DimLift: Interactive Hierarchical Data Exploration through Dimensional Bundling

Laura Garrison, Juliane Müller, Stefanie Schreiber, Steffen Oeltze-Jafra, Helwig Hauser, Stefan Bruckner

Abstract—The identification of interesting patterns and relationships is essential to exploratory data analysis. This becomes increasingly difficult in high dimensional datasets. While dimensionality reduction techniques can be utilized to reduce the analysis space, these may unintentionally bury key dimensions within a larger grouping and obfuscate meaningful patterns. With this work we introduce *DimLift*, a novel visual analysis method for creating and interacting with *dimensional bundles*. Generated through an iterative dimensionality reduction or user-driven approach, *dimensional bundles* are expressive groups of dimensions that contribute similarly to the variance of a dataset. Interactive exploration and reconstruction methods via a layered parallel coordinates plot allow users to *lift* interesting and subtle relationships to the surface, even in complex scenarios of missing and mixed data types. We exemplify the power of this technique in an expert case study on clinical cohort data alongside two additional case examples from nutrition and ecology.

Index Terms—Dimensionality reduction, interactive visual analysis, visual analytics, parallel coordinates.

1 Introduction

Dimensionality reduction techniques are frequently utilized to reduce the complexity of high dimensional data by projection to a lower dimensional space. However, when used alone and monolithically, these techniques can emphasize strong, uninteresting patterns in the data and hide important variations. For example, although cardiac risk is well-known to correlate with waist measurement, a more interesting, though subtle, relation to gender or smoking may be relevant for a clinician to see. Visual analytics leverages the strengths of powerful statistical tools, including dimensionality reduction, in tandem with user knowledge. However, while some visual analysis tools have been developed to create expressive dimensional groupings, they do not easily allow for incorporation of user knowledge for faceted hypothesis generation. Furthermore, connecting the results of the dimensionality reduction back to the original data for interpretation and relation to subsequent steps, e.g., decision making, can be difficult.

For instance, in clinical cohort studies medical experts are chiefly interested in untangling interesting and relevant measures of a given disease, e.g., cerebral small vessel disease (CSVD), for diagnostic purposes. Biomarker discovery is a complex and challenging process, and dimensionality reduction techniques provide a means to reduce the analysis space. However, these techniques may produce groupings that are not interesting to the expert for particular subcohorts, e.g., a specific gene expression level grouped with test results for young patients. Our method, which utilizes iterative dimensionality reduction to extract subsets of dimensions that contribute similarly to the variation of a dataset, allows for flexible user-driven restructuring of subcohorts and subsequent groupings to support exploratory hypothesis generation. For example, adjusting the previous subcohort to include middle-aged patients with high blood pressure may be done to support a new hypothesis that high expression of a particular gene in combination with a certain range of test scores, such as high blood pressure, can act as a set of indicators for CSVD in middle-aged patients. Similar such scenarios occur in many areas of science and engineering. These domains are interested in exposing patterns in subsets of large, complex populations, and benefit from this style of visual reasoning.

The rapid identification of interesting patterns and relationships is key to the analysis of complex high dimensional data. Achieving this requires effective integration of statistical methods with user knowledge to reduce the space to salient dimensions. Core to our approach is the concept of *dimensional bundles*: statistical- or user-driven groupings of dimensions that are accessible as a unit or at the component level. Our statistical approach utilizes factor analysis of mixed data (FAMD) [1], a dimensionality reduction technique applicable to complex, mixed-type data. We run this algorithm in multiple iterations over the data; each iteration captures and extracts a set of dimensions, so called *dimensional bundles*, which contribute similarly to the variance within the dataset. This avoids a monolithic treatment and instead produces hierarchical bundles of dimensions that retain the expressitivity of the original dataset. While previous approaches to dimensional grouping have focused on clustering or dimensionality reduction methods that

- *Laura Garrison is with Dept. of Informatics & Mohn Medical Imaging and Visualization Centre, Dept. of Radiology, Haukeland Univ. Hospital, University of Bergen, Norway. E-mail: laura.garrison@uib.no*
- *Juliane Müller is with Dept. of Neurology, Otto von Guericke University Magdeburg, Germany. E-mail: juliane.mueller@med.ovgu.de*
- *Stefanie Schreiber is with Dept. of Neurology & the Center for Behavioral Brain Sciences, Otto von Guericke University Magdeburg, Germany. E-mail: stefanie.schreiber@med.ovgu.de*
- *Steffen Oeltze-Jafra is with Dept. of Neurology & the Center for Behavioral Brain Sciences, Otto von Guericke University Magdeburg, Germany. E-mail: steffen.oeltze-jafra@med.ovgu.de*
- *Helwig Hauser is with Dept. of Informatics & Mohn Medical Imaging and Visualization Centre, Dept. of Radiology, Haukeland Univ. Hospital, University of Bergen, Norway. E-mail: helwig.hauser@uib.no*
- *Stefan Bruckner is with Dept. of Informatics & Mohn Medical Imaging and Visualization Centre, Dept. of Radiology, Haukeland Univ. Hospital, University of Bergen, Norway. E-mail: stefan.bruckner@uib.no*

Manuscript received September 4, 2020; revised January 5, 2021.

978-1-5386-5541-2/18/$31.00 ©2021 IEEE

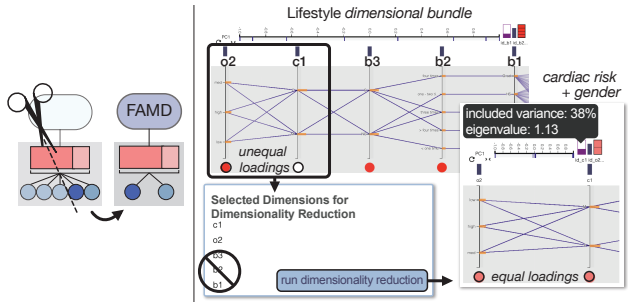

Fig. 1: We inspect a *dimensional bundle* comprised of lifestyle dimensions, e.g., education (b1), workout frequency (b2), smoking (b3), cardiac risk (o2), and gender (c1). We suspect a correlation between cardiac risk (o2) and gender (c1), so then *lift* these dimensions to better **target and test our hypothesis** by removing all other dimensions from this bundle. With an eigenvalue above 1 and changes in contributions/loadings indicated by hue at the bottom of the axes, we note a **subtle correlation** that was previously undetectable. This is conceptually illustrated on the left.

converge to an ideal representation of a high dimensional dataset [2], our approach facilitates dynamic visual navigation and composition of high dimensional data to *lift* subtle, interesting features to the surface. A simple example of *dimensional bundle* restructuring is shown in Fig. 1, where we explore the relationship between cardiac risk and gender.

Expressive visual mapping techniques from visual analytics have been developed to represent dimensional hierarchies, e.g., parallel coordinates plots with embedded cluster diagrams [3]. However, these solutions typically expect data in a pre-established hierarchical structure, or offer limited interaction methods for restructuring groups during analysis. By contrast, our method proposes to semantically connect and track data transformations through visual mappings and interactions that allow on-the-fly recomposition of *dimensional bundles*. This provides a flexible solution to swiftly adapt perspectives on high dimensional data with the potential to rapidly identify relevant relations, even when overshadowed by well known or less interesting trends. Our concept extends to complex mixed-type and incomplete data. Following a review of related work and a description of our methodology, we demonstrate the power of our approach in the context of three scientific datasets, one of which is a case study with domain experts in clinical neurology.

2 Related Work

Visual analysis of high dimensional data is a grand challenge in the visualization research community, with applications across numerous domains. Discussion of efforts in this general area are beyond the scope of this paper, but are detailed with a survey of advances in recent years by Liu et al [4]. Our work expands on the idea of simultaneous dimensions and items analysis for exploratory hypothesis generation, described as the Dual Analysis approach, by Turkay et al. [5]. This work, along with a follow-on clinical application study [6], describes a visual analysis workflow where users seamlessly move between analysis of dimensions through comparative descriptive statistics and item comparison to identify outliers and correlations of interest. Brushing and linking mechanisms provide clear visual feedback during the analysis process. This approach has since been extended to incorporate mixed data (continuous and categorical) with facilities for visualization of missing data by Müller et al. [7]. DimLift expands further on the reciprocity between dimension and item space by introducing *dimensional bundles* for analysis of high dimensional data. The SIRIUS system, presented by Dowling et al. [8], explores the interplay of dimension and item space while incorporating a nonlinear dimensionality reduction technique. This approach shows MDS projections for both dimension and item space in linked views to demonstrate correlations in high-dimensional data. Our approach similarly utilizes dimensionality reduction to aid correlation exploration of high-dimensional data, but adopts a linear technique to better track between the original and newly-produced dimensions.

Dimensionality reduction is used ubiquitously in visual analytics. Sacha et al. [9] provide an overview and classification of dimension reduction methods as used in visual analysis. Our work incorporates dimensionality reduction into a subset of the interaction scenarios they identified: data selection & emphasis, data manipulation, and feature selection & emphasis. Similar works in this space include the work of Tatu et al. [10], who utilize an interestingness-guided subspace search algorithm to identify subspace sets for subsequent visual analysis, although tools for user-driven subspace composition are limited. The DimStiller workflow by Ingram et al. [11] guides users through the dimensionality reduction process to find a single global optimal composition; our approach by contrast does not emphasize a single global optimum, and is designed for a variety of complementary perspectives onto relations between relevant subsets of the dimensions. Yuan et al. [12] combine a Dimension Projection Matrix, an extended scatterplot matrix, with a Dimension Projection Tree to explore data and dimension subspaces. Our approach tackles a similar goal of dimensional subspace analysis at both item and dimension levels with related interactions. However, our visual approach enables direct correlation comparison between multiple dimensions and items in a parallel coordinates view, and is targeted specifically at user-driven hypothesis exploration.

Although dimension reduction methods project relevant data features into low dimensional space, the results are often difficult to comprehend. Principal component analysis (PCA) [13], although a well-known and broadly applicable method utilized in dimensionality reduction, suffers from this interpretation gap. Müller et al. [14] present a general discussion of design solutions to clearly visualize the connection between data inputs and results from PCA. However, many of these solutions do not scale well with high dimensional data. Our visual interactive approach offers one method for bridging this intuitive gap into high dimensional data spaces. iPCA [15] is one other such solution designed to connect PCA results to source data. It uses multiple coordinated views to depict PCA results with interaction facilities for the user to adjust dimension contributions within any principal component—any adjustments

update visuals for the final PCA results. Our approach similarly uses visual elements and interactions to connect the raw data to the results of a linear dimensionality reduction method, but rather than using visualization to understand the semantics of PCA, our approach uses similar results as a tool in hypothesis formation.

Parallel coordinates are a well-known method for representing multidimensional data [16]. Nested or hierarchical plots, adapted from the traditional flat parallel coordinates plot, are used to visualize and evaluate structural relationships of the data. Numerous solutions present data aggregation by item relatedness into parallel coordinates as a means to reduce clutter and noise in the plot [17], [18], [19], [20], [21], [22], [23]. Each of these methods focuses on the hierarchical construction of sets of items, while our approach aims one level above this on the hierarchical construction of sets of dimensions. Several approaches have utilized parallel coordinates to visualize dimension-level aggregation, created either through algorithmic methods or pre-defined data hierarchies. These methods provide varying degrees of interaction to the user. For example, Wang et al. [24] and Dunica et al. [25] use parallel coordinates to visualize results of a single-run PCA, where each axis represents a different principal component. While we similarly incorporate principal components, we instead take an iterative algorithmic approach to produce principal components of selected subsets of particularly related dimensions. These subsets form the *dimensional bundles* in our method.

