A non-negative matrix factorization analysis of a multiple-choice education test

Kumer Pial Das[1], Jay Powell[2], Myron Katzoff[3], and S.Stanley Young[4]

[1]Lamar University [2]Better Schooling Systems [3]Center for Disease Control-Ret [4]National Institute of Statistical Sciences

Abstract

Educational questionnaires serve a number of purposes, one of which is to better understand the performance of students so that education delivery can be improved. Principal components analysis, PCA, is often used to help examine a large number of variables to better understand the relationships among them and to summarize into a relatively few "score" the information in the data set. Singular value decomposition, SVD, can also be employed to reduce the size of a big data matrix. Our idea is to contrast a relatively new matrix factorization method, non-negative matrix factorization, NMF, with PCA and SVD with the goal of pointing to individualized education. We show that NMF offers interpretive advantages.

Keywords: Principal component analysis, PCA, Singular value decomposition,SVD, Non-negative matrix factorization, NMF, Alternative scoring procedures, Performance scales.

1 Introduction

This study investigates the current testing practices with respect to using multiple-choice tests. Multiple-choice test result is designed to estimate students' knowledge. The items are derived from a representative sampling of course content and the scores are presumed to be proportional to the knowledge possessed by the examinees. The scores are the frequencies of the designated acceptable ("right") answers. This approach provides justification for assuming the remaining selections as unacceptable ("wrong") answers. These answers are assumed to contain no useful information and are converted to zero (0) during scoring. In this traditional system, the frequency of "right" answers provide a necessary and sufficient set of information about examinees subject-matter knowledge ([6]). However, it can be assumed that the selection of any right or wrong answer is mostly a thoughtful process instead of being random. Many researchers have been raising questions about whether "wrong" answer selection is entirely random. In 1965, Jay Powell started to explore "wrong" answers selection rationale. His interest in Piaget's work led him to use Gorham's ([3]) *Proverbs Test*. This study investigates all possible answers of the test not knowing which one is correct and by using two data reduction approaches it shows that NMF can interpret the data more efficiently than SVD can.

The rest of the article is organized as follows. In Section 2, we investigate a few statistical methods used in contingency data analysis including SVD and NMF. In Section 3, we investigate an education data set by using SVD and NMF with an aim of reducing the dimension of data. In Section 4, we give our results. We conclude the paper with a summary of main findings in Section 5.

2 Statistical Methods Used

Contingency tables of numeric data are often analyzed using dimension reduction methods like the singular value decomposition (SVD), and principal component analysis (PCA)([5]). This analysis produces score and loading matrices representing the rows and the columns of the original table and these matrices may be used for both prediction purposes and to gain structural understanding of the data. We provide a short introductory description of SVD and PCA.

2.1 Singular Value Decomposition

Singular Value Decomposition is based on a theorem from linear algebra. Matrix A can be broken down into the product of three matrices - an orthogonal matrix U, a diagonal matrix D, and the transpose of an orthogonal matrix V. Then the matrix A can be written as $A = UDV^T$.

SDV is a strong technique to reduce the dimension of a given matrix. If you take a picture with large pixel value (as an example 1200 x 800). Eventhough the picture has high quality, it takes a lot of storage. If the same picture can be represent as a low dimension (low pixel value), it is more memory efficient. If we think the pixel as a element of matrix A, SDV can be used to reduce the dimension of the size of the picture. Also, analysis can be faster if a low dimension representation of the information in the matrix is used. For each $A \in \mathcal{R}^{\text{mxn}}$ of rank r, there are orthogonal matrices $U_{m\times m}, V_{n\times n}$ and diagonal matrix $D_{r\times r} = Diag(\sigma_1, \sigma_2, \sigma_3, ..., \sigma_r)$ such that $A = U \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}_{nxm} V^T$ with $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq ... \geq \sigma_r \geq 0$ the σ_i's are call the nonzero singular values of A. When $r < p = min\{m, n\}$, A is said to have $p - r$ additional zero singular values. The above factorization is called a singular value decomposition of matrix A, and the column in U and V are called left hand and right hand singular vectors for A respectively.where $U^T U = I$, $V^T V = I$; (I is the identity matrix) the columns of U are orthonormal eigenvectors of AA^T , the columns of V are orthonormal eigenvectors of $A^T A$, and D is a diagonal matrix containing the square roots of eigenvalues from U or V in descending order.

