Computers & Graphics (2024)

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Interactive Tools for Explaining Multidimensional Projections for High-Dimensional Tabular Data

Julian Thijssen, Zonglin Tian, Alexandru Telea*

Department of Information and Computing Science, Utrecht University, Utrecht, 3584CC, The Netherlands

ARTICLE INFO

Article history:
Received July 4, 2024

Keywords: Multidimensional projections, Explanatory visualizations, User evaluations

ABSTRACT

We present a set of interactive visual analysis techniques aiming at explaining data patterns in multidimensional projections. Our novel techniques include a global value-based encoding that highlights point groups having outlier values in any dimension as well as several local tools that provide details on the statistics of all dimensions for a user-selected projection area. Our techniques generically apply to any projection algorithm and scale computationally well to hundreds of thousands of points and hundreds of dimensions. We describe a user study that shows that our visual tools can be quickly learned and applied by users to obtain non-trivial insights in real-world multidimensional datasets. We also show how our techniques can help understanding a real-world dataset containing quantitative, ordinal, and categorical attributes.

1. Introduction

High-dimensional data is present in many science and engineering fields and, thus, a key target for information visualization techniques. A main challenge in this respect is *scalability*, that is, how to visually depict datasets having hundreds of thousands of observations and tens to hundreds of dimensions. *Dimensionality reduction*, also called projection, techniques are one of the solutions of choice in this area [1, 2]. Compared to other high-dimensional visualizations such as table lenses [3], parallel coordinate plots [4], and scatterplot matrices [5], projections scale well on both sample and dimension counts. As such, projections have become the main technique for visualizing such data in *e.g.* biology, astronomy, chemistry, and machine learning.

A raw projection is, however, just a scatterplot which does not further help solving problems. As such, several methods have been proposed to *explain* the visual patterns present in projections. Simple brushing and color-coding allow one to see all dimensions of a single point, respectively one dimension over all points. Projections can also be explained globally by techniques such as biplot axes [6, 7, 8] and axis legends [9]. More recently, Da Silva *et al.* [10] proposed global explanations that encode how neighboring points in a projection are related to each other in terms of their dimension values. Neighborhood-based explanations are easy to interpret (as they use the original dimension names, color-coded in the projection), work with any projection technique, and provide information over all projected points. Yet, they also have important limitations [11]: They (1) do not *scale* to more than roughly 10-15 dimensions; and (2) do not explain *what* the patterns in the projection mean.

Recently, Thijssen *et al.* [12] extended the Da Silva *et al.* approach by observing that, for over roughly 10 dimensions, providing *global* explanations for an entire projection will not work – there are simply too many dimensions to color-code in the projection. They provided several mechanisms to overcome the above two problems (1,2) while keeping the computational scalability and genericity of Da Silva *et al.* More concretely,

*Corresponding author: Tel.: +31-30-253-4170;
e-mail: a.c.telea@uu.nl (Alexandru Telea)

they proposed to (1) globally explain projection patterns by the *values* of their contained points and (2) several interactive techniques that allow scaling explanations to tens of dimensions locally. They also presented preliminary evidence from a user study showing the effectiveness of their methods.

In this paper, we extend the work of Thijssen *et al.* in several directions:

- We present mechanisms that refine the explanatory capabilities of the original approach;
- We present a detailed analysis of a user study demonstrating the added-value of the aforementioned refinements for answering complex questions on tabular data;
- We show the added-value of our proposal by exploring a complex real-world dataset containing quantitative, ordinal, and categorical attributes.

We structure our paper as follows. Section 2 reviews related work on projection explanations. Sections 3 and 4 outline our explanation extensions. Section 5 details our study on the added value of our proposed mechanisms. Section 6 applies our techniques for the analysis of a real-world, complex, dataset. Section 7 discusses our proposal. Section 8 concludes our paper.

2. Related work

Let $D = \{\mathbf{p}_i\}$, $1 \leq i \leq N$, $\mathbf{p}_i = (p_i^1, \dots p_i^n) \in \mathbb{R}^n$ be a high-dimensional dataset with samples $\mathbf{p}_i$. The values $(p_i^k | 1 \leq i \leq N)$, for $1 \leq k \leq n$, form the dataset's k dimensions. We call D *tabular* when its n dimensions have well-understood semantics, *e.g.*, they represent the measurement of a specific property that D's analysts can reason about. Such datasets typically have a a few tens of dimensions [13].

A projection, or dimensionality reduction (DR) technique P, maps n-dimensional samples to q-dimensional ones, where $q \ll n$. When $q \in \{2, 3\}$, the projection of a dataset D, denoted $D^P = \{\mathbf{q}_i = P(\mathbf{p}_i) | \mathbf{p}_i \in D\}$, can be visualized as a scatterplot. If D^P preserves several aspects of D such as point relative distances or neighborhoods, then one can retrieve such *data structure* of D by assessing the *visual structure* of D^P. Several quality metrics have been proposed to gauge projection quality, such as trustworthiness and continuity [14], false and missing neighbors [15], normalized stress and Shepard correlation [16], neighborhood hit [17], and distance and class consistency [18, 19]. A recent survey [20] details how to measure and interpret such metrics.

A projection with high quelity-metric values is not *sufficient* to actually understand the projected data. Indeed, a 'raw' projection is just a scatterplot. Figure 1a shows this for a dataset containing $N = 6500$ wine samples, each having 11 measured physicochemical attributes and one additional dependent attribute (perceived quality) [21]. The dataset D is projected to 2D using the LAMP technique [16]. We see some structure in this projection; what this actually means, is yet unclear.

Projection explanations help users to assign meaning to patterns in a projection. The simplest such tool is color-coding points by the values of a given dimension. Figure 1b color codes

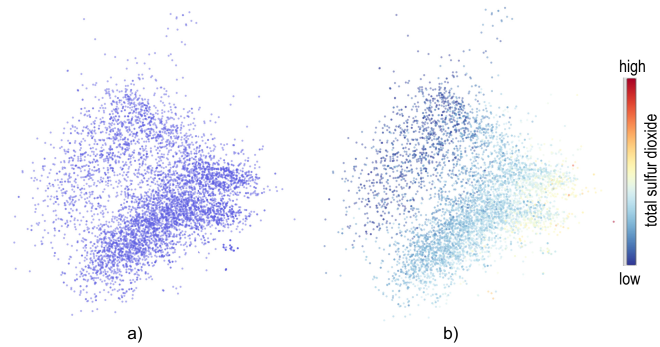


Fig. 1. Wine dataset projection (a) explained by color-coding (b). See Sec. 2.

the Wine projection by its *total sulfur dioxide* dimension, showing that the bottom-right projection area has relatively higher values of this dimension. This simple explanation however cannot consider multiple dimensions.

Several other explanatory techniques exist such as biplot axes [6, 7, 8], axis legends [9, 8], and error views [22, 23, 24, 15, 25, 26]. These techniques work *globally* – that is, the explanations they provide aim to characterize all points in a projection. This is challenging for local-and-nonlinear projection techniques [27], such as t-SNE [28] or UMAP [29], which exhibit strong variations between how they map different data-point nehighborhoods in D, meaning, they can hardly provide global, accurate, explanations anchored to the visual (2D) space. A different direction in explaining projection is given by RadViz [30] and related techniques [31]. These techniques force the projection to obey a given (typically circular) layout so one can relate samples to dimension values. Yet, issues concerning ordering of the dimensions and the global nature of the explanations persist with such methods.

Stahnke *et al.* [26] combined and extended several of the above techniques. They provided an interactive tool to explore projection errors, similar to [22, 24, 15], though using a different visual encoding. They also explained attribute values shared by a user-selected point set (similar to [10] and follow-ups, described below). However, they require users to specifically select a point set for explanation, whereas [10] and followers do the same for all projection points. Our local explanation techniques (Sec. 4) share many similarities with the selection-based mechanisms in Stahnke *et al.*, in particular our differential analysis tool, with the key difference that we show how the selected samples relate to the *entire* dataset, not just their local distribution. Pagliosa *et al.* [32] propose a related approach. Given a point set in the projection (via user interaction or data clustering), they show statistics that differentiate this set from the rest of the projection. Similar to [10], they consider variance of the selected attributes *vs* the rest for explanation; differently, and as in Stahnke *et al.*, the selection of the projection points to explain is done manually, so this approach cannot explain *all* points in a projection.

Joia *et al.* [33] proposed a strategy for text document projections. The projection is split into clusters of points having similar data values. Next, each cluster is labeled by a tag cloud formed by the most relevant keywords of the documents it contains. In contrast to Stahnke *et al.*, and similar to the approach of Da

Silva *et al.* (discussed below), this method explains an entire projection without requiring the user to select a subset of interest. However, setting clustering parameters to partition a projection into groups that next allow effective explanations can be tricky.

Da Silva *et al.* [10] explained projections by finding (and next color-coding) dimensions that contribute most to the similarity of neighbor points. In contrast to global explanations, this method adapts itself locally to show different dimensions that explain different point neighborhoods. Also, in contrast to Joia *et al.*, no explicit partitioning (clustering) of the projection is needed. Proposed explanations include dimension variance [10], local data dimensionality [34], strongest correlated dimensions [34, 11], and dimension values [12]. All these methods address the specific case of so-called *tabular* data, where the individual dimensions are (a) not too numerous and (b) hold specific semantics for the involved users. Yet, as Sec. 1 mentions, only very limited evidence is presented on how, and whether, such explanations work for real-world datasets and users. We address this in the remainder of this paper (specifically, Secs. 5 and 6).

3. Extending global explanations

Variance explanation: We first recall the variance-based explanation of Da Silva [10] which forms the basis of our extension.

Following the notations introduced in Sec. 2, let $\nu_i^P = \{\mathbf{q} \in D^P \| \|\mathbf{q}_i - \mathbf{q}\| \leq \rho\}$ be a neighborhood of radius ρ around projected point $\mathbf{q}_i \in D^P$. Points in ν_i^P come from the projection of a neighborhood $\nu_i = \{\mathbf{p} \in P | P(\mathbf{p}) \in \nu_i^P\}$ in the dataset D. They key idea of Da Silva's explanation – which we take over – is that close points have similar data values, so they can be explained in terms of such data similarities. For a projected point $\mathbf{q}_i$, one first computes the local variance of every dimension $1 \leq d \leq n$ over ν_i as

$$LV_i^d = \frac{1}{|\nu_i|} \sum_{\mathbf{p} \in \nu_i} \left(p^d - \frac{1}{|\nu_i|} \sum_{\mathbf{p} \in \nu_i} p^d \right)^2. \quad (1)$$

Next, a ranking of all n dimensions $\{\xi_i^d\}$, $1 \leq d \leq n$, is computed over ν_i as

$$\xi_i^d = \frac{LV_i^d / GV^d}{\sum_{j=1}^n LV_i^j / GV^j}, \quad (2)$$

where GV^d is the global variance of dimension d over the entire dataset D computed as

$$GV_i^d = \frac{1}{|D|} \sum_{\mathbf{p} \in \nu_i} \left(p^d - \frac{1}{|D|} \sum_{\mathbf{p} \in \nu_i} p^d \right)^2. \quad (3)$$

Intuitively, Eqn. 2 aims to capture how the variance of a dimension over a neighborhood *differs* from the global variance of that dimension. Intuitively put, low values ξ_i^d indicate dimensions d which vary very little over ν_i (as compared to their variance over D), and thus are a good way to explain why points in ν_i are similar. The normalization by GV in Eqn. 2 accounts for dimensions with different variances over D so that low-variance dimensions do not get a higher ranking than high-variance ones.

