Book Title
Book Editors
IOS Press, 2003

Investigating Generative Factors of Score Matrices

Titus Winters [a,1], Christian R. Shelton [a] and Tom Payne [a]
[a] *UC Riverside, Riverside CA, USA*

Abstract. An implicit assumption in psychometrics and educational statistics is that the generative model for student scores on test questions is governed by the topics of those questions and each student's aptitude in those topics. That is, a function to generate the matrix of scores for m students on n questions should rely on each student's ability in a set of t topics, and the relevance of each question to those topics. In this paper, we investigate score matrices from university-level computer science courses, and demonstrate that no such structure can be extracted from this data. Utilizing unsupervised machine learning techniques we provide evidence calling into question this fundamental assumption of educational statistics.

Keywords. Cognitive Science, Clustering, Dimensionality reduction, Educational Data Mining

1. Introduction

Over the past century, psychometrics has utilized mathematical techniques to extract meaning from collections of educational data. One of the most common data types in psychometrics is the score matrix, storing the scores for a collection of students on each question from a test. Commonly, educational researchers will apply algorithms developed in the early 20th century to discover the structure or generative model of these matrices. While advanced for their time, and sufficient for some problem types, the great advances made in machine learning may have surpassed these techniques, and may even invalidate some of the assumptions made by early psychometric techniques.

This paper presents an investigation into the generative model of score matrices recorded in computer science (CS) courses in our department. Here we apply machine learning techniques to this canonical psychometric problem and show that on some naturally-arising datasets, the assumptions made for the past century may not be valid.

Specifically, this paper investigates the assumption that a generative model for a score matrix should take into account the topics being covered on the test. From our investigations on course data, the factors that are considered most important by machine learning algorithms consistently have no relationship with human-identified topics.

2. Psychometrics

At its root, the quantitative branch of education that is most applicable to machine learning and educational data mining is *psychometrics*. Psychometrics focuses on measuring

[1]Correspondence to: Tel.: +1 909 262 0385; E-mail: titus@cs.ucr.edu.

Table 1. Descriptions of datasets. MC indicates multiple choice questions, FR indicates free-response questions subjectively graded, PC indicates partial-credit.

Data set	Course	m	n	Question Types	Scores
008	Intro Computing	269	80	MC	Binary
010	CS 1	190	68	MC	Binary
141	Algorithms	19	15	FR	PC
153	O. S.	66	34	MC,FR	Binary,PC
164	Networks	23	72	MC,FR	Binary,PC
Academic	Online	267	40	FR	Binary
Trivia	Online	467	40	FR	Binary

latent quantities of knowledge and ability in the human mind [1]. These are difficult values to quantify, as there is no direct way to measure them and no implicit units or dimensionality to such a measurement.

One of the most important developments in psychometrics was the development of common factor analysis (CFA), which became the primary area of research in psychometrics for half a century [2]. The dominant theory at the time was that intelligence was univariate (g-theory). However, some data simply did not match the proposed model. Utilizing then-new techniques for calculating correlations between measured values, Charles Spearman developed common factor analysis, which is still used today to understand large quantities of data by discovering the latent factors giving rise to that data. CFA assumes that a linear model of f factors can describe the underlying behavior of the system, in this case the ability of students to answer items correctly: $S_{i,j} = \mu_j + \sum_{k=1}^{f} W_{i,k}H_{k,j} + \epsilon$, where $S_{i,j}$ is the score for student i on item j, μ_i is a hidden variable related to the difficulty of the item, W is a matrix giving the ability of each student with respect to each factor, H is a matrix relating each factor to each item, and ϵ is the noise in score measurement. Generally, the factors in CFA are assumed to be the topics of the various questions. Other branches of psychometrics such as Item-Response Theory (IRT) have similar assumptions about the dominant importance of topic [3].

This paper is an attempt to verify the CFA assumption that the underlying factors in the generative model for a score matrix are the topics of the questions being tested.

3. Datasets

Each of the datasets evaluated in this paper has two components. Most important is the score matrix S, an $m \times n$ matrix of the scores for m students on n questions. We also have a collection of human-generated sets identifying which of the questions (columns of S) can be viewed as testing the same topic. We currently have two sources for datasets: real course data recorded at the per-question level, and an online quiz system. These datasets are summarized in Table 1.