2.2 Non-negative Matrix Factorization

The pervasive nature of nonnegative data is obvious in many applications. In many tables, the data entries are necessarily non-negative, and so the matrix factors meant to represent them should arguably also contain only non-negative elements. Lee and Seung in their seminal paper([8]) study in detail two numerical algorithms for learning the optimal nonnegative factors from data. Extensive literature reviews on NMF have been provided by Fogel et al. ([2])

We describe here the use of NMF, an algorithm based on decomposition by parts that can reduce the dimension of data. Let A be a $n \times p$ non-negative matrix, and $k > 0$ an integer. NMF consists in finding an approximation,

$$A \approx WH$$

where W, H are $n \times k$ and $k \times p$ non-negative matrices, respectively. We consider the rows to be cases and the columns to be variables, thus then the rows of H are the basis vectors and the rows of W say how they are added together to make a row of A. In practice, the factorization rank k is often chosen such that $k << min(n, p)$. In general, k can be bounded as $(n + p)k < np$. The objective behind this choice is to summarize and split the information contained in A into k factors: the columns of W. Depending on the application field, these factors are given different names: basis images, metagenes, source signals. We will be using the term *source signals* in this article. We study the use of PCA, SVD, and NMF to reduce the dimensionality of count data presented in a contingency table. Our primary goal is to remove noise and uncertainty by capturing the signal in the matrix. In theory, NMF can also be used for better interpretation of factoring matrices ([2]). We find that as the rank k increases the method uncovers substructures, whose robustness can be evaluated by a cophenetic correlation coefficient. These substructures may also give evidence of nesting subtypes. Thus, NMF can reveal hierarchical structure when it exists but does not force such structure on the data. The cophenetic correlation coefficient is based on the consensus matrix (i.e. the average of connectivity matrices) and was proposed by [1] to measure the stability of the clusters obtained from NMF. It is defined as the Pearson correlation between the samples distances induced by the consensus matrix, seen as a similarity matrix and their cophenetic distances from a hierarchical clustering based on these very distances, by default an average linkage is used. The elements of H can be used to cluster the objects. The cophenetic correlation measures the consistency of the clustering; bigger (maximum of 1.0) is better.

2.2.1 Statistical Software Used

We use both R and JMP to estimate the value of k. The SAS JMP scripts used in this study can be downloaded from www.niss.org/irMF.

3 Data

J. Powell ([6],[7]) and his colleagues designed various education studies. Most of these studies were motivated by an observation made in the early 1960's wherein some "wrong" answers given to multiple tests appeared to reflect systematic logical analysis by test subjects.One of these studies has used a multiple choice test (known as "the Proverbs Test") contains 40 items, each with 4 alternatives. It is interesting to note that many questions of "the Proverbs Test" are exemplified in a work of Pieter Brugel [Figure 1]. For example, question 32 can be explained by Figure 2.

Q32: Don't cast pearls before swines (pigs)
a. Put your efforts where they are appreciated
b. Don't give pearls to fools
c. Don't be wasteful
d. Don't always put yourself before everybody

Note that, "a" is the expected correct answer. However, "c" and "d" are also correct, but not a resatement of the proverb. The data is from Canada and at the time, Group 12 was college prep and qualitatively different from

Figure 1: Pieter Bruegel the Elder, The Folly of the World 1559.

the other groups. The students ranged in age from less that 8 to about 18. Raw percentage of student's response to Q32 is given in Table 1.

The test was given twice in each year. The data set which is also known as Windsor education data is represented by an expression matrix A of size 2293×160 whose rows contain 2,293 students and column contains 160 variables (40 questions with four alternatives). Matrix W has size $2293 \times k$, with each of the k columns defining a *source signal*. Matrix H has size $k \times 160$, with each of the 160 columns representing a *source signal* expression pattern of the corresponding sample. By grouping students of the similar age we end up with a contingency table and reduce the computations of NMF.

4 Results

The consensus matrices are computed at $k = 2$ to 10 for the Windsor education data (Figure 3,4,5). Samples are hierarchically clustered by using distances derived from consensus clustering matrix entries, colored from 0(deep blue, samples are never in the same cluster) to 1 (dark red, samples are always in the same cluster)

Figure 2: Q32: Don't cast pearls before swines (pigs).

