Factor Analysis of Incidence Data via Novel Decomposition of Matrices*

Radim Belohlavek[1,2] and Vilem Vychodil[1,2]

[1] State University of New York at Binghamton
PO Box 6000, Binghamton, NY 13902–6000, USA
rbelohla@binghamton.edu, vychodil@binghamton.edu
[2] Palacky University, Olomouc
Tomkova 40, CZ-779 00 Olomouc, Czech Republic

Abstract. Matrix decomposition methods provide representations of an object-variable data matrix by a product of two different matrices, one describing relationship between objects and hidden variables or factors, and the other describing relationship between the factors and the original variables. We present a novel approach to decomposition and factor analysis of matrices with incidence data. The matrix entries are grades to which objects represented by rows satisfy attributes represented by columns, e.g. grades to which an image is red or a person performs well in a test. We assume that the grades belong to a scale bounded by 0 and 1 which is equipped with certain aggregation operators and forms a complete residuated lattice. We present an approximation algorithm for the problem of decomposition of such matrices with grades into products of two matrices with grades with the number of factors as small as possible. Decomposition of binary matrices into Boolean products of binary matrices is a special case of this problem in which 0 and 1 are the only grades. Our algorithm is based on a geometric insight provided by a theorem identifying particular rectangular-shaped submatrices as optimal factors for the decompositions. These factors correspond to formal concepts of the input data and allow for an easy interpretation of the decomposition. We present the problem formulation, basic geometric insight, algorithm, illustrative example, experimental evaluation.

1 Introduction

1.1 Problem Description

Reducing data dimensionality by mapping the data from the space of directly observable variables into a lower dimensional space of new variables is of fundamental importance for understanding and management of data. Traditional approaches achieve dimensionality reduction via matrix decomposition. In factor analysis, a decomposition of an object-variable matrix is sought into an object-factor matrix and a factor-variable matrix with the number of factors reasonably

* Supported by research plan MSM 6198959214.

S. Ferré and S. Rudolph (Eds.): ICFCA 2009, LNAI 5548, pp. 83–97, 2009.

small. This way, objects can be represented in a lower dimensional space of factors from which their representation in the space of original variables can be retrieved by a linear combination.

Recently, new methods of matrix decomposition and dimensionality reduction have been developed. One aim is to have methods which are capable of discovering possibly non-linear relationships between the original space and the lower dimensional space [26,32]. Another is driven by the need to take into account constraints imposed by the semantics of the data. An example is Boolean factor analysis in which a decomposition of a binary matrix is sought into two binary matrices [10,21,25].

In this paper, we consider decompositions of matrices I with a particular type of ordinal data. Entries I_{ij} of I are grades to which the object corresponding to i-th row has, or is incident with, the attribute corresponding to the j-th row, e.g. to which a hotel is rated as a good hotel. Typical examples of such data are results of questionnaires where respondents (rows) rate services, products, etc. according to various criteria (columns); results of performance evaluation of people or machines (rows) by various tests (columns); or binary data in which case there are only two grades, 0 (no, failure) and 1 (yes, success). Our goal is to decompose an $n \times m$ object-attribute matrix I into a product

$$I = A \circ B$$

of an $n \times k$ object-factor matrix A and a $k \times m$ factor-attribute matrix B with a reasonably small number k of factors.

The scenario is thus similar to that of ordinary factor analysis but there are important differences. First, we assume that the entries of I, i.e. the grades, as well as the entries of A and B are taken from a bounded scale L of grades. Examples of such scales are the unit interval $L = [0, 1]$, the Likert scale $L = \{1, \ldots, 5\}$ of degrees of satisfaction, or other scales used in mathematical psychology and psychophysics [18]. Second, the matrix composition operation $\circ$ used in our decompositions is not the usual matrix product. Instead, we use a so-called t-norm-based product where a t-norm is a function which we use for aggregation of grades, cf. also [9]. A Boolean matrix product is a particular case of this product in which the scale has 0 and 1 as the only grades. It is to be emphasized that we attempt to treat graded incidence data in a way which is compatible with its semantics. This need has been recognized long ago in mathematical psychology, in particular in measurement theory [18]. For example, even if we represent the grades by numbers such as 0 $\sim$ strongly disagree, $\frac{1}{4}$ $\sim$ disagree, ..., 1 $\sim$ strongly agree, addition, multiplication by real numbers, and linear combination of graded incidence data may not have natural meaning. Likewise, decomposition of graded incidence matrix I into the ordinary matrix product of arbitrary real-valued matrices A and B suffers from a difficulty to interpret A and B, as well as to interpret the way I is reconstructed from, or explained by, A and B. This is not to say that the usual matrix decompositions of incidence data I are not useful. [22,31] report that decompositions of binary matrices into real-valued matrices may yield better reconstruction accuracies. Hence, as far as the dimensionality reduction aspect (the technical aspect) is concerned, ordinary decompositions may be favorable. However, when the knowledge discovery aspect plays a role, attention needs to be paid to the semantics of decomposition, and to the appropriate mathematical structure and geometry of the space of attributes, the space of factors, and the transformations between them.

Our paper is organized as follows. Section 1.2 provides and overview of related work. In Section 2, we define the decomposition problem and explain the factors we use for the decomposition and their role. Section 3 contains an illustrative example. An algorithm for decompositions of incidence matrices and its experimental evaluation is presented in Section 4.

1.2 Related Work

In case of matrices with real numbers as entries (sometimes referred to as continuous data), various methods for matrix decompositions have been developed. The best known include, in particular, factor analysis (FA), principal component analysis (PCA), and singular value decomposition (SVD) [2,13]. Results regarding optimality of such decompositions are available. However, these methods decompose a real-valued matrix into a product of matrices with possibly negative values which are generally difficult to interpret [22]. Non-negative matrix factorization [19] overcomes this problem at the expense of not minimizing the global reconstruction error. The advantage is that the decomposition describes the original data as additively composed of its easily interpretable parts. Restriction to certain values (non-negative ones) and the resulting gain in interpretability is related to our work.

There are several papers on decomposing binary matrices into non-binary matrices such as [20,27,28,30,36], see also [31] for further references.

