1 INTRODUCTION

High-dimensional datasets are becoming increasingly common in many application fields. Spectral imaging studies in biology and astronomy, omics data analysis in bioinformatics, or cohort studies of large groups of patients are some examples where analysts have to deal with datasets with a large number of dimensions. It is not even uncommon that such datasets have more dimensions than data items, which generally makes the application of standard methods from statistics substantially difficult (i.e., the “p >> n problem”). Most of the available analysis approaches are tailored for multidimensional datasets that consist of multiple, but not really a large number of dimensions and they easily fail to provide reliable and interpretable results when the dimension count is in the thousands or even hundreds [1].

In addition to the challenge that is posed by a truly large number of dimensions, it is often the case that dimensions have properties and relations that lead to structures between the dimensions. These structures make the space of dimensions heterogeneous and can have different causes. Dimensions can have difficult-to-relate scales of measure, such as categorical, discrete and continuous. Some can be replicates of other dimensions or encode exactly the same information acquired using a different method. There can be explicit relations in-between the dimensions that are known a priori by the expert. Some of these relations are likely to be represented as meta-data already. Very importantly also, there are usually inherent structures between the dimensions that could be discovered with the help of computational and visual analysis, e.g., correlation relations or common distributions types. Standard methods from data mining or statistics do not consider any known heterogeneity within the space of dimensions — while this might be appropriate for certain cases, where the data dimensions actually are homogeneous, it is obvious that not considering an actually present heterogeneity must lead to analysis results of limited quality.

A natural approach to understanding high-dimensional datasets is to use multivariate statistical analysis methods. These tools provide the analyst with the most essential measures that help with the extraction of information from such datasets. However, a major challenge with these tools is that their results are likely to become inefficient and unreliable when the dimension count gets substantially large [32]. Take, for instance, principal component analysis (PCA), i.e., a method that is a widely used for dimension reduction [21]. If we apply PCA to a dataset with, for example, 300 dimensions, understanding the resulting principal components is a big challenge, even for the most experienced analysts.

Exactly at this point, the exploitation of any known structure between the dimensions can help the analyst to make a more reliable and interpretable analysis. With an interactive visual exploration and analysis of these structures, the analyst can make informed selections of subgroups of dimensions. These groups provide sub-domains where the computational analysis can be done locally. The outcomes of such local analyses can then be merged and provide a better overall understanding of the high-dimensional dataset. Such an approach is very much in line with the goal of visual analytics [25], where the analyst makes decisions with the support of interactive visual analysis methods.

In this paper, we present an approach that enables a structure-aware analysis of high-dimensional datasets. We introduce the interactive visual identification of representative factors as a method to consider these structures for the interactive visual analysis of high-dimensional datasets. Our method is based on generating a manageable number of representative factors, or just factors, where each represents a subgroup of dimensions. These factors are then analyzed iteratively and together with the original dimensions. At each iteration, factors are refined or generated to provide a better representation of the relations between the dimensions.

To establish a solid basis for our method, we borrow ideas from factor analysis in statistics and feature selection in machine learning. Factor analysis aims at determining factors, representing groups of dimensions that are highly interrelated (correlated) [15]. These factors are assumed to be high-level structures of dimensions, which are not directly measurable. Similar to our motivation of an analysis of the structures in the dimensions space, factor analysis also assumes that there are inherent relations between the dimensions. However, factor...
analysis operates solely on the correlation relation between the dimen-
sions and does not allow the analyst to incorporate a priori information
on the structures. Moreover, similar to the other multivariate analysis
tools, the resulting factors become harder to interpret as the variable
count gets large [15]. A second inspiration for our approach are the
feature subset selection techniques, where variables (dimensions) are
ordered and grouped according to their relevance and usefulness to the
analysis [14]. Similarly, we interactively explore the set of dimensions
to extract sub-groups that are relevant for the generation of factors in
our method.

In order to visually analyze dimensions through the generation of
factors, we make use of visualizations where the dimensions are the
main visual entities. We analyze the generated factors together with
the original dimensions and make them a seamless part of the analy-
sis. Due to the iterative nature of our analysis pipeline, a number of
factors can be generated and refined as results of individual iterations.
We present techniques to compare and evaluate these factors in the
course of the analysis. Our factor generation mechanism can be both
considered as a method to represent the aggregated information from
groups of dimensions and a method to apply computational analysis
more locally, i.e., to groups of dimensions. Altogether, we present the
following contributions in this paper:

- Methods to create representative factors for different types of di-
mension groups
- A visual analysis methodology that jointly considers the repre-
sentative factors and the original dimensions
- Methods to assess and compare factors

2 Related Work

In many recent papers, it has been reported repeatedly that the integra-
tion of computational tools with interactive visual analysis techniques
is of key importance in extracting information from the nowadays
strongly challenging datasets. In that respect, Kecim [25] describes the
details of a visual analysis process, where the data, the visualization,
hypotheses, and interactive methods are integrated to extract relevant
information. Perer and Shneiderman [29] also discuss the importance
of combining computational analysis methods, such as statistics, with
visualization to improve exploratory data analysis.

