Factor Analysis of Sports Data via Decomposition of Matrices with Grades

Radim Belohlavek, Marketa Krmelova (Palacky University, Olomouc)

DAMOL
DATA ANALYSIS AND MODELING LAB
Palacky University, Olomouc, Czech Republic

Aim of This Paper

- in previous papers was presented method of factor analysis of Boolean data (Belohlavek, Vychodil, J. Comput. System Sci. 2010)
- it was also extended to data with graded attributes (Belohlavek, J. Logic Computation 2012 and Belohlavek, Vychodil, ICFCA 2009)
- in this paper we use this method
- we present the results of selected analyses
- demonstrate a usefulness of the method

Factor Analysis of Data with Fuzzy Attributes

- given an $n \times m$ object-attribute matrix I
- goal: find a decomposition

$$I = A \circ B \tag{1}$$

- A is $n \times k$ object-factor matrix
- B is $k \times m$ factor-attribute matrix
- k factors: possibly more fundamental attributes (or variables), which explain the original m attributes
- we want $k < m$ and, in fact, k as small as possible

- use the calculus of matrices over residuated lattices
- $\mathbf{L} = \langle L, \wedge, \vee, \otimes, \rightarrow, 0, 1 \rangle$, i.e. $I_{ij}, A_{il}, B_{lj} \in L$
- elements of L represent truth degrees, 0 and 1 are the smallest and largest one
- $\wedge$ and $\vee$ denote the infimum and supremum
- $\otimes$ and $\rightarrow$ denote the truth functions of many-valued logic connectives conjunction and implication
- the product $\circ$ in (1) is defined by

$$(A \circ B)_{ij} = \bigvee_{l=1}^{k} A_{il} \otimes B_{lj}. \tag{2}$$

- I_{ij} is the truth degree of the proposition "object i has attribute j"
- A_{il} is the truth degree of the proposition "factor l applies to object i"
- B_{lj} is the truth degree of "attribute j is one of the manifestations of factor l"

- "exists" and "and" are modeled by $\bigvee$ and $\otimes$:

$$\begin{aligned} &\text{object } i \text{ has attribute } j \text{ if and only if} \\ &\quad \text{there exists factor } l \text{ such that } i \text{ has } l \text{ (or, } l \text{ applies to } i) \\ &\quad \text{and } j \text{ is one of the particular manifestations of } l. \end{aligned} \tag{3}$$

- we compute from I, using a greedy approximation algorithm a set

$$\mathcal{F} = \{\langle C_1, D_1\rangle, \ldots, \langle C_k, D_k\rangle\} \subseteq \mathcal{B}(X, Y, I) \tag{4}$$

  of formal fuzzy concepts of I

- put

$$(A_{\mathcal{F}})_{il} = (C_l)(i) \quad \text{and} \quad (B_{\mathcal{F}})_{lj} = (D_l)(j), \tag{5}$$

- our algorithm (from Belohlavek, Vychodil, ICFCA 2009) computes suboptimal decomposition since the problem to compute optimal decomposition is an NP-hard optimization problem

Algorithm

- input: an $n \times m$ object-attribute matrix I
- output: minimal subset of $\{C \otimes D \mid \langle C, D\rangle \in \mathcal{B}(X, Y, I)\}$ which covers I
- put into $\mathcal{U}$ all $\langle i, j\rangle \mid I_{ij} > 0$
- let $\mathcal{F}$ is empty set
- repeat while $\mathcal{U}$ is not empty:
  - let D is empty
  - select $\langle j, a\rangle$ which maximizes

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

    where $D^+ = D \cup \{^a/_j\}$
  - repeat until exist such $\langle j, a\rangle$
  - $C = D^{\downarrow}$, $\mathcal{F} = \mathcal{F} \cup \{\langle C, D\rangle\}$, $\mathcal{U} = \mathcal{U} \setminus \{\langle i, j\rangle\}$, for all $\langle i, j\rangle$, where $C(i) \otimes D(j) \geq I_{ij}$

Examples

- five-element Łukasiewicz chain $L = \{0; 0,25; 0,5; 0,75; 1\}$
- operation $\otimes$ is given by $a \otimes b = \max(0, a + b - 1)$
- degrees are represented by shades of gray as follows

2004 Olympic Games Decathlon - Top 5

example from Belohlavek, Vychodil, ICFCA 2009

Score of top 5 athletes

	10	*lj*	*sp*	*hj*	40	*hu*	*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

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

The transformation from the table with scores to the matrix with degrees from $L = \{0, 0.25, 0.5, 0.75, 1\}$ is accomplished using functions