Decompositions of binary matrices into binary matrices have been studied in a number of papers. Early work was done by Markowsky et al., see e.g. [24,25,29] which already include complexity results showing the hardness of problems related to such decompositions. Decompositions of binary matrices into a Boolean product of binary matrices using Hopfield-like associative neural networks have been studied, e.g., by Frolov et al., see [10]. This approach is a heuristic in which the factors correspond to attractors of the neural network. Other heuristic approaches to Boolean factor analysis include [15,16]. [6] shows that the decomposition of a binary matrix I to a Boolean product of binary matrices is equivalent to covering the matrix by rectangular submatrices which contain 1s and shows that formal concepts of I [11] are optimal factors for such decomposition. The problem of covering binary matrices with their submatrices containing 1s was studied in [12]. [22] presents an algorithm for finding approximate decompositions of binary matrices into Boolean product of binary matrices which is based on associations between columns of I. [33] looks at the relationship between several problems related to decomposition of binary matrices.

2 Decomposition and Factors

2.1 Decomposition

Consider an $n \times m$ object-attribute matrix I with entries I_{ij} expressing grades to which object i has attribute j. We assume that the grades are taken from a bounded scale L. In general, we assume that L is equipped with a partial order $\leq$, is bounded from below and above by elements denoted 0 and 1, and conforms to the structure of a complete lattice, i.e. infima and suprema of subsets of L exist. We do not assume the scale to be linearly ordered although in practical applications this is usually the case. Grades of ordinal scales are conveniently represented by numbers, such as the Likert scale $\{1, \ldots, 5\}$. In such a case we assume these numbers are normalized and taken from the unit interval $[0, 1]$. As an example, the Likert scale is represented by $L = \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$. Due to the well-known Miller's 7 ± 2 phenomenon [23], one might argue that we should restrict ourselves to small scales but we consider arbitrary ones, including thus the unit interval $L = [0, 1]$ as well.

We want to decompose I into an $n \times k$ object-factor matrix A and a $k \times m$ factor-attribute matrix B which again have entries from the scale L. Entries A_{il} and B_{lj} are interpreted as degrees to which factor l applies to object i and to which attribute j is a manifestation of factor l, respectively. We assume that the object-attribute relationship is explained using the (hidden) factors as follows: object i has attribute j if there is a factor l which applies to i and for which j is one of its manifestations. Now, for a factor l there is a degree A_{il} to which l applies to i and a degree B_{lj} to which j is a manifestation of l. To obtain a degree a to which "l applies to i and j is a manifestation of l", we aggregate A_{il} and B_{lj} using an aggregation function $\otimes : L \times L \to L$ and put $a = A_{il} \otimes B_{lj}$, cf. [9]. This way, we obtain k degrees $A_{il} \otimes B_{lj}$, one for every factor $l = 1, \ldots, k$. Finally, we take the supremum $\bigvee$ of degrees $A_{il} \otimes B_{lj}$ (such supremum coincides with maximum if L is linearly ordered) as a result. That is, our composition operation for $I = A \circ B$ is defined by

$$(A \circ B)_{ij} = \bigvee\nolimits_{l=1}^{k} A_{il} \otimes B_{lj}. \tag{1}$$

Notice that if $L = \{0, 1\}$ and $\otimes$ is the truth function of conjunction, $A \circ B$ is the Boolean matrix product. We use t-norms for aggregation functions $\otimes$. T-norms originated in K. Menger's work on statistical metric spaces [17] and are used as truth functions of conjunctions in fuzzy logic [14]. Their properties make them good candidates for aggregating graded data [8,9,17]. Note that with $\otimes$ being a t-norm, (1) is used in fuzzy set theory to define compositions of fuzzy relations [35]. Examples of $\otimes$ include the Łukasiewicz t-norm on $L = [0, 1]$ or on an equidistant subchain of $[0, 1]$ defined by $a \otimes b = \max(0, a + b - 1)$, the minimum t-norm on $L = [0, 1]$ or on a subset of $[0, 1]$ defined by $a \otimes b = \min(a, b)$, and the product t-norm $a \otimes b = a \cdot b$ on $L = [0, 1]$. Using a decomposition $I = A \circ B$ with (1), attributes are expressed by means of factors in a non-linear manner:

Example 1. With Łukasiewicz t-norm, let $I = A \circ B$ be

$$\begin{pmatrix} 0.3 & 0.0 & 0.1 \\ 0.3 & 0.7 & 0.5 \\ 0.5 & 0.8 & 0.6 \end{pmatrix} = \begin{pmatrix} 0.2 & 0.8 \\ 0.9 & 0.8 \\ 1.0 & 1.0 \end{pmatrix} \circ \begin{pmatrix} 0.4 & 0.8 & 0.6 \\ 0.5 & 0.2 & 0.3 \end{pmatrix}.$$

Then for $Q_1 = (0.6\ 0.2)$ and $Q_2 = (0.4\ 0.3)$ we have $(Q_1 + Q_2) \circ B = (1.0\ 0.5) \circ B = (0.4\ 0.8\ 0.6) \neq (0.0\ 0.6\ 0.2) = (0.0\ 0.4\ 0.2) + (0.0\ 0.2\ 0.0) = Q_1 \circ B + Q_2 \circ B$. This demonstrates non-linearity of the relationship between factors and attributes.

2.2 Factors for Decomposition

Next, we describe the factors we use for decomposition of I. For this purpose, we make use of a so-called residuum induced by the t-norm $\otimes$ [14,17], i.e. a binary function $\rightarrow$ on L defined by

$$a \rightarrow b = \max\{c \in L \mid a \otimes c \leq b\}.$$

Residuum satisfies an important technical condition called adjointness, namely,

$$a \otimes b \leq c \text{ iff } a \leq b \rightarrow c.$$

L together with $\otimes$ and $\rightarrow$ forms a complete residuated lattice [34]. We leave out technical details including the properties of residuated lattices and refer to [14]. The residuum induced by the Łukasiewicz t-norm is defined by $a \rightarrow b = \min(1, 1 - a + b)$.

