Simple Structure in Boolean Matrix Factorization

Martin Trnecka[0000−0001−7770−2033] and
Marketa Trneckova[0000−0002−1311−7211]

Dept. Computer Science
Palacký University Olomouc, Olomouc, Czech Republic
martin.trnecka@gmail.com, marketa.trneckova@gmail.com

Abstract. We are going back to the root of Boolean matrix factorization (BMF)—to factor analysis—and we examine an implementation of the simple structure in BMF. We propose a Boolean counterpart of the simple structure, i.e. a novel way how to evaluate interpretability of factors. Moreover, we discuss the proposed approach from the formal concept analysis perspective, and we provide an analysis of the interpretability of results obtained via selected the state-of-the-art BMF algorithms on real-world data. Additionally, we propose a novel BMF algorithm utilizing a main criterion for the factor selection based on the simple structure.

Keywords: Boolean factor analysis · Interpretation of results · Simple structure.

1 Introduction

Boolean matrix factorization (BMF) is a natural and immensely popular way of summarizing binary (yes/no) data via new fundamental variables, called factors, that are implicitly hidden in the data. In the past, the BMF was successfully used in various fields, e.g. role-based access control [11], computational biology [19], recommender systems [14], logic circuit synthesis [13], classification [1], computer network analysis [15] and many other fields of study.

The general aim of BMF is to simplify complex data. More precisely, for a given Boolean matrix $\mathbf{I} \in \{0,1\}^{m\times n}$, with m objects and n attributes, to find matrices $\mathbf{A} \in \{0,1\}^{m\times k}$ and $\mathbf{B} \in \{0,1\}^{k\times n}$ for which

$$I \approx \mathbf{A} \circ \mathbf{B}.$$

Operator $\circ$ represents Boolean matrix multiplication, i.e.

$$(\mathbf{A} \circ \mathbf{B})_{ij} = \max_{l=1}^{k} \min(\mathbf{A}_{il}, \mathbf{B}_{lj}),$$

and $\approx$ represents approximate equality assessed by number of 1s. Finding the $\mathbf{A}\circ\mathbf{B}$ may be interpreted as a discovery of k factors that exactly or approximately describe the data, where $\mathbf{A}$ captures a relation between the original objects

Published in Francisco J. Valverde-Albacete, Martin Trnecka (Eds.): Proceedings of the 15th International Conference on Concept Lattices and Their Applications, CLA 2020, pp. 95–106, 2020.

and the factors, and $\mathbf{B}$ captures a relation between the factors and the original attributes.

While, BMF is widely used tool in general data-mining, the main motivation for the BMF comes from the psychology and the social sciences. In a broad sense, BMF is an implementation of the general idea of factor analysis introduced by psychologist Charles Spearman [21]. This is not surprising. In a fact BMF comes from the psychology, where Boolean data often occur (see e.g. [7]). Surprisingly, in BMF only small attention to the crucial aspect of the factor analysis—an interpretation of factors itself—is payed.

In the factor analysis, the interpretation of one particular factor is based on its loading scores, i.e. how much are the original attributes involved in the factor description. This view to factors is called a *Royce's model* [20] of the factor. The interpretation of a factorization, i.e. interpretation of all factors, is driven by the law of parsimony, well known as Occam's razor, i.e. we should pick the simplest explanation of facts. A solution selected via the parsimony law is called *simple structure*.

In BMF, the question of how to interpret one particular factor is usually answered with a help of formal concept analysis (FCA)[12], however the Royce model may also be adopted, especially in the case of general factorization. Before we explain this, let us remind the basic connection between BMF and FCA.

