Studentská vědecká soutěž O cenu děkana 2026

Visual Exploration of Boolean Factors

Mgr. Veronika Pavlíková

Školitel: doc. RNDr. Martin Trnečka, Ph.D.

1. ročník doktorského studia

Katedra informatiky

Visual Exploration of Boolean Factors

Veronika Pavlikova[0009−0001−5686−6044] and Martin Trnecka[0000−0001−7770−2033]

Dept. Computer Science, Palacký University Olomouc, Olomouc, Czech Republic
veronika.pavlikova@upol.cz, martin.trnecka@gmail.com

Abstract. Boolean matrix factorization (BMF) is a widely used technique for discovering latent patterns in binary data, offering high interpretability due to its Boolean nature. However, current analyses of BMF results are often limited to quantitative measures such as reconstruction error, leaving interpretability and relationships between factors largely unexplored. In this work, we investigate the qualitative aspects of the factors and propose a visualization-based approach for exploratory analysis that supports the interpretation of individual factors. Moreover, we introduce a novel method for visualizing factors that preserves their mutual relationships, facilitates the interpretation of individual factors, and provides a global view of the obtained factorization and, consequently, of the analyzed data. The method employs proximity-based representations derived from similarities over attributes, objects, or their combinations, together with multidimensional scaling to produce a low-dimensional overview of the factorization. We further introduce a software tool, BFLens, that enables interactive exploration of BMF results.

Keywords: Boolean matrix factorization · Visualization of factors · Factor evaluation · Factor exploration.

1 Introduction

Boolean matrix factorization (BMF) is a well-established and widely used data mining technique that reveals fundamental patterns implicitly present in data, thereby helping to explain the underlying structure of data.

The main aim of BMF is to describe the input Boolean matrix $\mathbf{I}$ with m rows $(x_1, \ldots, x_m)$, n columns $(y_1, \ldots, y_n)$, and potentially (if available) an additional column c representing the class label, via two Boolean matrices $\mathbf{A}$ and $\mathbf{B}$, as outlined in Fig. 1.

The matrix $\mathbf{A}$ represents an object–factor matrix and captures the relationship between the original objects of matrix $\mathbf{I}$ and the new variables called factors $(f_1, \ldots, f_k)$. Analogously, matrix $\mathbf{B}$ represents a factor–attribute matrix and captures the relationship between the factors and the original attributes of $\mathbf{I}$.

The Boolean nature of the method is particularly advantageous when it comes to data interpretation. Boolean factors are upon a permutation of rows and columns, rectangular areas—rectangles for short—in the input matrix containing ones and, if desired, a reasonably small number of zeros, which represent fundamental patterns in the data. Thus, a factor is defined by the attributes

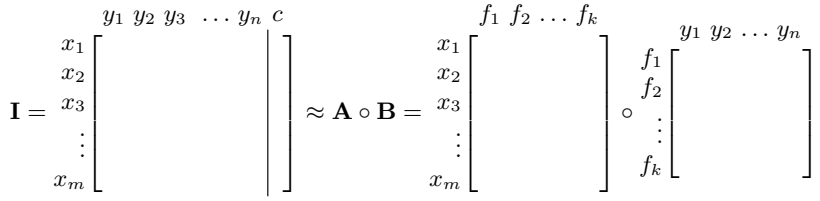


Fig. 1: Basic Boolean matrix factorization model considered in this paper. The operation $\circ$ denotes Boolean matrix multiplication. The symbol $\approx$ indicates that the factorization need not be exact, i.e., $\mathbf{I}$ and $\mathbf{A} \circ \mathbf{B}$ may differ in some entries.

that constitute its manifestation (this information is contained in the rows of matrix $\mathbf{B}$) and by the objects to which it applies (this information is contained in the columns of matrix $\mathbf{A}$). Boolean nature simplifies the interpretation to yes-no, i.e., a particular attribute is a manifestation of a given factor or not, similarly for the objects.[1]

Many BMF algorithms, as well as practical applications of BMF, build upon Formal Concept Analysis (FCA) [11]. In its simplest form, formal concepts can serve as factors, and FCA more generally provides a framework for BMF—most notably, factors can in certain situations be seen as formal concepts [1], which is advantageous from the viewpoint of interpretability.

Unfortunately, the actual interpretation of obtained factors—determining their meaning—is often reduced to a tedious manual inspection of the attributes and objects involved in a particular factor [4]. This process is particularly disadvantageous when BMF is used for exploratory data analysis. Moreover, relationships between factors—such as similarities—that may carry valuable information and facilitate the interpretation of individual factors are often overlooked.

Graphics provide an excellent approach for exploring data and are essential for presenting results. Although graphics—especially for result presentation—have been used in BMF for a long time, there is still no substantive theory on the topic. Moreover, visualization is typically limited to quantitative aspects. The most commonly presented characteristic is the so-called coverage quality curve [3]. There exist several variants of this curve. All of them share the property that they capture reconstruction error, i.e., the number of entries that are either not explained or incorrectly explained. A typical coverage quality curve is shown in Fig. 2.

From the coverage quality curve, one can observe how much of the data is explained by the factors, as well as the cumulative contribution of individual factors. Although coverage quality provides useful information—since BMF can essentially be viewed as a set cover problem [12]—important aspects remain hidden. For example, the curve does not reveal the number of attributes constituting individual factors; factors defined by a single attribute are clearly of

[1] Note that the model can also be extended to ordinal data, i.e., data containing degrees (or grades) on a given scale, without affecting interpretability; see, e.g., [5].