We are going to use formal concepts associated to I as factors for a decomposition of I. Formal concepts are particular pairs $\langle C, D\rangle$ of graded sets (fuzzy sets) C of objects and D of attributes, see [4]. That is, $C : \{1, \ldots, n\} \rightarrow L$ assigns to every object i a degree $C(i) \in L$ to which C applies to i. Likewise, $D : \{1, \ldots, m\} \rightarrow L$ assigns to every attribute j a degree to which D applies to j. Denote by L^U the set of all graded (fuzzy) sets in a set U, i.e. the set of all mappings from U to L, and put $X = \{1, \ldots, n\}$ (objects) and $Y = \{1, \ldots, m\}$ (attributes).

Definition 1. *[4] A formal concept of I is any pair $\langle C, D\rangle$ for which $C^{\uparrow} = D$ and $D^{\downarrow} = C$ where $^{\uparrow} : L^X \rightarrow L^Y$ and $^{\downarrow} : L^Y \rightarrow L^X$ are operators defined by*

$$C^{\uparrow}(j) = \textstyle\bigwedge_{i \in X}(C(i) \rightarrow I_{ij}),$$
$$D^{\downarrow}(i) = \textstyle\bigwedge_{j \in Y}(D(j) \rightarrow I_{ij}).$$

In the definition, $\bigwedge$ is the infimum in L (in our case, since X and Y are finite, infimum coincides with minimum if L is linearly ordered). The set $\mathcal{B}(X, Y, I)$ of all formal concepts of I is called the concept lattice of I. Formal concepts are simple models of concepts in the sense of traditional, Port-Royal logic. If I is (a characteristic function of) an ordinary binary relation (i.e. $L = \{0, 1\}$), formal concepts of I coincide with the ordinary formal concepts of Wille [11]. C and D are called the extent and the intent of a formal concept $\langle C, D\rangle$ and represent the objects and the attributes which fall under the concept. The graded setting takes into account that empirical concepts are graded rather than clear-cut. The concept lattice equipped with a subconcept-superconcept ordering $\leq$ defined by

$$\langle C_1, D_1\rangle \leq \langle C_2, D_2\rangle \text{ iff } C_1(i) \leq C_2(i) \text{ for all } i \in X,$$

which is equivalent to $D_2(j) \leq D_1(j)$ for all $j \in Y$, is indeed a complete lattice [4]. Note that since $\rightarrow$ can be interpreted as a truth function of implication, a formal concept $\langle C, D\rangle$ can be seen as a pair of graded sets C and D such that $D(j)$ is the degree to which j is shared by all objects to which C applies, and $C(i)$ is the degree to which i shares all attributes to which D applies [4].

We are going to use formal concepts of I in the following way. For a set

$$\mathcal{F} = \{\langle C_1, D_1\rangle, \ldots, \langle C_k, D_k\rangle\}$$

of formal concepts of I, we denote by $A_\mathcal{F}$ an $n \times k$ matrix in which the l-th column consists of grades assigned to objects by C_l. Likewise, we denote by $B_\mathcal{F}$ a $k \times m$ matrix in which the l-th row consists of grades assigned to attributes by D_l. That is,

$$(A_\mathcal{F})_{il} = (C_l)(i) \quad \text{and} \quad (B_\mathcal{F})_{lj} = (D_l)(j).$$

If $I = A_\mathcal{F} \circ B_\mathcal{F}$, $\mathcal{F}$ can be seen as a set of factors which fully explain the data. In such a case, we call the formal concepts from $\mathcal{F}$ *factor concepts*. Given I, our aim is to find a small set $\mathcal{F}$ of factor concepts. Using formal concepts of I as factors is intuitively appealing because, as mentioned above, the formal concepts are in fact, simple models of human concepts according to traditional logic approach. In fact, factors are often called "(hidden) concepts" in the ordinary factor analysis. In addition, the extents and intents of the concepts, i.e. columns and rows of $A_\mathcal{F}$ and $B_\mathcal{F}$, have a straightforward interpretation: they represent the grades to which the factor concept applies to particular objects and particular attributes.

Before we turn to an illustrative example, we provide a geometric interpretation of $A \circ B$. Let $I = A \circ B$. Denote by J_l the $n \times m$ matrix defined by $(J_l)_{ij} = A_{il} \otimes B_{lj}$, $l = 1 \ldots, k$. That is, $J_l = A_{\_l} \circ B_{l\_}$ is the $\circ$-product of the l-th column of A and the l-th row of B. (1) then yields that $I = J_1 \vee \cdots \vee J_k$, i.e. I is the $\bigvee$-superposition of $J_1, \ldots, J_k$. Matrices J_l are rectangular (rectangles) in that they result as the Cartesian products of graded sets. If $L = \{0, 1\}$, rectangular matrices are just matrices where the entries containing 1s form a submatrix, i.e. tiles in the sense of [12]. We thus have:

Theorem 1. *$I = A \circ B$ for an $n \times k$ matrix A and a $k \times m$ matrix B if and only if I is a $\bigvee$-superposition of rectangular matrices $A_{\_l} \circ B_{l\_}$, $l = 1, \ldots, k$.*

Remark 1. (1) Note that due to Theorem 1, tiling databases [12] means decomposing I into $A \circ B$ where columns of A and rows of B are the characteristic vectors of the sets of objects and attributes covered by the tiles.

(2) This remains true even for arbitrary scales L: Finding a decomposition of I is equivalent to covering I by rectangular submatrices, i.e. "graded tiles", which result by the Cartesian products of graded sets of objects and attributes, and are contained in I.

(3) Let $\mathcal{F}$ be a set of factor concepts, i.e. $I = A_\mathcal{F} \circ B_\mathcal{F}$. Due to Theorem 1, for any subset $\mathcal{F}'$ of $\mathcal{F}$ we have $(A_{\mathcal{F}'} \circ B_{\mathcal{F}'})_{ij} \leq I_{ij}$. That is, for any subset $\mathcal{F}'$ of $\mathcal{F}$, $A_{\mathcal{F}'} \circ B_{\mathcal{F}'}$ approximates I from below. We will see in Sections 3 and 4 that it is usually the case that even for a small subset $\mathcal{F}' \subseteq \mathcal{F}$, matrix $A_{\mathcal{F}'} \circ B_{\mathcal{F}'}$ is a good approximation of I.