We are primarily concerned with datasets from real courses. Over the past several terms we have recorded detailed score information for a number of courses across our curriculum, ranging from large non-major service courses to upper division electives.

For each of these courses, we have asked the instructor of the course, along with any other regular teachers of the course, to provide us with a set of question groups they would consider grouped by topic. To do this they are provided the text of the questions, but not the scores. They were allowed to place each question in as many groups as they

choose. This could be none if they felt the question is not connected to any other questions in the course, or more than one group if a question relates to more than one topic.

Both of the analysis experiments in this paper rely on the presence of this topic information provided by instructors. However, as it is unknown whether the score matrices for actual courses really contain topic information, we felt it prudent to develop datasets with clear dependence on topic. To this end we built two additional datasets, one derived from trivia questions from the popular game Trivial Pursuit [4], and one derived from questions from study guides for SAT Subject Tests [5]. Both quizzes are forty questions drawn from four topics, with ten questions in each topic.

The trivia quiz draws questions from Sports and Leisure, Science and Nature, Arts and Entertainment, and Literature. These topics were chosen from the six categories in Trivial Pursuit because they were felt to be the most independent from one another. The questions were drawn from the 20th Anniversary edition of Trivial Pursuit, which focuses primarily on the 1980s and 1990s. As our target audience for the quiz was college students, we felt this was more likely to result in correct answers than questions from other editions. Additionally, we chose only questions that had short one or two word answers, as the online quiz system we developed is a free response system and we wanted to minimize issues stemming from poor spelling.

The academic quiz draws questions from published study guides for the SAT Subject Tests. The SAT Subject Tests are taken by students in their final year of high school, and are used to demonstrate the depth of knowledge that a student has attained in particular areas. There are over twenty test subjects available. Again we chose subjects that we felt were maximally independent: Math, French, Biology, and World History.

Our assumption was that the trivia quiz could function as a baseline for topic extraction on data with little or no actual correlation within each topic. Examining the trivia questions, it is clear that they are indeed trivia: answering any given question correctly is a matter of isolated fact retrieval, and not a test of any deeper knowledge or skill. As such, even though there is an obvious "correct" answer when grouping these questions, we did not expect to be able to extract that answer with any of our candidate algorithms.

The academic quiz represents the opposite end of the spectrum: the questions that were included come from subjects that are generally studied for at least one academic year in high school, and possibly several more in college. The questions require a fair depth of knowledge, and are more topically connected than the trivia questions. Our expectation was that any algorithm that can succeed in topic clustering will have at least partial success on this dataset.

The quizzes were administered online over the course of about a week. The trivia quiz was completed by 467 different visitors, the academic quiz was completed by 267. Correct answers were immediately reported to the test taker, incorrect answers were reported only as "probably" incorrect. The full text input for any incorrect answers was recorded, allowing for post-testing processing and correction for unexpected formatting or spelling. This allowed us to focus on whether the student knew the answer.

4. Unsupervised Clustering

Although we performed several experiments on this data, due to space constraints we only present results for unsupervised clustering. Interested readers are invited to seek additional details on this work in [6]. This unsupervised clustering experiment is in some-

thing we have labelled *topic clustering*: given S and the human-generated groups, is there an unsupervised machine learning algorithm that can generate clusters of the questions in S similar to the human-generated answer?

4.1. Experiment

To evaluate the output of our candidate algorithms in topic clustering, we focus on question pairings. Each algorithm produces a set of groups of questions that are being claimed to belong to the same topic. Each dataset has a similar set of groups that are correct or acceptable answers. For the course datasets these were produced by the course instructors, for the quiz datasets these are the topics from which the questions were drawn.

To measure the similarity of the correct answer and the algorithm's results, we create the list of all pairs of questions that were placed in the same group by the algorithm, and the list of all pairs of questions that were grouped together by a human. Given these lists we can evaluate *precision* and *recall* for each algorithm. Precision is the percentage of the pairs reported by the algorithm that are present in the correct answer. Recall is the percentage of pairs in the correct answer that were also reported by the algorithm. Thus precision measures how accurate the answer is, and recall measures how complete the answer is. These are independent: grouping only two questions, and grouping those correctly, produces perfect precision but very low recall. Grouping all questions together into one group provides low precision (equal to random chance), but perfect recall.