For every $\mathbf{I} \in \{0,1\}^{n \times m}$ one may associate a pair $\langle ^{\uparrow}, ^{\downarrow} \rangle$ of arrow operators defined by

$$C^{\uparrow} = \{j \in Y \mid \forall i \in C : \mathbf{I}_{ij} = 1\} \text{ and } D^{\downarrow} = \{i \in X \mid \forall j \in D : \mathbf{I}_{ij} = 1\},$$

where C is a subset of all objects $X = \{1, \ldots, m\}$ of $\mathbf{I}$ and D is subset of all attributes $Y = \{1, \ldots, n\}$ of $\mathbf{I}$. A pair $\langle C, D \rangle$ for which $C^{\uparrow} = D$ and $D^{\downarrow} = C$ is called *formal concept*. The set of all formal concepts for $\mathbf{I}$ is defined in the following way

$$\mathcal{B}(\mathbf{I}) = \{\langle C, D \rangle \mid C \subseteq X, D \subseteq Y, C^{\uparrow} = D, D^{\downarrow} = C\}.$$

Now, we explain an important connection—originally described in a pioneer work [6]—between a set of formal concepts and the Boolean matrix factorization. Every set $\mathcal{F} = \{\langle C_1, D_1 \rangle, \ldots, \langle C_k, D_k \rangle\} \subseteq \mathcal{B}(\mathbf{I})$, where the indexing of the formal concepts $\langle C_l, D_l \rangle$ is fixed, induces the $m \times k$ and $k \times n$ Boolean matrices $\mathbf{A}_{\mathcal{F}}$ and $\mathbf{B}_{\mathcal{F}}$ by

$$(\mathbf{A}_{\mathcal{F}})_{il} = \begin{cases} 1, \text{if } i \in C_l, \\ 0, \text{if } i \notin C_l, \end{cases} \quad \text{and} \quad (\mathbf{B}_{\mathcal{F}})_{lj} = \begin{cases} 1, \text{if } j \in D_l, \\ 0, \text{if } j \notin D_l, \end{cases}$$

for $l = 1, \ldots, k$. That is, the lth column of $\mathbf{A}_{\mathcal{F}}$ and lth row $\mathbf{B}_{\mathcal{F}}$ are the characteristic vectors of C_l and D_l, respectively. The set $\mathcal{F}$ is also called a set of *factor concepts*. The advantage of this approach is that a factor can be seen as a formal concept, i.e. the interpretation of the factor is straightforward: factor = formal concept—an entity with extent and intent part.

The Boolean counterpart of simple structure or an elementary way how to evaluate whether the factorization is easily interpretable is missing.

In the paper we are going back to the root of BMF—to factor analysis—and we discuss an implementation of the simple structure introduced in [22] by L. L. Thurstone in BMF. The main contribution of this work is following: (i) we propose a Boolean counterpart of the simple structure, which may be utilized in BMF, i.e. a novel way how to evaluate a quality (interpretability) of factors and we discuss it from the formal concept analysis perspective; (ii) we evaluate the interpretability of factorization delivered by the state-of-the-art BMF algorithms; and (iii) we propose a novel BMF algorithm which uses the simple structure as a main criterion for the factor selection.

The rest of this paper is organized as follows. The next section provides basic insight to the simple structure and discusses its formalization that can be used in BMF. In Section 3, an evaluation of selected existing the state-of-the-art BMF algorithms from an interpretability viewpoint is performed. Section 4 describes a new BMF algorithm using our formalization of simple structure as a criterion for the factor selection as well as its experimental evaluation. The paper is concluded by Section 5.

2 Simple structure

In the factor analysis, the question of the interpretability of factors is decided by the law of parsimony, well known as Occam's razor. In other words the simplest explanation is the best one. Such explanation is selected via the parsimony law, and it is called a simple structure.

Even though, the BMF has motivation in the factor analysis, an interpretation of factors has only a small attention in contemporary BMF research. This is surprising because the interpretation of factors is a crucial part of factor analysis.

2.1 Thurstone's simple structure criteria

In 1947 Thurstone formalized in his work [22] five criteria of the simple structure. Later, Cattell in [9] claimed that the simple structure factors are (usually) simple to interpret. Unfortunately, the Thurstone's formalization of good factors is informal, vague and verbally described. According to Thurstone, interpretable factors satisfy the following five conditions:

1. Each row of the rotated matrix should contain at least one zero.
2. In each factor, the minimum number of zero loadings should be the number of factors in the rotation.
3. For every pair of factors there should be variables with zero loadings on one and significant loadings on the other.
4. For every pair of factors a large portion of the loadings should be zero, at least in a matrix with a large number of factors.
5. For every pair of factors there should be only a few variables with significant loadings on both factors.