2.3 Optimality of Formal Concepts as Factors

In this section we recall two important results from [5]. The first one says that formal concepts of I are universal factors.

Theorem 2. *For every I there is $\mathcal{F} \subseteq \mathcal{B}(X, Y, I)$ such that $I = A_{\mathcal{F}} \circ B_{\mathcal{F}}$.*

The second one says that, as far as exact decompositions of I are concerned, formal concepts are optimal factors in that they provide us with decompositions of I with the least number k of factors.

Theorem 3. *If $I = A \circ B$ for $n \times k$ and $k \times m$ binary matrices A and B, there exists a set $\mathcal{F} \subseteq \mathcal{B}(X, Y, I)$ of formal concepts of I with $|\mathcal{F}| \leq k$ such that for the $n \times |\mathcal{F}|$ and $|\mathcal{F}| \times m$ matrices $A_{\mathcal{F}}$ and $B_{\mathcal{F}}$ we have $I = A_{\mathcal{F}} \circ B_{\mathcal{F}}$.*

This means that in looking for decompositions of I, one can restrict the search to the set of formal concepts instead of the set of all possible decompositions.

3 Illustrative Example

Tab. 1 (top) contains the results of top five athletes in 2004 Olympic Games decathlon in points which are obtained using the IAAF Scoring Tables for Combined Events. Note that the IAAF Scoring Tables provide us with an ordinal scale and a ranking function assigning the scale values to athletes. We are going to look at whether this data can be explained using formal concepts as factors. We first transform the data to a five-element scale

$$L = \{0.00, 0.25, 0.50, 0.75, 1.00\}$$

by a natural transformation and rounding. As a consequence, the factors then have a simple reading. Namely, the grades to which a factor applies to an athlete

Table 1. 2004 Olympic Games Decathlon

Scores of Top 5 Athletes

	10	*lj*	*sp*	*hj*	40	11	*di*	*pv*	*ja*	15
Sebrle	894	1020	873	915	892	968	844	910	897	680
Clay	989	1050	804	859	852	958	873	880	885	668
Karpov	975	1012	847	887	968	978	905	790	671	692
Macey	885	927	835	944	863	903	836	731	715	775
Warners	947	995	758	776	911	973	741	880	669	693

Incidence Data Table with Graded Attributes

	10	*lj*	*sp*	*hj*	40	11	*di*	*pv*	*ja*	15
Sebrle	0.50	1.00	1.00	1.00	0.75	1.00	0.75	0.75	1.00	0.75
Clay	1.00	1.00	0.75	0.75	0.50	1.00	0.75	0.50	1.00	0.50
Karpov	1.00	1.00	0.75	0.75	1.00	1.00	1.00	0.25	0.25	0.75
Macey	0.50	0.50	0.75	1.00	0.75	0.50	0.75	0.25	0.50	1.00
Warners	0.75	0.75	0.50	0.50	0.75	1.00	0.25	0.50	0.25	0.75

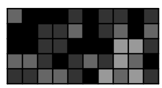
Legend: 10—100 meters sprint race; *lj*—long jump; *sp*—shot put; *hj*—high jump; 40—400 meters sprint race; 11—110 meters hurdles; *di*—discus throw; *pv*—pole vault; *ja*—javelin throw; 15—1500 meters run.

Table 2. Factor Concepts

F_i	*Extent*	*Intent*
F_1	$\{{}^{.5}/$Sebrle, Clay, Karpov, ${}^{.5}/$Macey, ${}^{.75}/$Warners$\}$	$\{$10, lj, ${}^{.75}/$sp, ${}^{.75}/$hj, ${}^{.5}/$40, 11, ${}^{.5}/$di, ${}^{.25}/$pv, ${}^{.25}/$ja, ${}^{.5}/$15$\}$
F_2	$\{$Sebrle, ${}^{.75}/$Clay, ${}^{.25}/$Karpov, ${}^{.5}/$Macey, ${}^{.25}/$Warners$\}$	$\{{}^{.5}/$10, lj, sp, hj, ${}^{.75}/$40, 11, ${}^{.75}/$di, ${}^{.75}/$pv, ja, ${}^{.75}/$15$\}$
F_3	$\{{}^{.75}/$Sebrle, ${}^{.5}/$Clay, ${}^{.75}/$Karpov, Macey, ${}^{.5}/$Warners$\}$	$\{{}^{.5}/$10, ${}^{.5}/$lj, ${}^{.75}/$sp, hj, ${}^{.75}/$40, ${}^{.5}/$11, ${}^{.75}/$di, ${}^{.25}/$pv, ${}^{.5}/$ja, 15$\}$
F_4	$\{$Sebrle, ${}^{.75}/$Clay, ${}^{.75}/$Karpov, ${}^{.5}/$Macey, Warners$\}$	$\{{}^{.5}/$10, ${}^{.75}/$lj, ${}^{.5}/$sp, ${}^{.5}/$hj, ${}^{.75}/$40, 11, ${}^{.25}/$di, ${}^{.5}/$pv, ${}^{.25}/$ja, ${}^{.75}/$15$\}$
F_5	$\{{}^{.75}/$Sebrle, ${}^{.75}/$Clay, Karpov, ${}^{.75}/$Macey, ${}^{.25}/$Warners$\}$	$\{{}^{.75}/$10, ${}^{.75}/$lj, ${}^{.75}/$sp, ${}^{.75}/$hj, ${}^{.75}/$40, ${}^{.75}/$11, di, ${}^{.25}/$pv, ${}^{.25}/$ja, ${}^{.75}/$15$\}$
F_6	$\{{}^{.75}/$Sebrle, ${}^{.5}/$Clay, Karpov, ${}^{.75}/$Macey, ${}^{.75}/$Warners$\}$	$\{{}^{.75}/$10, ${}^{.75}/$lj, ${}^{.75}/$sp, ${}^{.75}/$hj, 40, ${}^{.75}/$11, ${}^{.5}/$di, ${}^{.25}/$pv, ${}^{.25}/$ja, ${}^{.75}/$15$\}$
F_7	$\{$Sebrle, Clay, ${}^{.25}/$Karpov, ${}^{.5}/$Macey, ${}^{.25}/$Warners$\}$	$\{{}^{.5}/$10, lj, ${}^{.75}/$sp, ${}^{.75}/$hj, ${}^{.5}/$40, 11, ${}^{.75}/$di, ${}^{.5}/$pv, ja, ${}^{.5}/$15$\}\rangle$

can be described in natural language as “not at all”, “little bit”, “half”, “quite”, “fully”, or the like. Tab. 1 (bottom) describes the athletes’ performance using the five-element scale. In addition, we use the Łukasiewicz t-norm on L.