Further, all but one of the algorithms can produce a numeric certainty on each question's group membership. By varying the cutoff threshold on that certainty, we can adjust the size of the groups reported by each algorithm. If an algorithm's underlying model is accurate, then at least the most-certain questions should be correctly grouped. By evaluating precision and recall across the range of the certainty, we can generate precision-recall curves for each algorithm, demonstrating the accuracy of the algorithm across the certainty range.

4.2. Algorithms Surveyed

The algorithms considered in this study fall into three main groups: clustering algorithms, dimensionality reduction algorithms, and algorithms from educational statistics. Here we list the algorithms, how they are used to produce the certainty groups underlying the precision-recall curves, and describe those that readers may be unfamiliar with.

4.2.1. Clustering

We are evaluating k-means [7], spectral clustering [8] single-linkage [9], complete-linkage [10], and average-linkage [11] algorithms. All of these can give an ordering on which questions are most certainly grouped by sorting the questions by the distance to their cluster center.

Another set of algorithms that we are investigating is the use of single linkage, average linkage, or complete linkage agglomerative clustering on the *correlation matrix* of the questions, rather than working with the raw data itself. In this form, the pair of questions or question clusters to be merged at each step is given by the maximal entry in the correlation matrix. The three agglomeration techniques differ in how they merge the rows and columns for the chosen clusters for the next step.

Let M represent the correlation matrix, initially calculated such that the i, j^{th} entry of M is the correlation of question i with question j across all students. Each step of agglomeration will merge two questions or question clusters, reducing the total number of clusters. In our version of single-linkage clustering, when merging elements i and j, the resulting entry is given by the maximum of the two entries. Thus, when determining the l^{th} entry of the new vector, we use the maximum of $M_{i,l}$ and $M_{j,l}$. Average-linkage uses the average of the two elements, and complete-linkage uses the minimum.

Agglomeration continues until we have created the appropriate number of groups. Group membership is obvious, as this is a clustering algorithm. The ordering of which questions are most certain is given by the sum of the "correlations" used in merging together the clusters for each individual question. For example, a question that was merged twice, first because of an entry in M of 0.5 and then because of an entry of 0.35, the sum will be 0.85. Questions with a higher sum are considered more certain. This strongly favors questions that were clustered earlier.

This is somewhat analogous to our use of to spectral clustering. Here the "similarity" is correlation, the reduced dimensionality is n, and agglomerative clustering is used rather than k-means. A sufficiently general implementation of spectral clustering could directly support these alternate clustering forms.

4.2.2. Dimensionality Reduction

Dimensionality reduction algorithms can also be used to find the underlying structure of the data. For dimensionality reduction, we evaluated Singular Value Decomposition (SVD) [12], Independent Component Analysis (ICA) [13], and a non-negative matrix factorization (NNMF) [14].

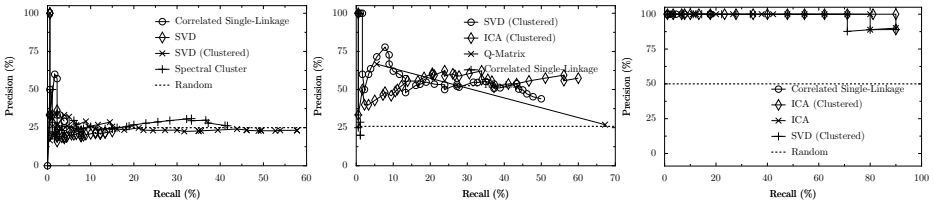
We also evaluate slight alterations of SVD and ICA. In the base versions of these algorithms group membership is determined by the maximum absolute value in the output vector corresponding to each question. In the modified versions we use k-means clustering on the reduced matrices, setting k to t (the number of factors) and assigning group membership for each question according to the resulting clustering.

4.2.3. Educational Statistics

We are also investigating two methods from educational statistics in this study: Common Factor Analysis [2], which we are using as the major point of comparison for educational statistics, and Q-Matrix [15]. Both of these have been specifically suggested by researchers in education for this problem. Q-Matrix is another method of capturing the underlying factors and abilities that give rise to a matrix of student scores. In the most simple case, Q-Matrix is applied to a binary $m \times n$ score matrix.