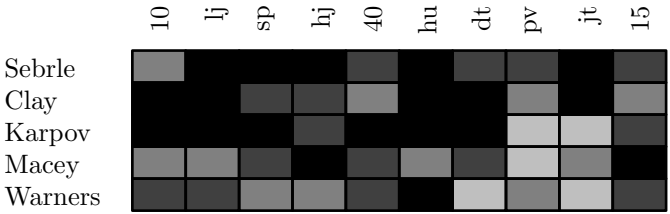
$$s_j : [0, \ldots, 1400] \to L \text{ defined by } s_j(p) = \text{round}\left(\frac{p - L_j}{H_j - L_j}\right)$$

where j is an attribute (discipline) and L_j and H_j are lowest and highest score achieved

Matrix I with Graded Attributes (input to the method)

	10	*lj*	*sp*	*hj*	40	*hu*	*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	1.00	0.50	1.00	0.50
Karpov	1.00	1.00	1.00	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	0.75	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

Graphical Representation of Matrix I



- algorithm found a decomposition of I using 6 factors

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

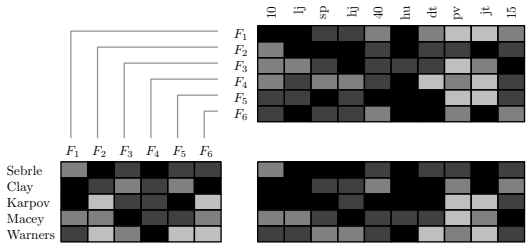


Figure : Decomposition $I = A_{\mathcal{F}} \circ B_{\mathcal{F}}$.

Interpretation of factor

- Factor 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
  - 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.
  - meaning: All the manifestations of this factor with degree 1 are 100 m, long jump, and 110 m hurdles
  - this factor can be interpreted as the ability to run fast for short distances—**speed**
  - factor applies particularly to Clay and Karpov which is well known in the world of decathlon
- Factor F_2:
  - 1 are long jump, shot put, high jump, and javelin
  - F_2 can be interpreted as the ability to apply very high force in a very short term—**explosiveness**
  - applies particularly to Sebrle, and then to Clay, who are known for this ability

- Factor F_3:
  - 1 are high jump and $1500\,\mathrm{m}$
  - this factor is typical for lighter, not very muscular athletes
  - Macey, who is evidently that type among decathletes ($196\,\mathrm{cm}$ and $98\,\mathrm{kg}$) is the athlete to whom the factor applies to degree 1

These are the most important factors behind data matrix I.
These factors make sense was confirmed to us by a decathlon coach, who particularly emphasized that factor F_2 is typical of the Czech decathlon school.

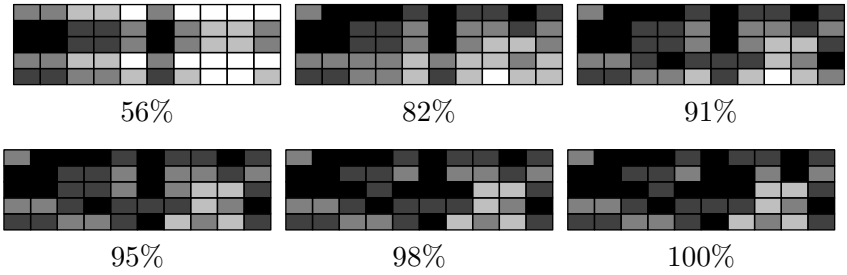


Figure : $\bigvee$-superposition of Factor Concepts

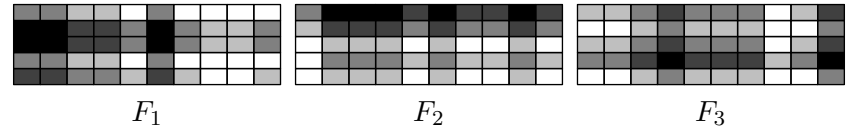


Figure : Factor Concepts as Rectangular Patterns.