The lowest-rank dimension $\lambda_i = \arg\min_{1 \leq d \leq n} \xi_i^d$ is picked to explain point $\mathbf{q}_i$. The C most-frequent such lowest-ranks λ_i over the whole projection D^P are mapped to a categorical colormap with C colors; Less-frequent ranks are mapped to a separate 'other dimensions' color. In our work, we use the $C = 20$ colormap of Kelly [35], excluding black and white. Finally, a *confidence* value C_i^d is computed for each $\mathbf{q}_i$ and each d, telling how well the chosen dimension λ_i explains point $\mathbf{q}_i$, as

$$C_i^d = \frac{1}{\sum_{\mathbf{q}_j \in \nu_i} \xi_j^d} \sum_{\mathbf{q}_j \in \nu_i} \begin{cases} \xi_j^d, & \text{if } d \text{ is top ranked for } \mathbf{q}_j \\ 0, & \text{otherwise} \end{cases}, \quad (4)$$

that is, the rank values ξ_j^d are summed up over all points $\mathbf{p}_j \in \nu_i^P$ having the *same top-ranked dimension* as $\mathbf{q}_i$, and the result is normalized by the ranks ξ_j^d summed over the entire ν_i^P. The confidence $C_i^{\lambda_i}$ for the lowest-rank dimension λ_i (color-mapped to explain point $\mathbf{q}_i$) is encoded in the point's luminance. So, bright areas show cases where the color-coded dimension explains well many points in those areas; and conversely for dark areas.

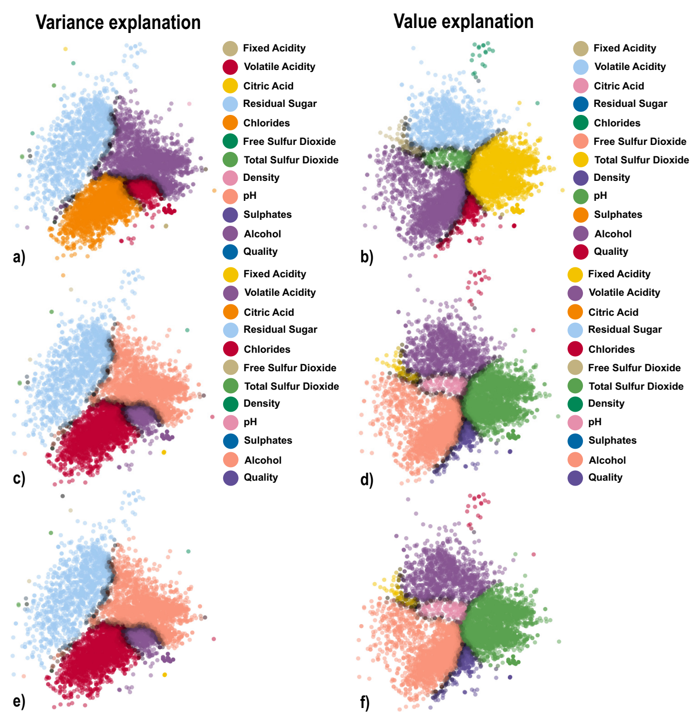


Fig. 2. Variance and value explanation of a projection. (a,b) Per-explanation coloring; (c,d) Consistent coloring; (e,f) Explanations in (c,d) using the Da Silva confidence.

Value explanation: Like for variance explanation, we also compute ranks of all dimensions $\{\xi_i^d\}$, $1 \leq d \leq n$, over each neighborhood ν_i. The key idea behind value ranking is to find dimensions which have *outlier* values over such neighborhoods. For this, we first compute the local average

$$LA_i^d = \frac{1}{|\nu_i|} \sum_{\mathbf{p} \in \nu_i} p^d \quad (5)$$

of dimension d over ν_i. We next compute the value ranking of dimension d as

$$\xi_i^d = \frac{(LA_i^d - GA^d)/GR^d}{\sum_{j=1}^n |LA_i^j - GA^j|/GR^j}, \quad (6)$$

where GA^d is the global average of dimension d over D as

$$GA_i^d = \frac{1}{|D|} \sum_{\mathbf{p} \in D} p^d \tag{7}$$

and $GR^d = \max_{1 \leq i \leq N} p_i^d - \min_{1 \leq i \leq N} p_i^d$ is the range of dimension d over D. Note how GR in Eqn. 6 has a similar normalization goal to GV in Eqn. 2. Dimensions d with positive ranks ξ_i^d are unusually high in neighborhood ν_i; dimensions with negative ranks are unusually low, respectively. The higher or lower the rank values are, the more unusual the dimension values are in a neighborhood as compared to their averages over D. Depending on the application, one can choose whether to highlight unusually high (or low) dimensions, or both. For simplicity, we next consider unusually high dimension values – that is, we pick the highest-rank dimension $\lambda_i = \arg\max_{1 \leq j \leq n} \xi_i^j$ to explain point $\mathbf{q}_i$. We color map these dimensions to show their identity, as for variance ranking.

Robust confidence: When the ranks ξ_j^d of a top-dimension are zero over an entire neighborhood, computing C_i^d will yield a division by zero (see Eqn. 4). Moreover, due to the summing of ranks in Eqn. 4, confidences are skewed in different directions based on the exact distribution of ranks in a neighborhood. Da Silva *et al.* [10] and subsequent work [34, 11] fixed these issues by evaluating Eqn. 4 on a neighborhood of larger radius $\rho_C > \rho$ than the radius ρ of the neighborhood ν_i used to compute ranks in Eqn. 2. The neighborhoods ρ_C work as a smoothing filter on the results of Eqn. 4 – this lowers, but does not fully remove, the chances of division-by-zero and skewness. Moreover, this additional parameter ρ_C brings extra complexity for users.

We remove these problems by computing the confidence as

$$C_i^{d,robust} = \frac{1}{|\nu_i|} \sum_{\mathbf{q}_j \in \nu_i} \begin{cases} 1, & \text{if } d \text{ is top ranked for } \mathbf{q}_j \\ 0, & \text{otherwise} \end{cases}. \tag{8}$$

Simply put, $C_i^{d,robust}$ tells how often a given top-ranked dimension d occurs over all points in a neighborhood ν_i, and has the same interpretation as Da Silva's original C_j^d. Our computation avoids the aforementioned division-by-zero and skewness problems.

Figure 2a shows the variance explanation on the Wine dataset introduced in Sec. 2. Variance ranking helps explaining *why* certain projection points are close to each other – for example, all red points have similar values of the *chlorides* dimension. Dark areas, close to the borders of same-color (same-explanation) regions, indicate points where the single-dimension explanation is less confident. However, the variance explanation does not tell us *what* close points represent. The value explanation addresses this (see Fig. 2b). We see, for instance, that most red points in the variance-explanation (a), *i.e.*, wines with similar *volatile acidity* values, are now yellow, *i.e.*, are wines with unusually high *total sulfur dioxide* values.

In the above scenario, the projection was recolored when switching explanations from variance to value. Recoloring also happens when any explanation is recomputed due to parameter changes, *e.g.* the radius value ρ used to compute the rankings in Eqns. 2 and 6. Recoloring can be confusing since the same color can be assigned different subsequent meanings. We solve this by keeping the color allocation as consistent as possible throughout such changes. At the start of the exploration, we compute an initial color allocation based on the ranking mode that is in effect (variance or value). Whenever the exploration triggers an update of the dimension ranks, we compute a new color allocation, but keep dimensions that were also part of the previous explanation assigned to their earlier colors. Newly-appearing dimensions in the new explanation get assigned the remaining available colors based on their frequency of being top-ranked as before.

Figure 2c,d show this process for the variance and value explanations depicted in Fig. 2a,b. When switching from variance to value explanation (or conversely), colors are now kept completely consistent. For example, the aforementioned *volatile acidity* dimension, which was red in the variance explanation (a), respectively light blue in the value explanation (b), is now consistently mapped to a purple color in both explanations (c,d).

In Figures 2a-d, brightness encodes our robust confidence $C_i^{d,robust}$. Figures 2e,f show the same dataset with brightness encoding the original Da Silva confidence C_i^d. Given that the results are practically identical, and the earlier-mentioned advantages of $C_i^{d,robust}$, we use our $C_i^{d,robust}$ further in this paper.

4. Local explanations addressing high dimension counts

Global explanations (Sec. 3) are limited by the size C of the categorical colormap used. That is, even if we can compute explanations for many dimensions via Eqns 2 and 6, we can only depict C of these *simultaneously*. Moreover, explaining projection patterns by a *single* dimension λ_i (whether via variance or values) only tells a small part of the full story. Indeed, in typical projections, close points are placed so because of *multiple* dimensions. Consider N different clusters of points in a projection. Barring any projection errors, this generally means that the dimension profiles, *i.e.*, the values that dimensions take on in those clusters, are sufficiently different from each other, otherwise their points would form a single cluster. Each such profile with D dimensions requires D colors to be explained. To fully explain the projection, all such N distinct dimension profiles would need to be explained simultaneously. As N increases, the number of dimensions that need to be explained increases.

We address these limitations by several mechanisms that explain the projection *locally*. As these points, selected for local explanation, are close in the projection, they are relatively similar in data values (assuming the projection is of good quality). Hence, the likelihood that they can be explained by a small number of dimensions increases. Moreover, by explaining *fewer* points, we can provide *more* details on these.

Figure 3 shows our local explanations, which we discuss next.