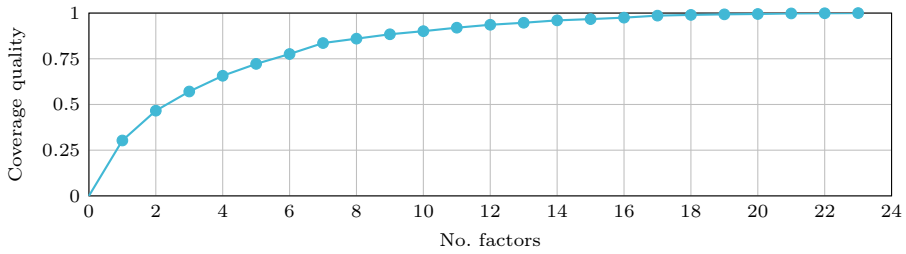


Fig. 2: An example of a coverage quality curve. The value indicates the proportion of the data explained by the factors; for example, 0.75 means that 75% of the input entries are explained.

limited importance. Similarly, the diversity of the associated objects affects interpretability, which varies depending on whether the objects are heterogeneous or homogeneous. Finally, similarities between factors are not captured: some factors may be similar, yet this potentially valuable information is not captured.

There are differences between graphics for presentation and graphics for exploration. Presentation graphics are generally used to summarize the information being presented and should be of high quality, including complete definitions and explanations of both the variables shown and the form of the graphic. In contrast, exploratory graphics are used to search for results. Many may be generated, and they should be quick to read and informative rather than precise. They are not intended for presentation; detailed legends and captions are therefore unnecessary. While a single presentation graphic may be viewed by thousands of readers, thousands of exploratory graphics may be produced to support the data investigation of a single analyst [8]. Another important difference is that presentation graphics are typically static, whereas exploratory graphics benefit greatly from interactivity.

Motivated by our experience with the practical use of BMF, the main aim of this work is to demonstrate how BMF results can be visualized to emphasize their informative value, primarily in exploratory tasks, but also—although this is not our primary focus—in presentation scenarios. We argue that suitable visualization can facilitate the interpretation of factors. Furthermore, we introduce a novel method for factor visualization that captures relationships between factors and provides a global view of the data. Finally, we present a software tool that enables interactive exploration of BMF results.

The rest of the paper is organized as follows. Section 2 discusses the analysis of individual factors and presents visualization techniques that facilitate their interpretation and reduce the number of factors considered for further analysis. Section 3 introduces a novel visualization approach. Section 4 provides illustrative examples of analyses utilizing the proposed approach. Section 5 presents the BFLens tool for exploratory analysis of factorization results. Finally, Section 6 concludes the paper and outlines directions for future research.

2 How to recognize a good factor

We consider the factorization model shown in Fig. 1. Royce [14] defines a factor in terms of the attributes constituting its manifestation. This idea, rooted in classical factor analysis, is justified because attributes play a key role in data explanation. The analyzed data capture only a part of the real world, represented by a limited set of observed objects. A factor that provides an explanation of the data is therefore applicable not only to the analyzed data but also to other data that are not present in analyzed data (generalization). From this perspective, BMF can be viewed as a projection from the attribute space to the factor space [1], where factor provides a different—more fundamental—description of observed world [4].

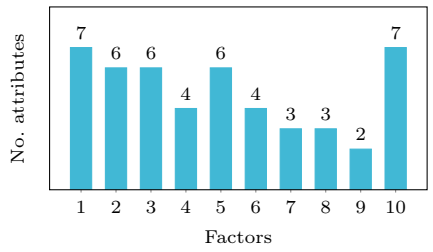
In some situations, however, the enumeration of attributes alone may not be sufficient for an intuitive explanation of a factor. In such cases, one may also consider the objects to which the factor applies. Although this set of objects is incomplete due to the data limitations, it can still be helpful for understanding the nature of the factor. This perspective is particularly justified when factorization is utilized as a form of clustering or in supervised machine learning tasks, where a class label (attribute c in Fig. 1) is present. Note that in such cases, the class label provides additional information that can be exploited in the interpretation of factors.

In general, factors can be evaluated in terms of (i) their attributes, (ii) their objects, possibly augmented by class labels, and (iii) both attributes and objects. To illustrate these perspectives, we present several ways of capturing and visualizing factorization, primarily for the exploratory task. In the following, we consider results obtained by analyzing the Zoo dataset [9] using the GreConD algorithm [1]. The Zoo dataset contains 101 animals described by 16 attributes and a single class attribute. The class attribute assigns each animal to one of seven categories: mammal, bird, reptile, invertebrate, bug, fish, and amphibian.

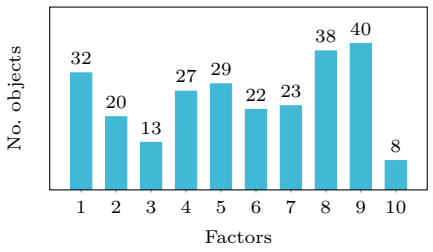
2.1 Attributes manifested by factors

A key characteristic of each factor is the number of attributes that constitute its manifestation. Clearly, factors with either too few or too many attributes are more difficult to interpret. Therefore, in exploratory analysis, it is highly useful to be able to assess this characteristic quickly. Following the basic principles of data visualization [16], a simple bar chart appears to be the most suitable choice, as illustrated in Fig. 3a. Based on this, one may filter out irrelevant concepts and thereby reduce the number of factors to be further analyzed.

