Computers & Graphics 98 (2021) 93–104

Contents lists available at ScienceDirect

Computers & Graphics

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

Special Section on EuroVA 2020

Using multiple attribute-based explanations of multidimensional projections to explore high-dimensional data

Zonglin Tian[a], Xiaorui Zhai[b], Daan van Driel[b], Gijs van Steenpaal[a], Mateus Espadoto[b,c], Alexandru Telea[a,*]

[a] Department of Information and Computing Sciences, Utrecht University, Utrecht, 3584 CC, Netherlands
[b] Bernoulli Institute, University of Groningen, Groningen, 9747 AG, Netherlands
[c] Institute of Mathematics and Statistics, University of São Paulo, São Paulo 05508-090, Brazil

ARTICLE INFO

Article history:
Received 14 November 2020
Revised 1 April 2021
Accepted 22 April 2021
Available online 7 May 2021

Keywords:
Dimensionality reduction
Explanatory techniques
High-dimensional data analysis

ABSTRACT

Multidimensional projections (MPs) are effective methods for visualizing high-dimensional datasets to find structures in the data like groups of similar points and outliers. The insights obtained from MPs can be amplified by complementing these techniques by several so-called explanatory mechanisms. We present and discuss a set of six such mechanisms that explain MPs in terms of similar dimensions, local dimensionality, and dimension correlations. We implement our explanatory tools using an image-based approach, which is efficient to compute, scales well visually for large and dense MP scatterplots, and can handle any projection technique. We demonstrate how the provided explanatory views can be combined to augment each other's value and thereby lead to refined insights in the data for several high-dimensional datasets, and how these insights correlate with known facts about the data under study.

1. Introduction

Multidimensional Projections (MPs) are among the methods of choice for visualizing high-dimensional data, as they scale well in terms of the number of data points and data dimensions that they can show on a given screen space. They are useful in exploring the data structure, specifically in identifying similar sets of points and outlier points. However, understanding what, in terms of data values, ranges, or relations between dimensions, makes these structures appear in the projection (and thus, in the data) is not trivial. Several mechanisms exist to this end, as follows. *Global* explanations, such as biplot axes [1,2] and axis legends [3,4] show how dimensions influence an entire projection, and as such cannot, in general, explain the formation of local patterns like clusters. Linked views and tooltips show *local* explanations, but require one to manually select structures of interest in the projection [5–7]. *Image-based* techniques [8–10] display local explanations everywhere on the projection, not requiring one to select specific point subsets. They scale well visually and computationally, are clutter-free, and can generically handle any high-dimensional dataset.

Da Silva *et al.* [11] proposed an image-based explanation that colors every projection point by the dimension that contributes most to the similarity of data points in that neighborhood. Previous work [12] extended this approach with additional explanations. First, principal component analysis (PCA) is used to analyze point neighborhoods to deduce and depict the local (intrinsic) dimensionality of the data. This allows users to separate regions of high intrinsic dimensionality in the projection (hard to explain by a few dimensions) from low-dimensionality regions where such explanations are feasible. Secondly, point neighborhoods are analyzed to detect and depict strong linear relationships between dimensions. These techniques complement existing mechanisms for projection explanation, can be computed efficiently on the GPU, and can be applied generically on any high-dimensional dataset visualized by any MP technique.

The joint work in [12] and [11] offers five explanatory views (distance contribution, variance, dimensionality, correlation) to explore MPs, arguing that more explanations would provide more insights in the data. Yet, the work in [12] offers a single example of a non-synthetic dataset where only two views are combined to extract insights. How the five views can be combined, in practice, to explore real-world data, and how the obtained findings match ground-truth information about such data, are open questions. Also, the parameters of the five views are not discussed in

* Corresponding author: Tel.: +31-30-253-4170.
E-mail address: a.c.telea@uu.nl (A. Telea).

https://doi.org/10.1016/j.cag.2021.04.034

detail. In this paper, we refine and extend this previous work with the following contributions:

- We provide additional examples of how the five explanatory views in [12] and [11] can be combined in a visual analytics fashion to find relevant insights in high-dimensional datasets that cannot be found using a single view;
- We illustrate the above process on five non-synthetic datasets, and correlate the obtained insights with ground-truth information independently extracted by other researchers from three of these datasets;
- We present a new method, variance ratio, for computing local dimensionality;
- We discuss how our explanatory views depend on their parameter settings and on the used projection techniques.