Approaches to **parallel coordinate dimension hierarchies** often incorporate other views on the data, integrated either separately or directly into the parallel coordinates. Huang et al. [3] create hierarchical clusters of dimensions in parallel coordinates using dendrograms that attach to each axis. DOFSA [26] and InterRing [27] are connected tools that allow interactive visual exploration and modification of hierarchical data. These modifications are made on InterRing and linked to other panels, e.g., parallel coordinates. By contrast, our method does not divide user attention over different graphical interfaces. Furthermore, the DOFSA hierarchy itself is flattened in parallel coordinates, and its order is informed by the hierarchy constructed in InterRing. Our approach does not flatten the hierarchy in this manner. Weidele [28] recently presented the conditional parallel coordinates method, which ties and reveals additional dimensions to the range of a given parent dimension only if certain conditions are met. Perhaps most similar in principle to our visual approach, Brodbeck & Girardin [29] and Andrews et al. [30] create aggregated dimension axes for parallel coordinates plots, which may then be expanded to reveal the contained dimensions. In contrast to these methods, we do not expect pre-defined hierarchies, instead allowing flexible regrouping as hypotheses evolve.

3 DimLift Approach

Key to complex high dimensional data analysis is the rapid identification of interesting dimensions. While dimensionality reduction is a core tool for high dimensional data analysis, nonlinear methods do not allow for an easy link back to the original dimensions, which Sedlmair et al. [31] identify as key tasks for users interested in finding important

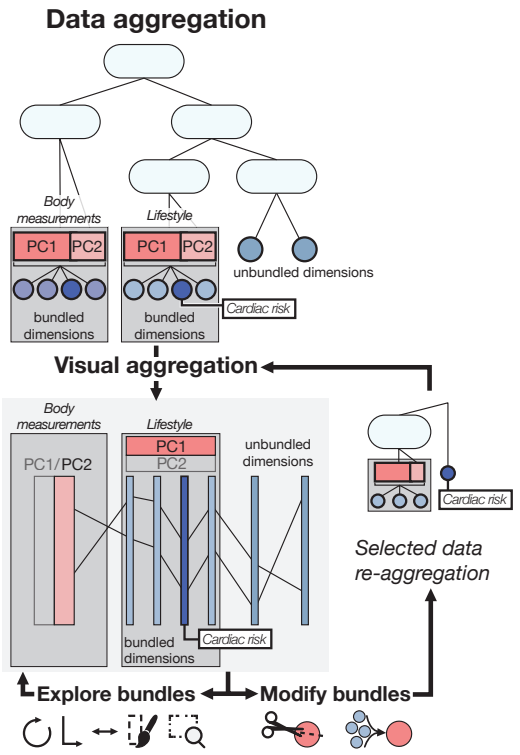

Fig. 2: Conceptual pipeline of *DimLift*. Factor analysis of mixed data (FAMD) is applied iteratively to produce *dimensional bundles* (body measurements, lifestyle). Data are mapped to a layered parallel coordinates plot for users to explore and structurally modify. FAMD is re-run on any structurally-altered *dimensional bundles* before visual remapping. We highlight the path of the dimension cardiac risk in one possible interaction flow in our approach.

original dimensions (as opposed to purely gaining insights on the dataset structure). Sedlmair et al. also identify the need of users to compare, or unmap, original dimensions to newly-created dimensions; nonlinear methods are also of limited use for this task. Lastly, Sacha et al. [9] identified user interactions as critical components of an exploratory visual analysis pipeline utilizing dimensionality reduction. While numerous solutions in this space have incorporated a human into the loop, many aim to help the user to better understand the results of the algorithm, or to guide the user to identify a single global optimum of reduced dimensions, as we discussed previously in Sec. 2. These solutions are less effective for an exploratory approach where the user develops multiple new hypotheses over a single session. For each newly-formed hypothesis, the user needs to identify interesting, important original dimensions.

In contrast, our *DimLift* approach utilizes dimensionality reduction and user-driven methods to produce similarly-contributing groups of dimensions, i.e., *dimensional bundles*, that serve as the primary unit of exploration and interaction. These bundles reduce the analysis space while allowing the user full control to discover interesting relationships that may otherwise go unnoticed. Inspired by Elmqvist and Fekete's [32] principles for the visualization of aggregate hierarchies, we show our analytical workflow in Fig. 2.

The user initiates algorithmic construction of *dimensional bundles* with a linear dimensionality reduction technique. Subsequent visual analysis allows the user to explore the degree of bundling of their data, which offers insights on the degree of correlation within the data. Our choice of a linear dimensionality reduction algorithm allows users to visually inspect bundle contents to identify important original dimensions that are now mapped to the new bundles. As new questions form, users may reconstruct bundles to emphasize and lift interesting patterns for detailed exploration. This series of steps may be repeated as new hypotheses and insights are continually formed.

In the remainder of this section, we present our methodological approach alongside a synthetic dataset containing human lifestyle and body measurements, organized as follows: In Section 3.1, we detail our method for automatically generating *dimensional bundles*, Section 3.2 discusses our visual encodings, and Section 3.3 describes our interaction facilities for lifting expressive dimensions. We conclude with a discussion of our treatment of mixed and missing data in Section 3.4.

3.1 Creating Expressive *Dimensional Bundles*

High dimensional data analysis typically involves producing a low dimensional projection of the data. Common dimensionality reduction techniques automatically treat a dataset monolithically, and may obfuscate subtle but relevant characteristics. In our simple example, smoking (b3) is an important indicator for cardiac risk (o2), but a standard dimensionality reduction does not easily show this relationship. It instead buries these dimensions in all five principal components (Fig. 3, top). In contrast, our iterative dimensionality reduction approach extracts subsets of dimensions that contribute similarly to the variance within the dataset. We define these subsets as *dimensional bundles*. For example, our approach places the lifestyle-related dimensions education level (b1), workout frequency (b2), and smoking (b3), together with the similarly-contributing variable cardiac risk (o2) (Fig. 3, bottom). In the following, we describe our algorithmic process to creating *dimensional bundles*. This is additionally described in pseudocode in Algorithm 1.

Algorithm 1: *Dimensional bundle* creation for two or more dimensions

```
1  initialize pool = all dimensions in dataset
2  do
3      mark all dims in pool as possibly contributing
4      initialize new bundle
5      perform FAMD on pool
6      for all dimensions in pool
7          if PC1 loading ≥ contribution threshold
8              move dimension from pool to new bundle
9          else
10             mark dimension as non-contributing
11 while pool contains dimensions marked as
       non-contributing
12 for all bundles
13     perform FAMD on bundle
14     store PC1 and PC2 for bundle
```

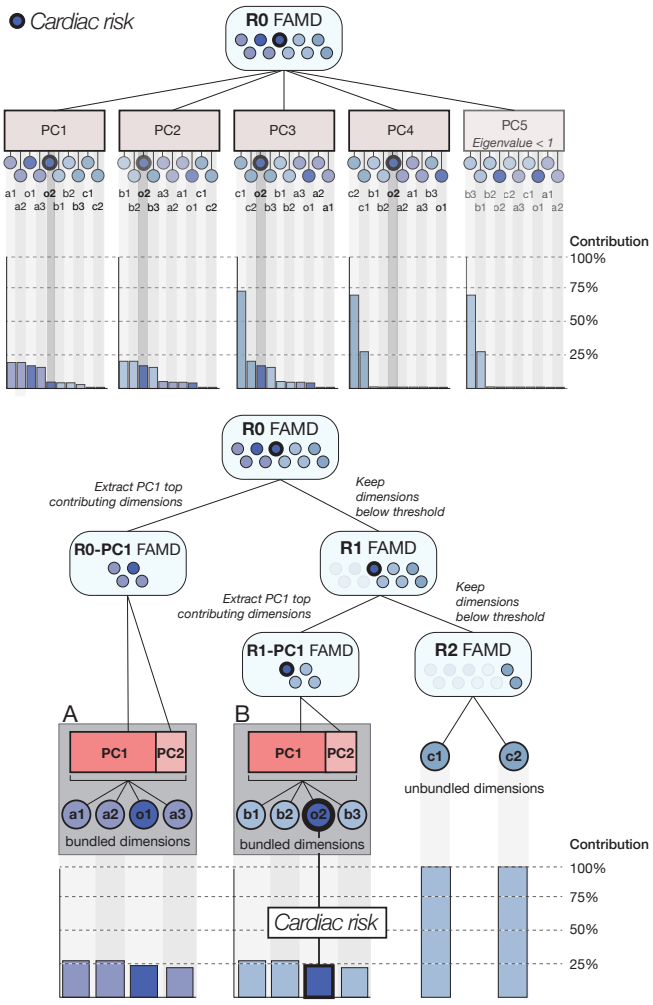

Fig. 3: We contrast our iterative algorithmic approach (bottom) with a standard approach (top) using a synthetic ten dimensional health and lifestyle dataset comprised of four quantitative [height (a1), weight (a2), waist circumference (a3), BMI (o1)] and six qualitative [education level (b1), workout frequency (b2), smoking (b3), gender (c1), eye color (c2), and cardiac risk (o2)] dimensions. A standard approach contains all ten dimensions in each principal component (PC), e.g., cardiac risk (o2) is present in all PCs. In contrast, our approach (bottom) produces a pair of *dimensional bundles* (**A**: body measurement, **B**: lifestyle) containing only dimensions with similar variance contributions, where cardiac risk is bundled into **B**. Dissimilarly contributing dimensions, i.e., c1 and c2, remain unbundled.

Bundle creation. Prior to analysis, we standardize all input dimensions; this ensures equal weighting between dimensions comprised of items on different scales. We then run a factor analysis of mixed data (FAMD) [33] on all dimensions, provided two or more dimensions are available in the pool (line 5). The resulting correlation matrix is used to determine principal components (PCs), their respective eigenvalues, and the contributing dimensions to each PC. We focus on the first principal component (PC1) for the creation of each bundle (line 7), as this captures the largest

variance within the data [13] and shows the most potential to create expressive bundles.

Formally, PC1 is defined to maximize the sum of squared correlation coefficients r^2 between itself and each dimension k:

$$\sum_{k} r^2(k, PC1) \tag{1}$$

Referencing the loading of each dimension, i.e., the correlation coefficient r that defines the factors by which the corresponding original attributes are multiplied so that they add up to the scores of PC1, we extract only those dimensions contributing above a threshold defined as $100\%/$*number of input dimensions* [34] (lines 6-8). We use this threshold as a baseline heuristic for creating bundles of similarly-contributing dimensions; it defines whether the contribution of a given dimension exceeds the average contribution to PC1. Thus, we formally define the initialization of a *dimensional bundle* as the set of all dimensions with loadings greater than or equal to this threshold, with respect to PC1.

On all *dimensional bundles* we then run another FAMD and save: (a) the principal component scores, which are the computed representations of the individual items for the bundled dimensions, and the (b) contribution and (c) loading of each dimension (lines 12-14). For this second run we do not use the threshold, and instead keep all contributing dimensions. We preserve the second principal component (PC2) in this second FAMD run to provide additional structural context for PC1, and to further indicate the quality of the bundling. We found diminishing returns for including any further PCs, particularly since the full dimensional information is already provided with the first FAMD sequence. Preserving PC1 and PC2 at the *dimensional bundle* level conforms to a manageable mental analysis model and avoids visually overwhelming the user. Thus, these three elements: PC1, PC2, and their contributing dimensions, comprise a complete organized *dimensional bundle* (Fig. 3A, B).