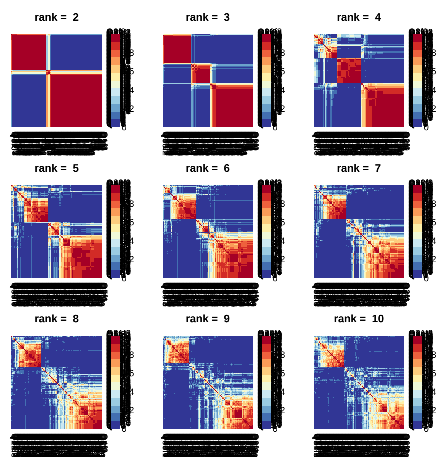

Figure 3: Consensus Map of all questions for rank 2-10

6

Table 1: Q32 Response in Percentage

Group	Age in months	$a(\%)$	$b(\%)$	$c(\%)$	$d(\%)$
1	$x \leq 96$	17.1	26.8	19.5	34.1
2	$96 < x \leq 108$	24.0	22.1	23.1	28.8
3	$108 < x \leq 120$	20.5	21.2	27.8	29.8
4	$120 < x \leq 132$	24.5	23.9	27.7	23.2
5	$132 < x \leq 144$	32.2	9.4	21.6	34.5
6	$144 < x \leq 156$	31.8	12.7	28.9	26.6
7	$156 < x \leq 168$	29.5	7.1	25.6	37.2
8	$168 < x \leq 180$	44.4	5.6	18.4	30.2
9	$180 < x \leq 192$	53.1	1.9	15.6	27.8
10	$192 < x \leq 204$	49.3	4.1	16.3	29.3
11	$204 < x \leq 216$	57.5	0.9	12.3	28.8
12	$216 < x$	68.8	0.6	13.4	16.6

Although a visual inspection is important, it is also important to have quantitative measure of the stability of clustering for each value of k. One measure proposed by Brunet et al. ([1]) is cophenetic coefficient, which indicates the dispersion of the consensus matrix, defined as the average connectivity matrix over many clustering runs. Observe how cophenetic coefficient changes as k increases. We select values of k where the magnitude of cophenetic coefficient begins to fall. Hutchins et al. ([4]) suggested to choose the first value where the RSS curve presents an inflection point. Even though both cophenetic coefficeint and RSS curve is suggesting as $k = 2$ the heatmap is not very supportive of the statement. Our impression is that there could be three distinct groups so we decided to set k at three.

Considering $k = 3$ we have observed that two groups of intermediate age students (Group 7 & 8) are in the small 1st group (Figure 4 and Figure 5). Four groups of older age students are the middle group. Six groups of younger students are the 3rd group. The variables are grouped into three groups. The first group is almost entirely the expected correct answers. The 2nd group consists of correct answers, but the answer is not related to the proverb. The 3rd group of answers is clearly wrong answers. We find that Powel ([6]) is correct that the students are using reasoning when they come up with a wrong answer.

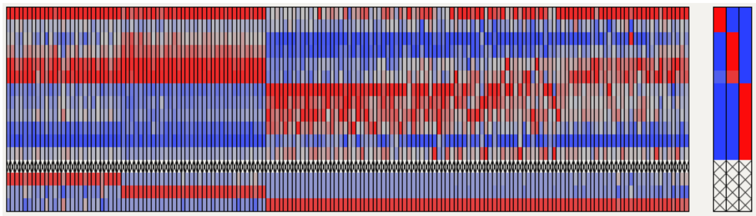

Figure 4: Combined heat map of NMF factorization, $k = 3$.

Row	ID	Age
1	08_354	168<x<=180
2	07_156	156<x<=168
3	12_157	216<x
4	11_212	204<x<=216
5	10_294	192<x<=204
6	09_320	180<x<=192
7	03_151	108<x<=120
8	04_155	120<x<=132
9	05_171	132<x<=144
10	02_104	96<x<=108
11	01_041	<=96
12	06_173	144<x<=156

Figure 5: NMF Organization of Age Groups. NMF with k=3 groups rows 1-2, rows 3-6 and rows 7-12.

Moreover, we have found that young students are on the lower left of the Figure 6 and as age increases the groups move from left to right. Group

8, marked with a "T" appears to be a transition group. The older groups, marked with "O" move from right to left. Component 2 may be assessing the students ability to distinguish the"expected correct answer" from the "correct, but not related to the proverb" answer.