Using shades of gray to represent grades from the five-element scale L, the matrix I corresponding to Tab. 1 (bottom) can be visualized in the following array (rows correspond to athletes, columns correspond to disciplines, the darker the array entry, the higher the score):



The algorithm described in Section 4 found a set $\mathcal{F}$ of 7 formal concepts which factorize I, i.e. for which $I = A_{\mathcal{F}} \circ B_{\mathcal{F}}$. These factor concepts are shown in Fig. 1 in the order in which they were produced by the algorithm. For example, factor concept F_1 applies to Sebrle to degree 0.5, to both Clay and Karpov to degree 1, to Macey to degree 0.5, and to Warners to degree 0.75. Furthermore, this factor concept applies to attribute 10 (100 m) to degree 1, to attribute lj (long jump) to degree 1, to attribute sp (shot put) to degree 0.75, etc. This means that an excellent performance (degree 1) in 100 m, an excellent performance in long jump, a very good performance (degree 0.75) in shot put, etc. are particular manifestations of this factor concept. On the other hand, only a relatively weak performance (degree 0.25) in javelin throw and pole vault are manifestations of this factor.

Therefore, a decomposition $I = A_{\mathcal{F}} \circ B_{\mathcal{F}}$ exists with 7 factors where:

$$A_{\mathcal{F}} = \begin{pmatrix} 0.50 & 1.00 & 0.75 & 1.00 & 0.75 & 0.75 & 1.00 \\ 1.00 & 0.75 & 0.50 & 0.75 & 0.75 & 0.50 & 1.00 \\ 1.00 & 0.25 & 0.75 & 0.75 & 1.00 & 1.00 & 0.25 \\ 0.50 & 0.50 & 1.00 & 0.50 & 0.75 & 0.75 & 0.50 \\ 0.75 & 0.25 & 0.50 & 1.00 & 0.25 & 0.75 & 0.25 \end{pmatrix},$$

$$B_{\mathcal{F}} = \begin{pmatrix} 1.00 & 1.00 & 0.75 & 0.75 & 0.50 & 1.00 & 0.50 & 0.25 & 0.25 & 0.50 \\ 0.50 & 1.00 & 1.00 & 1.00 & 0.75 & 1.00 & 0.75 & 0.75 & 1.00 & 0.75 \\ 0.50 & 0.50 & 0.75 & 1.00 & 0.75 & 0.50 & 0.75 & 0.25 & 0.50 & 1.00 \\ 0.50 & 0.75 & 0.50 & 0.50 & 0.75 & 1.00 & 0.25 & 0.50 & 0.25 & 0.75 \\ 0.75 & 0.75 & 0.75 & 0.75 & 0.75 & 0.75 & 1.00 & 0.25 & 0.25 & 0.75 \\ 0.75 & 0.75 & 0.75 & 0.75 & 1.00 & 0.75 & 0.50 & 0.25 & 0.25 & 0.75 \\ 0.50 & 1.00 & 0.75 & 0.75 & 0.50 & 1.00 & 0.75 & 0.50 & 1.00 & 0.50 \end{pmatrix}.$$

Again, using shades of gray, this decomposition can be depicted as:

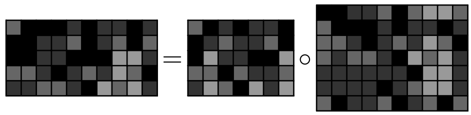


Fig. 1 shows the rectangular patterns corresponding to the factor concepts, cf. Theorem 1.

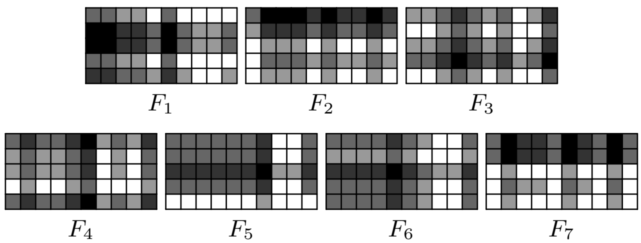


Fig. 1. Factor Concepts as Rectangular Patterns

Fig. 2 demonstrates what portion of the data matrix I is explained using just some of the factor concepts from $\mathcal{F}$. The first matrix labeled by 46% shows $A_{\mathcal{F}_1} \circ B_{\mathcal{F}_1}$ for $\mathcal{F}_1$ consisting of the first factor F_1 only. That is, the matrix is just the rectangular pattern corresponding to F_1, cf. Fig. 1. As we can see, this matrix is contained in I, i.e. approximates I from below, in that $(A_{\mathcal{F}_1} \circ B_{\mathcal{F}_1})_{ij} \leq I_{ij}$ for all entries (row i, column j). Note that Theorem 1 implies that this always needs to be the case, cf. Remark 1 (3). Label 46% indicates that 46% of the entries of $A_{\mathcal{F}_1} \circ B_{\mathcal{F}_1}$ and I are equal. In this sense, the first factor explains 46% of the data. Note however, that several of the $54\% = 100\% - 46\%$ of the other entries of $A_{\mathcal{F}_1} \circ B_{\mathcal{F}_1}$ are close to the corresponding entries of I, so a measure of closeness of $A_{\mathcal{F}_1} \circ B_{\mathcal{F}_1}$ and I which takes into account also close entries, rather than exactly equal ones only, would yield a number larger than 46%.