The dimensions that do not meet the contribution threshold remain in the original dimension pool (line 10). We recurse on this pool of dimensions (lines 2-11) until less than two dimensions remain, or until the eigenvalue of PC1 falls below 1, meaning that PC1 accounts for less variance than one of the original dimensions (Kaiser criterion [35]). These dimensions are left unbundled (Fig. 3, unbundled dimensions). This produces the branching structure as shown in the bottom diagram of Fig. 3. The results of this algorithm, both *dimensional bundles* and unbundled dimensions, serve as a meaningful basis for subsequent user-driven exploration and knowledge integration.

3.2 Visual Encodings

Projecting to a lower dimensional subspace in dimensionality reduction often creates a degree of disconnect from the source data [15]. If the analyst can identify interesting correlations leading to new discoveries in their bundles, but is ultimately unable to relate these correlations back to the original data, then this is not an actionable application of dimension grouping. To solve this issue, we preserve and map the semantics of *dimensional bundles* produced through dimensionality reduction directly to visual elements. Our visual aggregation utilizes a modified parallel coordinates plot that mirrors the results of the data aggregation step. Our approach is guided by principles of unambiguous data depiction and visual-data correspondence, inspired by the algebraic method of visualization design [36].

The basic unit of the *DimLift* method, *dimensional bundles*, consist of two principal components (first and second) and their constituent input dimensions. This composition forms a hierarchy of data representations. Usually, each *dimensional bundle* has a number of sibling bundles, which are other bundles produced by our iterative FAMD approach. Our visual design is based on the following requirements, which we draw from the basic high dimension data analysis tasks that we discussed at the beginning of Sec. 3:

R1 Support the creation of *dimensional bundles*
R2 Support the iterative modification of *dimensional bundles*
R3 Allow rapid retrieval of item values in a given *dimensional bundle*
R4 Lift dimensions of interest in a *dimensional bundle*
R5 Provide information on the quality of each *dimensional bundle*
R6 Allow for relation investigation between *dimensional bundles* and input dimensions

Parallel coordinates are a popular, generally applicable technique to visualize relationships and correlations in multidimensional datasets [16]. Furthermore, they have been shown as more effective in visual retrieval of data values relative to scatterplot matrices (SPLOMs) [37] (**R3**), and more performant than SPLOMs in solving tasks for higher dimensional data [38]. We utilize parallel coordinates but with some adaptations; although bifocal parallel coordinates presented by Kaur and Karki [39] visualize all dimensions simultaneously, this becomes overwhelming. We instead use an approach inspired by the perspective walls technique [40] to focus attention on bundles relevant to the user. Our modified plot further supports three layers of nested visual analysis within and between each *dimensional bundle*. This nested approach is inspired by model-based reasoning methods described by Liu et al., where deeper data insights can be obtained by presenting information sets and supersets [41]. The result is shown in Fig. 4 (**R1–R4, R6**). Beginning with the axes of a traditional parallel coordinates plot, we set the stroke-width of each axis relative to the number of contained dimensions, similar to Andrews et al. [30]. For each *dimensional bundle*, PC1 is depicted as the primary axis in the plot (Fig. 4A). Items are plotted by their scores (**R3**). PC2 is included on-demand as a secondary axis expanding horizontally from the primary axis; this scatter plot shows item scores for both PC1 and PC2 to inform on the similarity of dimensions included in these components (Fig. 4B). This approach is inspired by the natural orthogonality of the first and second principal components. Our nested plot approach is similar to previously suggested enhancements to parallel coordinates plots [3], [42], [43]. The innermost third level comprises all dimensions contributing to the principal components of the *dimensional bundle* (Fig. 4C) and plots the original item values (**R3, R6**). It resides conceptually within each *dimensional bundle* parallel coordinates axis as a second parallel coordinates plot that is made visible on-

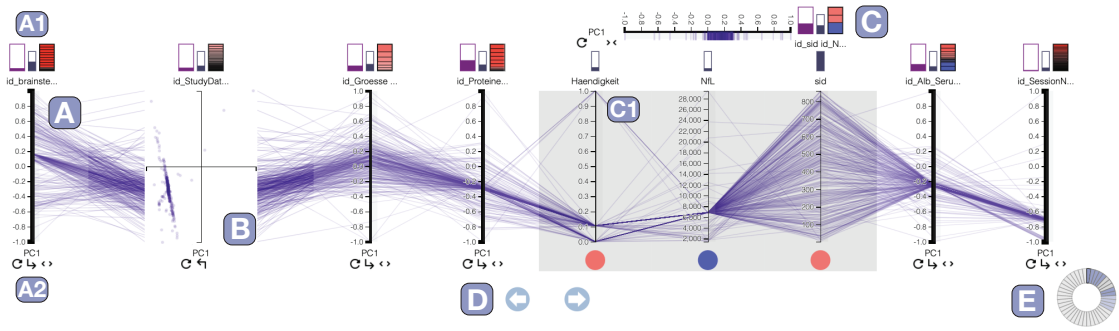

Fig. 4: *DimLift* is a mixed-initiative approach to creating and navigating *dimensional bundles*. They are defined as a subset of similarly contributing dimensions to the overall variation of a dataset, as computed from factor analysis of mixed data. Parallel coordinate axes (A) map to the first or second principal component (PC1, PC2) of a *dimensional bundle*. Glyphs (A1) provide feedback on variance contribution, missingness, and composition. View interactions (A2) allow users to pan (D) through the dataset, swap axes between PC1 and PC2, drill-down into a PC1 vs. PC2 score plot (B), or drill-down further to the *dimensional bundle* component dimensions (C) and their relationships (C1). A chart at the bottom right (E) provides an overview of all *dimensional bundles* and unbundled dimensions, a subset of which are visualized as plot axes.

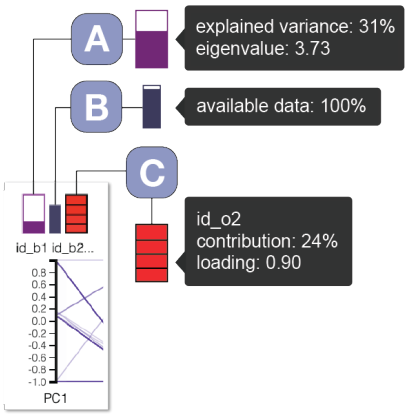

Fig. 5: Rectangular glyphs above each *dimensional bundle* axis provide information on bundle composition. These glyphs, with accompanying tooltips available on hover, display (A) the eigenvalue and explained variance, where height encodes the percent variance while the eigenvalue is revealed in the tooltip, (B) percent available, i.e., non-imputed, items, and (C) contributing dimensions and loadings in a given bundle, where bar height encodes the percent contribution while hue encodes the loading of each dimension.

demand (Fig. 4C1), as inspired by the approach by Andrews et al. [30].

Rectangular filled glyphs, positioned above each axis, provide information on *dimensional bundle* composition and variance contribution (Fig. 5) (**R5, R6**). Our glyph choice is driven by position-based principles from graphical perception research [44]. These glyphs display the eigenvalue and explained variance (Fig. 5A), available, i.e., non-imputed, data (Fig. 5B), and contributing dimensions with their respective loadings (Fig. 5C).

Understanding the explained variance alongside the eigenvalue is a critical aspect of determining the utility of a given bundle in its ability to explain properties of the dataset. Using the Kaiser criterion [35], if an eigenvalue is below 1, we can conclude that the contributing dimensions are more informative when unbundled. The variance contribution gives an indication of the type of relationship between dimensions—a low overall variance may indicate more complex, non-linear relations. We provide this information for each bundle as shown in Fig. 5A (**R5**).

Our approach also explicitly handles missing and imputed data. In particular, the proportion of non-imputed data items can provide feedback on the certainty of the bundles (**R5**). The amount of available, i.e., non-imputed data, is visualized by the glyph shown in Fig. 5B. A bundle containing primarily imputed items, e.g., a mostly white/unfilled glyph, offers less certainty than a fully-filled glyph for a bundle or single dimension. We provide further details on our approach to handling of missing data and imputation in Sec. 3.4.

To draw meaningful, actionable conclusions from an analysis the user must link back to the original data (**R3, R6**). The bundle composition glyph (Fig. 5C) shows the percent contribution and correlation direction, i.e., loading, of each dimension to the bundle. The glyph is broken into segments by each dimension's variance contribution. We encode correlation direction using a diverging red-blue colormap, where red indicates a positive correlation while blue indicates a negative correlation. These encodings are additionally present in the nested dimensions parallel coordinates plot for each bundle (Fig. 4C1) in circles placed under each dimension axis. This indicates to the user the relationship of item values to the principal component, and supports its interpretation. For instance, consider a synthetic dimensional bundle containing height, weight, and BMI: all dimensions are positively correlated and encoded with red at both plot levels. Brushing on any axis would highlight similarly high values in the principal component axis, showing that these values tend to increase and decrease together.

Both quantitative and qualitative data can be involved in relevant and interesting patterns. For instance, our synthetic dataset includes a cardiac risk outcome dimension which is comprised of quantitative body measurement and qualitative lifestyle dimensions (Fig. 3-o2). To help reduce visual clutter and clarify relative occurrences, we utilize a

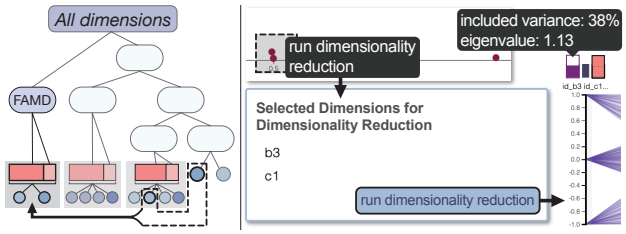

Fig. 6: *DimLift* structural interactions allow for the creation or modification of *dimensional bundles*. Using our synthetic health dataset we **create** a new bundle combining smoking (b3) with gender (c1); a resulting eigenvalue above 1 shows a fair grouping with equal dimension contributions. The left diagram provides a conceptual overview of this process.

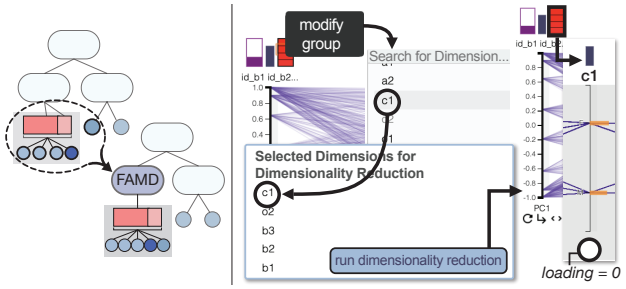

Fig. 7: In the analysis of a synthetic health dataset we may suspect gender (c1), to have interesting correlations with the lifestyle-related bundle, i.e., education level (b1), workout frequency (b2), cardiac risk (o2), and smoking (b3). We **add** gender (c1) to this bundle and observe that gender shows no contribution (loading = 0) to the bundle variance.

horizontal bar chart extending from each categorical parallel coordinate axis (Fig. 1), where bar length encodes the frequency of item occurrence in each category, as inspired by Hauser et al. [42].