Lens brushing: We select all projection points $\mathcal{S}$ in a given radius (adjustable via a GUI control) to the mouse pointer to be the focus of the detailed (local) explanations, see next. For these selected points, we compute the variance and value rankings as for the global explanations (Eqns. 2 and 6) by substituting ν_i with the user selection $\mathcal{S}$. Since $\mathcal{S}$ is fixed, in contrast to ν_i which are different for every projection point i, we now thus compute a single variance and value ranking for all points in $\mathcal{S}$ – that is, we

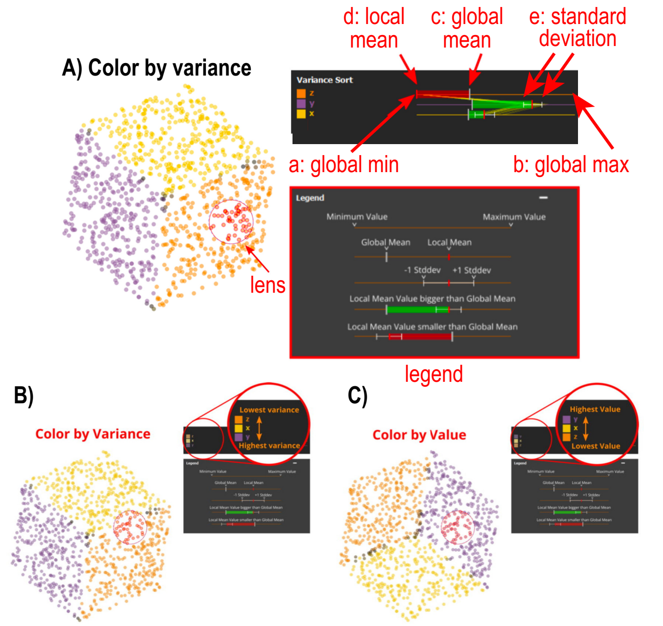


Fig. 3. Local explanation of lensed points (Sec. 4). A: Details of the explanation, including legend, for variance mode. B,C: Instances of the explanation for the variance, respectively value, modes.

explain the entire selection at once, rather than explaining every point i in the projection separately, as done earlier. Users can interactively switch between the variance explanation (which tells *why* points in $\mathcal{S}$ are close in the projection) and the value explanation (which tells *what* these points are, data-wise).

Local analysis: We display detailed explanations of the lensed points $\mathcal{S}$ in a widget right to the projection. Figure 3 shows this widget for a simple 3D axis-aligned cube dataset projected using PCA. The widget is structured as a table with one row per dataset dimension. For each dimension, we show its name, assigned color (by variance or value ranking, cf images (b) and (c)), and a set of statistics for that dimension, drawn right to the dimension name, described further below. In variance mode (Fig. 3B), dimensions are sorted top-to-bottom from lowest rank (lowest ratio of variance in the selected points $\mathcal{S}$ *vs* the whole projection) to highest rank (highest ratio of variance). In contrast to the Da Silva variance explanation (Eqn. 2), we not only show the least varying dimension (the one at the top) by color coding it in the projection, but *all* dimensions, sorted on variance over $\mathcal{S}$. In value mode (Fig. 3C), we sort dimensions top-to-bottom from highest mean value in $\mathcal{S}$ *vs* mean value over the whole projection to lowest mean value. In contrast to the global value explanation, this shows not only the most outlier-like dimension (at top, also color-coded in the projection), but all dimensions, sorted on their outlierness. In both modes (variance and value), we thus explain the lensed points not only by a *single* (color-coded) dimension, but by *all* dimensions, sorted top-to-bottom on how important they are for the chosen explanation mode.

Dimension statistics: The dimension sorting described above helps one find the most salient dimensions (in variance or value) but does not explain *how much* these contribute to the lensed points $\mathcal{S}$. That is, the sorting itself does not say much about the dimension variance or values themselves. For instance, a dimension listed at the top of the value ranking may have a relatively high value, or it may have a low value, as long as all other dimensions have even lower values. Hence, it is useful to show the values of the dimensions for the selected points.

We address this by showing both local and global statistics for each dimension d in the widget. We illustrate this next for the variance mode (Fig. 3A) – the same holds for the value mode. A *range line* (same categorical color as the dimension) indicates the full extent GR^d of dimension d over all projection points from the global minimum (Fig. 3a) to the global maximum (Fig. 3b). A large grey tick shows the dimension's global mean $\sum_{1\leq i\leq N} p_i^d/N$ (Fig. 3c). A similar red tick shows the dimension's local mean over the lensed points $\sum_{\mathbf{q}_i\in\mathcal{S}} p_i^d/|\mathcal{S}|$ (Fig. 3d) When the local mean is greater than the global mean, we draw a green bar between the two means to indicate a dimension having higher than usual (average) values over the lensed points. Similarly, when the local mean is smaller than the global mean, we draw a red bar between the two means, indicating a dimension having lower than usual values over the lensed points. The above visuals show the average value of a dimension but say nothing about how its values are *spread*. This spread is important as it tells whether the dimension has a big influence on the points being close together in the projection or not. Low-variance dimensions for a point set result in those points having small distances in the high-dimensional space and thus, typically, also small distances in the low-dimensional embedding (projection). To convey this, we show the standard deviation of each dimension over $\mathcal{S}$ with white whiskers drawn left and right of the local mean (Fig. 3e). Close whiskers indicate that the lensed points vary little over the analyzed dimension, thus the respective dimension is important for why the points are close in the projection. This is the same information as the top-to-bottom sorting in variance mode. However, in value mode, whiskers add the variance information which is not present in that mode. Note that, while our visualization is similar to a boxplot, it shows very different data: (1) our whiskers show a standard deviation, and not the minimum or maximum values or quartiles; (2) the (green or red) box we draw shows the difference between the global and local means of a dimension, and not quartile-related information, as in typical box plots [36].

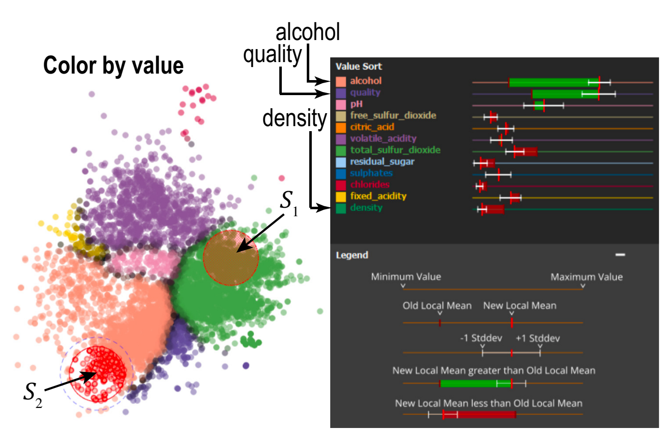


Fig. 4. Differential analysis of sets of points (Sec. 4).

Parallel coordinates plot: All statistics discussed above are aggregates over the selected points. This can be deceiving. For example, dimensions that have the same local mean over the selected points might have quite different value distributions over the samples in $\mathcal{S}$. The standard deviation whiskers show such differences but still work at an aggregated level and thus cannot convey skewed distributions or distributions with discrete value clusters. Figure 5 shows an example. The selected (red) points have two dimensions with the same local mean. If we showed only this mean (a), it would be unclear if the actual distributions of the dimension values over the red points are the same.

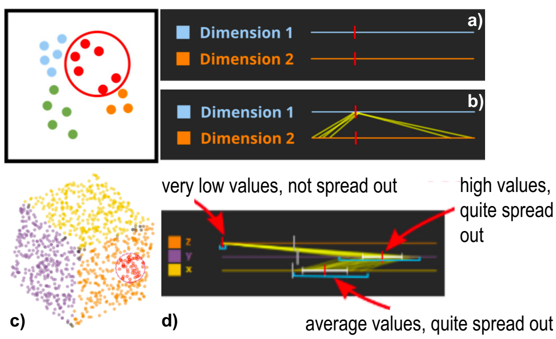


Fig. 5. Parallel coordinate plots for the selected points (Sec. 4).

We convey more detailed information over the selected points by drawing a PCP of all lensed points $\mathcal{S}$ atop of the horizontal range lines of all dimensions. To limit visual clutter, we draw the PCP half-transparent (see Fig. 5b). We now see that, while the local means of the two dimensions are the same, their value distributions are very different. Figure 5d shows the PCP lines in action for a selection of points on the already-explained cube projection (c). The x dimension (orange) shows near-zero values for all selected points – this is the dimension orthogonal to the cube's orange face. The y and z dimensions show, in contrast, high, respectively average, values, which are more spread out – these are the dimensions tangent to the orange face, over which the selected points have more variation and larger values.

Differential analysis: While local explanations show detailed information over a selected projection detail, one inherently needs to explore several such details in a sequence to understand a projection. This puts a certain burden on the user's memory. We alleviate this by offering a way to *compare* two different such user-selected details, as follows. The user selects a set of points $\mathcal{S}_1$, then presses a modifier key and selects a different set $\mathcal{S}_2$. The statistics that are normally shown in the analysis widget are now replaced by statistics showing the differences between $\mathcal{S}_1$ and $\mathcal{S}_2$. Figure 4 shows this for the Wine dataset using the value-ranking mode. The widget shows that the two top-most dimensions (*alcohol*, pink in the projection; and *quality*, dark purple in the projection) have long green bars, while the bottom-most dimension (*density*, dark green in the projection) has a red bar. This tells that wines in $\mathcal{S}_2$ have much higher alcohol and quality, but lower density, than wines in $\mathcal{S}_1$.

Dimension exclusion: Local analysis allows handling higher-dimensional data than global analysis as it shows details of all dimensions over a selected data subset. Still, datasets can contain dimensions that do not convey much information for a given analysis. These can take up valuable colors from our limited $C = 20$ categorical colormap and also clutter the explanation widget. Excluding them upfront from the entire analysis is undesirable as users may wish to examine different dimension sets – and keep the same projection – depending how the analysis unfolds. To address this, we allow users to click on dimensions in the widget to temporarily exclude them from the generated explanations. Doing so reassigns colors to the remaining dimensions and instantly re-creates the global and local explanations. Clicking on an excluded dimension adds it back to the generated explanations. Figure 6 illustrates this. In image (a), about half of the projection points are explained by unusual high values of the *diagnosis* dimension (yellow, top-most in the rank-by-value widget). To get more insight on what else makes these points different, we click on this dimension and disable it. The dimension turns white in the widget and moves to the bottom to indicate disabling. The regenerated explanation (Fig. 6b) splits the big yellow blob into differently-colored groups that provide more insights of how these points differ.

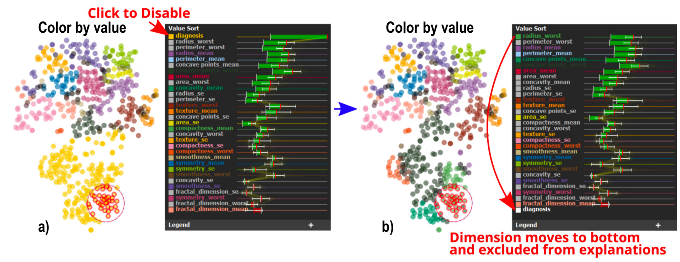


Fig. 6. Selective dimension disabling (Sec. 4).