The second matrix in Fig. 2, with label 72%, shows $A_{\mathcal{F}_2} \circ B_{\mathcal{F}_2}$ for $\mathcal{F}_2$ consisting of F_1 and F_2. That is, the matrix demonstrates what portion of the data matrix I is explained by the first two factors. Again, $A_{\mathcal{F}_2} \circ B_{\mathcal{F}_2}$ approximates I from below and 72% of the entries of $A_{\mathcal{F}_2} \circ B_{\mathcal{F}_2}$ and I coincide now. Note again that even for the remaining 28% of entries, $A_{\mathcal{F}_2} \circ B_{\mathcal{F}_2}$ provides a reasonable approximation of I, as can be seen by comparing the matrices representing $A_{\mathcal{F}_2} \circ B_{\mathcal{F}_2}$ and I, i.e. the one labeled by 72% and the one labelled by 100%.

Similarly, the matrices labeled by 84%, 92%, 96%, 98%, and 100% represent $A_{\mathcal{F}_l} \circ B_{\mathcal{F}_l}$ for $l = 3, 4, 5, 6, 7$, for sets $\mathcal{F}_l$ of factor concepts consisting of $F_1, \ldots, F_l$. We can conclude from the visual inspection of the matrices that already the two or three first factors explain the data reasonably well.

Let us now focus on the interpretation of the factors. Fig. 1 is helpful as it shows the clusters corresponding to the factor concepts which draw together the athletes and their performances in the events.

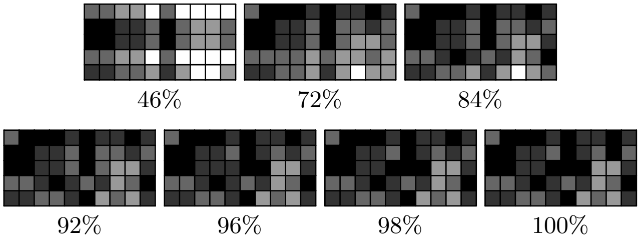


Fig. 2. $\bigvee$-superposition of Factor Concepts

Factor F_1: Manifestations of this factor with grade 1 are 100 m, long jump, 110 m hurdles. This factor can be interpreted as the ability to run fast for short distances. Note that this factor applies particularly to Clay and Karpov which is well known in the world of decathlon. Factor F_2: Manifestations of this factor with grade 1 are long jump, shot put, high jump, 110 m hurdles, javelin. F_2 can be interpreted as the ability to apply very high force in a very short term (explosiveness). F_2 applies particularly to Sebrle, and then to Clay, who are known for this ability. Factor F_3: Manifestations with grade 1 are high jump and 1500 m. This factor is typical for lighter, not very muscular athletes (too much muscles prevent jumping high and running long distances). Macey, who is evidently that type among decathletes (196 cm and 98 kg) is the athlete to whom the factor applies to degree 1. These are the most important factors behind data matrix I.

4 Algorithm and Experiments

In this section, we present a greedy approximation algorithm which takes a data matrix representing I as its input and produces a set $\mathcal{F}$ of formal concepts of I for which $I = A_{\mathcal{F}} \circ B_{\mathcal{F}}$. Due to Theorem 1, finding such $\mathcal{F}$ which is minimal in terms of the number of its elements is equivalent to finding a minimal subset of $\{C \otimes D \,|\, \langle C, D\rangle \in \mathcal{B}(X, Y, I)\}$ which covers I. Note that this can be seen as a graded version of a set covering problem which apparently has not been studied before. A further study of this problem including various versions of approximate coverings may yield useful results for processing of graded data. Now, a particular case of this problem for $L = \{0, 1\}$ is just the problem of covering a binary matrix with the smallest possible set of rectangles. This problem is known to be NP-hard, see [24,25,29] for early references, and also [6,12,33]. This indicates that we need an approximation algorithm for the problem of finding small $\mathcal{F}$ for which $I = A_{\mathcal{F}} \circ B_{\mathcal{F}}$.

An obvious approach to the design of such algorithm is to take an approximation algorithm for the (binary) set covering problem, such as the one described in [7], and modify it for the graded case. Such an algorithm would require us to compute first the set $\mathcal{B}(X, Y, I)$ of all formal concepts of I and then select candidates for factors from $\mathcal{B}(X, Y, I)$ using a greedy approach [7]. This would be time-demanding because $\mathcal{B}(X, Y, I)$ can be quite large.

Instead, we propose a different greedy algorithm. The algorithm generates maximal rectangles by looking for "promising columns". A technical property which we utilize is the fact that for each formal concept $\langle C, D\rangle$,

$$D = \bigcup_{j \in Y} \{^{D(j)}/j\}^{\downarrow\uparrow},$$

i.e. each intent D is a union of intents $\{^{D(j)}/j\}^{\downarrow\uparrow}$. Moreover, $C = D^{\downarrow}$. Here, $\{^{D(j)}/j\}$ denotes a graded singleton, i.e. the grade of j in $\{^{D(j)}/j\}$ is $D(j)$. As a consequence, we may construct any formal concept by adding sequentially $\{^{a}/j\}^{\downarrow\uparrow}$ to the empty set of attributes. Our algorithm follows a greedy approach that makes us select $j \in Y$ and degree $a \in L$ which maximize the size of

$$D \oplus_a j = \{\langle k, l\rangle \in \mathcal{U} \mid D^{+\downarrow}(k) \otimes D^{+\downarrow\uparrow}(l) \geq I_{kl}\}, \tag{2}$$

where $D^{+} = D \cup \{^{a}/j\}$ and $\mathcal{U}$ denotes the set of $\langle i, j\rangle$ of I (row i, column j) for which the corresponding entry I_{ij} is not covered yet. Note that the size of $D \oplus_a j$ is just the number of entries of I which are covered by formal concept $\langle D^{\downarrow}, D\rangle$. Therefore, instead of going through all possible formal concepts and selecting a factor from them, we just go through columns and degrees which maximize the value of the factor, i.e. the area covered by the factor, which is being constructed. The algorithm is summarized below.