There are interesting examples of works where such an integration
has been done. In MDSSteer [41], an embedding is guided with user
interaction leading to an adapted multidimensional scaling of multi-
variate datasets. A two-dimensional projection method, called the at-
tribute cloud, is employed in the interactive exploration of multivariate
datasets by Jänicke et al. [19]. Endert et al. [6] introduce observation
level interactions to assist computational analysis tools to deliver more
reliable results. Johansson and Johansson [20] enable the user to inter-
actively reduce the dimensionality of a dataset with the help of quality
metrics. In these works, interactive methods are usually used to refine
certain parameters for the use of computational tools. Our method,
differently, enables the integration of the computational tools by inter-
actively determining local domains where these tools are then applied
on. Fuchs et al. [13] integrate methods from machine learning with
interactive visual analysis to assist the user in knowledge discovery.
Oeltze et al. [28] demonstrate how statistical methods, such as correla-
tion analysis and principal component analysis, are used interactively
to assist the derivation of new features in the analysis of multivariate
data. With our work, we contribute to this part of the literature by hav-
ing the computational tools as inherent parts and integrating their
results seamlessly to the interactive visual analysis cycle. Moreover,
we bring together the local structures and the related analysis results
to construct a complete image of the relations in high-dimensional
datasets.

Multi-dimensional datasets, where the dimension count is a few
to several dozens approximately, have been studied widely in the vi-
sual analysis literature. Frameworks with multiple coordinated views,
such as XmdvTool [37] or Polaris [34], are used quite commonly by
now in visual multivariate analysis. Weaver [38] presents a method
to explore multidimensional datasets, where the analysis is carried out
by cross-filtering data from different views. Surveys by Wong and
Bergeron [42] and more recently Fuchs and Hauser [12] provide an
overview of multivariate analysis methods in visualization. Compared
to all these important related works there are however only few studies
published where really high-dimensional data are analyzed. One ex-
ample is the VAR display by Yang et al. [43], where the dimensions are
represented by glyphs on a 2D projection of the dimensions. In order
to lay out these glyphs in the visualization, multidimensional scaling
is used based on the distances between the dimensions. Fernstad et
al. [7] demonstrate their quality metric based reduction in the analysis
of high-dimensional datasets involving microbial populations.

Our new proposed method is realized through a visualization ap-
proach, where dimensions are the main visual entities and the analysis
is carried out together with the data items as recently presented by
Turkay et al. [36]. In this (dual analysis) approach, dimensions are an-
alyzed along with the data items in two dedicated linked spaces. This
concept enables us to include the representative factors, that we iden-
tify, tightly into the analysis. There are few other works where similar
dual analysis methods already proved to be useful, such as in param-
eeter space exploration [4], temporal data analysis [3], and multi-run
simulation data analysis [24]. Kehler et al. [23] integrate statistical
moments and aggregates to interactively analyze collections of multi-
variate datasets. Wilkinson et al. introduced graph-theoretic scagnos-
tics [39] to characterize the pairwise relations on multidimensional
datasets. In a later work [40], the same authors used these features to
analyze the relations between the dimensions. Similar to our work
where we analyze the feature space describing dimensions, Wilkin-
son et al. perform the analysis on the feature space that describes the
pairwise relations.

The structure of high-dimensional datasets and the relations be-
tween the dimensions have been investigated in a few studies, also.
Seo and Shneiderman devise a selection of statistics to explore the
relations between the dimensions in their Rank-by-Feature frame-
work [33]. They rank 1D or 2D visualizations according to statistical
features to discover relations in the data. However, in their method the
main focus is on the data items, not so much the dimensions. One very
relevant related work for us is the visual hierarchical dimension reduc-
tion method by Yang et al. [44]. They analyze the relations between
the dimensions to create a hierarchy that they later use to create lower-
dimensional spaces. In our method, we build upon this idea of con-
structing representative dimensions. However, their method mainly
involved an automatic derivation of the dimension hierarchy and the
representative dimensions were used as the new visualization domain.
In our approach, we treat the representative factors as objects of a dedi-
cated analysis by embedding them into the visualization together with
the original dimensions. Moreover, we provide different methods to
generate, compare and evaluate the representative factors. In a similar
work, Huang et al. [17] utilized the derived dimensions together with
the original dimensions. The authors used several dimension reduction
methods to derive new dimensions and observed how these dimen-
sions correlate with certain characteristics of the original dimensions.
In an interesting paper from the analytical chemistry field by Ivo
ev et al. [18], the authors present the idea to group variables according
to their inter-correlations and utilize them in dimension reduction and vi-
sualization. Although their method is applied only to principal compo-
nent analysis, it clearly demonstrates that grouping of variables indeed
improves the analysis of high-dimensional datasets.