Scalability: Our explanation system, implemented in C++ in the ManiVault framework [37], scales computationally well. It computes global explanations of datasets of hundreds of thousands of points and hundreds of dimensions in tens of seconds, and next interacts with these in real-time, on a commodity PC, and is openly available [38]. Figure 7 illustrates the visual scalability in sample (a) and dimension (b) counts. Image (a) shows a dataset consisting of 22 registered images of the same brain-cortex tissue patch, each image mapping a gene. Pixel brightnesses encode where in the tissue the gene is expressed. We treat each pixel as a sample having 22 dimensions, one from each image. This yields 115K 22-dimensional samples which we project with t-SNE [28] and next explain the projection. In Fig 7a, the global value explanation shows us how the projection is split into clearly separated point groups. We next lens over several points in the orange region, which corresponds to the Cux2 gene. The local explanation in the widget tells us that Cux2 is, indeed, unusually high in this region (see long green bar top of widget) and that only a few other dimensions have outlier values here (all other bars in widget are quite short). Figure 7b shows another dataset [39] of gene expressions in the brain cortex. This dataset has 2400 samples (cells from the analyzed brain region) each with 314 dimensions (gene expressions). The projection shows the spatial layout of these cells. Even though the dataset has hundreds of dimensions, the global value-ranking explanation is able to assign colors to unravel a salient band-like structure in the projection. Using the lens, we selected points in the purple band (bottom in the projection). The widget tells us that these have an unusually high expression of the Foxp2 gene (top-most bar in the widget), as well as showing other genes having high expressions in this area.

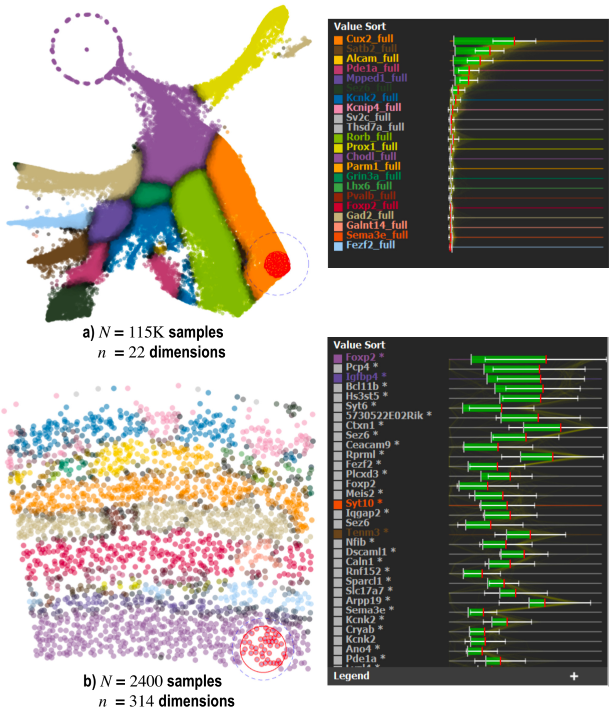


Fig. 7. Scalability of explanations in number of samples (a) and dimensions (b) (see Sec. 4).

5. Evaluation study

To evaluate the effectiveness and ease of use of our interactive system for projection explanations, we conducted a user study, which we describe next (see also Fig. 8).

5.1. Participants

We invited about 60 people to take part in the study (and/or further spread the invitation). Of these, 23 completed the study. Participation was fully anonymous, *i.e.*, we did not collect nor trace the participants' identities. Participants self-reported (at the end of the study) experience with multidimensional data between none and several years (see also Fig. 10a).

5.2. Study set-up

The participants were next asked to install our tool (Windows or Linux) and follow a tutorial (about 15 minutes) covering loading data, switching between variance and value explanations, and understanding the lens and local-explanation widget. Next, the participants were asked to analyze three multidimensional datasets and report answers via Google Forms. These datasets, all from the UCI repository [40] and well-known in projection evaluation literature, had increasing dimensionalities to test our system's scalability in this respect. The *Wine* dataset was described already in Sec. 2. The *Cancer* dataset ($N = 569, n = 31$) has 10 attributes describing the mean, max, and standard deviation of the size, shape, and texture values of cell nuclei in a lung tissue. The 10^{th} attribute tells whether the cells are benign or malignant. The *Spam* dataset ($N = 4601, n = 57$) contains frequencies of selected words aiming at classifying mails as spam or not, and also the classification result. The datasets were projected using LAMP [16] (*Wine*) and t-SNE [28] (*Cancer*, *Spam*).

5.3. Questions

For each dataset, participants had to answer four *control* (C) and three *live exploration* (LE) questions, as follows.

Control questions: The C questions involved studying screenshots of the application (produced by us) to select one of four multiple-choice answers. Answers were designed so that there was a single correct one. In each screenshot, different projection points were selected by the lens; images of both global and local explanations were also provided. The goal of the C questions was to see if the participants understood how to read a pre-computed visualization (without interaction), explained by the value mode, to come to a correct conclusion. Figure 9 shows the screenshots we provided for three such questions, one per studied dataset. The first question (a) shows a selection of points down in the projection; we tell users that, for this dataset, we know that higher attribute values mean a higher chance of malignancy, and conversely. Users are next told that the selected points are (obviously) malignant, as they have very high levels of the *diagnosis* attribute; we see this since (1) the points are yellow and (2) the yellow-labeled attribute in the widget, called *diagnosis*, shows a green bar. This means that *diagnosis* has higher values in the selection than the dataset's average. Next, users are asked which other attributes of the selected points suggest that the points are *benign*. The correct answer is one of the two *fractal dimension* attributes; these show red bars in the widget, so they have lower values in the selection than the dataset average. All other attributes are larger on average in the selection than in the dataset (see their green bars in the widget).

The second question (Fig. 9b) shows a selection in the Spam dataset. Users are told that the selected mails are mostly spam (see also the long green bar in the *spam* attribute, top in widget). They are asked to tell which of the topics are likely the content of these spam mail; answers include making money, advertising a product, improving credit scores, or none of the above. The correct answer is making money. Indeed, the widget shows that, for the above four attributes, only *money* (second-from-top in widget) has a significant green bar, *i.e.*, this attribute has higher values in the selection than overall in the dataset.

The third question (Fig. 9c) shows a selection in the Wines dataset. Users are told that the selected wines have unusually high levels of chloride (the points are red, which maps the *chloride* attribute; and this attribute, top in the widget, has a long green bar). Next, they are asked what can be said about the quality of the selected wines – if this is higher than average, lower than average, or nothing can be said about it. The correct answer is lower than average, since the *quality* attribute in the widget (third from bottom) has a sizeable red bar.

Live exploration questions: We asked participants to analyze the datasets interactively using the tool on their machines and

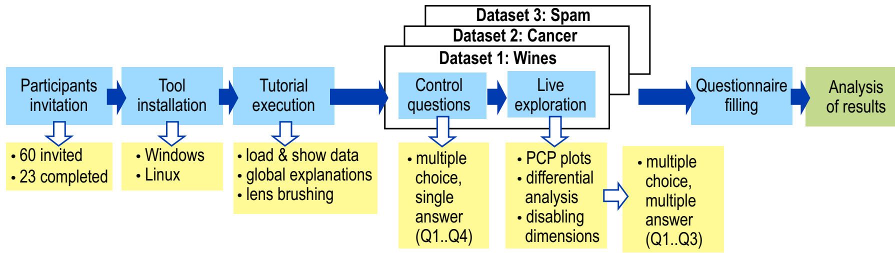


Fig. 8. Structure of the evaluation study (Sec. 5).

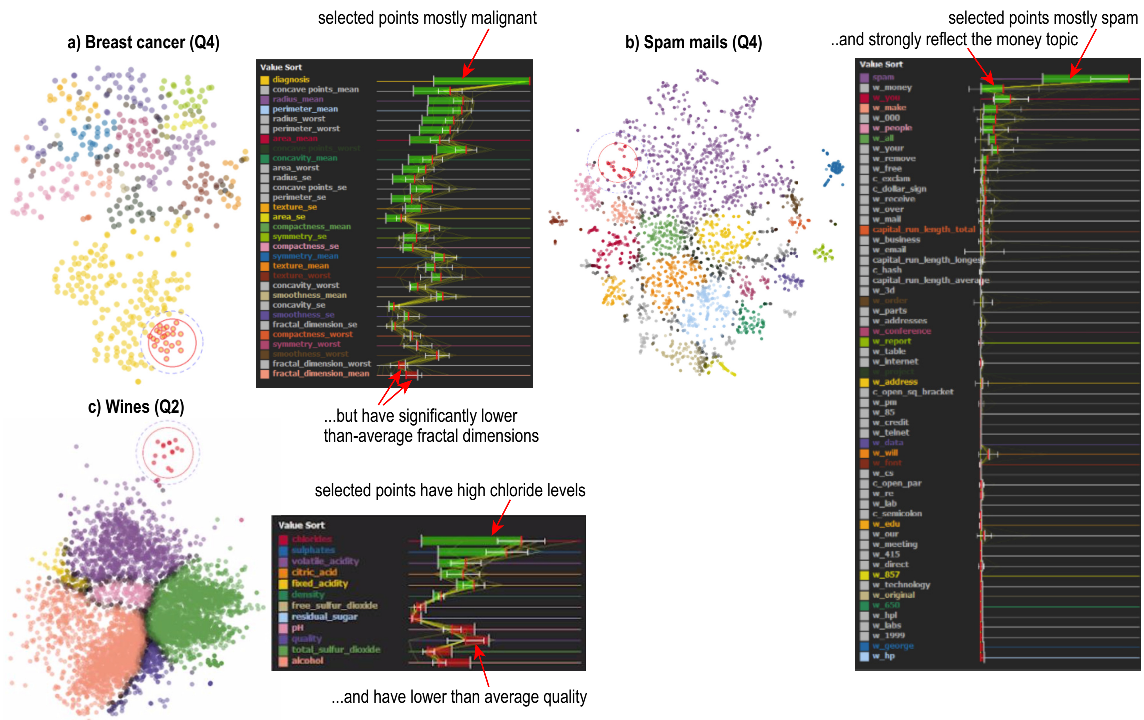
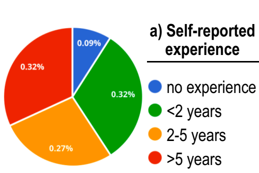


Fig. 9. Three control questions for the three studied datasets (see Sec. 5.4).

select one or more multiple-choice answers for several LE questions. We designed these questions to be harder and less clear-cut than the C ones. This, and the users' freedom to explore the visualization unconstrained, means that it is far harder to judge if an answer was 100% right or wrong. Hence, after having studied the respective datasets in depth, we ranked the LE questions' answers on an 4-point ordinal scale (very likely, likely, unlikely, very unlikely) telling how likely we ourselves would give that answer. Separately, we analyzed the coherence of the users' answers. High values tell that different people using our tool arrive at similar insights. When this occurs, we believe that the answer is likely correct since the chance that many users arrive at the same *wrong* answer is small, given their full freedom to explore the dataset.



b) Control question answer correctness

	Wine	Cancer	Spam
Q1.	100%	95.7%	100%
Q2.	100%	91.3%	82.6%
Q3.	69.6%	91.3%	95.7%
Q4.	100%	78.3%	100%

Fig. 10. Users' experience (a) and control question answering (b).

5.4. *Results*

The 12 control (C) questions were overwhelmingly correctly answered (Fig. 10b), suggesting that users were able to learn to correctly use our tool to perform low-to-medium difficulty tasks.