The structure of this paper is as follows. Section 2 presents related work. Section 3 details the five explanatory views [11,12] and presents a new method for computing local dimensionality. Section 4 shows how the total set of six views can shed insights on projections of non-synthetic datasets, which we next correlate with available ground-truth information. Section 5 discusses our techniques. Section 6 concludes the paper.

2. Related work

We start introducing a few notations. Let $D = \{\mathbf{x}_i\} \subset \mathbb{R}^n$, $1 \leq i \leq N$, be a n-dimensional dataset with points $\mathbf{x}_i = (x_i^1, \ldots, x_i^n)$, also called samples or observations. We call the vectors $\mathbf{X}_j = (x_1^j, \ldots, x_N^j)^T$, $1 \leq j \leq n$, the dimensions of D, also known as variables or attributes. Hence, D can be seen as a matrix of N rows (samples) and n columns (dimensions). A *projection* is a function $P : D \to \mathbb{R}^m$, $m \ll n$, which maps a high-dimensional point $\mathbf{x}$ to a low-dimensional one $P(\mathbf{x})$. In practice, $m \in \{2, 3\}$, so projecting an entire dataset D, denoted by $P(D) = \{P(\mathbf{x}) | \mathbf{x} \in D\}$, yields a 2D or 3D scatterplot. Projections aim to place points that are similar in D close to each other in $P(D)$ to enable users to recover the structure of D from the scatterplot $P(D)$. Similarity can be computed based on $\mathbb{R}^n$ distances [6,13,14] or $\mathbb{R}^n$ neighborhoods [15,16]. Recent surveys provide more details on the technicalities of MPs [17,18]. In our work next, P can be any projection technique chosen by the user as desired or demanded by one's application context.

Explanatory techniques for projections aim to enrich the bare scatterplot $P(D)$ with additional information that guides the user in interpreting $P(D)$. We classify such techniques in observation-centric, dimension-centric, and hybrid, as follows.

2.1. Observation centric explanations

These techniques aim to provide information about specific projection *observations* $P(\mathbf{x})$. Many such techniques aim to show the errors produced by the projection function P measured by *e.g.* normalized stress [6,10], correlation [19], Shepard diagrams [6], trustworthiness [20], continuity [20], neighborhood hit [21], distance consistency [22], ranking discrepancy [23,24], projection precision score [9], stretching and compression [8,25], and class consistency metrics [26]. Continuity and trustworthiness are closely related to the so-called missing neighbors, respectively false neighbors, of a projected point $P(\mathbf{x})$ [10]. For a recent survey that discusses most above metrics, we refer to [17].

Error metrics can be computed at three aggregation levels. *Global* errors generate a single (scalar) value for an entire scatterplot $P(D)$, so they help gauging the quality of such a scatterplot, but do little in explaining it. *Point pair* errors quantify the projection error of a point pair $(P(\mathbf{x}), P(\mathbf{y})) \in P(D) \times P(D)$ and can be rendered as Shepard diagrams [6] or line plots simplified by edge bundling [10]. *Point neighborhood* errors quantify the projection error of a point $P(\mathbf{x}) \in P(D)$ with respect to all its neighbors in $P(D)$ or, alternatively, all neighbors of $\mathbf{x} \in D$. These are further visualized using heatmaps [9,10] or Voronoi diagrams [8,25], thereby informing the user about projection problems at the location of every scatterplot point. This further assists one in determining where, and how much, one can trust a projection. However, such techniques cannot explain *why* certain points are projected close to each other (or not).

2.2. Dimension centric explanations

These techniques show how the *dimensions* $\mathbf{X}_j$ of a dataset D relate to the scatterplot. The simplest, and still most used, dimension centric explanation colors a scatterplot by the values of a selected dimension $\mathbf{X}_j$. This explains specific groups of points in the scatterplot by that dimension's values. Several dimensions can be used via interaction or small multiples. Yet, this approach cannot easily handle more than a few dimensions, leaving their selection to the user. Biplot axes [1,2] involve all dimensions in the explanation by drawing n lines atop of the scatterplot $P(D)$, each indicating the embedding of one of the dimensions $\mathbf{X}_j$ in the projection space $\mathbb{R}^m$. Axis legends [3,27] take a different route, by explaining how the n dimensions map to the 2D scatterplot's x and y axes using bar charts. Both biplots and axis legends have been generalized to explain also 3D projections and nonlinear projections [4].