3.3 Lifting Expressive Dimensions

A dimensionality reduction process that does not incorporate user interaction may overemphasize trivial aspects of the data. Key to the *DimLift* approach is *lifting*, an operation that changes the data hierarchy and structure of *dimensional bundles* for greater expressitivity. For instance, our synthetic grouping shows cardiac risk as bundled in the automatic process with the lifestyle bundle (Fig. 1). While useful for understanding that cardiac risk is, in our example, more closely correlated with lifestyle dimensions, we would like to lift this, and any other outcome-associated dimensions, to their own bundle for direct correlation assessment. Our method incorporating task-based user interactions allows for the flexibility to explore data at differing levels of granularity, and to reconstitute existing bundles to discover new and unexpected relationships. We divide the interaction techniques for our approach into two classes. For a demonstration of the following interactions, we refer readers to the video included in the supplementary materials.

Structural interactions are operations that alter *dimensional bundles* (Fig. 2E) by combining, adding to, or removing dimensions from these bundles. A linkage between layered parallel coordinates and a dimension scatterplot provides an easy mechanism for bundling interesting dimensions by similar statistical measures.

View interactions are inspired by the model for hierarchical aggregation interaction techniques proposed by Elmqvist & Fekete [32]. These do not change the fundamental structure of the data hierarchy (Fig. 2D), and include: pan, brush & subset, drill-down/roll-up, or swap levels in their exploration of the data space.

3.3.1 Structural Interactions

Dimensional bundles created automatically may not always be conducive to specific user analysis goals. As such, we introduce structural modifications that allow the user to create entirely new, or modify existing, *dimensional bundles* to lift interesting dimensions for analysis.

Creating new *dimensional bundles*. During the analysis a user may wish to visualize the degree that a group of conceptually-related dimensions, e.g., all lifestyle input variables in our synthetic human measurements dataset, are correlated. Similarly, seemingly conceptually-unrelated dimensions may exhibit similar descriptive statistics, e.g., similar mode or diversity measures, that would be interesting to apply dimensionality reduction to for deeper correlation assessment. Figure 6 demonstrates the workflow for creating a new *dimensional bundle* based on similar descriptive statistics, beginning with a marquee selection of a pair of dimensions positioned near each other. The user confirms their selection in the dimension grouping menu. Selection by descriptive statistics serves as a rough guide for the suitability of a possible bundle, which is then validated by applying the dimensionality reduction and visualizing the eigenvalue and contributing dimension attributes in the parallel coordinates plot. On creation, this bundle is briefly highlighted with a red underline in the parallel coordinates plot. When manually creating new bundles, redundant dimensions are not extracted from their original bundle to the new bundle—a single dimension can remain in multiple bundles. The reasoning is that it could be that one dimension is highly important in multiple bundles. For example, smoking (b3) is important semantically as a lifestyle variable and is logically bundled with other lifestyle variables, but it is additionally clinically interesting to bundle with, e.g., gender (c1), to assess for patterns or relationships between these two dimensions. Our approach allows the user to see this from both perspectives.

Modifying *dimensional bundles*. Algorithmically-created *dimensional bundles* may still bury an interesting dimension within, e.g., cardiac risk within a lifestyle bundle, or leave out a conceptually interesting dimension, e.g., gender from the lifestyle bundle. Rather than creating a new bundle, the user may simply modify the existing bundle and either add or remove dimensions in place. We demonstrate the workflow to add a dimension to an existing bundle in Fig. 7; a right-click on the contributing dimensions glyph for the bundle of interest opens the dimension selector panel where bundle membership may be updated. The user may search by name or explore the list to add dimensions. Similarly, dimensions may be removed by entering the same panel (Fig. 1). To remove a dimension, the user clicks on

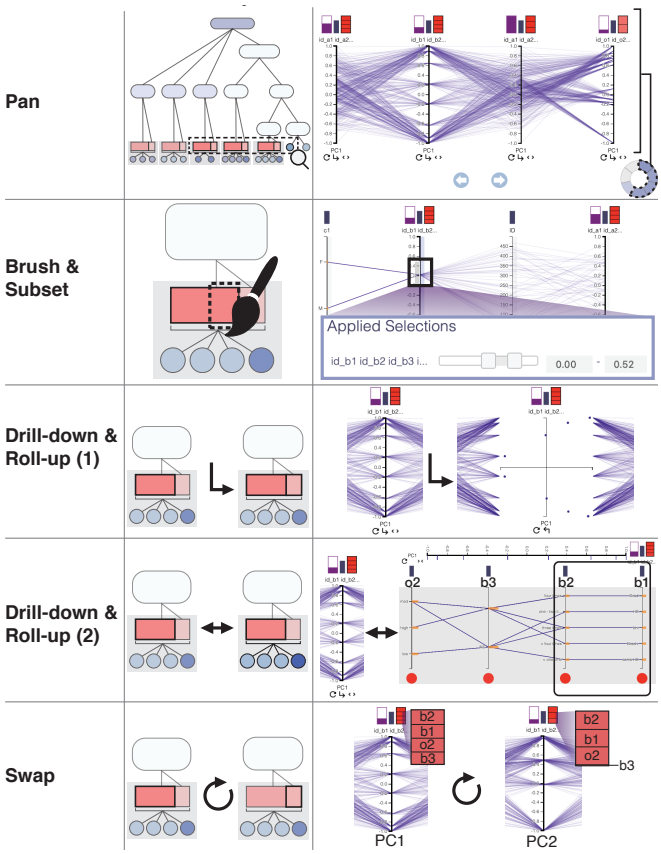

Fig. 8: *DimLift* view interactions allow iterative exploration of *dimensional bundles*. Panning through the parallel coordinates plot allows the user to explore correlations between all bundles. Brushing over a bundle axis, e.g., lifestyle, creates a subset of moderately active, university education level smokers with moderate cardiac risk. This selection is adjustable in an adjacent panel. Drilling down to a plot of PC1 vs. PC2 item scores shows a distribution shape that is interesting to explore further. Drilling further to the contributing individual dimensions shows a definite correlation between b1 (education level) and b2 (workout frequency) (black rectangle). Swapping the bundle axis from the first (PC1) to second (PC2) principal component shows that b3 (smoking) contributes no variation to this component, while it contributes similarly to other included dimensions in PC1.

the dimension in the selected dimension list for immediate removal. After dimension addition or removal, the user can choose to run the dimensionality reduction algorithm on the updated bundle. As with bundle creation, the updated group is highlighted briefly in the parallel coordinates plot. All contribution information is updated in the glyphs, and correlations in the parallel coordinates are updated automatically.

These structural modification tools empower the user to reconstruct the dimensional hierarchy for open exploration. With flexible bundle composition and modification, along with feedback on their suitability in the parallel coordinates plot, users may rapidly form new insights about their data by lifting dimensions of interest from their original bundles.

3.3.2 View Interactions

With the *DimLift* approach users may visually navigate *dimensional bundles* via the previously described layers: the top-level parallel coordinates axes for between-bundle navigation (A), or nested scatterplots (B) and further nested parallel coordinates (C) for within-bundle navigation (Fig. 4).

Pan. Pan operations are ubiquitous in visual analytics, particularly in aggregated datasets [32]. We utilize panning to bring *dimensional bundles* of interest into the field of view, as shown in Fig. 8. Arrow buttons allow incremental panning while a small donut chart, used as it mirrors the panning-carousel nature of the parallel coordinates plot, provides a quick overview of the created bundles and individual dimensions while acting as an additional navigational aid [7]. It additionally serves to spotlight those bundles with subsets applied. Within this glyph, bundles are denoted as purple, while individual dimensions are grey. In addition to these manual controls, panning can be facilitated by axis reordering based on descriptive statistics, i.e., variance, standard deviation (and their qualitative analogs), diversity, modality, and percent missing values [7], or by order of extraction via the iterative FAMD algorithm.

Brush & subset. Brushing and linking are commonly used in visual analytics to link data elements [5] across views. Our approach relies on this premise, but rather than only brushing and linking items or individual dimensions [7], we support brushing and linking of *dimensional bundles*. In our method the user may brush a dimension or dimension bundle in the dimensions overview plot (Fig. 6) or in the layered parallel coordinates plot (Fig. 8). In the latter, brushing creates subsets of *dimensional bundles*, as demonstrated in Fig. 8, which may be subsequently adjusted.

Drill-down and roll-up. Drill-down and roll-up are two primary methods for viewing data at multiple aggregation levels [32]. Since *dimensional bundles* comprise two PCs and raw dimensions, we utilize two different methods to access each data type in a bundle. The first branch explores PCs in a given *dimensional bundle*: with this method, the user may see the orthogonal axis of variation presented by the grouped data axis (Fig. 8); this provides a greater sense of the bundling strength and reasonability.

The second method accesses constituent dimensions of PCs within a given *dimensional bundle*. To differentiate from the first method we drill-down/roll-up on a horizontal axis, e.g., expand/collapse (Fig. 8). Expansion occurs in-place, and allows the user to assess correlations within and outside a given *dimensional bundle*.

Swap. Described as a flip operation by Elmqvist and Fekete, this allows the user to observe neighboring siblings in an aggregate hierarchy [32]. We can consider the first (PC1) and second principal components (PC2) as siblings in our aggregate structure. We view this operation as fundamentally different from drilling-down, as the swap does not add detail to the existing view but rather shifts to a different, related frame. Lifting the secondary axis of variation in the bundle to the surface permits visualization of interdimensional correlations through a different lens (Fig. 8). This

allows prioritization of second-level variation structures in the data to establish subsurface patterns.

3.4 Handling Mixed and Missing Data

Our algorithmic approach as described in Sec. 3.1 may furthermore handle complex data, as characterized by missing items and mixed data types. In the instance of a purely qualitative dataset we perform all steps previously described, with a few alterations: we first use multiple correspondence analysis (MCA) to convert all qualitative dimensions to quantitative dummy variables [45], [46]. Then, rather than the squared correlation coefficient criterion we instead use the squared correlation ratio to identify the PC1 leading to each *dimensional bundle*. In instances of mixed datasets, the algorithm simply uses the appropriate criterion to define each *dimensional bundle*.

Furthermore, data are often incomplete, as was true for one of our case studies which was 76% incomplete. While some solutions drop cases with missing data from the analysis, this can easily lead to an inaccurate picture of dimension correlations. Imputation of missing data is still a highly debated area of research, and is dependent on the analysis goals and the data itself. While our approach is flexibly designed to allow a variety of imputation methods, our default method is multiple imputation of chained equations (MICE) [47], a multiple imputation method, to minimize bias and reduce standard error. This default can be changed by the user. A key feature of MICE is that it can handle different variable types: quantitative continuous, binary, and ordered and ordered categorical data. As we aim our method to be broadly applicable to mixed datasets, this was a critical aspect of our decision process. It furthermore is widely used in epidemiology [48], a field known for its complex and highly missing data, and was chosen after discussions with our clinical collaborators on this paper.

MICE is applied by default to all dimensions with missingness of 78% or lower. We chose this default value experimentally, as this was the limit up to which MICE was typically still able to provide meaningful results. In the extreme case of dimensions with only a handful of total entries where multiple imputation produces meaningless results, e.g., 99% missing, we instead perform a single value imputation using the mean for quantitative variables and create a new "not defined" category for categorical data. The missing data glyphs serve as identifiers for the reliability of the data for patterns observed in these dimensions. We explore the impact of different imputation methods in the discussion section and supplementary material.