For the more complex live exploration (LE) questions, Fig. 11 shows the agreement scores. Long-and-bright bars in this figure tell consensus between users and also with our own assessment. Long and dark bars would indicate that many users would select an answer that we consider unlikely. As Fig. 11 shows, we see the former bars but not the latter, which indicates a strong agreement among users *and* with our assessment too. We detail these results next, grouping questions in terms of the type of analyses they implied. For all questions, we provide our own answers obtained using our tool (see Fig. 12).

Single cluster (Q1, Wine): This relatively simple analysis asks users to find very-low-density wines in the projection and find which other attribute is also out-of-proportion and thus likely causes the low density. This question can be easily answered using the lens and the value-ranking. Most subjects (52.2%) answered *alcohol*, which is also our pick. Yet, 30.4% of the subjects answered here *fixed acidity*. This is potentially due to

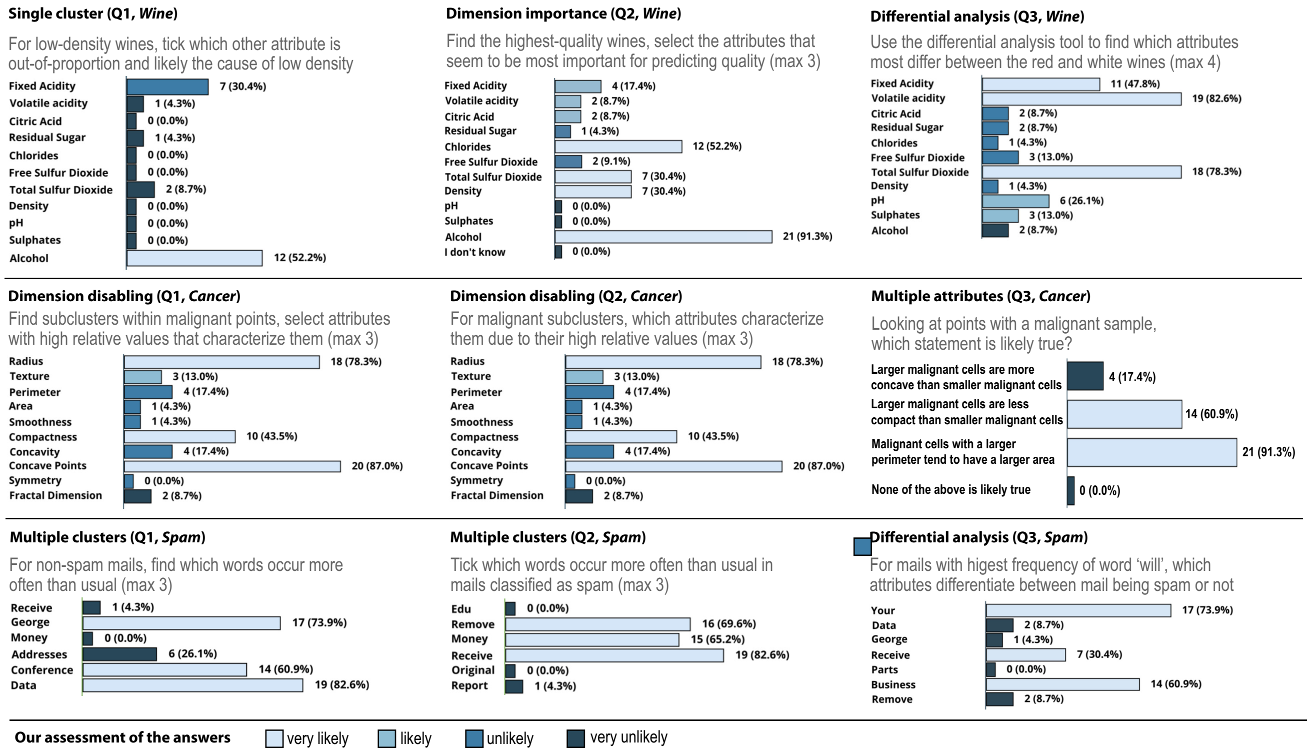


Fig. 11. Inter-user agreement (and our assessment of correctness likelihood) for answers of the 9 live exploration questions Q1-Q3 for all three datasets.

ambiguous phrasing of the question, which could be interpreted as having to find a dimension which deviates from the global mean in the same proportion as the density dimension. Figure 12a shows our analysis for this question. We see that, indeed, alcohol is significantly higher for the selected low-density points than all other points in the dataset.

Multiple clusters (Q1-Q2, Spam): Users were asked to find which words occurred more often in non-spam than in spam mails – thus, study at least 2 different clusters. This involved finding point clusters with spam, respectively non-spam, mails, via *e.g.* the variance global explanation, and then lensing in value-ranking mode to see which of the 6 words occurred there more often than elsewhere. Participants yielded very similar answers – and also similar to our own findings. Participants were v ery close to unanimous in their answers; answers with majority votes correspond exactly with our answers. On Q1, one answer (addresses) also has several votes. This is potentially due to confusion caused by the words 'addresses' and 'address' being dimensions in the dataset, the latter of which has unusually high values in the non-spam e-mails, whereas the first does not.

Multiple attributes (Q3, Cancer): This question – arguably the most complex we had – involved analyzing several attributes per point cluster. This requires interactively finding projection areas having low/high values of one attribute and then analyzing the other attributes in these areas. Again, we see strong inter-user agreement and also agreement with our own findings.

Differential analysis (Q3, Wine; Q3, Spam): Users were asked to tick up to four attributes that are most different between red and white wines. To answer this, they had to find both red and white wines using the global explanation, select points of these two types, and next use the differential analysis to find which attributes differ between these selections. We see again a strong agreement between users and also with ourselves. Figure 12c shows our own explanation for this question. We see that both *volatile acidity* and *total sulfur dioxide* have the largest differences followed by *fixed acidity* and *pH*. These results completely align with the responses of the participants.

Dimension disabling (Q1-Q2, Cancer): Questions 1 and 2 of the Breast Cancer dataset asked the participants to find point clusters in the projection where particular attributes had higher values than all other attributes, and to note down which attributes these were. Such clusters had to be found for points that were completely dominated by a malignant diagnosis (high value) in the diagnosis dimension, meaning all points were assigned the same color (of the diagnosis dimension, see Fig. 12 d1). In our analysis, we found three major distinct subclusters within the point cluster with a malignant diagnosis. These were characterized by high values of the *radius*, *concave points*, and *compactness* dimensions.

As Fig. 11 shows, participants most commonly answered *concave points* (87.0%), *radius* (78.3%), and then *compactness* (43.5%), which matches our analysis. Before going to Q2, participants were briefed on how they can disable and re-enable dimensions and were told to disable the diagnosis dimension, thereby uncovering the colors of the subclusters (see Fig. 12 d2, color: value mode). We see that the *compactness* cluster is quite small and was thus harder to find for Q1. Q2 then next asked participants to repeat the task of Q1 with the newly revealed colour groups. In this second task, we expected participants to have an easier time finding the specific clusters as the assigned point colors are indicating them. Given the relative small size of the *compactness* cluster, making it hard to find in the first task

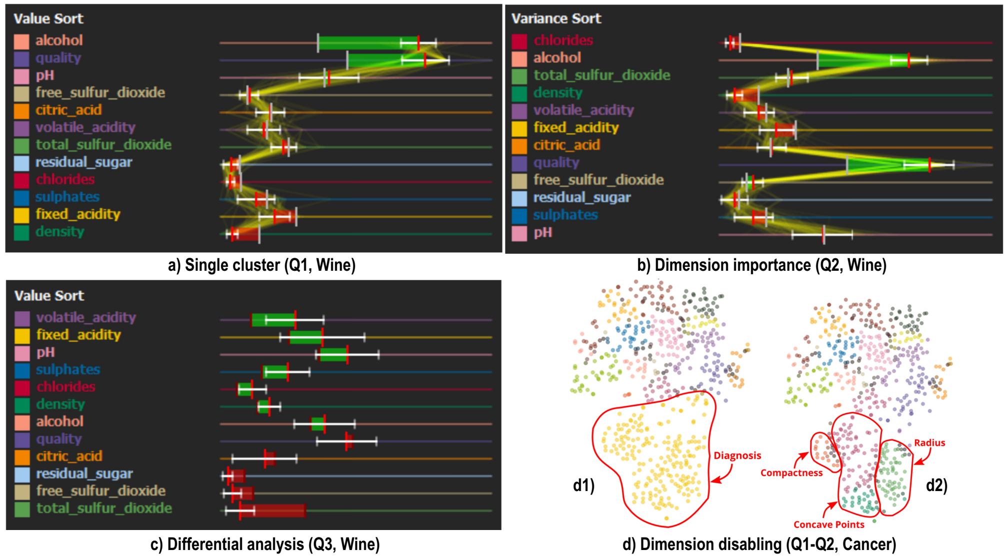


Fig. 12. Our analysis supporting the answers of the live exploration (LE) questions. See Sec. 5.4.

without being able to see the colors, we expected it to be found much more often in the second task, as well as a lesser increase in the other cluster attributes. Participant responses (Fig. 11) show the *compactness* dimension increased from selected ticked by 43.5% of participants to 60.9% between Q2 an Q3 for the Cancer dataset, which matches our expectations.

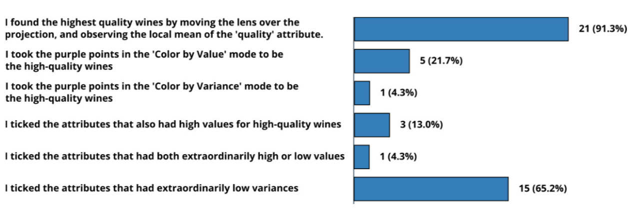


Fig. 13. Tool mechanisms used to answer Q2, Wine. See Sec. 5.4.

Dimension importance (Q2, Wine): A common scenario in the analysis of real-world datasets is finding variables influencing a dependent variable. Question 2 of the Wine dataset asks the subjects to perform such an analysis by finding the region in the projection with the highest-quality wines and ticking the dimensions they believe to influence quality. Figure 11 shows the recorded answers. Again, we see a good agreement of the users with our own explanation (large light bar for dimension *alcohol*). Figure 12b shows our own answer for this question. From the dimensions ranking, we see that *chlorides* has the least variance for the selected high-quality wines (since it is the top dimension in variance mode), telling that having this particular value of chlorides may be important for the high quality of the wines. Next in the ranking comes *alcohol*, and then *total sulphur dioxide* and *density*. These four dimensions are given the most votes by participants.

Compared to the other LE questions, this question is open up to interpretation and personal judgement – finding how variables influence each other can be interpreted quite broadly. As such, we asked a follow-up question to find out how participants used our tool to reach their conclusion. Participants could report the usage of any of six predefined solutions (selected by us) or additionally report a different solution via free text. Figure 13 shows the recorded answers. Interestingly, no 'other' solution was reported apart of our six options. We see the most users answered the question by moving the lens over the projection and keeping track of the local mean shown for the *quality* dimension. Once they found some high-quality wines, most users indicated next that they ticked the dimensions that had very low variances. Our own solution to answer this question was practically identical.