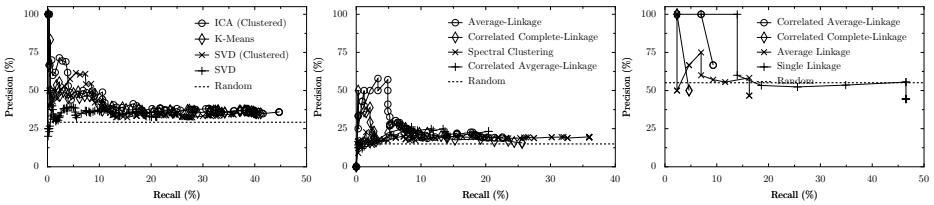
Q-Matrix operates on this matrix along with the assumed number of underlying abilities or factors, t. Each question is assumed to have a binary relationship with each of the t factors. Each student is similarly assumed to have a binary ability in those factors. Thus Q-Matrix is decomposing the input matrix into a binary $m \times t$ matrix of student abilities in each of those factors, as well as a binary $t \times n$ matrix of question relevance. A student is assumed to answer a question correctly if and only if the factors relevant to that question are a subset of their own abilities.

Most or all Q-Matrix solvers utilize gradient descent on the reconstruction error of the matrix in order to find the optimal model to fit the data. Given an initial random



(a) Trivia (b) SAT (c) SAT: Math & French

Figure 1. Precision-recall curves for best four algorithms on baseline datasets



(a) 008 (b) 010 (c) 141

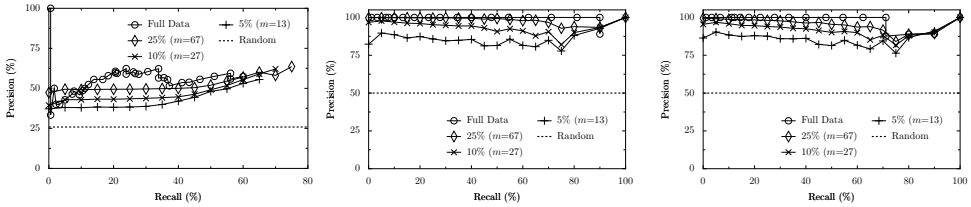
Figure 2. Precision-recall curves for best four algorithms on course datasets, four factors

configuration, the algorithm finds the single best entry to toggle in the student matrix or the question matrix in order to best reproduce the input matrix. As a discrete optimization problem with a search space growing as $O(2^{t*(m+n)})$, this is computationally intensive.

4.3. Results

First we need to confirm our assumptions about the ability for topic information to be extracted from the baseline data. We would like to confirm that the algorithms we are surveying do *not* extract clusters related by topic on the trivia data. This is confirmed in Figure 1a: even the four best algorithms applied to this dataset fail to significantly outperform random chance. Similarly, we would like to see that the questions in the academic dataset do cluster by topic. This is partially confirmed, as seen in Figure 1b. However, even the highest-performing algorithms appear to only be performing at around 50% precision. Further investigation demonstrates that the Math and French portions of the academic dataset can be correctly clustered by almost all of our surveyed algorithms, as shown in Figure 1c. The Biology and World History questions behave like the trivia data. Since SAT-level questions on Biology and World History are generally retrieval of isolated facts, more than testing of deeply learned skills, this is less surprising.

On data where topic is clearly a factor (the Math and French) we have shown that almost any machine learning technique will acceptably separate the data. For data where topic is not expected to be a factor (the trivia data), all of the algorithms surveyed perform only as well as random chance. One important question is, "Where on this spectrum do course data fall?" Figure 2 demonstrates this for three of our course datasets: although there is possibly some dependence on topic, in general even the best algorithms are performing no better than chance. Given this evidence, topic has little to do with scores on the course datasets.