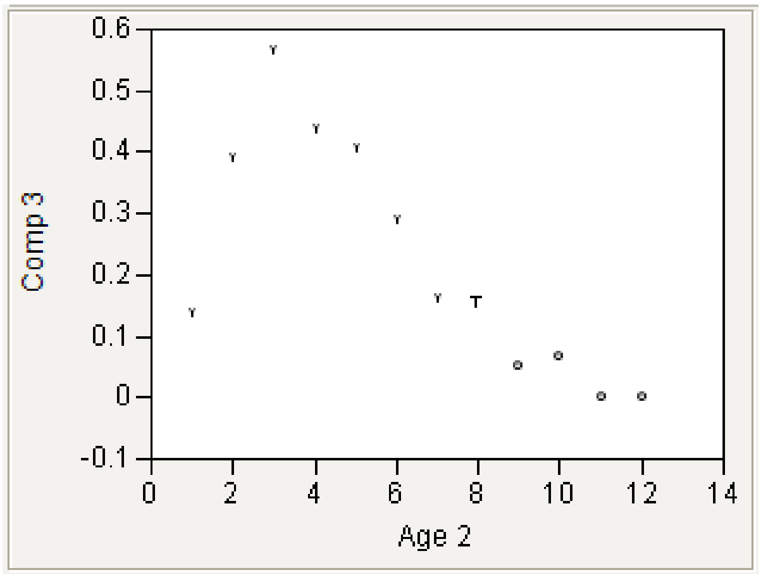

Figure 6: A plot of Component 2 versus Component 1.

The reasoning of choosing $k = 3$ fits with subject matter expert opinion and can be observed in Figure 7. The opinion is that young students use logic to try and figure out the world. Older students shift to memory and authority. In summary, we think non-negative matrix factorization should be considered as an analysis tool when examining contingency tables. Here, as in other settings, the method often suggests relationships are appealing to subject matter experts.

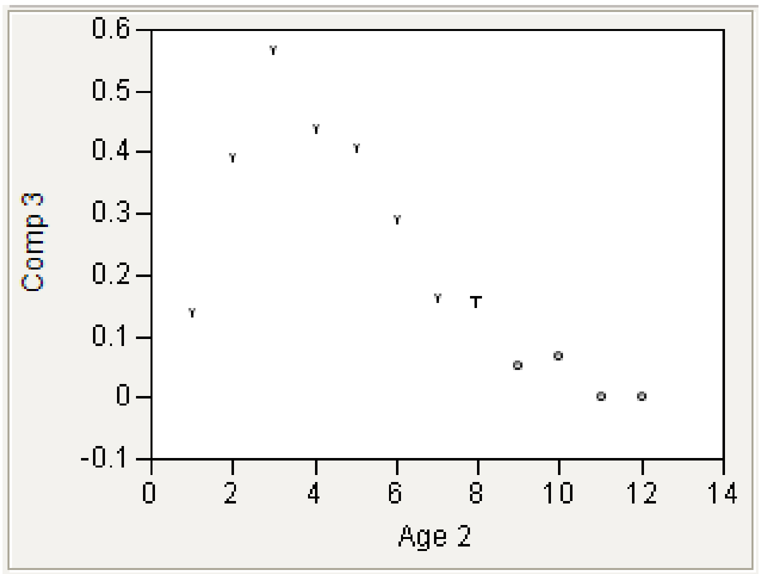

Figure 7: Component 3 versus the student age.

References

[1] Brunet, J.P., Tamayo, P., Gould, T.R., and Mesirov, J.P.(2004) Metagenes and molecular pattern discovery using matrix factorization, Proceedings of the National Academy of Sciences. 101,4164-4169.

[2] Fogeel, P., Hawkins, D.M., Beecher, C., Luta, G. and Young, S.S.(In Press) A Tale of two matrix factorization, The American Statistician.

[3] Gorham, D. R. (1957). Proverbs test. Missoula, MT: Psychological Test Specialists.

[4] Hutchins, L.N.(2008) Position-dependent motif characterization using non-negative matrix factorization, Bioinformatics. 1;24(23),2684-901.

[5] Good, I.J.(1969) Some applications of the singular decomposition of a matrix, Technometrics. 11(4),823-831.

[6] Powell, J. C. (1968). The interpretation of wrong answers from a multiple-choice test. Educational and Psychological Measurement, 28, 403-412.

[7] Powell, J. C. (2010). Testing as feedback to inform teaching. In J. Michael Spector, et al. Learning and Instruction in the Digital Age: Making a Difference through Cognitive Approaches. New York: Springer. Chapter 3.

[8] Lee, D.D., and Seung, H.S.(1999) Algorithms for non-negative matrix factorization, Nature. 401,788-793.