Summarizing the above, we found that participants who used our tool independently and not supervised by us arrived at very similar answers of the posed questions. We deem these answers to be correct, given our own independent analysis of the same datasets. While not a formal proof, we argue that this is evidence that our tool can help obtaining valuable insights in high-dimensional data in a predictable way.

5.5. *Overall feedback*

Figure 14 shows extra details from the participants' feedback. Image (a) shows the opinions on the variance ranking. The top three bars show the answers to our questions on the usefulness of this explanation (see figure for questions). Most users found the variance ranking useful for finding important dimensions and clusters to further explore. Yet, 13% of them found the variance ranking of no extra value. The free answers provided by the users mentioned various issues such as the ranking yielding 'nice' visualizations and structure to the projection; and being overall interesting to explore.

Image (b) shows opinions on the value ranking. As for variance ranking, most users found this mode useful to find important dimensions, clusters to explore, and extremal values. Only one user stated that this mode has no extra value; none found the red-green and standard deviations bars (Sec. 4) confusing. Free answers mentioned that this mode brings additional insights; one user said they would confuse this mode with variance ranking.

Image (c) shows opinions on the PCP plot. Most users found the plot useful to help them gauge the distribution of values in the selection and, overall, providing additional explanatory value. Yet, 2 users found the plot having no extra value and 4 users that the plot makes the explanatory widget more confusing. Free answers mentioned that the PCP plot provides ‘faint’ but useful cues of the data distribution; one user, though, mentioned he/she ‘hates’ this plot (but did not further explain why).

Images (d-g) show how users evaluated the usefulness of all our proposed mechanisms – variance ranking, value ranking, differential analysis, and disabling dimensions, on a 7-point Likert scale ranging from not very useful to very useful. Most users found overall all mechanisms useful. On the above-mentioned Likert scale, we have variance mode: mean score 4.83 (SD=1.63); value mode: mean score 6.52 (SD=0.77); differential analysis: mean score 5.74 (SD=1.03); dimension exclusion: mean score 5.74 (SD=1.42).

6. Evaluation on Real-World Data

To bring more insights in the added-value of the proposed projection exploratory techniques, we use them next to analyze a more complex real-world dataset.

Dataset: The European Values Study (EVS) dataset was created following a large-scale, cross-national and longitudinal survey, which includes a large number of questions on moral, religious, social, political, occupational, and family values that have been replicated since the early eighties [41]. The survey goals are to measure how groups of people in Europe have similar (or different) so-called *value systems* and thereby better understand which aspects unite, respectively divide, people. This can help decisional factors at various levels to devise policy instruments to foster convergence along desirable values. The survey has 111 main questions (some with sub-questions) leading to 282 answers per participant. The survey which we used in our analysis was answered by 56491 citizens from 34 European countries.

Scalability-wise, projections can easily handle this dataset ($N = 56491$ samples, $n = 282$ dimensions). Yet, *preprocessing* all 282 dimensions to make them ‘compatible’ for dimensionality reduction is in itself a challenge, since the dimensions are of different types (quantitative, ordinal, categorical using many different category scales); some questions allow multiple-choice answers and others not; and several questions exhibit a high frequency of missing answers. Separately, *interpreting* such projections – even with our explanatory techniques – would be very challenging since the 111 questions address widely different topics – religion, welfare, politics, role of the state, elections, education, EU enlargement, living standards, economy, and more. As mentioned earlier, our explanatory mechanisms are designed to handle tens, but not hundreds, of dimensions.

As such, we chose the less ambitious but more focused goal of studying only one aspect of the EVS dataset, namely questions about *religious beliefs*. Table 1 shows the 21 questions on this topic and their possible answers (for full details, see [41]). From the $N = 56491$ samples, we kept to further project the $N' = 22532$ ones which contain no missing (NA) values for any of the selected 21 dimensions. We refrained from standard techniques for imputing missing values (*e.g.* based on averages or most-frequent values) as domain specialists involved with this dataset advised us against such options which, in their experience, could introduce significant biases. However, for question v53 (’Did you ever belong to a religious denomination?’), we also kept samples having NA answers since this indicates people who do not describe themselves as belonging to a religious denomination. Next, we converted categorical data to numerical data via one-hot encoding [42]. Finally, we normalized all quantitative variables to the range [0, 1] by standardization (subtract the mean, divide by standard deviation); and weighed the sets of one-hot-encodings that map one categorical variable by $1/\sqrt{2}$, so they have a proportional contribution to the total similarity function as the quantitative variables.

Results: Figure 15 shows the t-SNE projection of the EVS dataset colored by variance. Image (a) shows the overview. The projection consists of well-separated point clusters which suggest a clear grouping of the respondents based on their religion-related answers. We see some coarse-level structure: Several central groups (light blue) indicate people with no religious denomination. We also see several light-purple groups at different places on the outskirts the projection. These are people who answered similarly to *v9* (are you in a church/religious organization?) Since there are several such groups, the answers to *v9* are different (some are and some are not in such organizations) and/or other factors exist that differentiate them.

To get more insight in the projection, we select a few groups for further analysis. Image (b, red points) shows such a group to the bottom. The widget tells us that these are, compared to the dataset’s average, people more present in church organizations, who more often believe in God, heaven, hell, and the afterlife, and go to church more often. Interestingly, they have a wide spectrum of beliefs concerning the form God takes (*v62*). We can cautiously describe them as ‘institutionally religious’ people. Image (c) selects a cluster top-left in the projection. Its widget, and the earlier-observed purple color in image (a), tell us that these are also people in religious organizations. Yet, the top green bars in the widget show that, compared to the dataset average, they don’t believe in afterlife, God, heaven, and hell, but strongly believe children should have religious faith and overall believe God is important. We can describe such people as formally non-religious but supporting the ethical importance of religion. Finally, image (d) explores a cluster just right to the one in image (c). Comparing its widget with that of (c) we see that the second bar from the top (believe in reincarnation, *v61*) changes a lot: These are people who do not believe in reincarnation, while those selected in (c) did, with all their other attributes being roughly similar.

Figure 16 shows the projection explained by outlier values. Image (a) uses the same colormap as Fig. 15a. We get more insights into the projection structure: The right yellow groups share outlier answers to *v6* (whether religion is important). The middle purple groups, overlapping many of the light-blue groups in Fig. 15a, have outlier answers to *v9* (whether in a religious organization). Bottom-left, brown groups have outlier answers to prayer frequency outside religious services (*v64*). Finally, the top-left green groups have outlier answers to whether religious faith is desirable for children (*v93*).

Let us re-examine the same selected groups as in Fig. 15b-d via outlier values. Figure 16b shows that people in the bottom group pray (outside of religious services) significantly less, and

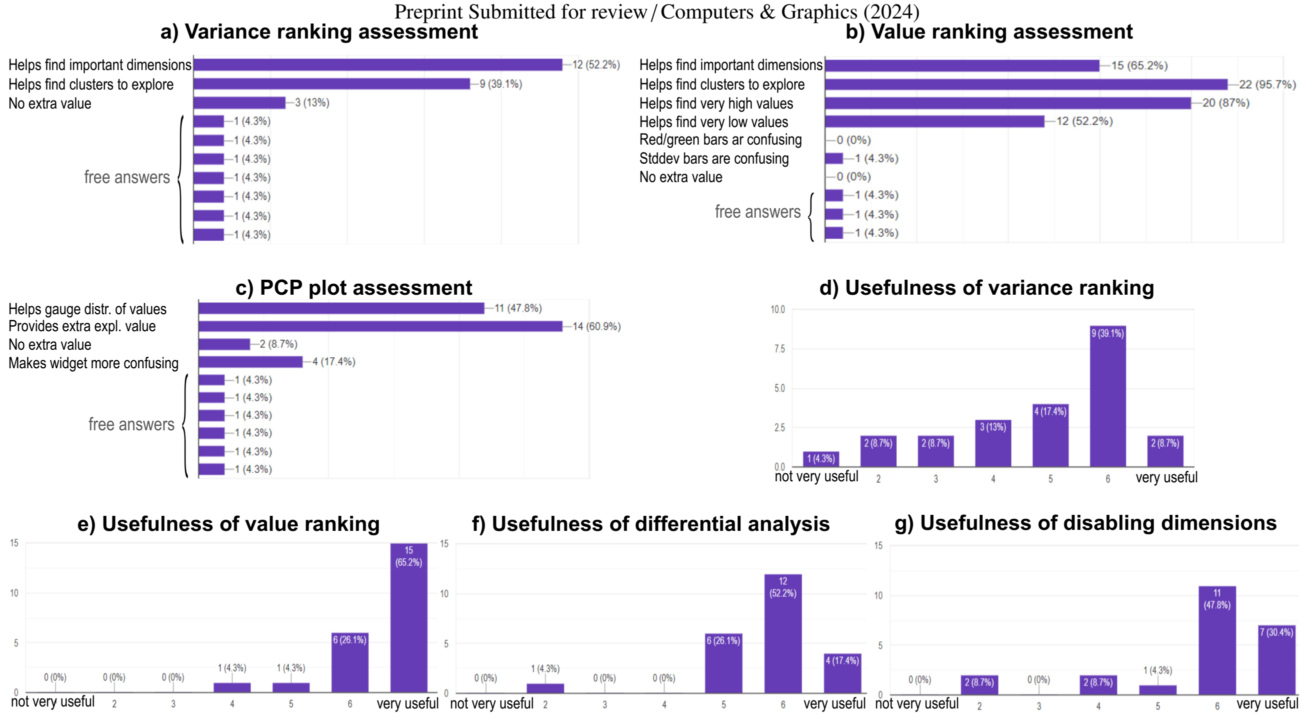


Fig. 14. Details of our user evaluation concerning questions about our techniques' overall perceived added-value. See Sec. 5.5.

Table 1. Questions and representations of answers of 21 religion-related opinions from the EVS dataset. See Sec. 6.