Our work now contributes to the literature with a structure-aware in-
teractive visual analysis scheme for high-dimensional datasets. More-
over, we demonstrate that the visually-guided use of computational
analysis tools can provide more reliable and interpretable results.

3 Representative Factors

With our method, we explore and consider the structures in the di-
mensions space during the high-dimensional data analysis. In order to
achieve a structure-aware analysis of the data, we represent the under-
lying structures with representative factors, or factors, for short. We
then analyze and evaluate these factors together with the original data
to achieve a more informed use of the computational analysis tools.

A conceptual illustration of our approach is presented in Figure 1. Here, we start by computing statistics $s_1$ and $s_2$, e.g., mean and standard deviation, for each of the dimensions in the dataset. We analyze the dimensions by visualizing them in a $s_1$ vs. $s_2$ scatterplot, where each visual entity (i.e., point) is a dimension (1). We notice some structure (a cluster in the lower right), which we then represent with a factor (2). With the help of a computational method, e.g., PCA, we generate the representative factor for the selected group of dimensions and replace these dimensions with the generated factor (3). We continue the analysis by exploring the relations between the factor and the represented dimensions, as well as the other dimensions (4). The analysis continues iteratively with the generation of new factors and/or the refinement of the existing ones.

Our method operates (in addition to the original dataset) on a data table dedicated specifically to the dimensions. We construct this dimensions-related data table by combining a set of derived statistics with available meta-data on the dimensions. In order to achieve this, we assign a feature vector to each dimension, where each value is a computed statistic/property or some meta-data about this dimension. If we consider the original dataset to consist of $n$ items (rows) and $p$ dimensions (columns), the derived data table has a size of $p \times k$, i.e., each dimension has $k$ values associated to it. The set of dimensions is denoted as $D$ and the new dimensions properties table as $S$.

Through a visual analysis of $S$, we determine structures within the dimensions that then result in a number of sub-groups. We represent these sub-groups of dimensions with representative factors and assign feature vectors to these factors by computing certain features, e.g., statistics. Since factors share the same features as the original dimensions, this enables the inclusion of the factors into the visual analysis process. Moreover, these factors are also used to visually represent the associated sub-group of dimensions. Factors serve both as data aggregation and as a method to apply computational tools locally and represent their results in a common frame together with the original dimensions.

As an illustrative example, we analyze an electrocardiography (ECG) dataset from the UCI machine learning repository [9] in the following sections. The dataset contains records for 452 participants, some of whom are healthy and others with different types of cardiac arrhythmia. There are 16 known types of arrhythmia and a cardiologist has indicated the type of arrhythmia for all the records in the dataset. This dataset is analyzed to determine the features that are helpful in discriminating patients with different arrhythmia types.

The raw ECG measurements are acquired through 12 different channels, and for each single channel 22 different features (a mixture of numerical and nominal attributes) are calculated (leading to $12 \times 22 = 264$ values per participant). Already this description reveals an important inherent structure within all dimensions, i.e., that they form kind of a 2D array of dimensions (channels vs. features). In addition to the above ECG measurements, 11 additional ECG-based features are derived and 4 participant specific pieces of information are included. The result is a $452 \times 279$ table ($n = 452$ and $p = 279$).

3.1 Computational and Statistical Toolbox

In order to generate and integrate representative factors into the visual analysis process, we need methods to visually determine the factors and to analyze them together with the other dimensions in $D$. The dual analysis framework as presented by Turkay et al. [36] provides us with the necessary basis to visually analyze the dimensions together with the data items. We make use of visualizations, where the dimensions are the main visual entities, as well as (more traditional) visualizations of the data items. In order to make the distinction easier, the visualizations with a blue background are visualizations of data items and those with a yellow background are visualizations of the dimensions. For the construction of the factors, we determine a selection of computational tools and statistics that can help us to analyze the structure of the dimensions space.

As one building block, we use a selection of statistics to populate several columns of the $S$ table. In order to summarize the distributions of the dimensions, we estimate several basic descriptive statistics. For each dimension $d$, we estimate the mean ($\mu$), standard deviation ($\sigma$), skewness ($skew$) as a measure of symmetry, kurtosis ($kurt$) to represent peakedness, and the quartiles ($Q_{1-4}$) that divide the ordered values into four equally sized buckets. We also include the robust estimates of the center and the spread of the data, namely the median ($med$) and the inter-quartile range ($IQR$). Additionally, we compute the count of unique values ($uniq$) and the percentage of univariate outliers ($%out$) in a dimension. $uniq$ values are usually higher for continuous dimensions and lower for categorical dimensions. We use a method based on robust statistics [23] to determine $%out$ values. In order to investigate if the dimensions follow a normal distribution, we also apply the Shapiro-Wilk normality test [31] to the dimensions and store the resulting p-values ($pVal_{ab}$) in $S$. Higher $pVal_{ab}$ indicate a better fit to a normal distribution. In the context of this paper, we limit our interest to the normal distribution due to its outstanding importance in statistics [21].