Find-Factors(I)
1 $\mathcal{U} \leftarrow \{\langle i, j\rangle \mid I_{ij} \neq 0\}$
2 $\mathcal{F} \leftarrow \emptyset$
3 **while** $\mathcal{U} \neq \emptyset$
4   **do** $D \leftarrow \emptyset$
5     $V \leftarrow 0$
6   **select** $\langle j, a\rangle$ **that maximizes** $|D \oplus_a j|$
7   **while** $|D \oplus_a j| > V$
8   **do** $V \leftarrow |D \oplus_a j|$
9         $D \leftarrow (D \cup \{^{a}/j\})^{\downarrow\uparrow}$
10   **select** $\langle j, a\rangle$ **that maximizes** $|D \oplus_a j|$
11     $C \leftarrow D^{\downarrow}$
12     $\mathcal{F} \leftarrow \mathcal{F} \cup \{\langle C, D\rangle\}$
13   **for** $\langle i, j\rangle \in \mathcal{U}$
14   **do if** $I_{ij} \leq C(i) \otimes D(j)$
15   **then**
16             $\mathcal{U} \leftarrow \mathcal{U} \setminus \{\langle i, j\rangle\}$
17 **return** $\mathcal{F}$

The main loop of the algorithm (lines 3–16) is executed until all the nonzero entries of I are covered by at least one factor in $\mathcal{F}$. The code between lines 4 and 10 constructs an intent by adding the most promising columns. After such an intent D is found, we construct the corresponding factor concept and add it to $\mathcal{F}$. The loop between lines 13 and 16 ensures that all matrix entries covered by the last factor are removed from $\mathcal{U}$. Obviously, the algorithm is sound and finishes after finitely many steps with a set $\mathcal{F}$ of factor concepts.

Table 3. Exact Factorizability

k	Łukasiewicz ⊗ no. of factors	minimum ⊗ no. of factors
5	5.205 ± 0.460	6.202 ± 1.037
7	7.717 ± 0.878	10.050 ± 1.444
9	10.644 ± 1.316	13.379 ± 1.676
11	13.640 ± 1.615	15.698 ± 1.753
13	16.423 ± 1.879	17.477 ± 1.787
15	18.601 ± 2.016	18.721 ± 1.863

Experimental Evaluation. We now present experiments with exact and approximate factorization of randomly generated matrices and their evaluation. First, we observed how close is the number of factors found by the algorithm FIND-FACTORS to a known number of factors in artificially created matrices. In this experiment, we were generating 20×20 matrices according to various distributions of 5 grades. These matrices were generated by multiplying $m \times k$ and $k \times n$ matrices. Therefore, the resulting matrices were factorizable with at most k factors. Then, we executed the algorithm to find $\mathcal{F}$ and observed how close is the number $|\mathcal{F}|$ of factors to k. The results are depicted in Tab. 3. We have observed that in the average case, the choice of a t-norm is not essential and all t-norms give approximately the same results. In particular, Tab. 3 describes results for Łukasiewicz and minimum t-norms which can be seen as two limit cases of t-norms [17]. Rows of Tab. 3 correspond to numbers $k = 5, 6, \ldots, 15$ denoting the known number of factors. For each k, we computed the average number of factors produced by our algorithm in 2000 k-factorizable matrices. The average values are written in the form of "average number of factors ± standard deviation".

As mentioned above, factorization and factor analysis of binary data is a special of our setting with $L = \{0, 1\}$, i.e. with the scale containing just two grades. Then, the matrix product $\circ$ given by (1) coincides with the Boolean matrix multiplication and the problem of decomposition of graded matrices coincides with the problem of decomposition of binary matrices into the Boolean product of binary matrices. We performed experiments with our algorithm in this particular case with three large binary data sets (binary matrices) from the Frequent Itemset Mining Dataset Repository, see `http://fimi.cs.helsinki.fi/data/`. In particular, we considered the CHESS, CONNECT, and MUSHROOM data sets. The results are shown in Tab. 4. The columns labeled by n and m show the numbers of rows and columns of the matrices (e.g., MUSHROOM is a 8124×119 binary matrix). The column labeled by 50% says the following: The first number is the number of factors sufficient to explain 50% of the data entries. For example, the first 5 factors explain 50% of data for CHESS data, i.e. $A_{\mathcal{F}} \circ B_{\mathcal{F}}$ covers 50% of entries of matrix I with $|\mathcal{F}| = 4$. The second number is the ratio

$$\text{number of attributes/number of factors}$$

Table 4. Factorization of Boolean Matrices

Input Database	Dimensions n	m	Portion of data explained 50%		70%		90%	
CHESS	3196	75	5	15.00	13	5.77	33	2.27
CONNECT	67557	129	4	32.25	10	12.90	39	3.31
MUSHROOM	8124	119	7	17.00	19	6.26	46	2.59

which can be regarded as the coefficient of reduction of dimensionality. For example, for the MUSHROOM data set, the first 7 factors produced by our algorithm explain 50% of data and the corresponding coefficient of reduction is $119/7 = 17.00$. The columns labeled by 70% and 90% have analogous meaning.

5 Conclusions and Future Work

We presented a novel approach to factor analysis of matrices with ordinal data. The factors in this approach correspond to formal concepts in the data matrix and the relationship between the factors and original attributes is a non-linear one. One feature of the model is a transparent way of treating the grades which results in good interpretability of factors. Another feature is its feasibility regarding theoretical analysis. As an example, the factors we use are optimal in terms of their number. Furthermore, we proposed a greedy approximation algorithm for the problem of finding a small set of factors and provided results of experiments demonstrating its behavior. Future research will include the following topics:

- Comparison, both theoretical and experimental, to other methods of matrix decompositions.
- Approaches to the problem of approximate factorization of I, continuing our experiments with approximate factorization presented in this paper.
- Development of further theoretical insight focusing particularly on reducing further the space of factors to which the search for factors can be restricted. Note that decompositions of a matrix with grades into a binary matrix and a matrix with grades was studied in [3].
- Study the computational complexity aspects of the problem of approximate factorization, in particular the approximability of the problem of finding decompositions of matrix I [1].
- Explore the applications of the decompositions studied in this paper. One application area is factor analysis. The usefulness of the decompositions in this area was illustrated by the example in Section 3. Another topic which needs to be explored is the possible utilization of the dimensionality reduction provided by the decompositions.