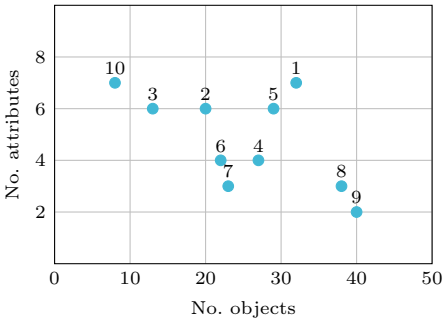
Closely related is another key indicator reflecting how large a portion of the occurrences of a given attribute is represented by a factor. For example, the significance of a factor differs substantially depending on whether an attribute occurs in 75% of objects and the factor covers 90% of these occurrences, or only 10%. This can be quickly observed from the chart depicted in Fig. 3e.



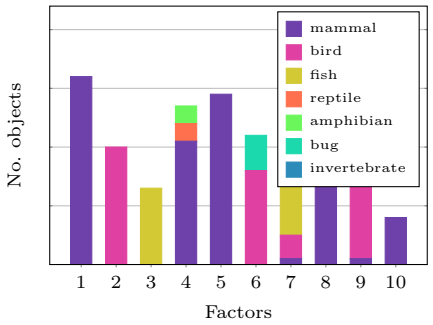
(a) Number of attributes for each factor.



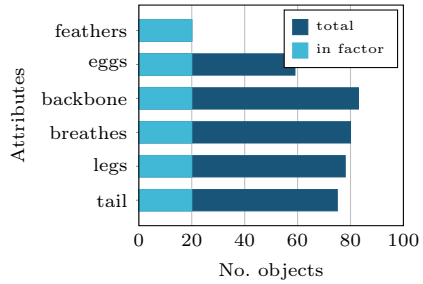
(b) Number of objects for each factor.



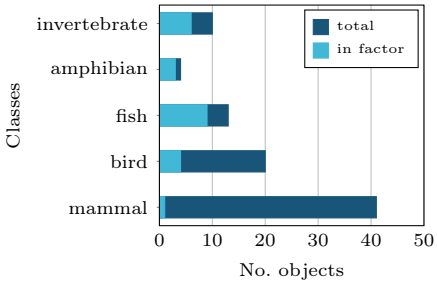
(c) Size of factors in scatter plot.



(d) Distribution of classes in factors.



(e) Distribution of attributes in one factor.



(f) Distribution of classes in one factor.

Fig. 3: Different characteristics of the BMF of the Zoo dataset. For readability, only the top 10 factors are shown.

2.2 Objects to which factors apply

Analogously to attributes, one may wish to quickly assess the number of objects to which each factor applies. Clearly, factors applying to either too few or too many objects are difficult to interpret. Again, a simple bar chart (Fig. 3b) is a suitable choice.

The number of attributes and the number of objects involved in a factor can be examined jointly. A comprehensive overview can be obtained using a

scatter plot that displays the number of objects on one axis and the number of attributes on the other; this visualization is shown in Fig. 3c. From this plot, one can easily identify factors that cover a large portion of the data (points located in the upper-right part of the graph), as well as factors that exhibit sufficiently large numbers of objects or attributes from the perspective of analysis. This plot, together with the charts in Fig. 3a and Fig. 3b, can also be presented with axes showing relative (percentage) proportions instead of absolute counts.

The interpretation of a factor can be significantly facilitated by analyzing the classes, if available, to which the corresponding objects belong. Clearly, factors dominated by a single class are much easier to interpret than very general factors that cover multiple classes. For a quick overview, a bar chart showing the class distribution within individual factors is again useful, as illustrated in Fig. 3d. Also useful is information indicating how large a portion of the occurrences of a given class is represented by a factor (see Fig. 3f).

Overall, Fig. 3 provides a concise overview of the analyzed factorization. For example, the second extracted factor contains a relatively large number of attributes and objects and thus belongs to the larger factors (see Figs. 3a, 3b, and 3c). It is exclusively associated with birds and includes all objects possessing the attribute *feathers*, as well as approximately one third of those with the attribute *eggs* (see Fig. 3e). Similarly, factor 7 is comparatively smaller and predominantly describes amphibians, fish, and invertebrates (see Fig. 3d). An analysis of its attributes (not shown in Fig. 3 due to space limitations) suggests that this factor corresponds to a predator that lives in or hunts near water.

3 Big picture

So far, we have considered only the properties of individual factors. In BMF, however, multiple factors form a whole. When interpreting the factorization as a whole, not only the aforementioned characteristics of individual factors play a role, but also properties of the factorization itself.

In classical factor analysis, the interpretability of factorization is guided by the law of parsimony, commonly known as Occam's razor—namely, that the simplest explanation is preferred. Such an explanation is selected according to the parsimony principle and is referred to as a simple structure. In 1947, Thurstone formalized five criteria for simple structure in his work [15]. Later, Cattell [7] argued that factors exhibiting simple structure are usually easy to interpret. Unfortunately, Thurstone's formalization of good factors is informal, vague, and described only verbally.

According to Thurstone, interpretable factors satisfy the following five conditions, which we translate into BMF terminology:

1. Each row of matrix $\mathbf{B}$ should contain at least one zero.
2. Each row of matrix $\mathbf{B}$ should contain at least as many zeros as there are factors.
3. For every pair of rows of matrix $\mathbf{B}$, there should be attributes with zeros in one factor and ones in the other.