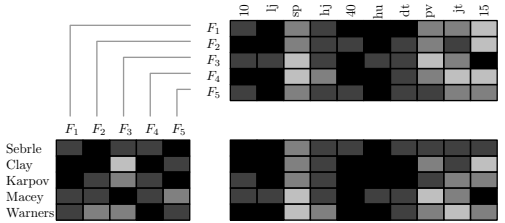
2004 Olympic Games Decathlon - Top 5 By Their Best Results

- take the top 5 athletes of the 2004 Olympic Decathlon but we take their best performances during all of their decathlon competitions
- it is reasonable if we want to avoid a possible bad luck in a particular discipline such as a bad start in 100 m
- for discipline j, we put

$$s_j(p) = \begin{cases} 1 & \text{for } p \in [H_j, H_j - 100), \\ 0.75 & \text{for } p \in [H_j - 100, H_j - 200), \\ 0.5 & \text{for } p \in [H_j - 200, H_j - 300), \\ 0.25 & \text{for } p \in [H_j - 300, H_j - 400), \\ 0 & \text{for } p \leq H_j - 400, \end{cases}$$

  where H_j is the highest score ever achieved during a decathlon competition for discipline j
- note that $H_{10} = 1042$; $H_{lj} = 1117$; $H_{sp} = 1048$; $H_{hj} = 1061$; $H_{40} = 1025$; $H_{hu} = 1064$; $H_{di} = 993$; $H_{pv} = 1152$; $H_{ja} = 1040$; $H_{15} = 963$.

	10	lj	sp	hj	40	hu	di	pv	ja	15
Sebrle	942	1089	880	944	921	1002	859	972	907	798
Clay	1010	1050	868	887	944	1022	993	941	920	670
Karpov	931	1073	910	915	968	984	929	1004	743	729
Macey	940	1002	841	944	998	931	836	849	799	990
Warners	947	1022	800	831	978	973	824	886	692	693



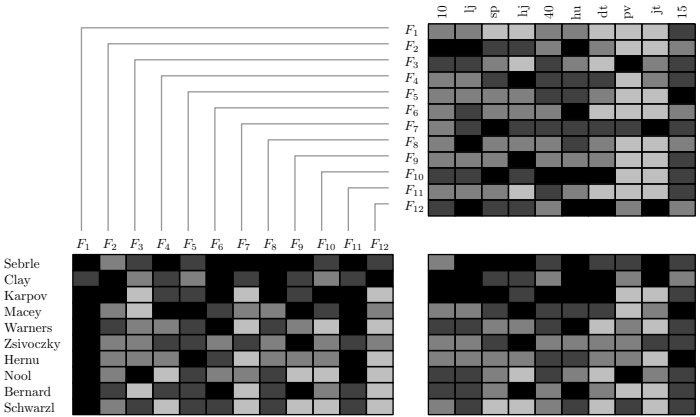
- factors in this case are different from those in the previous example but a reasonable similarity is still apparent
- F_1 applies to degree 1 to Clay and Karpov in both examples and applies to the other athletes to similar degrees in both examples as well
- intents of the first factor are different although a reasonable similarity is apparent as well (presence of long jump and hurdles to degree 1, presence of 100 m and high jump to high degrees)
- a similar observation can be made on F_2 (connects Sebrle and Clay) and F_3 which is typical of Macey.

2004 Olympic Games Decathlon - Top 10

- the matrix I corresponding to the top 10 athletes at Olympic Games decathlon in 2004
- it was used same transformation function like in 2004 Olympic Games Decathlon example

Scores of the 5th–10th Athletes

	10	*lj*	*sp*	*hj*	40	*hu*	*di*	*pv*	*ja*	15
Zsivoczky	881	847	809	915	842	856	780	819	790	748
Hernu	867	859	768	831	874	942	761	849	704	782
Nool	906	942	744	698	870	874	706	1035	758	704
Bernard	931	930	777	915	855	953	762	731	667	704
Schwarzl	865	932	729	749	826	942	714	941	683	721

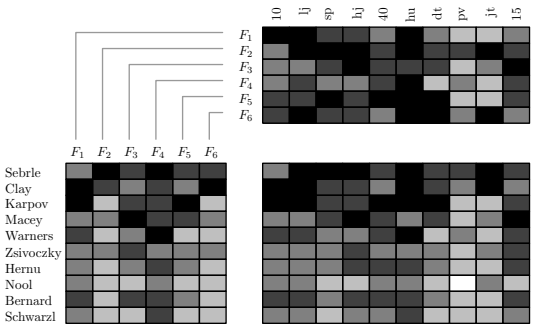


- compared to the factors from first example, the factors in this example are generally different although some similarities are apparent
- factor F_2 here is exactly the same (has same intent) as F_1
- F_{12} is the same as F_6 and F_4 is almost the same as F_3