One common measure to study the relation between dimensions is the correlation between them. We compute the Pearson correlation between the dimensions to determine how the values of one dimension relate to the values of another dimension. Correlation values are in the range $[-1, +1]$ where -1 indicates a perfect negative and +1 a perfect positive correlation.

Additionally, we use multidimensional scaling (MDS) to help us to investigate the structure of the dimensions space. MDS is a method that projects high-dimensional data items usually to a 2D space by preserving the distances between them as good as possible. Here, we use MDS directly on the dimensions, similar to the VAR display by Yang et al. [43]. We use the correlations between the dimensions to compute a distance matrix, where this distance information is used as an input to MDS. As a result, MDS places the highly inter-correlated groups close to each other. All these computational analysis tools are available through the integration of the statistical computation package R [35]. This mechanism enables us to easily include a variety of tools in the analysis.
3.2 Factor Construction

Constructing factors that are useful for the analysis is crucial for our method. Since factors are representatives for sub-groups of dimensions, they are constructed to preserve different characteristics of the underlying dimensions. The machine learning and data mining literature provides us with valuable methods and concepts under the title of feature (generally called an attribute in data mining) selection and extraction [14]. Feature extraction methods usually map the data to a lower dimensional space. On the other hand, feature subset selection methods try to find dimensions that are more relevant and useful by evaluating them with respect to certain measures [5].

Here, we introduce three different methods to construct representative factors using a combination of feature extraction and selection techniques. Each factor construction method is a mapping from a subset of dimensions \( D' \) to a representative factor \( D_R \). The mapping can be denoted as \( f : D' \rightarrow D_R \), where \( D' \subseteq 2^D \). The \( t \) dimensions that are represented by \( D_R \) are denoted as \( d_{R0}, \ldots, d_{Rt} \). Each factor creation is followed by a step where we compute a number of statistics for \( D_R \) and add these values to the \( S \) table. In other words, we extend the \( D \) table with a \( D_R \) column and the \( S \) table with a row associated with \( D_R \). Notice that each \( D_R \) column consists of \( n \) values similar to the other columns of the \( D \) table.

3.2.1 Projection Factors

The first type of representative factor is the projection factors. Such factors are generated using the output of projection-based dimension reduction methods that represent high-dimensional spaces with lower dimensional projections. Projection factors are preferred when we want the resulting factor(s) to represent most of the variance of the underlying dimensions [21]. In order to determine structures that are suitable to be represented via this type of factors, we analyze the correlation relations between the dimensions. Subsets of dimensions that are highly inter-correlated are good candidates to be represented by a projection factor.

In the context of this paper, we use principal component analysis as the underlying reduction method. However, depending on the nature of the data and the analysis, different reduction methods [21] could be employed here, too.

During each projection-factor generation we create two factors, being the first two principal components here. We choose to include two components in order to be able to visualize also the data items in a scatter plot when needed. For \( D' \), where the variance structure cannot be well captured by two components, we suggest two options. The first option is to apply PCA to several subsets of \( D' \) and create factors for each of these subsets. These subsets can be determined by observing the inter-correlations between the dimensions in \( D' \) and separating the sub-groups with stronger inter-correlations. The second option is to use more components (factors) than two where a more accurate number can be determined by certain methods suggested in the literature, such as observing a scree-plot [21]. In our analysis, we prefer the first method instead of creating a larger number of factors per \( D' \), since it creates easier to interpret factors.

In order to determine sub-groups of dimensions that are suitable to be represented with projection factors, we can make use of MDS. If we apply MDS on the dimensions using the correlation matrix as the distance function and visualize the results, the clusters in such a view corresponds to highly inter-correlated sub-groups, i.e., suitable for a projection factor. In Figure 2-a, we see such a sub-group of dimensions (consisting of 10 dimensions) that is suitable to be represented with a projection factor. We then apply PCA to these 10 selected dimensions and store the first two principal components as the representative factors for these 10 dimensions.

Projection factors are the most suitable factors when the goal of the analysis is dimension reduction. Since different dimension reduction methods have different assumptions regarding the underlying data, evaluating these assumptions leads to more reliable results. In that respect, dimensions can be analyzed in terms of their descriptive statistics, normality test scores and \( \text{unig} \) values to determine their suitability.