Let us note, that the loading matrix is the factor-attribute matrix. Basically, the criterion of the simple structure is derived from properties of the loading matrix. This approach adopts the Royce definition of factor, namely "factor is a construct operationally defined by its factor loadings". In other words, factors are represented via attributes that are manifestation of them. In our terminology, factors are represented via rows of matrix $\mathbf{B}$.

The third and the fifth criterion are of overriding importance, namely the loading matrix is the simple structure if each pair of factors have a few high loadings.

2.2 Simple structure for BMF

There have been many attempts—without noticeable success—to formalize the simple structure (see e.g. [8]) in factor analysis. In [23] a formalization of the Thurstone's five criteria via logical formulas for a decomposition of matrices over a finite scale is proposed. Despite the fact, that the BMF is a special case of the decomposition over a finite scale [23], in the following part we discuss that the approach in [23] cannot be straightforwardly adopted in the Boolean case.

In BMF, the first Thurstone's criterion is always met, with an exception where the input data contain an object which have all attributes. The second criterion may not be easy to satisfy, especially if an exact decomposition is required. In such case BMF algorithms may produce greater number of factors than attributes (for more details see e.g. [3]). The third criterion can be understood as follows: for every pair of factors, there should be an attribute which is a manifestation of one of them and is not a manifestation of the other one. As a consequence of this, the third criterion is always met in BMF. The penultimate criterion is in conflict with in FCA well-known fact that the maximal rectangles are the best choice from the interpretability point of view. Namely, let us consider a concept "dog". If we cut some attributes from them, the interpretation of this concept suffers. The last criterion says, that the number of common attributes for each pair of factors should be minimal.

Since the first and the third criterion are meaningless in BMF and the fourth criterion violates a basic view on Boolean factors, we can conclude, that the well-interpretable factorization satisfies the condition that all pairs of factors have the number of common attributes as small as possible. This is consistent with an observation presented [4], where a kind of dependencies between factors in a presence of noise in data was observed. Obviously, if the sets of attributes, that describe factors, are very similar, such factorization is hard to interpret.

We define a measure, adopting the Royce model of factor, which calculate a diversity (interpretability) of two sets of attributes A and B:

$$div(A,B) = 1 - max\left(\frac{|A \cap B|}{|A|}, \frac{|A \cap B|}{|B|}\right).$$

A high value of $div(A,B)$ indicates that sets A and B have a small number of common attributes. $div(A,B) = 1$ means that the sets have no common attributes at all.

For the set of factors $\mathcal{F}$, we may compute the diversity as an average diversity of each pair of factors, i.e.

$$div(\mathcal{F}) = \sum_i \sum_{j>i} div(D_i, D_j)/number\_of\_pairs$$

for D_i, D_j such that $\langle C_i, D_i\rangle, \langle C_j, D_j\rangle \in \mathcal{F}$. Note, the diversity is equal to 1 for the set of factors that have no common attributes.

We employed the diversity measure to evaluate the quality (interpretability) of factorization. Note that a different approach to the quality was proposed in [2]. In what follows we emphasize a fact that the BMF problem is a set covering problem [3], i.e. the number of entries that are covered by a particular factor is an important measure.

3 Comparison of existing methods

In the following section we address the question of the interpretability of factors that are obtained via existing BMF algorithms. We compare—by means of the diversity measure—results delivered by selected state-of-the-art BMF algorithms, namely GreConD [6], Asso [18], GreEss [3], PaNDa [17], Hyper [24], 8M [10], GreConD+ [5]. An overview of these algorithms can be found e.g. in [2–4].

3.1 Datasets

We used several, in BMF standard, real-world datasets from [3, 11, 16] that are listed, together with their basic characteristics and the numbers of factors that are required by selected algorithms to achieve an exact decomposition in Table 1.

Moreover we build three new datasets CLA, Rio medal and Rio participation.[1] CLA data include as objects 128 regular contributors (authors that have more than one contribution) on CLA conference. The attributes of CLA data consist of 13 individual conferences (years). Both Rio datasets include 207 objects representing countries participated on Olympic games in Rio in 2016 and 28 attributes that represent sport disciplines such as aquatic sports, cycling, gymnastics, etc. We used two relations between objects and attributes. The first one reflects that a particular country had a representative in a sport discipline (Rio participation). The second one reflects if a particular country wins some medal in sport discipline (Rio medal).