4 Case Studies

We implemented our approach as a web application using Javascript and D3.js [49]. Descriptive statistics computations and dimension groupings are performed in a Flask Python back end, and we use FactoMinR [1] to perform the FAMD in R. The full source code is available at *https://github.com/lauragarrison87/DimLift*.

Following initial analysis of the data via our iterative FAMD algorithm and visual aggregation, the user may explore the resulting dimension hierarchy. As part of the exploratory analysis process, users may flexibly adjust membership of *dimensional bundles* and construct a new dimensional hierarchy to lift interesting dimensions to the surface.

In the following, we demonstrate the value and versatility of *DimLift* applied to three data scenarios, one of which corresponds to an ongoing clinical collaboration. Analysis of data in these domains typically utilizes statistical analysis packages which are unwieldy when applied to open-ended data exploration [50]. Before discussing our clinical cohort case study, we introduce nutrient and plant ecology datasets to demonstrate our method's general applicability by reproducing insights from existing domain literature.

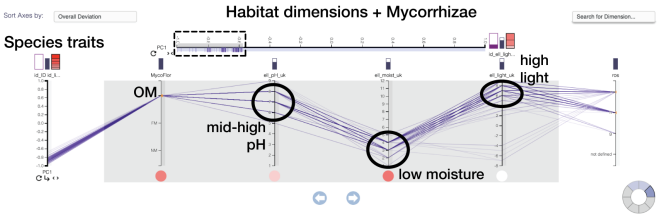

Fig. 9: Analysis of BiolFlor-MycoFlor dataset confirms Hempel et al. findings [51]. An initial *dimensional bundle* contains mycorrhizae, i.e., fungi symbiotically-associated with plant roots, and light preference, showing a clear correlation between these dimensions. We add pH and moisture dimensions and run a FAMD for this bundle. We then brush low scores on the axis to subset only obligate mycorrhizal (OM) i.e., need symbiotic fungi relationships to survive, plant species, and find that these species tend to favor environments with higher soil PH, drier habitat, and more light. This corroborates the study findings.

4.1 Plant Ecology

Plant traits are frequently used in large-scale ecological studies to describe species distribution in plant communities [51]. Mycorrhizae are fungi that form symbiotic associations with roots of certain plant species; these fungi serve a key role in helping ecologists understand plant species characteristics and their distribution. Mycorrhizal plants in this study are classified in three groups: (1) obligate mycorrhizal (OM), i.e., always requiring fungi, (2) facultatively mychorrhizal (FM), i.e., occasionally requiring fungi, and (3) nonobligate mycorrhizal (NM), i.e., never requiring fungi. We pattern our analysis after a study by Hempel et al. [51] which analyzed these relationships in large plant communities through mixed PCA and linear correlation methods. The data for this study are extracted from BiolFlor [52], a database containing biological and ecological information on vascular plants in Germany. We explore mycorrhizal status and plant trait data, totaling 13 dimensions, for 1758 plant species, following the data selection procedure as described by Hempel et al. for habitat characteristics, species traits, and mycorrhizal status. Our goal was to corroborate a subset of study findings relating mycorrhizal status to habitat characteristics and species traits using our approach.

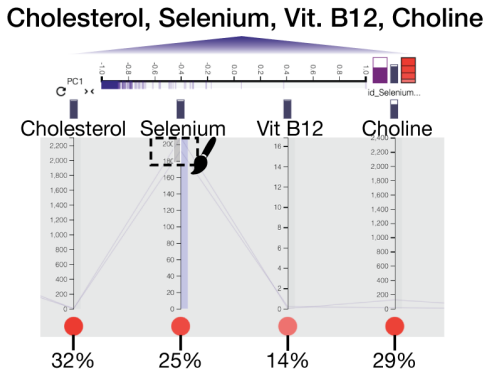

Fig. 10: Analysis of FDA nutrient dataset using our approach. Selenium, vitamin B12, cholesterol, and total choline are bundled together automatically in our approach; this corroborates a known link in clinical literature between Selenium and Cholesterol. Subsetting to only the high values of selenium shows low values of cholesterol, although with values near 0 this finding needs confirmation with a larger dataset.

Hempel et al. [51] demonstrated that OM species tend to be positively associated with higher temperature, drier habitats and higher soil pH; and negatively associated with moist, acidic and fertile soils. We can confirm these positive associations in our method, noting the red contribution glyph bars for the bundle comprised of these dimensions (Fig. 9, top right glyph). Interestingly, by simply brushing a range [–1, –0.6] in PC1 of the bundle we are able to identify this relationship for all dimensions without creating a subset of any individual dimension (Fig. 9, rectangular marquee). By comparing inter-axes correlations and by drilling-down into this habitat bundle, we corroborate their PCAmix finding that FM plant species are associated with differing plant traits and habitat characteristics relative to OM/NM species (Fig. 9). These findings, generated in a very short session, show promise for our approach in quickly lifting and establishing relationships that corroborate results from a complex plant ecology study.

4.2 USDA National Nutrient Data

We next demonstrate insights generated with our method using data from the USDA SR28 National Nutrient Database [53]. This database is the standard reference for food nutritional content in the United States; many of these variables correlate and thus this dataset lends itself well to dimensionality reduction. The subset we analyzed is predominately comprised of quantitative data and consists of 899 data items in 53 dimensions. Selenium is an essential micronutrient for effective thyroid hormone and reproductive function; when levels are sufficient in the body it has been shown to provide antioxidant and anti-inflammatory effects [54]. Cholesterol plays a known role in cardiovascular health; high total cholesterol levels are strongly linked to higher cardiovascular risk [55]. Chen et. al. [56], in their seven-year longitudinal nutritional cohort study, found that participants with higher levels of selenium exhibited a greater decrease in total cholesterol over the course of the study; this offers insights on selenium's possible mitigating effect of cardiovascular risk in elderly populations. Our goal in this exploratory analysis of nutrient data was to establish the possible ease and clarity of discovering this known clinical link.

Our approach rapidly lifts Selenium to the surface, placing it in a *dimensional bundle* alongside Vitamin B12, Cholesterol, and Total Choline; each contribute approximately 25% to the bundle variance (Fig. 10). This immediate insight corroborates the correlation between these nutrients, while also providing an interesting line of inquiry on the additional relatedness of vitamin B12 and Choline. Creating a subset of high selenium values, our results are not as conclusive since we have a small population sample in our dataset, but the results indicate the same link as shown in the clinical literature that we discovered in a short period of exploration. If we then remove VB12 and Choline from the bundle to hone in on the relationship between Selenium and Cholesterol, we find equivalent positive loading values for each. This further supports a positive colinear relationship between these two nutrients. If performed in a standard FAMD this link would have been difficult to identify, as the results would show all dimensions in each principal component. The subtle link between selenium and cholesterol would be buried beneath the stronger variance contributions of other nutrients, e.g., magnesium, folate, and calcium, to the data. Our approach allows this dimensional relationship to be immediately apparent.

4.3 Clinical Cohort: Cerebral Small Vessel Disease

The ultimate analysis goal for any clinical cohort dataset is to develop testable hypotheses that can lead to better treatment options and outcomes for the patient. One of the great difficulties with clinical datasets lies in the successful identification of interesting measures and patterns, particularly in diseases where the etiology is not entirely clear, e.g., in cerebral small vessel disease (CVSD). The current standard for analysing this type of data consists of queries with complex statistical analysis packages. Of such tools, our expert participants most frequently use SPSS. This and similar applications typically require extensive processing times, and are not generally conducive to an iterative, exploratory approach to the data. Having previously analyzed these data in SPSS, one expert noted that for an efficient analysis with SPSS they need to already have in mind the variables of interest and be familiar with the data characteristics prior to their assessment.

The data consist of 307 patients collected from clinical routine in the university hospital data management system. The data are mixed, consisting of 168 dimensions containing demographic, laboratory, education, and lifestyle information. 24 additional dimensions describe the volume of 24 brain structures, e.g., hippocampus and caudate, as derived from T1-weighted magnetic resonance imaging data. As is typical with this type of data, 76% of entries are missing due to, e.g., missed appointments, not all patients needing the same tests, and other criteria.

We performed two joint analyses of a clinical cohort for the study of CSVD. After a short presentation explaining the method and the prototype application, the experts explored the application themselves using a "think-aloud" protocol.

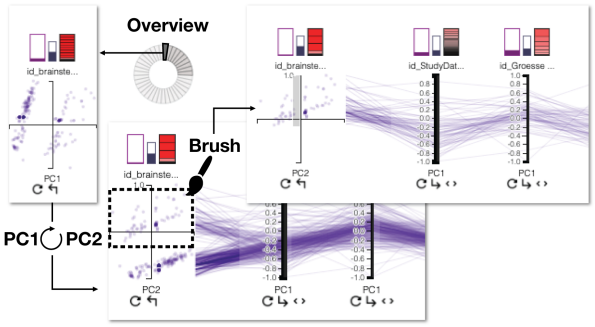

Fig. 11: In exploring a clinical cohort dataset for cerebral small vessel disease (CSVD), experts select a bundle of primarily imaging data for closer examination and drill-down to observe two distinct clusters. Swapping the axis to PC2 allows subset creation of the top cluster; this corresponds to selection of non-imputed items within the bundle. Addition of APOE-related dimensions to the bundle allows for correlation assessment of these interesting dimensions within a single bundle.

Our primary goal with this study was to allow domain experts to freely explore their data in *DimLift* to assess ease and speed of iterating and forming new insights into possible CSVD-related measures and patterns. We highlight key aspects of their respective analyses; for further demonstration we refer the reader to the supplementary video.

Analysis 1. The first analysis was performed alongside one MD/PhD clinical neurologist specializing in cognitive aging and mixed cerebral pathologies, who is also a co-author of this paper, and one master-level engineer in neuroscience in a paired analysis session for one hour and 30 minutes. Both are experts in advanced statistical analysis of CSVD data; their workflow was particularly interested in the bundle contents and loadings. On loading the data, experts first browsed the number and contents of the created *dimensional bundles*. Noting from the glyphs above each axis that many bundles suffer from missing data, the experts used hover features to assess bundle variance and dimension contributions. The experts were surprised that *lacunes* and *microbleedings* were bundled along with two Boston/STRIVE criteria dimensions and thus, lifted together. However, their bundling makes sense since these have been shown to correlate [57]. From this, they can hypothesize that lacunes and microbleeds in certain regions of the brain could be associated with a certain subset of Boston criteria. This has implications on bleeds in certain areas of the brain being indicators for aspects of CSVD. They stated, *"We would have probably not seen this in another framework."*

Locating another interesting bundle containing primarily imaging data, as well as diagnosis and sex, they then drilled-down for further exploration. In this second level they observed two clusters. While variables contributing to PC1 are mostly imaging-related, PC2 contributing variables include Boston/STRIVE criteria at lumbar puncture, group, and sex. They swapped the axes and observed how the item distribution (Fig. 11) and dimension statistical distributions are affected. They noted that brushing the top cluster selects individual dimension values that are complete, i.e., Boston/STRIVE criteria, Sex, and Group axes now exclude "not defined" items through this subset selection. Experts then added APOE-related dimensions to investigate the relationships within this bundle. However, they noted that the APOE dimensions do not provide strong contributions (the loading glyph on the original dimension is white in color)—this implies that these are not particularly correlated, and may indeed be better treated as separate dimensions or *dimensional bundles* (Fig. 11). This allows them to reject their hypothesis that APOE genotypes are highly correlated to Boston/STRIVE criteria at lumbar puncture, group, and sex for this dataset. However, they note that this would be more interesting to explore in a larger cohort before fully rejecting this. They further noted that this subcohort is characterized by generally mid-to-high range white matter and CSF volume values, but a broad range of volume data for other regions. From this, they hypothesized that these volume ranges of white matter and CSF can act as potential biomarkers for CSVD. This requires additional followup with a larger cohort and additional cognitive and clinical tests.