4. For every pair of rows of matrix **B**, a large proportion of attributes should be zero in both rows, particularly when the number of factors is large.
5. For every pair of rows of matrix **B**, there should be only a few attributes with ones in both rows.

Thurstone's definition of good factorization adopts the Royce model of a factor. Among the five criteria, the third and the fifth are of overriding importance, i.e., in a good factorization, factors share only a small number of attributes.[2]

To provide a big picture—a global view of the data and the factorization—we propose the following approach based on Multi Dimensional Scaling (MDS) [13]. MDS is a dimensionality reduction technique that represents objects in a low-dimensional space while preserving their pairwise dissimilarities as faithfully as possible. By projecting factors into a two- or three-dimensional space, we obtain a visual representation reflecting their mutual similarities and differences, thereby providing an intuitive overview of the factorization structure. In our case, the specific meaning of the dimensions is not essential, as the projection primarily serves as a map of relationships between factors. Although the dimensions do carry meaning, their interpretation is non-trivial and difficult to determine explicitly.

Note that the idea of visualizing factors was already outlined in [10], where, similarly to classical factor analysis, results were visualized for two factors without additional quantitative measures. In contrast, the method proposed in this paper captures relationships among all factors and is based on similarity measures.

MDS takes as input a matrix of pairwise dissimilarities or similarities—commonly referred to as proximities—between factors and computes their coordinates in a low-dimensional space such that the distances between points approximate the original proximities as closely as possible. Consequently, similar factors are placed close to each other, whereas dissimilar factors appear further apart. The quality of the low-dimensional embedding is typically assessed using a stress function

$$\text{stress} = \sqrt{\frac{\sum_{i<j} \left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i<j} d_{ij}^2}}$$

where d_{ij} denotes the original distances (or dissimilarities) between factors i and j, and $\hat{d}_{ij}$ denotes the corresponding distances between the same factors in the low-dimensional embedding. The stress function quantifies how well the embedding preserves the original relationships in the data, with lower values indicating better preservation.

Although general guidelines for acceptable stress values exist, in our case MDS is employed primarily for visualization. The stress value therefore indicates whether a two- or three-dimensional representation is sufficient or whether the

[2] Note that the first and second criteria take into account the number of attributes that constitute the manifestation of a factor. This aspect was addressed in Section 2.1.

data cannot be adequately represented in such low-dimensional spaces. The latter case is itself informative, as high stress values may reflect the intrinsic complexity of the factorization rather than merely poor visualization quality.

3.1 Similarity matters

The proximity matrix represents relationships between factors, thereby providing a richer characterization of them. Accordingly, the MDS outcome is determined by the similarity or dissimilarity measure used to construct the matrix.

In general, a suitable similarity measure should increase with the number of shared attributes and decrease with increasing differences, while appropriately emphasizing shared positive features. For binary data, it is often desirable to downweight shared absences. Moreover, psychological studies suggest that human similarity judgments are feature-based and need not satisfy metric axioms [17]. Although MDS can, in principle, handle various types of similarities, in standard settings it typically assumes metric properties.

In [2], 69 practically applicable similarity measures for Boolean data were identified. In the following, we consider three variants for assessing similarity between factors, namely similarity based on their attributes, similarity based on their objects, and similarity based on both their attributes and objects. Each variant allows the use of various similarity measures.

For the first two variants, one may employ the standard scheme: for two binary vectors x and y representing sets of attributes or objects, let a denote the number of elements common to both x and y, b the number of elements present in x but not in y, c the number of elements present in y but not in x, and d the number of elements absent from both x and y.

In the case of the last variant, one may compare the rectangles corresponding to the compared factors; that is, let a denote the number of overlapping entries, b the number of entries present in the first rectangle but not in the second, c the number of entries present in the second rectangle but not in the first, and d the number of entries absent from both rectangles. Furthermore, relationships between factors may also be considered. For example, two factors can be regarded as similar even without overlapping entries, provided they share certain objects or attributes.

The selection of a similarity measure is a key step in data analysis. However, since our primary goal is to illustrate the usefulness of the proposed approach, we employ a basic measure satisfying the previously stated properties, namely the Jaccard distance

$$D_J(x, y) = 1 - \frac{a}{a + b + c}$$

for first two variants and an the overlapping distance

$$D_O(x, y) = 1 - \frac{a}{\min(a + b,\ a + c)}$$

for the last variant.

4 Illustrative examples of analysis

We select three datasets to demonstrate how the proposed approach can facilitate analysis of factors: the Zoo dataset, already described in Section 2; the Congressional voting records dataset [18]; and the Pokémon dataset.[3] Prior to applying our method, all datasets were binarized. In the Zoo dataset, the categorical attribute *legs* was converted into a binary distinction, namely *has legs* versus *does not have legs*. In the Congressional voting records dataset, missing values were handled by duplicating each attribute into separate *yes* and *no* variants. For the Pokémon dataset, we adopted the binarization proposed in [6].

We applied MDS to the factors obtained by the GreConD algorithm, using the Jaccard-based dissimilarity D_J, computed over both intents and extents, and the overlap-based dissimilarity D_o.

4.1 Zoo data

For this dataset, we obtained 22 factors. Given this relatively small number, all factors were retained for subsequent visualization.