3.2 Experimental evaluation

We computed factorizations for each dataset in Table 1 via above mentioned BMF methods. Table 2 shows the interpretability (the diversity measure) for sets

[1] All datasets as well as implementation of below presented algorithm GreDConD are available online at https://github.com/Marketa-Trneckova/CLA2020-Simple-Structure-BMF.

Table 1: Datasets

dataset	dimensions	$\|\|\mathbf{I}\|\|$	$\|\|\mathbf{I}\|\|/(m \cdot n)$	GreEss	GreConD	GreDConD
Breast	699 × 20	6974	0.50	19	19	21
CLA	128 × 13	448	0.27	14	16	16
DBLP	6980 × 19	17173	0.13	19	21	22
Domino	231 × 79	730	0.04	20	21	22
Ecoli	336 × 34	2688	0.23	40	41	40
Chess	3196 × 76	118252	0.49	113	124	114
Iris	150 × 19	750	0.26	20	20	19
Mushroom	8124 × 119	186852	0.19	105	120	130
Paleo	501 × 139	3537	0.05	145	151	152
Post	90 × 25	720	0.32	29	32	30
Rio participation	207 × 28	1776	0.31	28	44	37
Rio medal	207 × 28	285	0.05	29	32	30
Zoo	101 × 28	862	0.30	25	30	28
Zoo extended	144 × 21	1019	0.34	26	27	29

of factors that achieve a prescribe coverage of input data (column c). Namely, 75%, 90%, 95% and 100% coverage is shown. The value N/A in the table means, that particular algorithm is not able to achieve prescribed coverage. The coverage is computed as a percentage of nonzero entries in data that are covered by some factors (for more details see e.g. [3]).

The highest values of the diversity are obtained via algorithm Hyper, which for almost all datasets returns value 1 or a value close to 1. The main reason is that Hyper usually produces factors including only one attribute—such factors are really easy to interpret. This can be seen as a drawback of the simple structure, because such factors are usually not useful in practice.

Algorithms GreEss and GreConD return comparable results. Both of them use formal concepts as factors and as a result of this both produce so-called from-below decomposition, i.e. no zero entries in input data are covered by some factor. In almost all cases GreEss slightly outperform GreConD.

GreConD+, PaNDa and Asso return sets of factors that are not able to cover the whole data. For example in many cases factors computed via PaNDa cover less that 70% of input data. Differently from GreEss and GreConD, factors produced by GreConD+, PaNDa and Asso are rectangles in data that may contain zeros. The Royce's model of the factor enables to evaluate such factors. Surprisingly, 8M—the oldest BMF method—produces very good factors, however each factor contains a large number of zero entries.

Table 2: Diversity of factorizations.