![Fig. 2. Groups of dimensions that are suitable to be represented by different types of factors. a) MDS is applied to the dimensions using the correlation information. A highly inter-correlated group is selected to be represented by a projection factor. b) A group of dimensions that are likely to come from a normal distribution (skew and kurt \( \sim 0 \)) is to be represented by a distribution model factor. c) Meta-data is utilized to select a group of dimensions (same channel, different features) that then can be represented by a medoid factor.](image-url)
representative, we employ an idea from partitioning around medoids (PAM) clustering algorithm \[22\]. In this algorithm, cluster centers are selected as the most central element, we choose the dimension \( d \) selected as the most central element of the cluster. Similarly, to find the most central element, we choose the dimension \( d \in D' \) that has the minimum total distance to the other dimensions, computed as:

\[
\text{argmin}_d \left( \sum_{j=1}^{t-1} \text{dist}(d, d_j), d \neq d_j, (d, d_j) \in D' \right)
\]

where \( \text{dist} \) is chosen as the Euclidean distance and \( t \) is the total number of dimensions in \( D' \). This dimension \( d \) is then selected as the representative. In Figure 2-c, we make use of the meta-data information to determine a group that is suitable to be represented via a medoid factor. Here, we plot the channel codes and the feature codes on a scatterplot. The first five features associated with a channel are known to be associated with the width of sub-structures in the channel, thus they can be represented by a medoid factor.

### 3.2.3 Medoid Factors

The third type of representative factor is the medoid factors, that are generated by selecting one of the members of \( D' \) as the representative of \( D' \). Such factors are preferred when the dimensions in \( D' \) are known to share similar contextual properties or some of the dimensions could be filtered as redundant. The user may prefer to select one of the dimensions and discard the rest due to redundancy. Meta-data on the dimensions can be visualized either over descriptive statistics or fitness scores to known distributions.

**3.3 Integrating Factors in the Visual Analysis**

In order to include the factors into the dimensions visualizations, we compute all the statistics that we already computed for the original dimensions also for the representative factors. We add these values on \( D_R \) as a row to the table \( S \). This enables us to plot the factors together with the original dimensions.

Figure 3-a shows the dimensions in a plot of \( \text{med} \) vs. \( \text{IQR} \). We then select all the continuous dimensions that are related to the first channel \( D_I \) and apply a local PCA to the selected dimensions. We leave out the categorical data dimensions since they are not suitable to be included in PCA calculations. We perform the same operation also for the other 11 channels. This leaves us with a total of 12 representatives, each of which represents 16 dimensions. We compute the \( \text{med} \) and \( \text{IQR} \) values also for the \( D_R \) and replace the original dimensions with their representatives in Figure 3-b. The representatives are colored in shades of green to distinguish them from the original data dimensions. Here, we see the relation between different channels through the distribution of the factors over the \( \text{med} \) vs. \( \text{IQR} \) plot. In order to see how a single factor relates to the represented dimensions over the \( \text{med} \) and \( \text{IQR} \) values, the factor is expanded and connected with lines to the represented dimensions (Figure 3-c). The relations between the factor and the represented dimensions are also observed on a skew vs. kurt view (Figure 3-d).

**Brushing representative factors:** Representative factors require a different way of handling in the linking and brushing mechanism. When the user selects a representative factor \( D_R \) in a view, all the dimensions \( d_{i}^{p} \) that are represented by \( D_R \) in the other views are highlighted. Similarly, when the user selects one of the \( d_{i}^{p} \) dimensions, the related \( D_R \) is highlighted in the other views. Figure 4 illustrates how the selections of factors are linked to the other views. Here, for each factor selected in the \( \text{med} \) vs. \( \text{IQR} \) view, 6 associated dimensions are selected in the second skew vs. kurt view. Therefore there are 21 selected dimensions in total in the right view. This mechanism enables us to interact with information at both the original dimension level and the aggregated level.

### 3.4 Evaluation of the representatives

The evaluation and a more quantitative comparison of the factors is an essential part of a representative factor based analysis pipeline as presented here. We provide two different mechanisms to evaluate the factors using quantitative measures.
The first method is related to the correlation based coloring of the factors and the represented dimensions. As an inherent part of the factor generation, we compute the Pearson correlation between $d_R$ and the dimensions that it represents $d_R^i$. The result is a set of $t$ values $corr_R$, where each value is in the range [-1, 1] as described already. We color-code these pieces of correlation information in the views using two different color maps (Figure 3-e). Firstly, we represent the aggregated correlation values as shades of green. For each $D_R$, we find the average of the absolute values of $corr_R$. More saturated green represent higher levels of correlation (either positive or negative) and paler green represent lower levels. Secondly, we encode the individual values of $corr_R$ when a factor is expanded. Each represented dimension $d_R^i$ is colored according to the correlation with $D_R$. Here, we use a second color map where negative correlations are depicted with blue and positive correlation with red.