After projection into two-dimensional space (Fig. 4), several clusters of factors can be observed across all visualizations, easily identifiable even without detailed inspection of the coordinates. We refer to these as reference clusters. This is consistent with the underlying logic of MDS algorithms, which aim to position the most distinctive entities first.

Analysis of the clusters reveals clear patterns corresponding to major animal classes. Specifically, clusters represent mammals (cluster C_1), birds (cluster C_2), and fish (cluster C_3). Remaining factors are positioned according to their proximity to these reference clusters, and the clusters remain recognizable across all similarity measures.

Cluster C_1 contains factors 1, 5, 10, and 21, all describing mammals exclusively. These factors share defining characteristics, notably *hair* and *milk production*.

The bird-dominated cluster C_2 contains factor 2, describing only birds. The main shared attribute is *eggs*, although a few insects also appear due to traits such as *eggs*, *legs*, *breathing*, or *flight*. These inclusions reflect functional rather than phylogenetic similarities.

The last cluster C_3 is dominated by *fish* and associated with traits such as an *aquatic habitat*, *eggs*, *backbone*, *fins*, *tail*, *teeth*, and *predator*. Atypical aquatic mammals and birds, including *dolphins* and *penguins*, also appear nearby. Their relative rarity within their classes places them closer to *fish* in the embedding, reflecting shared ecosystemic and physiological traits.

The strong correspondence between clusters, animal classes, and defining physiological traits is exemplified by factor 17, representing aquatic mammals, which consistently occupies an intermediate position between the *fish* and *mammal* clusters across all visualizations.

[3] `https://www.kaggle.com/datasets/rounakbanik/pokemon`.

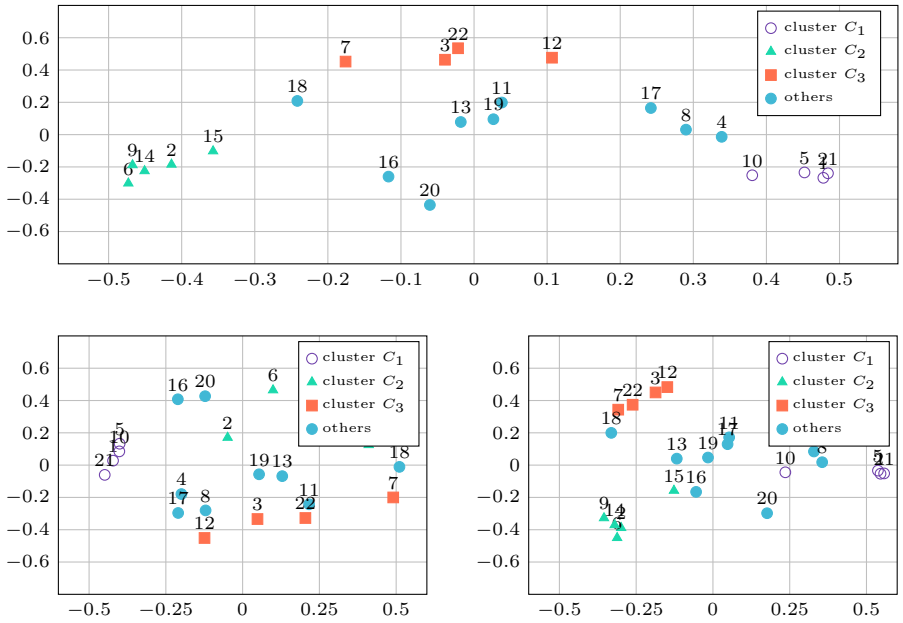


Fig. 4: MDS projections of the Zoo dataset for: overlapping coefficient (top), Jaccard over intents (bottom left), and Jaccard over extents (bottom right).

In all cases, the stress value was low, suggesting that the factors are well differentiated and that the obtained embedding is informative.

4.2 Pokémon data

The Pokémon dataset comprises 801 Pokémon from seven generations, each described by 73 binary attributes encoding type membership (allowing multiple types), effectiveness against other types at different intensity levels, and legendary status. We restricted our analysis to the first generation (151 Pokémon). The resulting factorization produced 51 factors; however, this number proved impractical for visual interpretation. Therefore, only the first 25 factors, covering 90% of the data, were retained for subsequent analysis.

The resulting projections reveal less distinct cluster structures compared to the other datasets. Across the considered similarity measures, clusters are generally weaker and less clearly separated.

For the overlapping coefficient (see Fig. 5, top), two large clusters can be observed in the central part of the graph. Although these clusters contain several overlapping factors, their internal cohesion is not particularly strong. What these factors share is that they are highly dissimilar to most of the remaining factors; however, each of them differs in a distinct way.

Cluster C_1 consisting of factors 4, 9, and 23 appears consistently across all visualizations; this cluster represents *ground-* and *rock-type* Pokémon.