dataset	c	GreDConD	GreConD	GreEss	Hyper	8M	GreConD+	Asso	PaNDa
Breast	75	0.331	0.245	0.328	0.840	1	0.2	0.278	0.333
	90	0.329	0.284	0.322	0.913	1	0.272	0.299	N/A
	95	0.331	0.275	0.316	0.937	1	0.284	0.3	N/A
	100	0.3	0.263	0.284	0.95	0.872	0.25	N/A	N/A
CLA	75	1	1	1	1	0.733	1	1	N/A
	90	0.97	0.95	1	1	0.813	1	1	N/A
	95	0.956	0.956	1	1	0.826	1	1	N/A
	100	0.95	0.95	0.978	1	0.845	1	1	N/A
DBLP	75	0.833	1	0.6	1	0.25	0.616	1	N/A
	90	0.691	0.611	0.704	1	0.823	0.650	0.611	N/A
	95	0.617	0.593	0.675	1	0.823	0.651	0.53	N/A
	100	0.683	0.667	0.638	1	0.868	0.662	N/A	N/A
Domino	75	0.833	1	0.6	0.633	0.25	0.616	1	N/A
	90	0.691	0.611	0.704	0.974	0.823	0.650	0.611	N/A
	95	0.617	0.593	0.675	0.990	0.823	0.650	0.53	N/A
	100	0.683	0.667	0.638	0.996	0.868	0.662	N/A	N/A
Ecoli	75	0.678	0.386	0.461	0.447	0.5	0.483	0.362	0.458
	90	0.672	0.470	0.453	0.718	0.55	0.563	0.365	N/A
	95	0.699	0.569	0.600	0.818	0.714	0.668	0.335	N/A
	100	0.639	0.578	0.556	0.903	0.945	0.571	N/A	N/A
Chess	75	0.758	0.550	0.618	0.922	1	0.323	0.195	0.178
	90	0.740	0.612	0.617	0.956	0.867	0.588	0.122	N/A
	95	0.734	0.621	0.642	0.966	0.908	0.592	0.140	N/A
	100	0.680	0.615	0.602	0.972	0.959	0.598	N/A	N/A
Iris	75	0.933	0.933	0.933	1	0.933	0.95	1	N/A
	90	0.956	0.944	0.944	1	0.821	0.929	0.982	N/A
	95	0.962	0.955	0.936	1	0.865	0.918	0.967	N/A
	100	0.933	0.926	0.926	1	0.892	0.838	N/A	N/A
Mushroom	75	0.456	0.407	0.376	0.960	0.668	0.390	0.246	N/A
	90	0.470	0.417	0.390	0.981	0.851	0.410	0.282	N/A
	95	0.466	0.416	0.387	0.987	0.894	0.428	0.299	N/A
	100	0.469	0.435	0.405	0.996	0.951	0.436	N/A	N/A
Paleo	75	0.993	0.991	0.998	1	0.957	0.997	0.999	N/A
	90	0.994	0.994	0.998	1	0.978	0.997	0.998	N/A
	95	0.995	0.995	0.998	1	0.982	0.997	0.998	N/A
	100	0.994	0.994	0.997	1	0.983	N/A	N/A	N/A
Post	75	1	1	1	1	0.81	0.881	1	N/A
	90	0.972	0.887	0.968	1	0.855	0.901	0.995	N/A
	95	0.968	0.901	0.966	1	0.863	0.903	0.987	N/A
	100	0.949	0.897	0.946	1	0.888	0.895	0.968	N/A
Rio medal	75	1	1	1	1	0.81	0.881	1	N/A
	90	0.974	0.887	0.968	1	0.855	0.901	0.995	N/A
	95	0.968	0.901	0.966	1	0.863	0.903	0.987	N/A
	100	0.950	0.897	0.946	1	0.888	0.895	0.968	N/A
Rio participation	75	0.781	0.411	0.933	0.979	1	0.817	0.089	N/A
	90	0.786	0.555	0.852	0.988	0.792	0.741	0.132	N/A
	95	0.746	0.547	0.838	0.982	0.836	0.736	N/A	N/A
	100	0.689	0.625	0.774	0.980	0.901	N/A	N/A	N/A
Zoo	75	0.805	0.571	0.576	0.947	0.889	0.653	0.545	0.621
	90	0.795	0.714	0.675	0.967	0.867	0.696	0.538	N/A
	95	0.778	0.732	0.760	0.978	0.923	0.677	0.620	N/A
	100	0.765	0.713	0.734	0.983	0.945	0.667	N/A	N/A
Zoo extended	75	0.789	0.604	0.78	0.82	0.457	0.754	0.586	N/A
	90	0.805	0.780	0.818	0.904	0.631	0.779	0.545	N/A
	95	0.801	0.760	0.874	0.916	0.642	0.779	0.594	N/A
	100	0.792	0.766	0.809	0.918	0.838	N/A	N/A	N/A

4 Improving existing BMF algorithms

Almost all BMF algorithms—directly or indirectly—utilize the coverage as main criterion for factor selection (for more details see e.g. [3]). Despite the fact that the coverage is an important criterion, it does not reflect the interpretability of factors. Reasonable question is if we can improve existing BMF algorithms to give a more interpretable factorization that still covers a large part of data.