- we might be interested in the question of how well the factors from top 5 athletes explain the new dataset regarding the top 10 athletes
- from the factors of top 5 athletes ($F_l = \langle C_l, D_l \rangle$), we compute candidate factors of the top 10 athletes ($G_l = \langle P_l, Q_l \rangle$) as follows:

$$P_1 = D_1^{\downarrow}, Q_1 = P_1^{\uparrow}, \ldots, P_6 = D_6^{\downarrow}, Q_6 = P_6^{\uparrow},$$

- in general we do not have $I = A_{\mathcal{G}} \circ B_{\mathcal{G}}$
- nevertheless, the first factor G_1 explains 50% of the data, the first two factors 69%, the first three factors 80%, the first four factors 86%, the first five factors 89%, and all factors in $\mathcal{G}$ explain 91% of the data


10
ij
sp
bj
40
hu
dt
pv
jt
15
F_1
F_2
F_3
F_4
F_5
F_6
F_1 F_2 F_3 F_4 F_5 F_6
Sebrle
Clay
Karpov
Macey
Warners
Zsivoczky
Hernu
Nool
Bernard
Schwarzl

2004 Olympic Games Modern Pentathlon

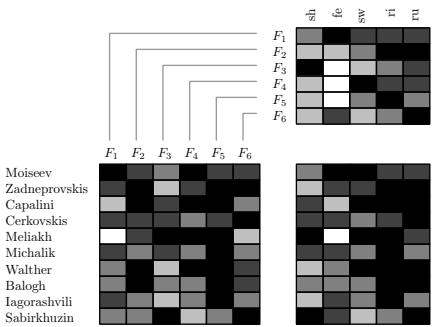
- consists of pistol shooting (sh), fencing (fe), 200 m freestyle swimming (sw), show jumping (ri), and a 3 km cross-country run (ru)

	sh	*fe*	*sw*	*ri*	*ru*
Moiseev	1036	1000	1376	1032	1036
Zadneprovskis	1000	916	1308	1088	1116
Capalini	1084	776	1336	1116	1080
Cerkovskis	1096	916	1252	1004	1088
Meliakh	1168	692	1332	1144	1004
Michalik	1108	888	1260	1144	932
Walther	952	832	1336	1116	1084
Balogh	1036	804	1240	1172	1044
Iagorashvili	988	916	1252	1172	948
Sabirkhuzin	1156	888	1216	908	1034

To transform the scores of discipline j to degrees, we used the function

$$s_j(p) = \begin{cases} 1 & \text{for } p \in [H_j, H_j - \frac{1}{5}(H_j - L_j)), \\ 0.75 & \text{for } p \in [H_j - \frac{1}{5}(H_j - L_j), H_j - \frac{2}{5}(H_j - L_j)), \\ 0.5 & \text{for } p \in [H_j - \frac{2}{5}(H_j - L_j), H_j - \frac{3}{5}(H_j - L_j)), \\ 0.25 & \text{for } p \in [H_j - \frac{3}{5}(H_j - L_j), H_j - \frac{4}{5}(H_j - L_j)), \\ 0 & \text{for } p \leq H_j - \frac{4}{5}(H_j - L_j), \end{cases}$$

where H_j and L_j are the highest and the lowest score achieve in discipline j in the 2004 Olympic Games modern pentathlon. Note that $H_{sh} = 1168$, $L_{sh} = 892$; $H_{fe} = 1000$, $L_{fe} = 664$; $H_{sw} = 1376$, $L_{sw} = 1140$; $H_{ri} = 1172$, $L_{ri} = 584$; $H_{ru} = 1116$, $L_{ru} = 752$.



- F_2's manifestations are riding and cross-country run which is typical for athletes who are in a good physical shape and have good endurance
- each of the other factors more or less corresponds to a single discipline which corresponds to the intuitive idea that the disciplines are diverse and require diverse skills (this is how pentathlon was designed)

Conclusions, Further Issues and Future Work

- we presented several examples of factor analysis of sports data
- we only presented selected examples, further are available in proceedings
- we refrained from formalizing some of the issues involved, such as "explanation of data by factors", "similarity of factors", how well the factors of one dataset serve as good factors of another dataset
- we also skipped theoretical results such as the influence of the choice of the scale of degrees, the operation $\otimes$, the influence of the transformation from scores to degrees
- more examples with detailed descriptions as well as formal treatment of some of issues mentioned above will appear in the full version of this paper
- future research: a comparison, experimental and possibly also theoretical, of relationships of the presented method with related methods that involve matrix decomposition, notably the non-negative matrix factorization