(a) SAT (b) SAT: Math & French (ICA) (c) SAT: Math & French (SVD)

Figure 3. Precision-recall curves on for best algorithm(s) on baseline data, varying population sizes

Table 2. Average Precision within given Recall Ranges on Course Datasets, 8 factors

Algorithm	0 to 20	20 to 40	40 to 60	60 to 80	80 to 100	AVG
NNMF	**33.9**	**41.2**				**33.9**
Complete-Link (corr)	33.8	17.5				33.1
Random	30.2	30.2	30.2	30.2	**30.2**	30.2
Complete-Link	28.8	36.1	**37.4**			29.8
ICA (cluster)	28.0	38.0				29.5
SVD (cluster)	27.0	37.7				28.9
ICA	28.4					28.4
Average-Link (corr)	28.3					28.3
Spectral	28.1	23.4		26.0	27.2	27.6
k-means	26.4	30.0				27.3
Average-Link	26.8					26.8
CFA	27.6	20.8				25.9
Single-Link	24.4	26.4	22.6	23.1	22.0	24.4
Single-Link (corr)	24.3					24.3
SVD	23.4					23.4
Q-Matrix	19.1	21.4		**31.8**		22.4

An important difference between the course data and the baseline data is the size of the datasets: the baseline data has significantly larger student populations than the course data. It is possible that topic clusters can be more accurately extracted if student populations were larger. Since it is not possible to increase the student population in a real course, instead we test the converse of this hypothesis: "Can we still extract topic information from reduced student populations on the academic dataset?" This is shown in Figure 3. For student populations as small as about 20 students, topic information can be extracted well from the Math and French data. This strongly suggests that it is not the size of the input that causes the poor performance in Figure 2.

Summarizing all of the algorithms on all the course datasets, and setting $t = 8$ (the largest workable value for some of the smaller datasets), we get Table 2. All of the precision values falling within the given range for recall across all course datasets are averaged together and provided here. A blank entry indicates no evaluation of the output of that algorithm had a recall in that range. By setting t as large as possible, we minimize issues of forcing multiple topics into the same group, at the expense of lower recall values. Here we see that nothing significantly outperforms random chance, with the possible exception of the Non-Negative Matrix Factorization. On the course data, it does not appear that topic is a prime factor in the generative model for student scores.

5. Conclusions

Neither experiment presented here successfully shows the dependence of scores on topic information for course data. This contradicts some of the fundamental assumptions in psychometrics. We have provided strong evidence that on common types of data (namely, scores collected in university courses), the generative model for those scores is, in fact, *not* based on the topics for those questions. In our informal evaluation of the extracted groupings we have found anecdotal evidence of the importance of time (which questions were asked on the same day), trick questions (which intro CS questions the authors have difficulty answering), and many groups whose relation is completely unclear. This means that for individual questions in a such a course, it appears to be more important that the student is having a good day, or is good at trick questions, or are generally good in the course, rather than whether they know a lot about the topic of a particular question.

The datasets used in these experiments are publicly available on the lead author's website. We encourage other researchers to explore their own techniques for extracting information from such course matrices.

References

[1] Karl Pearson. *The life, letters and labours of Francis Galton*. University Press, 1914.
[2] Charles Spearman. General intelligence, objectively determined and measured. *American Journal of Psychology*, 15:201–293, 1904.
[3] Frank B. Baker. *Fundamentals of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, 2001.
[4] Trivial pursuit. http://www.trivialpursuit.com, 2006.
[5] SAT subject tests. http://www.collegeboard.com/student/testing/sat/about/SATII.html, 2006.
[6] Titus Winters. *Educational Data Mining: Collection and Analysis of Score Matrices for Outcomes-Based Assessment*. PhD thesis, UC Riverside, May 2006.
[7] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkely Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.
[8] I. Fischer and J. Poland. Amplifying the block matrix structure for spectral clustering. In M. van Otterlo, M. Poel, and A. Nijholt, editors, *Proceedings of the 14th Annual Machine Conference of Belgium and the Netherlands*, pages 21–28, 2005.
[9] R. Sibson. Slink: An optimally efficient algorithm for the single link cluster methods. *The Computer Journal*, 16(1):30–34, 1973.
[10] D. Defays. An efficient algorithm for a complete link method. *The Computer Journal*, 20(4):364–366, 1977.
[11] R. R. Sokal and C. D. Michener. Statistical method for evaluating systematic relationships. *University of Kansas science bulletin*, 38:1409–1438, 1958.
[12] J. C. Nash. The singular-value decomposition and its use to solve least-squares problems. In Adam Hilger, editor, *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation, 2nd ed*, pages 30–48, 1990.
[13] Aapo Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.
[14] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2001.
[15] M. Birenbaum, A. Kelly, and C. Tatsuoka. Diagnosing knowledge state in algebra using the rule-space model. *Journal for Research in Mathematics Education*, 24(5):442–459, 1993.