We choose GreConD—one of the most successful BMF algorithms—as a base for a new algorithm. GreConD works as follows. To produce matrices $\mathbf{A}_{\mathcal{F}}$ and $\mathbf{B}_{\mathcal{F}}$ GreConD uses a particular greedy search for factor concepts which allows to compute factor formal concepts "on demand". The algorithm constructs the factor concepts by adding sequentially "promising columns" to candidate $\langle C, D\rangle$ to factor concept. A new column j that maximize the number of newly covered entries of the input data is added to $\langle C, D\rangle$. This is repeated until no such columns exist. If there is no such column, the $\langle C, D\rangle$ is added to the set $\mathcal{F}$.

Algorithm 1: GreDConD algorithm

```
Input: Boolean matrix $\mathbf{I}$.
Output: Set $\mathcal{F}$ of factor concepts.
$\mathcal{F} \leftarrow \emptyset$
while $(\mathbf{A}_{\mathcal{F}} \circ \mathbf{B}_{\mathcal{F}}) \neq \mathbf{I}$ do
    $\langle C, D\rangle \leftarrow \langle\emptyset, \emptyset\rangle$
    $total\_cost \leftarrow 0$
    while $\langle C, D\rangle$ is changing do
        $total\_cost' \leftarrow total\_cost$
        foreach $j \notin D$ do
            $D' \leftarrow (D \cup \{j\})^{\downarrow\uparrow}$
            $C' \leftarrow D'^{\downarrow}$
            $\mathcal{F}' \leftarrow \mathcal{F} \cup \langle C', D'\rangle$
            $cost \leftarrow ||\mathbf{A}_{\mathcal{F}'} \circ \mathbf{B}_{\mathcal{F}'}|| \cdot$ Diversity$(\mathcal{F}, D')$
            if $cost > total\_cost'$ then
                $total\_cost' \leftarrow cost$
                $C'' \leftarrow C'$
                $D'' \leftarrow D'$
            end
        end
        $total\_cost \leftarrow total\_cost'$
        $C \leftarrow C''$
        $D \leftarrow D''$
    end
    $\mathcal{F} \leftarrow \mathcal{F} \cup \langle C, D\rangle$
end
return $\mathcal{F}$
```

We use an architecture of GreConD and we modify the computation of the cost value. A pseudocode of the modification, which we called GreDConD[2] is depicted in Algorithm 1.

A value of the *cost* (line 11) is computed as the number of newly covered entries of the input matrix **I** multiplied by a value of diversity obtained via procedure Diversity depicted in Algorithm 2.

Algorithm 2: Diversity procedure

```
Input: Set $\mathcal{F}$ of factor concepts, candidate $D$ to the set $\mathcal{F}$
Output: The average diversity $d$.
foreach $\langle A_i, B_i \rangle \in \mathcal{F}$ do
    $s_i \leftarrow 1 - \max(\frac{|B_i \cap D|}{|B_i|}, \frac{|B_i \cap D|}{|D|})$
end
$d = \frac{\sum_{i=1}^{|\mathcal{F}|} s_i}{|\mathcal{F}|}$
return $d$
```

Such setting of the factor selection criteria tends to prefer large factors—factors that cover a large part of the data—that are more diverse and thus potentially easily interpretable. Note, that the others BMF algorithms may be improved in a similar way. In what follows we provide an experimental evaluation of the new algorithm.

4.1 Experimental evaluation

We perform with GreDConD experiments described in Section 4.1. From Table 2 one may clearly observe that GreDConD algorithm outperform GreConD as well as GreEss (with an exception of CLA, Paleo, Rio participation and Zoo extended). Both are its main competitors.

Additionally we observed, in the case of GreConD, GreEss, GreDConD, that the interpretability of factors tends to decrease when an exact coverage is achieved. This points to a well-known problem, regarding the deciding what is a good coverage for data. The decrease indicates, that factors, that are not simply interpretable were added to factorization. As a very promising further research seems to be investigating, if the diversity measure can be used as a stopping criterion for a factor enumeration.

Moreover, we deeply analyzed the first three factors. Figures 1, 2 and 3 represent attributes of datasets in order Zoo extended, Rio participation and Breast dataset and show how many of the first three factors involves this attribute. The darker square represents the higher number of factors having this attribute in its loading. White color means, that no factor explain this attribute, black color means, that all three factors have this attribute.

[2] GreDConD is the abbreviation for Greedy Diversity Concepts on Demand.