All above dimension centric explanations act as generalizations of the classical axis labels present in 2D Cartesian scatterplots – that is, they allow users to see which are the values of one or multiple dimensions that determine the *overall* projection shape. However, they do not *explicitly* connect the explanations to individual scatterplot points or point groups, leaving this to be done (visually) by the user. In contrast, observation centric techniques explicitly mark individual points by the provided explanations (*e.g.*, errors); however, such techniques do not involve dimensions in the explanation.

2.3. Hybrid explanations

Hybrid techniques aim to join the strengths of observation centric and dimension centric ones. The simplest form involves brushing points to show their attributes in a tooltip. More involved techniques involve interactively selecting and/or modifying specific *points* S in the projection. By next arranging $P(D) \setminus S$ around S, one can explain $P(D) \setminus S$ in terms of (known) attribute values of S. The VIBE system [28] allows selecting and placing points of interest (POIs) in the 2D projection space according to one's mental map of how the respective data samples relate to each other. The remaining data points are projected based on their similarity to POIs. A similar approach is proposed in [6] and by the Force-SPIRE text visualization system [29]. The "dust and magnets" technique [30] extends these interaction metaphors by allowing users to interact with both POIs and data points, using animation to map the data-to-POI similarities. Interaction also supports navigating through a space of 2D scatterplots (whose axes are directly explained by their dimensions) created from the high-dimensional data [31,32]. Pagliosa *et al.* propose a 'projection inspector' that offers several such interactive exploratory mechanisms. Interactive techniques are very powerful in providing 'details on demand' (on both observations and dimensions) to the user. However, they require interaction effort, and also cannot explain an *entire* projection, but rather the point(s) interacted with.

Image-based techniques, also known as dense maps, are a different hybrid approach. These rasterize the 2D projection space $\mathbb{R}^2$ and synthesize, for each pixel $\mathbf{p}$, an explanation based on the points in $P(D)$ nearest to $\mathbf{p}$. This space-filling approach allows a

large amount of information to be conveyed; and removes issues of observation-centric techniques caused by overlapping points in $P(D)$. Da Silva *et al.* [11] create dense maps where pixel hues encode the dimension that best explains the similarity of points in $P(D)$ close to each pixel, and brightness encodes the explanation confidence. Van Driel *et al.* [12] extend this technique with explanations of the local dimensionality of data and dimension correlations. We detail both above techniques in Section 3.

Dense maps have been used to explain projection errors [9,10,25]). Rodrigues *et al.* used dense maps to visualize the decision zones of classifiers of high-dimensional data [33]. Like us and [11,12], they also use pixel hues and luminances to encode a classifier's decision, respectively decision confidence, at a data point $\mathbf{x}$ mapping to a pixel $P(\mathbf{x})$. Our goals are different, as we aim to explain a dataset in terms of its *dimensions*, rather than a classifier in terms of its *decisions*.

3. Explanatory mechanisms

The image-based explanatory techniques introduced in Section 2.3 exploit the distance or neighborhood preservation property of MPs: Let $\nu_i \subset P(D)$, $\nu_i = \{\mathbf{y} \in P(D) | \ \|\mathbf{y} - \mathbf{y}_i\| \leq \rho\}$, be a neighborhood of size ρ of scatterplot points $\mathbf{y}$ centered at $\mathbf{y}_i$. Since points in ν_i are, by construction, close, and since P is expected to (reasonably) preserve similarities, the points $\mu_i \in D$ that project to ν_i are expected to be similar. Hence, it makes sense to compute an *explanation* of μ_i and next visually encode this on all scatterplot points $\mathbf{y}_i$.