The second mechanism to evaluate the factors is called profile plots. When the set of statistics associated with dimensions is considered, factors do not represent all the properties equally. If we consider again how the same factor relates to the represented dimensions over med and IQR in Figure 3-c and skew vs. kurt, in Figure 3-d, we see different levels of similarity between $D_R$ and the represented dimensions. Since these relations for all the statistics, i.e., columns of $S$, are different, we build profile plots to visually represent this difference information. In order to find the similarity between $D_R$ and $d_R^i$ with respect to the statistic $s$, we compute the following value:

$$sim_s = 1 - \frac{1}{d} \sum_{i=0}^{d-1} |s(D_R) - s(d_R^i)|$$

The $sim$ values are in the range [0, 1] where higher values indicate that the representative has similar $s$ values as the represented dimensions. We present the $sim_s$ values for all the different statistics in a histogram-like view called profile plots as seen in Figure 5-right. Here, each bin of the plot corresponds to a different $s$ (as listed in the figure) and the $sim_s$ value determines the height of the bin. Additionally, we color-code the average of $sim$ values as the background to the profile plots, with the color map (marked 1) in Figure 3. In Figure 5, we see two examples of factors where the profile plot for the first factor preserves most of the features of the underlying dimensions. However, the second profile plot shows that the factor has different values for most of the features of the underlying dimensions.

4 Analytical Process

The structure-aware analysis of the dimensions space through the use of these factors involves a number of steps. In the following, we go through the steps and exemplify them in the analysis of the ECG data.

Still, these steps are general enough to provide a guideline for the analysis of heterogeneous high-dimensional data using the representative factors.

Step 1: Handling missing data – Missing data are often marked prior to the analysis and available as meta-data. It is important to handle missing data properly and there are several methods suggested in the corresponding literature [15]. We employ a simple approach here and replace the missing values with the mean value of continuous dimensions prior to the normalization step. Similarly, in the case of categorical data, we replace the missing values with the mode of the dimension, i.e., the most frequent value in the dimension. Moreover, we store the number of missing values per each dimension in $S$ for further reference.

Step 2: Informed normalization – Normalization is an essential step in data analysis to make the dimensions comparable and suitable for computational analysis. Different data scales require different types of normalization (e.g., for categorical variables scaling to the unit interval can be suitable, but not $z$-standardization) and different analysis tools require different normalizations, e.g., $z$-standardization is preferred prior to PCA. We enable three different normalization options, namely, scaling to the unit interval [0,1], $z$-standardization, and, robust $z$-standardization. In the robust version, we use $med$ as the robust estimate of the distribution’s center and $IQR$ for its spread. In order to determine which normalization is suitable for the dimensions, we compute certain statistics, namely $uniq$, $pVal_{shp}$, and $%out$, prior to normalization. We visualize $uniq$ vs. $%out$ (Figure 6-a) to determine the groups of dimensions that are suitable for different types of normalizations. Dimensions with low $uniq$ values (marked with 1 in figure) are usually categorical and scaling to the unit interval is suitable. Dimensions with higher $uniq$ values (marked 2) are more suitable for $z$-standardization and, for those dimensions that contain larger percentage of one dimensional outliers (marked 3), a robust normalization is preferable. We normalize the same sub-group of dimensions using all the three methods and apply PCA separately on the three differently normalized groups. Figure 6-b shows the first two principal components factors. We observe that non-robust and robust normalizations resulted in similar outputs, however the unit scaling resulted in PCs that carry lower variance.

Step 3: Factor generation – In this step, we analyze the structures in the dimensions space firstly through the help of meta-data information. We choose to represent each channel only by the first principal component. Each channel in the ECG data has 22 dimensions associated, however, we select a sub-group of these features (the continuous
features (dimensions) that have larger uniq values and then construct projection factors for each channel. The resulting groups are now displayed on a uniq vs. %out plot (Figure 7).

Step 4: Evaluating and refining factors iteratively – In figure 7-1 we notice that the factor that is representing the V2 channel (denoted as D1′), has a higher percentage of 1D outliers. This is interpreted as a sign of an irregular distribution of items in this factor and we decide to analyze this factor further. First, we have a look at the items in a scatterplot of the first two components of D1′ and we clearly see that there are two separate groups (Figure 7-2). However, when we expand the selected factor to see its relation with the underlying dimensions, we observe that there are dimensions that the factor has strong correlations (D1″) and some other that have weak correlations (D2″). We decide to refine this factor further by creating two smaller groups D1 and D2 and visualize the new factors in the same view (Figure 7-3). When we observe the items in visualizations of the first two components of the new factors (Figure 7-4,5), we see that the grouping is solely due the dimensions in D1. The dimensions in D2 carry no significant information.

In order to analyze the separated group of patients in Figure 7-5, we observe the arrhythmia class label column in a histogram. We find out that the selected group accounts for almost all the patients with coronary artery disease (Figure 7-6). This shows that these three dimensions associated with the V2 channel are distinctive features for coronary artery disease.

Here, we present a step-by-step iterative analysis where at each iteration we refine the factors and dig deeper into the data. The above example demonstrates how the representative factors enables a more controlled use of computational tools and a better understanding of the relations in-between the dimensions.