One may see from Figures 1, 2 and 3, that the first three factors obtained via GreDConD share less attributes in comparison to other methods. Almost the same results may be observed in the case of GreEss but these factors are narrower and explained by smaller number of attributes (see Figure 2).

Since the number of factors is an important characteristic, we should mention that the number of factors (see Table 1) required for an exact decomposition via GreDConD is comparable—sometimes even smaller—to the number of factors delivered by GreConD algorithm.

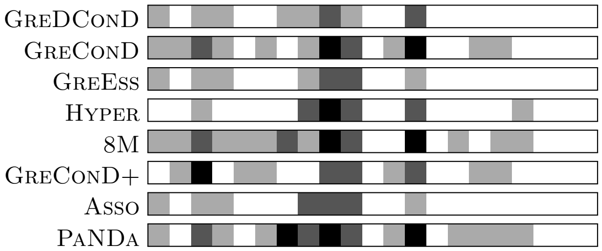

Fig. 1: Zoo extended: 3 factors

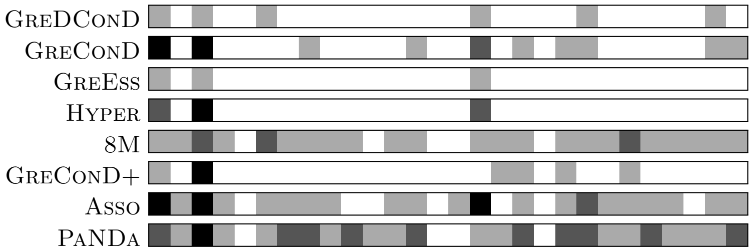

Fig. 2: Rio participation: 3 factors

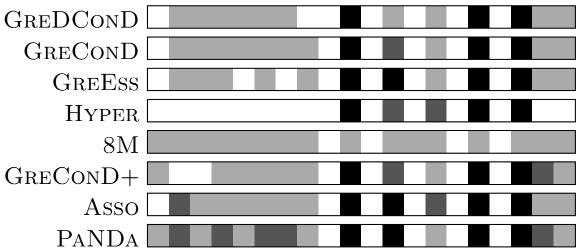

Fig. 3: Breast: 3 factors

5 Conclusion

We propose a Boolean counterpart of the simple structure—which provides a measure of interpretability of results in factor analysis—that can be utilized in BMF. We presented a new measure that reflects the Thurston's criterion for the simple structure, and we evaluate factorizations delivered by the state-of-the-art BMF algorithms via the presented measure. Additionally we propose an improvement one of existing BMF algorithm which uses the simple structure as a main criterion for the factor selection. The new presented algorithm returns slightly better results in a comparison to similar methods.

Acknowledgments

Supported by Junior research Grant No. JG_2020_003 of the Palacký University Olomouc. Support by Grant No. IGA_PrF_2020_019 of IGA of Palacký University is also acknowledged.

References

1. Belohlavek, R., Grissa, D., Guillaume, S., Nguifo, E.M., Outrata, J.: Boolean factors as a means of clustering of interestingness measures of association rules. Annals of Mathematics and Artificial Intelligence **70**(1-2), 151–184 (2014)
2. Belohlavek, R., Outrata, J., Trnecka, M.: Toward quality assessment of Boolean matrix factorizations. Inf. Sci. **459**, 71–85 (2018). https://doi.org/10.1016/j.ins.2018.05.016
3. Belohlavek, R., Trnecka, M.: From-below approximations in Boolean matrix factorization: Geometry and new algorithm. J. Comput. Syst. Sci. **81**(8), 1678–1697 (2015). https://doi.org/10.1016/j.jcss.2015.06.002
4. Belohlavek, R., Trnecka, M.: Handling noise in boolean matrix factorization. Int. J. Approx. Reasoning **96**, 78–94 (2018). https://doi.org/10.1016/j.ijar.2018.03.006
5. Belohlavek, R., Trnecka, M.: A new algorithm for boolean matrix factorization which admits overcovering. Discrete Applied Mathematics **249**, 36–52 (2018). https://doi.org/10.1016/j.dam.2017.12.044
6. Belohlavek, R., Vychodil, V.: Discovery of optimal factors in binary data via a novel method of matrix decomposition. J. Comput. Syst. Sci. **76**(1), 3–20 (2010). https://doi.org/10.1016/j.jcss.2009.05.002
7. Boeck, P.D., Rosenberg, S.: Hierarchical classes: Model and data analysis. Psychometrika **53**, 361–381 (1988)
8. Carroll, J.B.: An analytical solution for approximating simple structure in factor analysis. Psychometrika **18**(1), 23–38 (1953)
9. Cattell, R.B.: The scientific use of factor analysis in behavioral and life sciences. Springer US (1978)
10. Dixon, W.: Bmdp statistical software manual to accompany the 7.0 software release, vols 1–3. (1992)
11. Ene, A., Horne, W.G., Milosavljevic, N., Rao, P., Schreiber, R., Tarjan, R.E.: Fast exact and heuristic methods for role minimization problems. In: Ray, I., Li, N. (eds.) 13th ACM Symposium on Access Control Models and Technologies, SACMAT 2008, Estes Park, CO, USA, June 11-13, 2008, Proceedings. pp. 1–10. ACM (2008). https://doi.org/10.1145/1377836.1377838