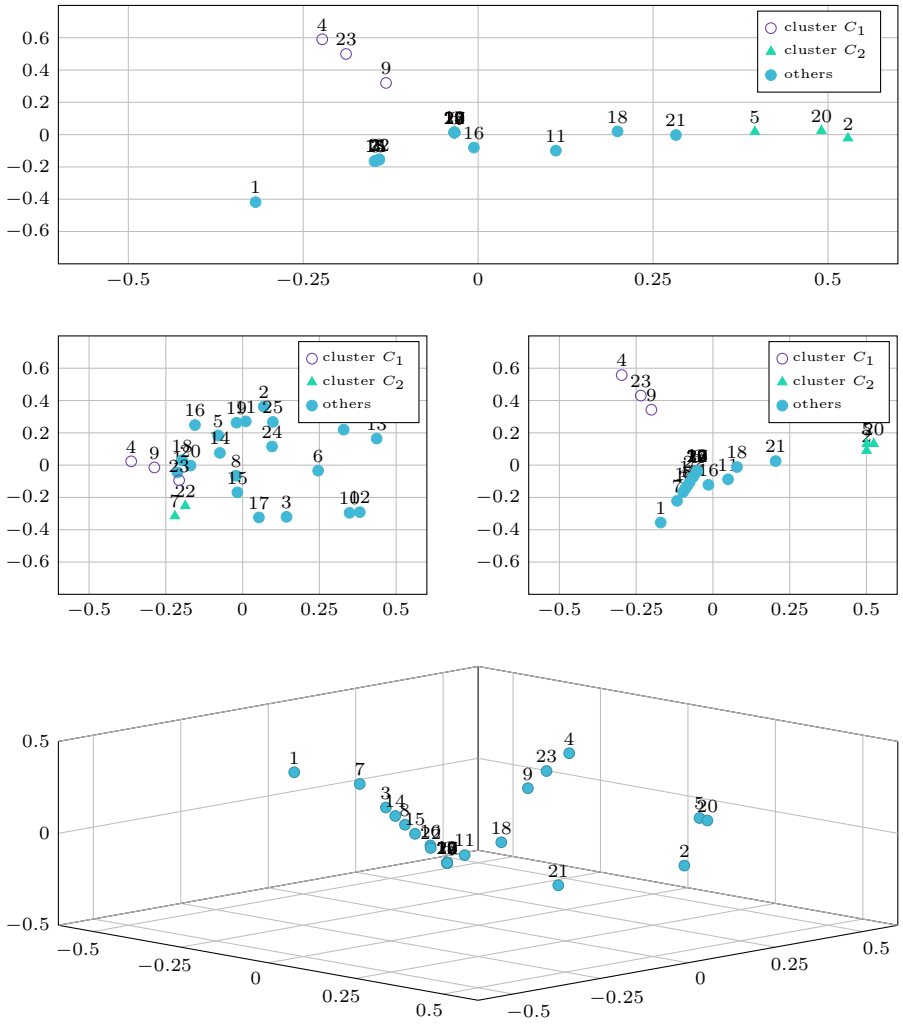


Fig. 5: MDS projections of the Pokémon dataset for: overlapping coefficient (top), Jaccard over intents (middle left), Jaccard over extents (middle right) and Jaccard over extents in 3D.

When using the overlapping coefficient and Jaccard similarity over extents, another cluster emerges, grouping factor 5 (*grass-type* Pokémon) with factors 2 (*poison-type* Pokémon) and 20 (mainly *grass-* and *poison-type*), see C_2 in Fig. 5 (top and middle right). This indicates that *grass-type* Pokémon share similarities with a subset of *poison-type* Pokémon. This is consistent with the well-known overlap between *grass-* and *poison-type* Pokémon in the first generation, where many *grass-type* also possess the *poison-type*.

A similarly interesting cluster appears when using Jaccard over intents. In the lower-left region of the graph, factors 7 and 22 form a cluster, see C_2 in Fig. 5 (middle left). Factor 7 corresponds to a subset of *water-type* Pokémon, while factor 22 represents *ice-* and *water-type* Pokémon. This grouping indicates that *ice-type* Pokémon share notable similarities with many *water-type* Pokémon. This reflects the underlying design of Pokémon type system as in early generations where most *ice-type* Pokémon simultaneously possess the *water-type*.

Using Jaccard similarity over extents results in a projection with a relatively high stress value (0.72). Increasing the embedding dimension to three, see Fig. 5 (bottom), improves the representation, as reflected by a lower stress value (0.64).

Overall, the Pokémon dataset exhibits relatively high structural complexity. As a result, cluster structures are less pronounced; nevertheless, meaningful relationships between factors can still be identified in the visualizations.

4.3 Congressional voting records data

The Congressional voting records data captures the voting behavior of 435 members of the U.S. House of Representatives on 16 key legislative issues in 1984. Each congressman is described by binary attributes indicating their votes, together with a class label specifying party affiliation (Democrat or Republican).

The BMF of the voting data yielded 76 factors. For subsequent analysis, we retained the first 33 factors, covering 90% of the data.

After plotting the projections, three clearly distinguishable clusters emerge across all graphs (see Fig. 6). Factors in cluster C_1 predominantly contain objects labeled Democrat, whereas factors in cluster C_2 group objects labeled Republican. Cluster C_3 comprises factors whose objects belong to both the Democrat and Republican classes, with class ratios varying but typically close to 50:50.

Within clusters C_1 and C_2, the factors exhibit high mutual similarity. In contrast, the factors in cluster C_3 are not necessarily similar to one another; however, they share some degree of similarity with other factors.

Democratic representatives in cluster C_1 are characterized by factors with "yes" votes on the adoption of the budget resolution and aid to the Nicaraguan Contras. These votes reflect core Democratic positions of the era, emphasizing government spending and selective foreign intervention.

Republican representatives in cluster C_2 are characterized by voting in favor of crime-related measures, approval of foreign aid to El Salvador, and specific positions on education spending. They overwhelmingly voted "yes" on these issues. In contrast, Democrats generally voted oppositely, highlighting a clear separation between the parties.

