to figures 3 and 2.

For the Trivia data, the accuracy is comparable to the random, no topic skill condition. This results concurs with the conclusion of Winters et al. [2005], namely that topic subject is not a determining factor that affects

[2] If we had a very large number of items, this number, 0.36, would be close to 0.25, the theoretical accuracy of a random match in a 4×4 contingency table. However, the 40 items distribution in this table create an opportunity of over fit for the algorithm that decides which cluster is assigned to which skill. The difference of 0.11 $(0.36 - 0.25)$ can be attributed to this over-fitting.

Table I.
Experiments over the parameter space of skills, items, and examinee
(respectively β_q, β_n, and β_m in equation (2)).

	Parameter space								
	Topic skill (β_q)		Item (β_n)		Examinee (β_m)			Accuracy	
	Mean	S.d.	Mean	S.d.	Mean	S.d.	N folds	Mean acc.	S.d. acc.
1*	0	0	0	1	0	1	20	0.36	0.05
2	0	0.10	0	1	0	1	20	0.48	0.07
3	0	0.25	0	1	0	1	20	0.60	0.11
4	0	0.50	0	1	0	1	20	0.93	0.08
5	0	1	0	1	0	1	20	1	0
6	0	1	0.50	1	0.50	1	20	1.00	0.01
7	0	1	1	1	1	1	20	0.98	0.07
8	0	1	1.50	1	1.50	1	20	0.90	0.12
9	0	1	2	1	2	1	20	0.81	0.16
Trivia data parameters									
10	0	0.12	-1.05	0.73	-1.05	0.45	20	0.75	0.12
11**	*n.a.*	*0.12*	*-1.05*	*0.73*	*-1.05*	*0.45*	*20*	*0.35*	*0.03*
SAT data parameters									
12	0	0.24	-0.33	0.86	-0.33	0.50	20	0.98	0.05
13**	*n.a.*	*0.24*	*-0.33*	*0.86*	*-0.33*	*0.50*	*20*	*0.72*	*0.02*
14***	*n.a.*	*0.24*	*-0.33*	*0.86*	*-0.33*	*0.50*	*20*	*0.96*	*0.05*

* No topic skill effect conditions
** Real data
*** Real data and scoring for the Mathematics and French topics only

test performance. Considering that they obtained similar results for topics from academic computer science courses, these results are disconcerting.

However, we conjectured earlier that the low success rate of the Trivia data could explain the low accuracy results obtained. This is only partly the case. When the simulations parameters are set to the same values as the Trivia data, the accuracy obtained is 0.75 (experiment 10[3]) whereas the real data results are 0.35 (experiment 11). Therefore, results of experiment 10 suggest that the gap between 0.75 and 0.35 is attributable to the lack of skill effect in this data.

Comparing the results to the accuracy reported on experiments 11 and 13 for real data, we observe that for SAT data, the accuracy is lower than experiment 12 and somewhere between experiments 3 and 4, which corresponds to a standard deviation of topic skill between 0.25 and 0.5 when β_n and β_m have a (0,1) standard distribution. In other words, the skill effect is a little less than half the item and examinee effects.

If we look only at the clustering for Mathematics and French (experiment 14) which are the most separable topics, then the accuracy goes up to 0.96, which is much closer to experiment 12. In terms of relative effect, the skill effect between Mathematics and French is close to the 0.93 accuracy obtained in 4, for which the standard deviation of skill effect is 0.50 of the item and examinee parameters.

In summary, the Trivia data shows negligible effect of topic skill, whereas the SAT data shows an effect that is essentially attributable to the Mathematics and French topics that can be clearly distinguished in the Q-matrix derived with NMF. The topic skill effect can be quantified as somewhere between 1/4 to 1/2 of the

[3] Experiment 10 has a relative skill-item s.e. of 0.12/0.73 = 0.16, standing between experiments 2 and 3, and a relative skill-examinee s.d. of 0.12/0.45 = 0.27, standing close to experiment 3. If the performance followed some additive function of each of these ratios, we would expect the performance to be no better than that of experiment 3, 0.60. Given that it stands higher at 0.75, we have to conclude that the effect of s.d. ratios over the performance is more complex, possibly a ratio of s.d. such as topic/(item × examinee).

item and examinee effect as measured by the standard deviation, and over 1/2 if we only take Mathematics and French effects alone.

6. DISCUSSION

In undertaking this exploratory work, we were hoping to show that the failure to find an effective Q-matrix from some data sets, such as the Trivia data set, was due to highly skewed tests scores: either the scores are too high or too low, and the raw data becomes too sparse of successes or failures to allow the NMF algorithm to derive a reliable Q-matrix. Results from our experiments suggest that, in fact, this is only partly the case. It still leaves open the suggestion that the topic skill factor has sometimes a negligible effect on performance, or at least a much lower effect than we are generally are inclined to believe. From Winters et al.'s [2005] previous results, we can expect this to be the case for many courses that divide their content according to sub-topics.

Our results further indicate that for well delineated topic skills like Mathematics and French, the effect is relatively strong, in a range around half that item difficulty and examinee ability according to the results in table I, at least for highly separable topics like Mathematics and French. In this case, the accuracy of matching items to skills with NMF is well in the range of 90%, which confirms the effectiveness of this technique under these conditions.

This study was conducted under the assumption that we know the number of skills for the clustering and for building the Q-matrix. This is not the case in general. However, the visualization technique used throughout this paper shows that for well delineated topic skills, clustering with NMF is easily perceived through the human eye.

REFERENCES

Barnes, T. 2006. Evaluation of the q-matrix method in understanding student logic proofs. In *Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference, Melbourne Beach, Florida, USA, May 11-13, 2006*, G. Sutcliffe and R. Goebel, Eds. AAAI Press, 491–496.

Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P., and Plemmons, R. J. 2007. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis 52,* 1, 155 – 173.

Cen, H., Koedinger, K. R., and Junker, B. 2005. Automating cognitive model improvement by A* search and logistic regression. In *Educational Data Mining: Papers from the 2005 AAAI Workshop.*, J. Beck, Ed. Technical Report WS-05-02. Menlo Park, California: AAAI Press, 47–53.

Cen, H., Koedinger, K. R., and Junker, B. 2006. Learning factors analysis — A general method for cognitive model evaluation and improvement. In *Intelligent Tutoring Systems, 8th International Conference, ITS 2006, Jhongli, Taiwan, June 26-30, 2006, Proceedings.* 164–175.

CollegeBoard. 2011. Sat subject tests practice questions. `http://sat.collegeboard.com/practice/sat-subject-test-preparation` (consulted on April 2, 2011).

Desmarais, M. C. and Pelczer, I. 2010. On the faithfulness of simulated student performance data. In *3rd International Conference on Educational Data Mining EDM2010*, R. S. J. de Baker, A. Merceron, and P. I. Pavlik, Eds. www.educationaldatamining.org, 21–30.

Lee, D. D. and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature 401,* 6755, 788–791.

Schachtner, R., Poppel, G., and Lang, E. 2010. A nonnegative blind source separation model for binary test data. *Circuits and Systems I: Regular Papers, IEEE Transactions on 57,* 7, 1439 –1448.

Tatsuoka, K. K. 1983. Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement 20*, 345–354.

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