Analysis 2. Our second analysis session also lasted one hour 30 minutes, with a medical expert who has one year of experience in CSVD research and who is less experienced in statistical analysis. This user was primarily interested in generating a picture of the typical patient for each diagnostic group in the dataset. As such, they were less focused on the bundle loadings and differences in principal component loadings for each bundle, and rather interested in the composition and linear correlations within bundles. Their workflow generally went as follows: *(1) Bundle overview, (2) Subset within bundle to explore correlations, (3) Modify bundle contents, and (4) Repeat steps 1-3.*

In their overview of bundles they observed that a high proportion of the data was missing. They noted that this information is helpful because they know the statistics they explore have reduced power in hypothesis generation. In the bundle comprised largely of imaging data they were initially surprised that these were bundled, but reasoned that this was logical since these were tied to Boston/STRIVE criteria, which was also grouped in this bundle. The user then created a subset of CAA/HA/Mixed patients in the Group dimension, hypothesizing that if a patient has CAA that they are more likely to also suffer from seizures, stroke, and dementia in the pathology dimension. Direct correlation visualization between these parallel coordinate dimension axes allowed them to confirm this; such a finding also corroborates clinical literature findings. However, they stated that this would need to be verified in a larger and more complete clinical dataset.

On exploring other dimensions in the same bundle, the user indicated that some dimensions were not, in fact, particularly interesting to analyze, e.g., all of the imaging dimensions except for the hippocampal and white matter volume measurements. The ability to easily modify this bundle to remove these uninteresting dimensions for their current hypothesis was extremely helpful for them. In doing so, they were able to quickly note a slight positive correlation between these two dimensions that also related to the diagnostic group subset; this allowed formation of a second

hypothesis: that hippocampal and white matter volumes correlate to this group of diagnoses in CSVD.

Having previously analyzed the data in SPSS, they noted that these preliminary trends and relationships they found interesting with over one hour of using SPSS could be found in five minutes using our approach, simply by drilling down into a bundle that contained already most of the dimensions of interest and subsetting to a certain diagnostic group, i.e., the CAA/HA/Mixed category.

Expert Feedback. Although clinical experts noted that our visual approach appears very complex in the beginning, they were able to operate it independently after ten minutes. They stated its usability to be very intuitive due to the visualizations, hovering facilities, and interactions. However, they felt that dimensions with freeform or prose text entry were not ideal for exploratory analysis with this tool. Not only is the variance on these dimensions massive because there are no defined categories (every entry can be unique), but readability may be problematic in the parallel coordinates view. They felt also that tracking patients and variable changes over a longitudinal study would not be easy with this system, although they felt this approach serves a different purpose. Experts felt that SPSS provided means for a more direct and targeted analysis method, while DimLift takes a more open, discovery-oriented approach. As such, DimLift may be unnecessary to use if one has already identified target variables and wishes to perform specific statistical analyses of significance. However, they felt that for open exploration DimLift is a faster (e.g., Analysis 2 required 1.5 hours in SPSS compared to ten minutes in DimLift to identify a new hypothesis) and easier-to-use solution with visual aids that are neither readily or easily available with SPSS and R. To use the DimLift approach to its full potential experts agreed it is important to have a basic knowledge of statistics and dimensionality reduction techniques, otherwise the rationale for the algorithmic bundles may be difficult to appreciate. This level of statistical knowledge is common in clinical research. With our approach to high dimension space exploration and modification, all three experts were able to rapidly gain new insights into the data via the *dimensional bundles*, and to easily reflect the principal components back to the original dimensions. They stated that this tool is especially helpful for hypothesis generation, and recommended its usage within clinical research.

5 Discussion

Although we explored a number of possible algorithms to drive our technique, we ultimately chose factor analysis of mixed data as it is quite general and allows the analysis of mixed data by combining PCA and MCA. The broad applicability of this algorithm makes it a clear first choice for exploring this type of hierarchical creation for our visualization. However, this comes with an expectation for normally-distributed data, which is not always the case. An interesting avenue for future investigation is how our approach could be integrated with nonlinear dimensionality reduction techniques, although this presents other challenges in mapping back to the original dimensions.

Our algorithmic approach furthermore treats all dimensions as active, i.e., all dimensions are used in FAMD, and excludes the possibility for supplementary dimensions, i.e., dimensions that are not used directly in FAMD. While supplementary dimensions do not impact the eigenvalue of a bundle or dimension contributions, these can provide further insights by distinguishing correlations between active vs. supplementary dimensions. This is particularly interesting to explore further in user-driven bundle creation.

Our treatment of *dimensional bundle* labeling concatenates the names of all input dimensions, producing long names which are not fully visible at a glance. The ordering does not mirror the contributions of the dimensions, since we preserve the label for the first and second principal component to avoid confusion. Descriptive labeling of the new dimensions produced by dimensionality reduction poses an ongoing challenge in the community, and our approach could benefit from more advanced solutions.

Handling of missing data is an active area of research, and imputation methods are highly dependent on the particulars of the dataset. Our exploration of imputation methods involved a literature search, discussion with clinical collaborators, and testing of four selected imputation methods in our clinical cohort dataset, chosen as the test imputation dataset for its high proportion of missingness. The imputation methods we tested included overall mean imputation, hot deck imputation, principal components method, using the missMDA R package [46], and multiple imputation of chained equations. In our tests we found that each imputation method created 10-11 bundles, with big trends or correlations generally preserved between each method, i.e., lacunes and microbleeds from various regions of the brain were mostly bundled together. Although naturally some differences were present between each method, the differences and bundling for these generally followed an outcome that made conceptual sense. Although we ultimately chose MICE as our default imputation method for its popularity in epidemiology studies, which are known for their complexity with mixed data types and missing elements [48], we have available as options the ability for the user to switch to any other imputation method as necessitated by the characteristics of their data and their analysis goals. We document in supplementary material details of our testing of these different methods. A strong benefit that we found in this exploration of the effects of imputation methods on our bundling is that it allowed us to discover more robust patterns within data we analyzed. The exploration of different types of imputation presents an exciting and challenging topic of research in visual analytics.

An additional challenge in missing data imputation is that there is not an established threshold of missingness in literature for which statistical analyses become no longer relevant [58]; this instead is highly dependent on the data itself. This problem was particularly relevant to our clinical case study, which in several dimensions were only 3% complete. We experimented with threshold settings for degree of completeness for each dimension, and used this threshold to determine whether we applied MICE or a more simple single-imputation method.

Currently, data preprocessing uses the existing implementation of FAMD in R, which provides adequate performance for moderately-sized datasets. We include Table 1, which lists the case study dataset number of items, dimen-

TABLE 1: Processing times required (MacBook Pro quad-core i5 processor) for three datasets using our approach.

Dataset	Items	Dimensions	Processing time (sec)
Plant	1758	13	132.4s
Nutrient	899	53	24.5s
Clinical	307	193	6.8s

sions, and processing time. We tested all cases on a MacBook Pro quad-core i5 processor. As we can see in Table 1, processing time is more sensitive to the number of items, rather than the number of dimensions. In order to more efficiently process high item datasets, a more optimized custom implementation would be beneficial. However, wide and shallow datasets, i.e., low item but high dimensionality, are processed relatively quickly.

Parallel coordinates as the base design for bundles may begin to suffer with a very high number of independent, uncorrelated dimensions in a dataset, as this would introduce a high number of axes that would then be prone to issues already known with visualization in parallel coordinates. Although we have explored the utility of our approach in datasets numbering into hundreds of dimensions, as befitting the data of interest for our clinical partners, we imagine a future work exploring the extensibility of this approach to even higher dimensional data. User interactions to structure their own bundles by conceptual relatedness paired with the described view interactions may still mitigate this dimensionality challenge; through a relatively small feature set we allow a comprehensive analysis of the structure of the data with enough flexibility to explore and generate new hypotheses from this starting point. More complex interaction facilities that could perform a combination of steps in one would save the user time, but then run the risk of losing track of the semantics for the user to fully understand the consequences of their adjustments in the visualization.

6 Conclusions

We presented *DimLift*, a novel approach to creating and interacting with *dimensional bundles* that lifts interesting relationships to the user's attention. While prior approaches allow exploration of data in both item and dimension space, *dimensional bundles* provide an additional layer that reduces the analysis space in an expressive manner. Our method is driven by an iterative factor analysis of mixed data (FAMD) that produces expressive subsets of dimensions contributing similarly to the overall variance of a dataset. We provide a means to more transparently link data inputs and track transformations of *dimensional bundles* during the exploratory process through visual and interaction design elements grounded in a layered parallel coordinates plot. Through these interactions, expert users are able to explore possible dimensions of interest in the context of the structural hierarchy, and then proceed to dismantle and rebuild this hierarchy through different views and levels in the hypothesis generation process to meet their own hypotheses for bundle expressivity. We demonstrated our workflow in a study of ecological and nutrient data and in a paired clinical case study with medical experts. With each of these cases we were able to both corroborate existing findings, and establish new insights.

While statistics remains a necessary tool in high dimensional data analysis, statistical strength cannot itself dictate feature importance. User knowledge and semantics remain critical elements of this process. Furthermore, we draw a distinction between *analysis* and *exploration*. While analysis requires specific questions to leverage statistical techniques, exploration proceeds and utilizes statistics in a stepwise fashion, allowing the user to disregard irrelevant information and lift relevant items to the surface.

Dimensional bundles are a useful concept for interacting with high dimensional data. This opens the door for a number of areas of future research, including their possible connections to edge bundling for graph and network data visualization, as described by Holten and Van Wijk [59]. While *DimLift* primarily focuses on formation and interactions with *dimensional bundles*, a logical next step may allow for selections made in dimension subspace to additionally drive dimensional bundle formation. Similar to lifting of interesting dimensions, this could allow for interesting subspaces to be lifted to a primary view level. Other areas of future work may focus on evaluating *dimensional bundles* in other domains as a field study. This may bring further insights on the broad utility of such bundles in subspace exploration. We may also explore this in a controlled study uniquely adapted for exploratory tasks, although by nature this is quite challenging. Additional areas of future exploration includes the applicability of our approach in non-linear dimensionality reduction methods, where the interpretation gap presents a major hurdle to understanding the data, as well as integration of *dimensional bundles* with other interaction and visualization techniques.

Acknowledgments

We thank Frank Schreiber and Philipp Ulbrich for their invaluable feedback in our clinical case study. Parts of this work have been done in the context of the Center for Data Science (CEDAS) at the University of Bergen. This research is supported by the University of Bergen and the Trond Mohn Foundation in Bergen (#813558, Visualizing Data Science for Large Scale Hypothesis Management in Imaging Biomarker Discovery (VIDI)), and by the Federal State of Saxony-Anhalt, Germany (FKZ: I 88).