Da Silva *et al.* [11] propose two such explanations. Let $\lambda^j_{\mathbf{x},\mathbf{x}'} = \|\mathbf{x}^j - \mathbf{x}'^j\|_1^2 / \|\mathbf{x} - \mathbf{x}'\|_n^2$ be the contribution of dimension j to the distance between two points $\mathbf{x}$ and $\mathbf{x}'$ in D, where $\|\cdot\|_k$ is Euclidean distance in $\mathbb{R}^k$. This point-pair contribution is extended to neighborhoods μ_i by averaging the local contributions of $\mathbf{x}_i$ and all its neighbors, as $\overline{\lambda}^j_i = \sum_{\mathbf{x} \in \mu_i} \lambda^j_{\mathbf{x},\mathbf{x}_i} / |\mu_i|$, where $|\cdot|$ denotes set size. These average contributions are next normalized as

$$\lambda^j_i = \frac{\overline{\lambda}^j_i / \gamma^j}{\sum_{j=1}^n \left(\overline{\lambda}^j_i / \gamma^j \right)}, \tag{1}$$

where the normalization γ^j is the contribution $\overline{\lambda}^j$ of dimension j of the full dataset D with respect to its centroid. Since normalized, $\lambda^j_i \in [0, 1]$, with lower values telling dimensions that contribute *little* to distances in μ_i, *i.e.*, explain well why points in μ_i are *similar*. An alternative to Eq. 1 is to compute the relative variance ω^j_i of dimension j over the neighborhood μ_i as

$$\omega^j_i = \frac{LV^j_i / GV^j}{\sum_{j=1}^n (LV^j_i / GV^j)}, \tag{2}$$

where LV^j_i is the variance of dimension j for all points in μ_i, normalized by the variance GV^j of the same dimension j over all points in D. Just as λ^j_i, $\omega^j_i \in [0, 1]$, with lower values telling dimensions that vary little in a neighborhood.

The scatterplot $P(D)$ is explained by color-coding its points by the C dimensions that have overall low values of λ^j_i (or ω^j_i, depending on the user's choice) over all points. C is set to a low value, *e.g.* 8, since categorical colormaps should be small. Luminance is used to encode the *confidence* in the visual explanation: If j is the dimension picked to color point i, confidence κ is computed as the sum of λ^j_i (or ω^j_i) values for all points in the neighborhood μ_i, normalized by the sum of the same terms over *all* dimensions over μ_i. If neighbors of point i are best explained by the same dimension j as i, the color will appear bright, and conversely. We render the scatterplot by drawing radial splats of R pixels radius, textured with color and luminance computed as above, and using a opacity (alpha) varying from fully opaque in the center to slightly transparent at the borders, to smoothly blend neighbor splats. Setting R is discussed further in Section 5.

Table 1
Definitions of local dimensionality and confidence.

Definition	Dimensionality δ	Confidence κ
Total variance	$\min \delta \left\| \frac{\sum_{i=1}^{\delta} \alpha_i}{\sum_{i=1}^{n} \alpha_i} \geq \theta \right.$	$1 - \frac{\sum_{i=1}^{\delta} \alpha_i - \overline{\alpha}}{\sum_{i=1}^{n} \alpha_i}$
Minimal variance	$\left\| \left\{ \frac{\alpha_i}{\sum_{j=1}^{n} \alpha_j} \geq \theta, 1 \leq i \leq n \right\} \right\|$	$\frac{\sum_{i=1}^{\delta} \alpha_i}{\sum_{i=1}^{n} \alpha_i}$
Variance ratio	$1 + \min \delta \left\| \frac{\sum_{i=1}^{\delta} \Delta\lambda_i}{\sum_{i=1}^{n} \Delta\lambda_i} \geq \theta \right.$	$1 - \frac{\sum_{i=1}^{\delta} \Delta\lambda_i}{\sum_{i=1}^{n} \Delta\lambda_i}$

Fig. 1a,b show a 3K point dataset spread over three faces of an axis-aligned cube (with added noise), projected with PCA to 2D, explained by dimension contribution, respectively variance. Points on each cube face share very similar values of a dimension, so are bright and colored by the respective dimension. It is important to see that these are the original data dimensions (x, y, z), and not latent dimensions synthesized by PCA (eigenvectors). Points along cube edges are dark, since two (or three, for the cube corner) dimensions are needed to explain their similarity with neighbors. Hence, their color coding in the visualization and corresponding legend. Although these two explanations are practically identical for the cube dataset, we will see later on that they can subtly differ, thus both bringing in added value in the projection understanding process.

3.1. Adding dimensionality explanation

Da Silva *et al.*'s explanations (Eqs. 1 and 2) cannot provide full insights into the structure of high-dimensional data. Take *e.g.* a non-axis-aligned cube like in Fig. 1a and embed it into a high-dimensional space. While the data structure stays the same, both distance contributions and variances cannot select a single dimension to explain the cube's faces, since all dimensions contribute to the data structure.