No.	Summary of questions in the EVS survey	Answer values	Meaning of the answer values
v6	How about the importance of religion in your life?	[1,2,3,4], [8,9]	[*very*, *quite*, *not*, *not at all*], [don't know (DK), no answer (NA)]
v9	Do you belong to a religious or church organization?	[1,2], [8,9]	[*yes*, *no*], [DK, NA]
v36	How much do you trust people from another religion?	[1,2], [8,9]	[*completely*, *somewhat*, *very much*, *trust at all*], [DK, NA]
v51	Do you belong to a religious denomination?	[1,2], [8,9]	[*yes*, *no*], [DK, NA]
v52	Which denomination do you belong to?	[1-17], [88,99,77]	[a set of 17 *denominations*], [DK, NA, not applicable]
v53	Did you ever belong to a religious denomination?	[1,2], [8,9]	[*yes*, *no*], [DK, NA]
v54	How often do you attend religious services these days?	[1-7], [8,9]	[7 degrees from *more than once a week* to *never*], [DK, NA]
v55	How often did you attend religious services when you were 12 years old?	[1-7], [8,9]	[7 degrees from *more than once a week* to *never*], [DK, NA]
v56	Would you say you are a ... person? (read out)	[1,2,3], [8,9]	[*religious*, *not religious*, *convinced atheist*], [DK, NA]
v57	Do you believe in God?	[1,2], [8,9]	[*yes*, *no*], [DK, NA]
v58	Do you believe in Life after death?	[1,2], [8,9]	[*yes*, *no*], [DK, NA]
v59	Do you believe in Hell?	[1,2], [8,9]	[*yes*, *no*], [DK, NA]
v60	Do you believe in Heaven?	[1,2], [8,9]	[*yes*, *no*], [DK, NA]
v61	Do you believe in reincarnation?	[1,2], [8,9]	[*yes*, *no*], [DK, NA]
v62	Which form do you think God takes?	[1,2,3,4], [8,9]	[*person*, *sort of spirit*, *think nothing*, *no God*], [DK, NA]
v63	How important is God in your life?	[1-10], [88,99]	[10 degrees from *not at all* to *very important*], [DK, NA]
v64	How often do you pray outside of religious services? (read out)	[1-7], [8,9]	[7 degrees from *everyday* to *never*], [DK, NA]
v93	Do you think religious faith is desirable for a child to have?	[1,2], [8,9]	[*yes*, *no*], [DK, NA]
v115	How much confidence do you have in the Church?	[1,2,3,4], [8,9]	[*great*, *quiet a lot*, *not very much*, *none at all*], [DK, NA]
v134	Democracy needs that religious authorities ultimately interpret the law.	[0,1-10], [8,9]	[*against democracy*, 10 degrees from *not at all* to *essential*], [DK, NA]
v196	To be a Christian is important for being an European person.	[1,2,3,4], [8,9]	[4 degrees from *very important* to *not at all important*],[DK, NA]

believe in spirits significantly less, than the dataset average. This matches well our earlier description of 'institutionally religious' people. Figure 16c confirms our earlier findings from Fig. 15c. The bars in the widgets of these two figures are the same. What differs is the sorting order: In variance mode (Fig. 15c), bars are sorted from low to high variance, allowing us to find the least varying, thus most homogeneous, dimensions over a selection; in value mode (Fig. 16c), bars are sorted from high to low outlierness, allowing us to find dimensions having unusually high (or low) values in a selection. The added-value of the two modes becomes clear when we examine Fig. 16d, where we selected the same group as in Fig. 15d: As explained earlier, the difference of this group and the one left of it is immediate when we compare the widgets in Figs. 16c,d – the variance sort shows the belief in reincarnation (orange dimension, second-top) changes a lot between the two widgets, telling what makes the groups different. In value mode, this dimension is the one-but-last in Fig. 16c but pops second-to-top in Fig. 16d. Hence, variance sort helps more to explain the differences of these two groups.

The scenarios involving images (c,d) in Figs. 15 and 16 aim to find what differentiates two point groups. We can complete this task also by the differential analysis tool (Sec. 4). Consider the three small groups selected in red at the center of Fig. 16e. The widget tells us that these are people not in church organizations (long green bar at top) but who, interestingly, do believe in hell and reincarnation much more than the dataset average (long red bars at the bottom). What differentiates these three groups *from each other*? To answer this, we select first the top two groups (A and B) and use the differential tool (Fig. 16f). The widget now shows a single long red bar at the bottom, telling that group B has people who believe far more in heaven than the ones in group A. All other bars are relatively short, so this belief in heaven is *the* main differentiator of these two groups. Next, we select groups B and C and use again the differential tool (Fig. 16g). The widget shows a long green and a long red bar, telling that people in cluster C believe far less in the afterlife, but believe far more often in heaven, than people in group B. Finding such differentiators between point groups would have been significantly harder without the differential tool that shows what makes them, pair-wise, different.

Assessment: We ran our findings with an expert who has a strong background in both infovis and the social sciences domain from which the dataset emerges, and was not involved in the development or testing of our tool. Our questions were (a) whether our explanatory techniques have the potential to show currently-unknown insights on these data; and (b) whether our visualization (projection plus interactive explanations) do make sense and are superior to the common tools known by experts in their domain. The answers to both questions were clearly positive: (a) The findings we highlighted earlier in this section were unknown to researchers in the field *and* were also deemed interesting and worthy of further analysis; (b) there were no similar tools known in the expert's domain which could allow researchers to explore the EVS data in the way we did – the closest tool they would know of is a (t-SNE) projection annotated by the values of a *single* dimension selected by users (which, as shown in Sec. 2 and Fig. 1, clearly does not scale to more than a few dimensions). While this evidence is not enough to draw strong conclusions, we believe it offers sufficient ground to assert that our proposal is of potential added-value to scientists aiming to explore high-dimensional datasets via explained projections.

7. Discussion

We next discuss several key aspects of our proposal.

Genericity: Our proposed explanatory methods are generic – they work for any projection technique P and high-dimensional dataset D, including data having quantitative, ordinal, and categorical attributes (see Sec. 6), as long as one has a (good quality) projection of the data to explore.

Scalability: Our explanatory methods only require the computation of variance-and-value metrics over relatively small point neighborhoods in the projection (Eqns. 1 and 5). These are $O(\kappa N n)$ for N dataset points having n dimensions and κ points in the local neighborhood of radius ρ in a projection (see Sec. 3) – and trivially to parallelize in a SIMD manner.

Ease of use: Using our explanatory techniques is easy as outlined by the presented study in Sec. 5. All our users, having quite diverse backgrounds, were able to understand our techniques and apply them to find correct results on three relatively complex datasets and questions in several tens of minutes.

Limitations: Our proposal has several limitations. First, as stated in Secs. 1 and 2, we only address *tabular* data, which contains a limited number of dimensions n (roughly, tens) that all have clear semantics for the user. If dimensions do not have a clear meaning for users, using them to explain a projection does not make much sense. A related limitation is that we cannot handle data with *missing values*. This can significantly decrease the applicability of our method to the full extent of information present in real-world datasets (see Sec. 6). While we can argue that handling missing values is out of the scope of our explanatory techniques for projections, it is definitely interesting to think how one could meaningfully 'insert' such values into a projection or, alternatively, complete the statistics shown by our explanatory widgets by all valid attributes present in a dataset. Secondly, our local explanations (Sec. 4) are also limited in showing statistics over the brushed selection – averages, ranges, standard deviations, and parallel coordinate plots. These simple to interpret signals are by no means exhaustive. Finding more involved (summarized) descriptions of what makes a neighborhood 'particular' is an open research topic. Finally, our differential tool allows comparing two neighborhoods at a time (Sec. 4). It is definitely interesting to extend this to compare multiple such neighborhoods.

8. Conclusion

We have presented a set of interactive visual techniques for the exploration and explanation of multidimensional projections. Our techniques include local and global value-based explanations, detailed statistics on all dimensions, comparing projection regions, and dimension filtering. Our techniques can generically handle any projection algorithm and scale computationally and visually to datasets of over 100K samples and over 300 dimensions. A user study showed that our techniques can be quickly learned, are found useful, and can be applied to answer non-trivial questions involving real-world multidimensional datasets, and lead to similar findings from different users for the same datasets and questions. We also showed that our techniques can be applied to complex, real-world, datasets containing attributes of mixed type – ordinal, categorical, and quantitative – to unravel hitherto unknown insights from the respective datasets.

Several directions can be explored next. Global explanations, although useful, are still limited as they inherently show a *single* dimension. Further studying the original idea proposed – but not elaborated – by Da Silva [10] to use dimension-sets, possibly complemented by dimension-value-ranges, has strong potential to improve the added value of such explanations. Separately, we could incorporate knowledge on the specific projection method used to make the explanatory metrics more insightful than using generic variance and outlier-value computations. Also, both our global and local analyses can be enhanced to support more targeted queries, *e.g.* 'show me other projection regions similar to this selected one'. Finally, deploying our tool in a long-term analysis scenario involving a real use-case and domain experts would bring additional evidence for its practical value.

Acknowledgments

We thank dr. Tamara Mchedlidze for sharing with us the EVS dataset and insights thereon. Zonglin Tian was supported in his work by the China Scholarship grant number 201906080046.

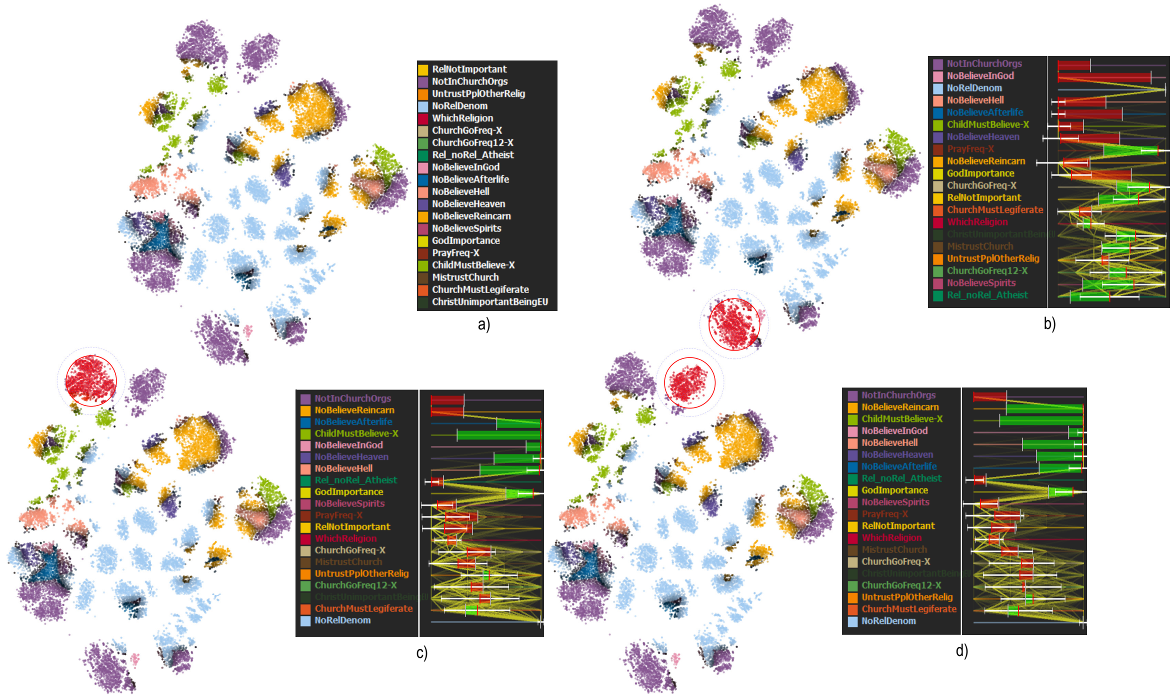


Fig. 15. Variance explanation of the EVS dataset. (a) Overview showing the main variables that explain the projection clusters. (b-d) Details for three selected clusters. See Sec. 6.