References

[1] S. Lê, J. Josse, F. Husson *et al.*, "Factominer: an R package for multivariate analysis," *Journal of Statistical Software*, vol. 25, no. 1, pp. 1–18, 2008.

[2] H. Kriegel, P. Kröger, and A. Zimek, "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 3, no. 1, pp. 1–58, 2009.

[3] M. L. Huang, T.-H. Huang, and X. Zhang, "A novel virtual node approach for interactive visual analytics of big datasets in parallel coordinates," *Future Generation Computer Systems*, vol. 55, pp. 510–523, 2016.

[4] S. Liu, D. Maljovec, B. Wang, P. Bremer, and V. Pascucci, "Visualizing high-dimensional data: Advances in the past decade," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 3, pp. 1249–1268, 2017.

[5] C. Turkay, P. Filzmoser, and H. Hauser, "Brushing dimensions—a dual visual analysis model for high-dimensional data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2591–2599, 2011.

[6] C. Turkay, A. Lundervold, A. Lundervold, and H. Hauser, "Hypothesis generation by interactive visual exploration of heterogeneous medical data," in *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, vol. 7947, 2013, pp. 1–12.

[7] J. Müller, L. Garrison, S. Schreiber, S. Bruckner, H. Hauser, and S. Oeltze-Jaffra, "Integrated dual analysis of quantitative and qualitative high-dimensional data," 2021.

[8] M. Dowling, J. Wenskovitch, J. Fry, L. House, and C. North, "Sirius: Dual, symmetric, interactive dimension reductions," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 172–182, 2018.

[9] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim, "Visual interaction with dimensionality reduction: A structured literature analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 241–250, 2017.

[10] A. Tatu, F. Maaß, I. Färber, E. Bertini, T. Schreck, T. Seidl, and D. Keim, "Subspace search and visualization to make sense of alternative clusterings in high-dimensional data," in *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2012, pp. 63–72.

[11] S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Möller, "Dimstiller: Workflows for dimensional analysis and reduction," in *2010 IEEE Symposium on Visual Analytics Science and Technology*, 2010, pp. 3–10.

[12] X. Yuan, D. Ren, Z. Wang, and C. Guo, "Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2625–2633, 2013.

[13] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[14] W. Müller, T. Nocke, and H. Schumann, "Enhancing the visualization process with principal component analysis to support the exploration of trends," in *Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation - Volume 60*, 2006, pp. 121–130.

[15] D. H. Jeong, C. Ziemkiewicz, B. Fisher, W. Ribarsky, and R. Chang, "iPCA: An interactive system for PCA-based visual analytics," *Computer Graphics Forum*, vol. 28, no. 3, pp. 767–774, 2009.

[16] J. Heinrich and D. Weiskopf, "State of the art of parallel coordinates," in *Eurographics 2013 - State of the Art Reports*, 2013, pp. 96–116.

[17] Y. H. Fua, M. Ward, and E. Rundensteiner, "Navigating hierarchies with structure-based brushes," in *Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis'99)*, 1999, pp. 58–64.

[18] Y. Fua, M. Ward, and E. Rundensteiner, "Hierarchical parallel coordinates for exploration of large datasets," in *Proceedings Visualization '99 (Cat. No.99CB37067)*, 1999, pp. 43–508.

[19] A. O. Artero, M. C. F. de Oliveira, and H. Levkowitz, "Uncovering clusters in crowded parallel coordinates visualizations," in *Proceedings of the IEEE Symposium on Information Visualization*, 2004, pp. 81–88.

[20] T. Van Long and L. Linsen, "Multiclustertree: interactive visual exploration of hierarchical clusters in multidimensional multivariate data," *Computer Graphics Forum*, vol. 28, no. 3, pp. 823–830, 2009.

[21] K. Candan, L. D. Caro, and M. L. Sapino, "PhC: Multiresolution visualization and exploration of text corpora with parallel hierarchical coordinates," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 2, p. 22, 2012.

[22] G. Richer, J. Sansen, F. Lalanne, D. Auber, and R. Bourqui, "Enabling hierarchical exploration for large-scale multidimensional data with abstract parallel coordinates," in *International Workshop on Big Data Visual Exploration and Analytics 2018*, vol. 2083, 2018, pp. 76–83.

[23] Z. Vosough, M. Hogräfer, L. A. Royer, R. Groh, and H. J. Schulz, "Parallel hierarchies: A visualization for cross-tabulating hierarchical categories," *Computers & Graphics*, vol. 76, pp. 1–17, 2018.

[24] X. Z. Wang, S. Medasani, F. Marhoon, and H. Albazzaz, "Multidimensional visualization of principal component scores for process historical data analysis," *Industrial & Engineering Chemistry Research*, vol. 43, no. 22, pp. 7036–7048, 2004.

[25] R. Dunia, T. F. Edgar, and M. Nixon, "Process monitoring using principal components in parallel coordinates," *AIChE Journal*, vol. 59, no. 2, pp. 445–456, 2013.

[26] J. Yang, W. Peng, M. Ward, and E. Rundensteiner, "Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets," in *IEEE Symposium on Information Visualization 2003*, 2003, pp. 105–112.

[27] J. Yang, M. Ward, and E. Rundensteiner, "InterRing: an interactive tool for visually navigating and manipulating hierarchical structures," in *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002.*, 2002, pp. 77–84.

[28] D. K. I. Weidele, "Conditional parallel coordinates," in *2019 IEEE Visualization Conference (VIS)*, 2019, pp. 221–225.

[29] D. Brodbeck and L. Girardin, "Visualization of large-scale customer satisfaction surveys using a parallel coordinate tree," in *IEEE Symposium on Information Visualization 2003 (IEEE Cat. No. 03TH8714)*, 2003, pp. 197–201.

[30] K. Andrews, M. Osmić, and G. Schagerl, "Aggregated parallel coordinates: integrating hierarchical dimensions into parallel coordinates visualisations," in *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*, 2015, pp. 1–4.

[31] M. Sedlmair, M. Brehmer, S. Ingram, and T. Munzner, "Dimensionality reduction in the wild : Gaps and guidance," 2012.

[32] N. Elmqvist and J. Fekete, "Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 3, pp. 439–454, 2009.

[33] J. Pagès, "Analyse factorielle de données mixtes," *Revue de Statistique Appliquée*, vol. 52, no. 4, pp. 93–111, 2004.

[34] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[35] K. A. Yeomans and P. A. Golder, "The guttman-kaiser criterion as a predictor of the number of common factors," *The Statistician*, pp. 221–229, 1982.

[36] G. Kindlmann and C. Scheidegger, "An algebraic process for visualization design," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2181–2190, 2014.

[37] X. Kuang, H. Zhang, S. Zhao, and M. McGuffin, "Tracing tuples across dimensions: A comparison of scatterplots and parallel coordinate plots," *Computer Graphics Forum*, vol. 31, no. 3pt4, pp. 1365–1374, 2012.

[38] R. Netzel, J. Vuong, U. Engelke, S. O'Donoghue, D. Weiskopf, and J. Heinrich, "Comparative eye-tracking evaluation of scatterplots and parallel coordinates," *Visual Informatics*, vol. 1, no. 2, pp. 118–131, 2017.

[39] G. Kaur and B. B. Karki, "Bifocal parallel coordinates plot for multivariate data visualization," in *VISIGRAPP (3: IVAPP)*, 2018, pp. 176–183.

[40] G. G. Robertson, J. D. Mackinlay, and S. Card, "The perspective wall: Detail and context smoothly integrated," in *Proceedings of ACM CHI*, vol. 91, 1991, pp. 173–179.

[41] Z. Liu and J. Stasko, "Mental models, visual reasoning and interaction in information visualization: A top-down perspective," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 999–1008, 2010.

[42] H. Hauser, F. Ledermann, and H. Doleisch, "Angular brushing of extended parallel coordinates," in *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002.*, 2002, pp. 127–130.

[43] H. Janetzko, M. Stein, D. Sacha, and T. Schreck, "Enhancing parallel coordinates: Statistical visualizations for analyzing soccer data," in *IST Electronic Imaging Conference on Visualization and Data Analysis*, 2016.

[44] W. S. Cleveland and R. McGill, "Graphical perception: Theory, experimentation, and application to the development of graphical methods," *Journal of the American statistical association*, vol. 79, no. 387, pp. 531–554, 1984.

[45] G. Saporta, "Simultaneous analysis of qualitative and quantitative data," in *XXXV RIUNIONE SCIENTIFICA Societa Italiana di Statistica*, vol. 1. CEDAM-CASA EDITRICE, MILANO, 1990, pp. 62–72.

[46] V. Audigier, F. Husson, and J. Josse, "A principal component method to impute missing values for mixed data," *Advances in Data Analysis and Classification*, vol. 10, no. 1, pp. 5–26, 2016.

[47] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: issues and guidance for practice," *Statistics in medicine*, vol. 30, no. 4, pp. 377–399, 2011.

[48] A. R. T. Donders, G. J. Van Der Heijden, T. Stijnen, and K. G. Moons, "A gentle introduction to imputation of missing values," *Journal of clinical epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006.

[49] M. Bostock, V. Ogievetsky, and J. Heer, "D3 data-driven documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.

[50] S. B. Green and N. J. Salkind, *Using SPSS for Windows and Macintosh, books a la carte*. Pearson, 2016.

[51] S. Hempel, L. Götzenberger, I. Kühn, S. G. Michalski, M. C. Rillig, M. Zobel, and M. Moora, "Mycorrhizas in the central european flora: relationships with plant life history traits and ecology," *Ecology*, vol. 94, no. 6, pp. 1389–1399, 2013.

[52] I. Kühn, W. Durka, and S. Klotz, "BiolFlor: a new plant-trait database as a tool for plant invasion ecology," *Diversity and Distributions*, vol. 10, no. 5/6, pp. 363–365, 2004.

[53] A. R. S. US Department of Agriculture, "USDA national nutrient database for standard reference, release 28," http://www.ars.usda.gov/nea/bhnrc/mafcl, may 2016.

[54] M. P. Rayman, "Selenium and human health," *The Lancet*, vol. 379, no. 9822, pp. 1256–1268, 2012.

[55] H. C. Stary, A. B. Chandler, S. Glagov, J. R. Guyton, W. Insull Jr, M. E. Rosenfeld, S. A. Schaffer, C. J. Schwartz, W. D. Wagner, and R. W. Wissler, "A definition of initial, fatty streak, and intermediate lesions of atherosclerosis. a report from the committee on vascular lesions of the council on arteriosclerosis, american heart association." *Circulation*, vol. 89, no. 5, pp. 2462–2478, 1994.

[56] C. Chen, Y. Jin, F. W. Unverzagt, Y. Cheng, A. M. Hake, C. Liang, F. Ma, L. Su, J. Liu, J. Bian *et al.*, "The association between selenium and lipid levels: a longitudinal study in rural elderly chinese," *Archives of Gerontology and Geriatrics*, vol. 60, no. 1, pp. 147–152, 2015.

[57] M. Pasi, G. Boulouis, P. Fotiadis, E. Auriel, A. Charidimou, K. Haley, A. Ayres, K. M. Schwab, J. N. Goldstein, J. Rosand *et al.*, "Distribution of lacunes in cerebral amyloid angiopathy and hypertensive small vessel disease," *Neurology*, vol. 88, no. 23, pp. 2162–2168, 2017.