We improve this by explaining the data's *local* (or intrinsic) *dimensionality*. For each neighborhood μ_i of a point $\mathbf{x}_i \in D$, we compute the n eigenvalues α_i of its covariance matrix, sorted decreasingly. From these, we compute the local dimensionality δ of μ_i and its confidence κ in three different ways (see also Table 1).

Total variance (TV): We define dimensionality δ as the minimal number of largest eigenvalues $\alpha_1 \geq \ldots \geq \alpha_\delta$ needed to explain a user-set fraction θ of the data variance in μ_i. The confidence κ equals how much the sum of these largest δ eigenvalues deviates from the mean of all n eigenvalues.

Minimal variance (MV): The TV model works well when eigenvalues significantly drop. However, take the (limit) case where all eigenvalues are equal. TV then computes $\delta = \theta / n$, even though locally the data is truly n-dimensional. To capture this, we define δ as the number of eigenvalues larger than a minimal user-set variance θ, and confidence κ as the sum of these divided by TV, similar to Kaiser's criterion used in explanatory factor analysis [34,35].

Variance ratio (VR): Several metrics are known in 3D diffusion tensor analysis to describe the shape of local neighborhoods [36]. We generalize these to nD data and compute dimensionality δ by summing differences of consecutive eigenvalues $\Delta\lambda_i = \lambda_i - \lambda_{i+1}$ normalized by the largest one, λ_1. Each difference captures a significant 'drop' in consecutive eigenvalues, and the sum accounts for the effect of all drops. Thresholding this sum by a user-set θ yields the local dimensionality. Besse and Falguerolles [37] and North *et al.* [38] save described similar models of local dimensionality. Note that, in the definition of δ for VR (Table 1), we define

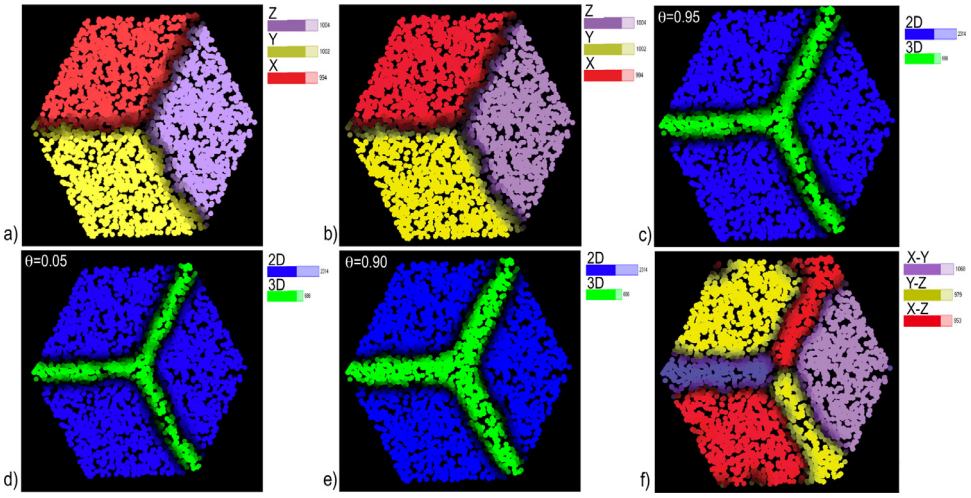

Fig. 1. Explanatory techniques illustrated on a synthetic cube dataset. The (a) dimension contribution and (b) variance color points by the dimension (X, Y, Z) that makes them most similar to their neighbors. The local dimensionality with total (c), minimal (d), and variance ratio (e) color points by their local intrinsic dimensionality (2D or 3D). The (f) dimensions correlation colors points to indicate the strongest-correlated dimension pair (X-Y, Y-Z, X-Z) close to each point. Bars in the legends show the number of points explained by each dimension (a,b), dimensionality (c,d,e), and dimension pair (f). See Section 3.

$\lambda_{n+1} = 0$. Also, if $\lambda_1 > \theta$, we set $\delta = 1$; if $\lambda_1 < \theta$, we set $\delta = 0$ (the whole dataset is concentrated in a single n-dimensional point).