4.4 Discussion

The resulting projections provide a concise overview of the factorization as a whole. This overview makes it possible to identify groups of related factors, as well as factors that are clearly separated from the others, thereby highlighting

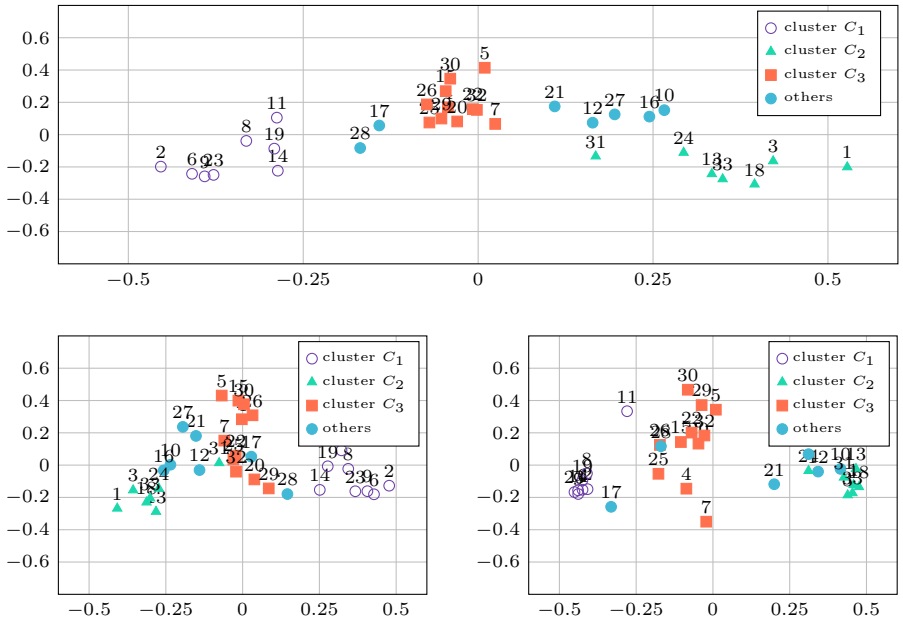


Fig. 6: MDS results on the Congressional Voting Records dataset for three similarity measures: overlap coefficient (top), Jaccard over intent (bottom left), and Jaccard over extent (bottom right).

potentially significant and interpretable factors (cf. the second and third Thurstone criteria). Moreover, the similarity between certain factors significantly facilitates their interpretation. Once one factor has been interpreted, the meaning of related factors can be inferred more easily from their mutual relationships.

During subsequent analysis, attention may be restricted to selected clusters, or a representative factor may be chosen from each cluster. This strategy is particularly suitable when working with many factors or heterogeneous data. It is especially useful when a class attribute is available, as objects from the same class may naturally form subcategories not explicitly encoded in the original attributes and not immediately apparent in the data. In the visualization, factors associated with a single class may therefore appear in distinct clusters, revealing meaningful differences between these subcategories. This reduction simplifies interpretation by enabling analysis of individual substructures before integrating them into the broader context. Importantly, the same approach remains applicable even without an explicit class attribute. Overall, the resulting visualization provides a valuable tool for analyzing and interpreting BMF results.

4.5 Practical notes

The proposed approach is limited by the properties of MDS. As the number of factors increases, projection into two- or three-dimensional space becomes

less accurate. Since we employ classical MDS, it is also possible in our case to assess how well the original data are approximated in a lower-dimensional space. In general, this issue can be addressed by focusing only on the first l factors instead of considering all factors.

5 BFLens

BFLens[4] is a web-based application developed as part of this work that enables interactive exploration of BMF results through the visualization techniques presented in this paper, as well as additional views such as the coverage quality curve, similarity matrices, and interactive inspection of individual factors. The tool aims to simplify the otherwise time-consuming process of factor exploration, which has so far largely relied on manual analysis.

The software is implemented as a web-based visualization tool using the Next.js framework. Its monolithic design allows execution as a standalone application without reliance on external services, enabling local deployment. In addition, the application can be deployed on a server, facilitating presentation and sharing of results through a web interface. The use of Next.js supports efficient rendering, flexible deployment strategies, and a unified codebase for both local execution and web-based presentation.

The user interface, optimized for desktop use, is divided into four sections, each highlighting a different aspect of the factorization results: (i) the general overview, presenting the data and extracted factors, including factor size and coverage; (ii) the class distribution, summarizing class occurrences across factors when available; (iii) the factor overlap, examining shared attributes and classes between factors and the original data; and (iv) the factor similarity section, visualizing relationships and similarities among factors. An example of the user interface is shown in Fig. 7.

For performance reasons, we chose to precompute all statistics using an external script rather than calculating them within the application. The script takes the factorization as input—namely the input matrix $\mathbf{I}$ and matrices $\mathbf{A}$ and $\mathbf{B}$—and outputs the processed data in JSON format, which are subsequently visualized by the application. This design choice is motivated by computational efficiency, as computing all characteristics for larger datasets can be time-consuming and is better handled by technology optimized for numerical processing. In future work, we plan to integrate this preprocessing step directly into the application.

6 Conclusion

In this paper, we addressed the problem of interpreting BMF results from a qualitative and exploratory perspective. While existing approaches typically emphasize quantitative measures such as reconstruction error, we focused on structural

[4] The source code of the application is publicly available on GitHub at `https://github.com/PavlikovaVeronika/BFLens`.