[58] Y. Dong and C.-Y. J. Peng, "Principled missing data methods for researchers," *SpringerPlus*, vol. 2, no. 1, p. 222, 2013.

[59] D. Holten and J. J. Van Wijk, "Force-directed edge bundling for graph visualization," in *Computer graphics forum*, vol. 28, no. 3, 2009, pp. 983–990.

Laura Garrison joined the Visualization Research Group in the Department of Informatics at the Univ. of Bergen, Norway as a doctoral researcher in 2018. She received her M.Sc. in biomedical visualization in 2012 from the Univ. of Illinois. Her research focuses medical visualization and visual analytics, drawing from her background as a medical illustrator.

Juliane Müller joined the *Medicine and Digitalization* (MedDigit) group at the Department of Neurology, Univ. of Magdeburg, Germany as a doctoral researcher in 2018. She received her M.Sc. in Computer Science in 2016 from TU Braunschweig. Her research interests include information visualization, the visual analysis of medical data, and visual explainability of model-based clinical decision support.

Stefanie Schreiber is a professor at Department of Neurology, Univ. of Magdeburg, Germany since 2018. In 2014, she received a habilitation (venia legendi) in Neurology and in 2007 a Ph.D. in Medicine from Univ. of Magdeburg. Her research interests include Cerebral Small Vessel Disease (CSVD) and Neuromuscular Disorders.

Steffen Oeltze-Jafra heads the working group *Medicine and Digitalization* (MedDigit) at the Department of Neurology, Univ. of Magdeburg, Germany. From 2016 to 2018, Steffen was Group Leader at the Innovation Center Computer Assisted Surgery (ICCAS), Univ. of Leipzig, Germany. In 2016, he received a habilitation (venia legendi) in Computer Science, in 2010 a Ph.D. in Computer Science, and in 2004, a diploma in Computational Visualistics from Univ. of Magdeburg. His research interests are in the quantitative analysis of clinical routine data, the visual analysis of medical and biological data and in model-based clinical decision support.

Helwig Hauser received his MSc (1995) and Ph.D. (1998) in Computer Science from TU Wien, Austria. From 1994–2000 he worked as an assistant professor at the Institute of Computer Graphics at TU Wien, before joining the VRVis Research Center (Vienna, Austria), becoming scientific director in 2003. In 2004, he was entitled a "Privatdozent" at TU Wien after his successful Habilitation. Since 2007, he a is professor in Visualization at the Department of Informatics of the Univ. of Bergen, Norway. His interests include interactive visual analysis, illustrative visualization, and the combination of scientific and information visualization, with particular focus in the application of visualization to the fields of medicine, geoscience, climatology, biology, engineering, and others.

Stefan Bruckner is a full professor in Visualization at the Department of Informatics of the Univ. of Bergen, Norway. He received his master's degree (2004) and Ph.D. (2008), both in Computer Science, from the TU Wien, Austria, and was awarded the habilitation (venia docendi) in Practical Computer Science in 2012. Before his appointment in Bergen in 2013, he was an assistant professor at the Institute of Computer Graphics and Algorithms of the TU Wien. His research interests include all aspects of data visualization, with a particular focus on interactive techniques for the exploration and analysis of spatial data.

Supplementary Material: *DimLift* imputation methods for missing data

Laura Garrison, Juliane Müller, Stefanie Schreiber, Steffen Oeltze-Jafra, Helwig Hauser, Stefan Bruckner

To understand the effects of different imputation methods on our *DimLift* approach, we performed a small-scale study of four different imputation methods, which we narrowed down from a review of the state-of-the-art of data imputation [5, 3] and discussions with our clinical collaborators. Each of these imputation methods are implemented and available for for use based on the unique characteristics of a given dataset. These methods include:

- Overall mean value for quantitative data/"not defined" for qualitative data [5, 4]
- Hot-deck imputation [1]
- Multiple imputation of chained equations (MICE) [6]
- Principal components imputation [2]

Although we initially investigated cold-deck imputation [1] as well, this was not possible with our clinical dataset, as there was no suitable dataset which we had access to for comparison. We additionally note that our goal for this was not to identify the most accurate or best method of imputation–this determination is not possible without a ground truth, and for the clinical dataset we tested this was unavailable. This is reflective of typical imputation challenges in the wild. Our general goal was to establish the degree to which our bundling approach is preserved in spite of different imputation methods, and to explore the semantic relevance of differences we observed in the bundling.

We show the top-level bundle results from each imputation method in Fig. 1; these are sorted by their order of extraction from the main pool of dimensions. We can see that each method produces 10-11 bundles; the hot deck and MICE methods produced 10 bundles, while overall mean and principal components methods produced 11 bundles. The bundles generally maintain the broad semantic themes of tests, lifestyle, and measurements; the core difference in the imputation methods reside largely in the granularity of their bundling. For example, the first extracted bundle is generally consistent between imputation methods; it is comprised mostly of lacune and microbleed measurement dimensions. This bundle shows only 17% complete data, so demonstrates a solid use case for the robustness of our bundling with differing imputation methods. As an example of the difference in bundling with different imputation methods, we observed that the classification of smoker/nonsmoker is bundled differently in each. However, it is interesting to note that each of these bundles still make semantic sense; this dimension has myriad effects on other dimensions, and in each bundle its relations to companion dimensions present interesting lines of further inquiry depending on the specific questions the analyst develops in their exploration. For example, MICE and principal components imputation are quite similar, and include smoking bundled with alcohol.

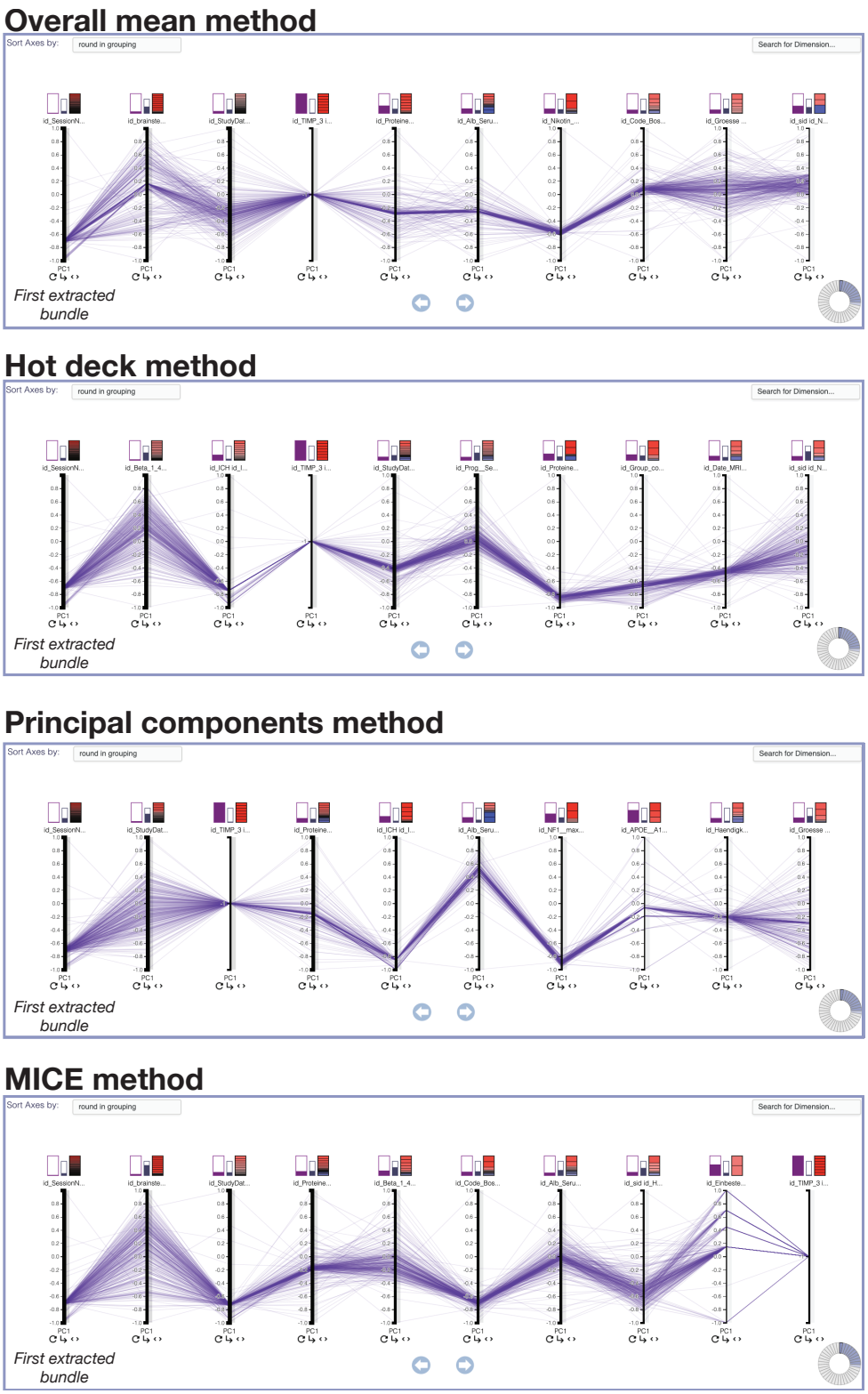

Figure 1: Visual output of top-level *dimensional bundles* produced with each of the four imputation methods tested: overall mean, hot deck, principal components, and MICE.

However, hot deck imputation leads to a bundling of smoking and alcohol along with others including education level, gender, and blood pressure/cholesterol level measurements. In overall mean imputation, smoking is bundled with EPVS, which presents an interesting connection our clinical collaborators noted for further inquiry.

In our qualitative assessment of the effects of imputation methods on our approach, we have found that in general the primary patterns of bundles are preserved; nuanced differences are apparent in each approach, but on inspection of these we found that they made sense semantically. Although no imputation method is ever entirely ideal and its utility is highly dependent both on the specifics of the dataset and the goals of the analyst, our findings indicate that our approach can flexibly accommodate and visualize the bundle results of different imputation methods with a degree of robustness. This presents an exciting area of further inquiry in exploring the results of imputation methods in visual analytics research.

References

[1] Rebecca R Andridge and Roderick JA Little. "A review of hot deck imputation for survey non-response". In: *International statistical review* 78.1 (2010), pp. 40–64. DOI: 10.1111/j.1751-5823.2010.00103.x.

[2] V. Audigier, F. Husson, and J. Josse. "A principal component method to impute missing values for mixed data". In: *Advances in Data Analysis and Classification* 10.1 (2016), pp. 5–26. DOI: 10.1007/s11634-014-0195-1.

[3] A Rogier T Donders, Geert JMG Van Der Heijden, Theo Stijnen, and Karel GM Moons. "A gentle introduction to imputation of missing values". In: *Journal of clinical epidemiology* 59.10 (2006), pp. 1087–1091. DOI: 10.1016/j.jclinepi.2006.01.014.

[4] J. Pagès. "Analyse factorielle de données mixtes". In: *Revue de Statistique Appliquée* 52.4 (2004), pp. 93–111.

[5] Joseph L Schafer and John W Graham. "Missing data: our view of the state of the art." In: *Psychological methods* 7.2 (2002), p. 147. DOI: 10.1037/1082-989X.7.2.147.

[6] I. R. White, P. Royston, and A. M. Wood. "Multiple imputation using chained equations: issues and guidance for practice". In: *Statistics in medicine* 30.4 (2011), pp. 377–399. DOI: 10.1002/sim.4067.