5 USE CASE: ANALYSIS OF HEALTHY BRAIN AGING STUDY DATA

In this case we analyze the data related to a longitudinal study of cognitive aging [2, 46]. The participants in the study were healthy individuals, recruited through advertisements in local newspapers. Individuals with known neurological diseases were excluded before the study. All participants took part in a neuropsychological examination and a multimodal imaging procedure, with about 7 years between the first and third wave of the study. One purpose of the study was to investigate the association between specific, image-derived features and cognitive functions in healthy aging [46]. In the study, 3D anatomical magnetic resonance imaging (MRI) of the brain has been complemented with diffusion tensor imaging (DTI) and resting state functional MRI [16, 45]. Here we are interested in the analysis of the anatomical MRI recordings. These recordings are segmented automatically [10], and statistical measures, such as surface area, thickness and volume (among several others) are computed for each of the segmented cortical and subcortical brain regions. The neuropsychological examination covered tests of motor function, attention/executive function, visual cognition, memory- and verbal function. The participants’ results on these tests are evaluated by a group of neuropsychologists.

The dataset covers 83 healthy individuals with the measurements from the first wave of the study in 2005. For each subject, a T1-weighted image was segmented into 45 anatomical regions, and 7 different measures were extracted for each region. For a complete list of brain regions, refer to the work by Fischl et al. [8]. These computations are done automatically using the software called Freesurfer [10]. The 7 features associated with each brain region are number of voxels, volume and mean, standard deviation, minimum, maximum and range of the intensity values in the region. This information on the brain regions and the features is represented in the meta-data file, which is then used in the analysis. The above operation creates $45 \times 7 = 315$ dimensions per subject. In addition, details about each individual, such as age and gender, and the results of the neuropsychological examination are added to this dataset. With this addition, the resulting dataset has 357 dimensions. In other words, the resulting table’s size is $83 \times 357$ – a great challenge for visual as well as computational analysis. Such a high dimensionality usually requires analysts to delimit the analysis to a selected subset of segments, based on an a priori specified hypothesis. Our aim here is to discover different subsets of individuals and brain regions that are relevant for building new hypotheses.
We start our analysis with the missing value handling and the normalization step. Missing values in the dataset are identified with different strings in different columns of our dataset. And these identifiers (specific for each dimension) are recorded in the meta-data file. We replace the missing values with the mean (or mode) of each column. (specific for each dimension) are recorded in the meta-data file. We decide that one of these features needs to be left out from the hypothesis. However, clear instructions and a video demonstration of this system is equal to the volume. However, such relations may not be always easily derived from the names of the features and require visual feedback to be discovered. Moreover, the profile plot reveals that the range of intensity feature (7) preserves most of the statistics in the underlying dimensions. We also mark the standard deviation of intensities as interesting, since the underlying dimensions have different correlation relations with the representative factor. This indicates that this feature is likely to show differences between the brain regions.

We continue by delimiting the feature set for the brain regions to those two selected features. This means that we delimit the operations to $45 \times 2$ dimensions and apply MDS on these 90 dimensions using the correlation matrix as the distance values. We identify a group of dimensions that are highly correlated in the MDS plot (Figure 9-c). We find out that this group is associated with the sub-structures in the Cerebellum Cortex (CerCtx) and CerCtx is represented with 5

![Image](Image 54x597 to 199x742)

**Fig. 8.** Missing values are handled automatically in our system, and the effects of this transformation is observed here. Normality test scores before and after the transformation are to the left. For a large number of dimensions, the normality test scores improved. On the right, the dimension Cerebellum Cortex middle Posterior is inspected before and after missing values are replaced.

As the next step, we create factors to represent each brain-region (not CerCtx, since it is already represented by a medoid factor). We compute a PCA locally for each brain region and create representative factors. In Figure 9-d, we see the factors (using only the first component) over a normality score vs. %out plot. Here, each factor represents a single brain region. We select the brain regions, where the representative shows a normal distribution. Such a normally distributed subset provides a reliable basis to apply methods such as PCA on the participants. From this analysis, the regions of interest are right and left lateral ventricle, brain stem, left and right choroid plexus and right inferior lateral ventricle. Using only the selected regions, we apply PCA on the subjects (Figure 9-e). We select a group of outlier participants and visualize them on a scatterplot of birth year vs. gender. We observe that this group is mainly composed of older participants. This observation leads to the hypothesis that the selected brain structures are affected by aging.

Here, we comment on the findings related to the selected brain regions. Right and left lateral ventricle are part of the ventricular system that are filled with cerebrospinal fluid (CSF). These regions are interesting and expected findings, and they are known to increase with age (since the brain tissue parenchyma shrinks and the intracranial volume remains constant). Brain stem image information might not be so reliable in the periphery of the core magnetic field homogeneity of the scanner, thus needs to be left out from the hypothesis. Left and right choroid plexus are small protuberations in the ventricles’ walls/roof that produces CSF. It is unexpected for these structures to influence interesting age-related associations. However, this is an unexpected and important finding that our analysis can provide and can be subject to further investigation.