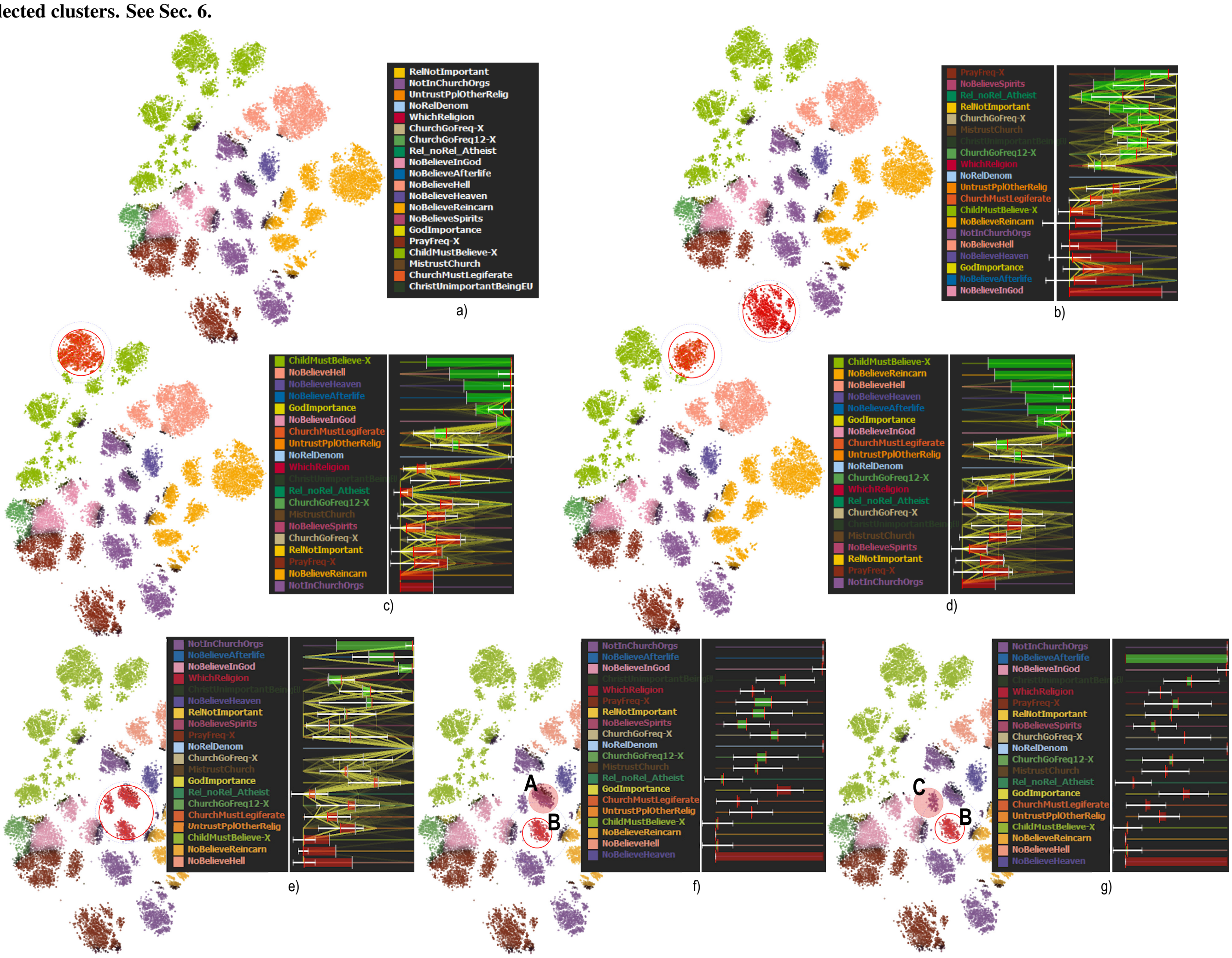


Fig. 16. Value explanation of the EVS dataset. (a) Overview showing the main variables that explain the projection clusters. (b-d) Details for three selected clusters. (e-g) Differential analysis of three small clusters in the center. Insets right of each projection show our local explanation widgets. See Sec. 6.

References

[1] van der Maaten, L, Postma, E. Dimensionality reduction: A comparative review. Tech. Rep.; Tilburg University, Netherlands; 2009. Tech. report TiCC TR 2009-005.

[2] Nonato, L, Aupetit, M. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. IEEE TVCG 2018;25(8):2650–2673.

[3] Rao, R, Card, SK. The table lens: Merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. In: Proc. ACM SIGCHI. 1994, p. 318–322.

[4] Inselberg, A, Dimsdale, B. Parallel coordinates: A tool for visualizing multi-dimensional geometry. In: Proc. IEEE VIS. 1990, p. 361–378.

[5] Elmqvist, N, P, PD, Fekete, JD. Rolling the dice: multidimensional visual exploration using scatterplot matrix navigation. IEEE TVCG 2008;(14):1141–1148.

[6] Greenacre, M. Biplots in practice. Fundacion BBVA, Bilbao; 2010.

[7] Gower, J, Lubbe, S, Roux, N. Understanding biplots. Wiley; 2011.

[8] Coimbra, D, Martins, R, Neves, T, Telea, A, Paulovich, F. Explaining three-dimensional dimensionality reduction plots. Information Visualization 2016;15(2):154–172.

[9] Broeksema, B, Baudel, T, Telea, A. Visual analysis of multidimensional categorical datasets. Computer Graphics Forum 2013;32(8):158–169.

[10] da Silva, R, Rauber, P, Martins, R, Minghim, R, Telea, AC. Attribute-based visual explanation of multidimensional projections. In: Proc. EuroVA. 2015, p. 97–101.

[11] Tian, Z, Zhai, X, van Driel, D, van Steenpaal, G, Espadoto, M, Telea, A. Using multiple attribute-based explanations of multidimensional projections to explore high-dimensional data. Computers & Graphics 2021;98:93–104.

[12] Thijssen, J, Tian, Z, Telea, A. Scaling up the explanation of multidimensional projections. In: Proc. EuroVA. 2023,.

[13] Munzner, T. Visualization Analysis and Design: Principles, Techniques, and Practice. CRC Press; 2014.

[14] Venna, J, Kaski, S. Visualizing gene interaction graphs with local multidimensional scaling. In: Proc. ESANN. 2006, p. 557–562.

[15] Martins, R, Coimbra, D, Minghim, R, Telea, AC. Visual analysis of dimensionality reduction quality for parameterized projections. Computers & Graphics 2014;41:26–42.

[16] Joia, P, Coimbra, D, Cuminato, JA, Paulovich, FV, Nonato, LG. Local affine multidimensional projection. IEEE TVCG 2011;17(12):2563–2571.

[17] Aupetit, M. Sanity check for class-coloring-based evaluation of dimension reduction techniques. In: Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization. ACM; 2014, p. 134–141.

[18] Sips, M, Neubert, B, Lewis, J, Hanrahan, P. Selecting good views of high-dimensional data using class consistency. Comp Graph Forum 2009;28(3):831–838.

[19] Tatu, A, Bak, P, Bertini, E, Keim, D, Schneidewind, J. Visual quality metrics and human perception: An initial study on 2D projections of large multidimensional data. In: Proc. AVI. ACM; 2010, p. 49–56.

[20] Espadoto, M, Martins, R, Kerren, A, Hirata, N, Telea, A. Toward a quantitative survey of dimension reduction techniques. IEEE TVCG 2019;27(3):2153–2173.

[21] Cortez, P, Cerdeira, A, Almeida, F, Matos, T, Reis, J. Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems 2009;47(4):547–553. https://archive.ics.uci.edu/ml/datasets/wine+quality.

[22] Aupetit, M. Visualizing distortions and recovering topology in continuous projection techniques. Neurocomputing 2007;10(7–9):1304–1330.

[23] Schreck, T, von Landesberger, T, Bremm, S. Techniques for precision-based visual analysis of projected data. Information Visualization 2010;9(3):181–193.

[24] Lespinats, S, Aupetit, M. CheckViz: Sanity check and topological clues for linear and nonlinear mappings. Computer Graphics Forum 2011;30(1):113–125.

[25] Martins, R, Minghim, R, Telea, AC. Explaining neighborhood preservation for multidimensional projections. In: Proc. CGVC. Eurographics; 2015, p. 121–128.

[26] Stahnke, J, Dork, M, Muller, B, Thom, A. Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions. IEEE TVCG 2016;22(1):629–638.

[27] Yin, H. Nonlinear dimensionality reduction and data visualization: A review. Intl Journal of Automation and Computing 2007;4(3):294–303.

[28] van der Maaten, L, Hinton, GE. Visualizing data using t-sne. JMLR 2008;9:2579–2605.

[29] McInnes, L, Healy, J, Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. 2018. ArXiv:1802.03426v2 [stat.ML].

[30] Hoffman, P, Grinstein, G, Marx, K, Grosse, I, Stanley, E. DNA visual and analytic data mining. In: Proc. IEEE Visualization. 1997, p. 437–441.

[31] Angelini, M, Blasilli, G, Lenti, S, Palleschi, A, Santucci, G. Effectiveness error: Measuring and improving radviz visual effectiveness. IEEE TVCG 2022;28(12):4770–4786.

[32] Pagliosa, L, Pagliosa, P, Nonato, LG. Understanding attribute variability in multidimensional projections. In: Proc. SIBGRAPI. 2016,.

[33] Joia, P, Petroneto, F, Nonato, LG. Uncovering representative groups in multidimensional projections. Comp Graph Forum 2015;34(3):281–290.

[34] van Driel, D, Zhai, X, Tian, Z, Telea, A. Enhanced attribute-based explanations of multidimensional projections. In: Proc. EuroVA. Eurographics; 2020, p. 37–41.

[35] Kelly, KL. Twenty-two colors of maximum contrast. Color Eng 1965;3(26):26–27.

[36] Tukey, JW. Exploratory Data Analysis. Addison-Wesley; 1977.

[37] Vieth, A, Kroes, T, Thijssen, J, van Lew, B, Eggermont, J, Basu, S, et al. Manivault: A flexible and extensible visual analytics framework for high-dimensional data. IEEE Transactions on Visualization and Computer Graphics 2023;.

[38] Thijssen, J, Tian, Z, Telea, A. Visual explanation system for multidimensional projections. 2023. https://github.com/JulianThijssen/ProjectionExplorer.

[39] Zhang, Y, Miller, JA, Park, J, Lelieveldt, BP, et al. Reference-based cell type matching of spatial transcriptomics data. In: bioRxiv. 2022,https://doi.org/10.1101/2022.03.28.486139.

[40] Dua, D, Graff, C. UCI machine learning repository. 2022. http://archive.ics.uci.edu/ml.

[41] EVS/WVS, . Joint EVS/WVS 2017-2022 dataset. GESIS, Köln. ZA7505 Datenfile Version 4.0.0, https://doi.org/10.4232/1.14023; 2022. doi:10.4232/1.14023.

[42] Hancock, JT, Khoshgoftaar, TM. Survey on categorical data for neural networks. Journal of Big Data 2020;7(1):1–41.