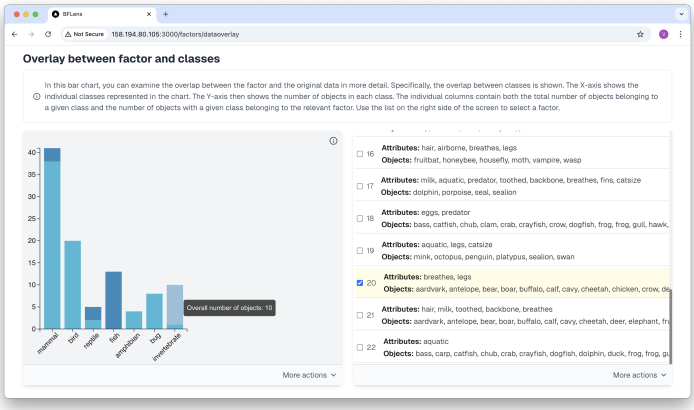


Fig. 7: BFLens tool.

relationships between factors and their role in interpretation. We discussed key characteristics of individual factors and proposed visualization techniques that support exploratory analysis, facilitate interpretation, and reduce the number of factors considered in subsequent investigation.

To capture the overall structure of a factorization, we introduced a similarity-based approach combined with multi dimensional scaling, enabling simultaneous visualization of relationships among all factors. The resulting projections provide an intuitive global overview, reveal clusters of related factors, and highlight distinctive or particularly interpretable factors. The approach is designed primarily as an exploratory tool, supporting interactive investigation rather than formal statistical inference.

We demonstrated the proposed methodology on datasets with different structural properties and showed that it uncovers meaningful patterns even in structurally complex data. Finally, we presented BFLens, a web-based tool for interactive exploration of BMF results that integrates the proposed visualizations into a unified analytical framework.

Future research will pursue several directions. First, the results of the proposed analysis depend on the chosen similarity measure. Although preliminary experiments with alternative similarities revealed substantial differences, some clusters remained stable across measures. A natural continuation of this work is a more extensive experimental study of the similarity measures discussed in [2], including their systematic comparison with respect to suitability for visualization and exploratory analysis. Furthermore, the proposed approach can be extended to support direct comparison of different factorizations, a task frequently required in practice. Another promising direction is the incorporation of interactive zooming mechanisms. In particular, a selected region of the projection could be re-visualized using an alternative similarity measure, potentially revealing finer-grained relationships between factors and facilitate their interpretability.

Acknowledgments. This work was supported by Grant No. IGA_PrF_2025_18 and IGA_PrF_2026_XX of the Internal Grant Agency of Palacký University Olomouc.

References

1. 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
2. Belohlavek, R., Mikula, T.: Similarity metrics vs human judgment of similarity for binary data: Which is best to predict typicality? Appl. Soft Comput. 153, 111270 (2024). https://doi.org/10.1016/j.asoc.2024.111270
3. 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
4. Belohlavek, R., Trnecka, M.: Semantic explorations in factorizing Boolean data via formal concepts. Int. J. Approx. Reasoning 173, 109247 (2024). https://doi.org/10.1016/j.ijar.2024.109247
5. Belohlavek, R., Krmelova, M.: Factor analysis of ordinal data via decomposition of matrices with grades. Ann. Math. Artif. Intell. 72, 23–44 (2014). https://doi.org/10.1007/s10472-014-9398-6
6. Brazdilova, K., Trnecka, M., Trneckova, M.: Data preprocessing using banded structure and image morphology enhancing Boolean matrix factorization. Knowl.-Based Syst. 329(B), 114366 (2025). https://doi.org/10.1016/j.knosys.2025.114366
7. Cattell, R.B.: The scientific use of factor analysis in behavioral and life sciences. Springer US (1978)
8. Chen, C.-h., Härdle, W.K., Unwin, A. (eds.): Handbook of Data Visualization. Springer, Berlin, Heidelberg (2008)
9. Forsyth, R.: Zoo [Dataset]. UCI Mach. Learn. Repos. (1990). https://doi.org/10.24432/C5R59V
10. Ganter, B., Glodeanu, C.V.: Ordinal factor analysis. In: Domenach, F., Ignatov, D.I., Poelmans, J. (eds.) Formal Concept Analysis (ICFCA 2012), Lect. Notes Comput. Sci., vol. 7278, pp. 128–139. Springer, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29892-9_15
11. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations, 2nd edn. Springer (2024)
12. Miettinen, P., Neumann, S.: Recent developments in Boolean matrix factorization. In: Proc. IJCAI 2020, Survey Track, pp. 4922–4928 (2020). https://doi.org/10.24963/ijcai.2020/685
13. Borg, I., Groenen, P.J.F.: Modern Multidimensional Scaling: Theory and Applications, 2nd edn. Springer, New York, NY (2005)
14. Royce, J.R.: Factors as theoretical constructs. Am. Psychol. 18(8), 522 (1963)
15. Thurstone, L.L.: Multiple-factor analysis: A development and expansion of the vectors of mind. University of Chicago Press (1947)
16. Tufte, E.R.: The Visual Display of Quantitative Information, 2nd edn. Graphics Press, Cheshire, CT (2001)
17. Tversky, A.: Features of similarity. Psychol. Rev. 84(4), 327–352 (1977). https://doi.org/10.1037/0033-295X.84.4.327
18. Congressional Voting Records [Dataset]. UCI Mach. Learn. Repos. (1987). https://doi.org/10.24432/C5C01P