In order to validate the significance of our findings, we focused on the nine participants that we selected in Figure 9-e. As mentioned above, we analyzed the data from 2005, i.e., when all the participants are known to be healthy. Since the data is from a longitudinal study, there are internal reports on how the cognitive function of the participants evolved over time in the next waves of the study. Through these reports, we observe that one of the nine participants is described as showing an older infarct (through MRI scans) and six of the remaining participants (75%) showed declining cognitive function during the study period. The percentage (of cognitive function decline) in the other participants is 28%. This shows a clinical importance of the selected participants. Moreover, this result supports the above hypothesis that the selected brain regions are related to age-related disorders.

All in all, the above observations clearly suggest that the interactive visual analysis of the MRI dataset leads to significant and interesting results that are very unlikely to be achieved using conventional analysis methods.

Above, we have presented only a subset of the analytical studies that we performed on this dataset. The overall analysis benefits highly from the comparison and the evaluation of the computational analysis results that are performed locally. We demonstrate that our methods are helpful in exploring new relations that provide a basis for building new hypotheses.

6 DISCUSSIONS

To adopt our approach, the experts need to have a deep understanding of the statistics and computational tools that are employed in the analysis. This makes the learning curve of our system steeper than classical visual analysis systems. However, we observed that our tool could easily be integrated into the working pipeline of neuroinformaticians and neuropsychologists. These experts who analyze such complex datasets normally make use of computational analysis tools such as Matlab or R [35] and have an overall understanding of computational analysis. And compared to these systems, our solution is much more intuitive thanks to the support from interactive visual methods in the use of computational tools. We even state that such a tool can easily serve as an educative tool to train scientists in multivariate computational analysis. However, clear instructions and a video demonstration of
an analysis of a simple dataset is regarded as highly important. One suggestion to improve the usability of the system is to further exploit the integration of R and develop a modular system that is accessible also for the domain experts. In order to get a clearer image of the requirements, a formal user study is needed. Such a study could lead to simplifications in the analysis process. To make the high-level operations more accessible and traceable, we need to devise special methods where the outcomes of the iterative steps are visually abstracted through a work-flow like interface. Such abstractions can also play a role in the presentation of the results and improve the usability of our system.

Different visualization methods such as parallel coordinate plots could also be incorporated to visualize the factors together with the original dimensions. One possible method to achieve this is to use hierarchical parallel coordinates, suggested by Fua et al. [11]. At several stages in our analysis, we are building new factors using a subset of factors, which implies that we are creating a hierarchy of factors. In our present realization, we only visualize the relations between the factors and the raw dimensions. Augmenting the visualization with such a hierarchy can likely lead to additional insight. Hierarchical differences in scatterplots, as introduced by Piringer et al. [30], is a powerful technique to visualize such hierarchies.

Apart from the present case of healthy aging, the applicability of our tool could also be explored in the broader context of open access brain mapping databases such as BrainMap [26] and NeuroSynth [27]. These databases provide imaging data and meta-data from several thousand published articles available for meta-analyses and data mining, and thus are suitable for visual and explorative analysis methods.

7 Conclusion
With our method, we present how the structures in high-dimensional datasets can be incorporated into the visual analysis process. We introduce representative factors as a method to apply computational tools locally and as an aggregated representation for sub-groups of dimensions. A combination of the already available information and the derived features on the dimensions are utilized to discover the structures in the dimensions space. We suggest three different approaches to generate representatives for groups with different characteristics. These factors are then compared and evaluated through different interactive visual representations. We mainly use dimension reduction methods locally to extract the information from the sub-structures. Our goal is not to solely assist dimension reduction but rather to enable an informed use of dimension reduction methods at different levels to achieve a better understanding of the data. In both of the analysis examples, we observe that the results of the analysis become much more interpretable and useful when the analysis is carried iteratively on local domains and the insights are joined at each iteration.

The usual work flow when dealing with such complex datasets is to delimit the analysis based on known hypotheses and try to confirm or reject these using computational and visual analysis. With the advent of data generation and acquisition technologies, new types of highly complex datasets are produced. However, when these datasets are considered, little is known a priori, thus data driven, explorative methods are becoming more important. Our interactive visual analysis scheme proved to be helpful to explore new relations between the dimensions that can provide a basis for the generation of new hypotheses.

Acknowledgments
We would like to thank Peter Filzmoser for the discussions and comments on the project. We would also like to thank Paolo Angelelli for helping with the communication and preparation of the dataset. The aging study was supported by grants from the Western Norway Regional Health Authority (# 911397 to AJL and # 911593 to AL).