References

1. Ausiello, G., et al.: Complexity and Approximation. Combinatorial Optimization Problems and Their Approximability Properties. Springer, Heidelberg (2003)
2. Bartholomew, D.J., Knott, M.: Latent Variable Models and Factor Analysis, 2nd edn., London, Arnold (1999)
3. Bartl, E., Belohlavek, R., Konecny, J.: Optimal decompositions of matrices with grades into binary and graded matrices. In: Proc. CLA 2008, The Sixth Intl. Conference on Concept Lattice and Their Applications, Olomouc, Czech Republic, pp. 59–70 (2008) ISBN 978–80–244–2111–7
4. Belohlavek, R.: Concept lattices and order in fuzzy logic. Annals of Pure and Applied Logic 128(1–3), 277–298 (2004)
5. Belohlavek, R.: Optimal decompositions of matrices with grades. IEEE IS 2008, Proc. Intl. IEEE Conference on Intelligent Systems, Varna, Bulgaria, pp. 15-2–15-7 (2008) IEEE Catalog Number CFP08802-PRT, ISBN 978-1-4244-1740-7
6. Belohlavek, R., Vychodil, V.: Discovery of optimal factors in binary data via a novel method of matrix decomposition. J. Computer and System Sciences (to appear)
7. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 2nd edn. MIT Press, Cambridge (2001)
8. Fagin, R.: Combining fuzzy information from multiple systems. J. Computer and System Sciences 58, 83–99 (1999); Preliminary version in PODS 1996, Montreal, pp. 216–226 (1996)
9. Fagin, R., Lotem, A., Naor, M.: Combining fuzzy information: an overview. SIGMOD Record 31(2), 109–118 (2002)
10. Frolov, A.A., Húsek, D., Muraviev, I.P., Polyakov, P.A.: Boolean factor analysis by Hopfield-like autoassociative memory. IEEE Transactions on Neural Networks 18(3), 698–707 (2007)
11. Ganter, B., Wille, R.: Formal Concept Analysis. Mathematical Foundations. Springer, Berlin (1999)
12. Geerts, F., Goethals, B., Mielikäinen, T.: Tiling Databases. In: Suzuki, E., Arikawa, S. (eds.) DS 2004. LNCS, vol. 3245, pp. 278–289. Springer, Heidelberg (2004)
13. Golub, G., Van Loan, C.: Matrix Computations. Johns Hopkins University Press (1996)
14. Hájek, P.: Metamathematics of Fuzzy Logic. Kluwer, Dordrecht (1998)
15. Keprt, A., Snášel, V.: Binary factor analysis with help of formal concepts. In: Proc. CLA, pp. 90–101 (2004)
16. Keprt, A., Snášel, V.: Binary Factor Analysis with Genetic Algorithms. In: Proc. IEEE WSTST, pp. 1259–1268. Springer, Heidelberg (2005)
17. Klement, E.P., Mesiar, R., Pap, E.: Triangular Norms. Kluwer, Dordrecht (2000)
18. Krantz, H.H., Luce, R.D., Suppes, P., Tversky, A.: Foundations of Measurement. vol. I (Additive and Polynomial Representations), vol. II (Geometric, Threshold, and Probabilistic Represenations), vol. III (Represenations, Axiomatization, and Invariance). Dover Edition (2007)
19. Lee, D., Seung, H.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
20. Leeuw, J.D.: Principal component analysis of binary data. Application to roll-call analysis (2003), `http://gifi.stat.ucla.edu`
21. Mickey, M.R., Mundle, P., Engelman, L.: Boolean factor analysis. In: Dixon, W.J. (ed.) BMDP statistical software manual, vol. 2, pp. 849–860. University of California Press, Berkeley (1990)

22. Miettinen, P., Mielikäinen, T., Gionis, A., Das, G., Mannila, H.: The discrete basis problem. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS (LNAI), vol. 4213, pp. 335–346. Springer, Heidelberg (2006)
23. Miller, G.A.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychol. Rev. 63, 81–97 (1956)
24. Nau, D.S.: Specificity covering: immunological and other applications, computational complexity and other mathematical properties, and a computer program. A. M. Thesis, Technical Report CS–1976–7, Computer Sci. Dept., Duke Univ., Durham, N. C (1976)
25. Nau, D.S., Markowsky, G., Woodbury, M.A., Amos, D.B.: A Mathematical Analysis of Human Leukocyte Antigen Serology. Math. Biosciences 40, 243–270 (1978)
26. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326 (2000)
27. Sajama, O.A.: Semi-parametric Exponential Family PCA. In: NIPS 2004 (2004)
28. Schein, A., Saul, L., Ungar, L.: A generalized linear model for principal component analysis of binary data. In: Proc. Int. Workshop on Artificial Intelligence and Statistics, pp. 14–21 (2003)
29. Stockmeyer, L.J.: The set basis problem is NP-complete. IBM Research Report RC5431, Yorktown Heights, NY (1975)
30. Tang, F., Tao, H.: Binary principal component analysis. In: Proc. British Machine Vision Conference 2006, pp. 377–386 (2006)
31. Tatti, N., Mielikäinen, T., Gionis, A., Mannila, H.: What is the dimension of your binary data? In: The 2006 IEEE Conference on Data Mining (ICDM 2006), pp. 603–612. IEEE Computer Society, Los Alamitos (2006)
32. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323 (2000)
33. Vaidya, J., Atluri, V., Guo, Q.: The Role Mining Problem: Finding a Minimal Descriptive Set of Roles. In: ACM Symposium on Access Control Models and Technologies, pp. 175–184 (June 2007)
34. Ward, M., Dilworth, R.P.: Residuated lattices. Trans. Amer. Math. Soc. 45, 335–354 (1939)
35. Zadeh, L.A.: Fuzzy sets. Inf. Control 8, 338–353 (1965)
36. Zivkovic, Z., Verbeek, J.: Transformation invariant component analysis for binary images. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), vol. 1, pp. 254–259 (2006)