12. Ganter, B., Wille, R.: Formal Concept Analysis Mathematical Foundations. Springer-Verlag, Berlin, Heidelberg (1999)
13. Hashemi, S., Tann, H., Reda, S.: Approximate logic synthesis using Boolean matrix factorization. In: Approximate Circuits, pp. 141–154. Springer (2019)
14. Ignatov, D.I., Nenova, E., Konstantinova, N., Konstantinov, A.V.: Boolean matrix factorisation for collaborative filtering: An fca-based approach. In: Agre, G., Hitzler, P., Krisnadhi, A.A., Kuznetsov, S.O. (eds.) Artificial Intelligence: Methodology, Systems, and Applications - 16th International Conference, AIMSA 2014, Varna, Bulgaria, September 11-13, 2014. Proceedings. Lecture Notes in Computer Science, vol. 8722, pp. 47–58. Springer (2014). https://doi.org/10.1007/978-3-319-10554-3_5, https://doi.org/10.1007/978-3-319-10554-3_5
15. Kocayusufoglu, F., Hoang, M.X., Singh, A.K.: Summarizing network processes with network-constrained Boolean matrix factorization. In: 2018 IEEE International Conference on Data Mining (ICDM). pp. 237–246. IEEE (2018)
16. Lichman, M.: UCI machine learning repository (2013), http://archive.ics.uci.edu/ml
17. Lucchese, C., Orlando, S., Perego, R.: Mining top-k patterns from binary datasets in presence of noise. In: Proceedings of the SIAM International Conference on Data Mining, SDM 2010, April 29 - May 1, 2010, Columbus, Ohio, USA. pp. 165–176. SIAM (2010). https://doi.org/10.1137/1.9781611972801.15
18. Miettinen, P., Mielikäinen, T., Gionis, A., Das, G., Mannila, H.: The discrete basis problem. IEEE Trans. Knowl. Data Eng. **20**(10), 1348–1362 (2008). https://doi.org/10.1109/TKDE.2008.53
19. Nau, D.S., Markowsky, G., Woodbury, M.A., Amos, D.B.: A mathematical analysis of human leukocyte antigen serology. Mathematical Biosciences **40**(3-4), 243–270 (1978)
20. Royce, J.R.: Factors as theoretical constructs. American Psychologist **18**(8), 522 (1963)
21. Spearman, C.: "General intelligence," objectively determined and measured. The American Journal of Psychology **15**(2), 201–292 (1904)
22. Thurstone, L.L.: Multiple-factor analysis; a development and expansion of the vectors of mind. (1947)
23. Trnecka, M., Trneckova, M.: Can we measure the interpretability of factors? In: Proceedings of the 14th International Conference on Concept Lattices and Their Applications. pp. 179–190 (2018)
24. Xiang, Y., Jin, R., Fuhry, D., Dragan, F.F.: Summarizing transactional databases with overlapped hyperrectangles. Data Min. Knowl. Discov. **23**(2), 215–251 (2011). https://doi.org/10.1007/s10618-010-0203-9