Figs. 1c-e show the total, minimal, and variance ratio explanations for the noisy cube. The thresholds θ are listed in the figure and discussed next in Section 5. The explanations are color-coded on the projection points, as detailed in the legends. The legend bars' sizes tell how many points are assigned a given explanation (dimensionality). The cube's faces are blue, meaning that these points are locally in $\delta = 2$ dimensional neighborhoods embedded in nD. Close and on the cube edges, green tells that $\delta = 3$ dimensions are needed to explain the data here. The blue and green area are separated by (thin) dark bands, indicating projection areas where the confidence of assigning a dimensionality of $\delta = 2$ or $\delta = 3$) is low – these are the transition areas between the blue ($\delta = 2$) and green ($\delta = 3$) areas. The three local dimensionality explanations are very similar to each other, indicating that the PCA-based analysis underlying all three computations makes sense. For more complex datasets, the explanations can slightly differ and convey interesting insights, similar to the differences between the distance contribution and variance explanations discussed earlier (see Fig. 1a,b).

3.2. Adding correlation explanation

High-dimensional data is often explained by how its dimensions *correlate*. Yet, assessing *global* correlation over an entire dataset is of limited value when the underlying phenomenon is a mix of local (linear) patterns. To address this, we compute and depict correlations over neighborhoods. For each point neighborhood μ_i, we compute the $K = n(n+1)/2$ Pearson or Spearman correlations between all dimension-pairs $(j,k) \in [\![1,n]\!] \times [\![1,n]\!]$. We sort these pairs in descending correlation-strength order, and select the C top-ranked pairs that are most frequent over all points i. This resembles selecting the explaining dimensions in [11], but now we select dimension-pairs rather than individual dimensions. We show these C pairs via a categorical colormap, using luminance to map the absolute correlation values. Fig. 1f shows this for the noisy cube. The legend tells that the three faces map to strong correlations of the three dimensions x, y, and z, as expected. The edges orthogonal to faces show the same correlation. Indeed, for the face xy, for instance, the orthogonal edge has near-constant x and y, and strongly varying z, values, so x and y are correlated along it.

This visualization can only show the C top-ranked, most frequent, correlations from all possible K ones. However, users may want to examine the presence (or absence) of *specific* correlations. For this, we show the entire set of K dimension-pairs using a *matrix view*. To illustrate how this works, we consider next a real, non-synthetic, dataset example.

3.3. Concrete dataset

This dataset [39,40] has 1030 samples measuring how 8 ingredients influence concrete strength. The independent dimensions are cement, blast furnace slag (*BFSlag*), fly ash water (*FlyAsh*), superplasticizer (*Splastic*), coarse aggregate (*Caggr*), fine aggregate (*Faggr*), each in kg per cubic meters; and the concrete age, measured in days. One is interested to understand which independent dimensions influence concrete strength.

Fig. 2 a shows the matrix view next to the t-SNE projection of this dataset. Matrix cells are colored by the same colormap as used in the projection. Dark blue tells all dimension-pairs whose correlations have a frequency higher than zero but lower than the C top-ranked pairs. To see where, on the projection, a pair correlates, the user clicks a dark blue cell, *e.g.* the *FlyAsh-Caggr* one in Fig. 2a. The color used for the C^{th} top dimension-pair (*Water-Caggr*, cyan) is then used for the clicked pair and the C^{th} pair is made dark blue. Doing this shows a single cyan spot in the projection (Fig. 2b, dashed circle) – the only place where *FlyAsh* and *Caggr* strongly correlate.

The matrix view supports two other tasks. The cells of the top C (strongest correlated) dimension-pairs are outlined in white, helping one to easily return to the original color mapping after having selected some other dimension-pairs to explain. Rows and columns having many cells with the non-default (dark blue) color tell *groups* of strongly correlated variables. For instance, the second top row in Fig. 2a, for the *Faggr* dimension, shows four such cells that indicate *Faggr*'s strong correlation with *Cement* (yellow), *BFSlag* (green), *FlyAsh* (orange), and *Caggr* (purple), respectively.

Da Silva [41] also used this dataset, also projected with t-SNE, to find attributes that predict high concrete strength. For this, they colored the projection by each of the 8 independent dimensions, and next by the dependent dimension (concrete strength). Fig. 2b (same as Fig. 5.10 in [41]) shows the dependent dimension, allowing one to find two high-concrete-strength clusters. By manually

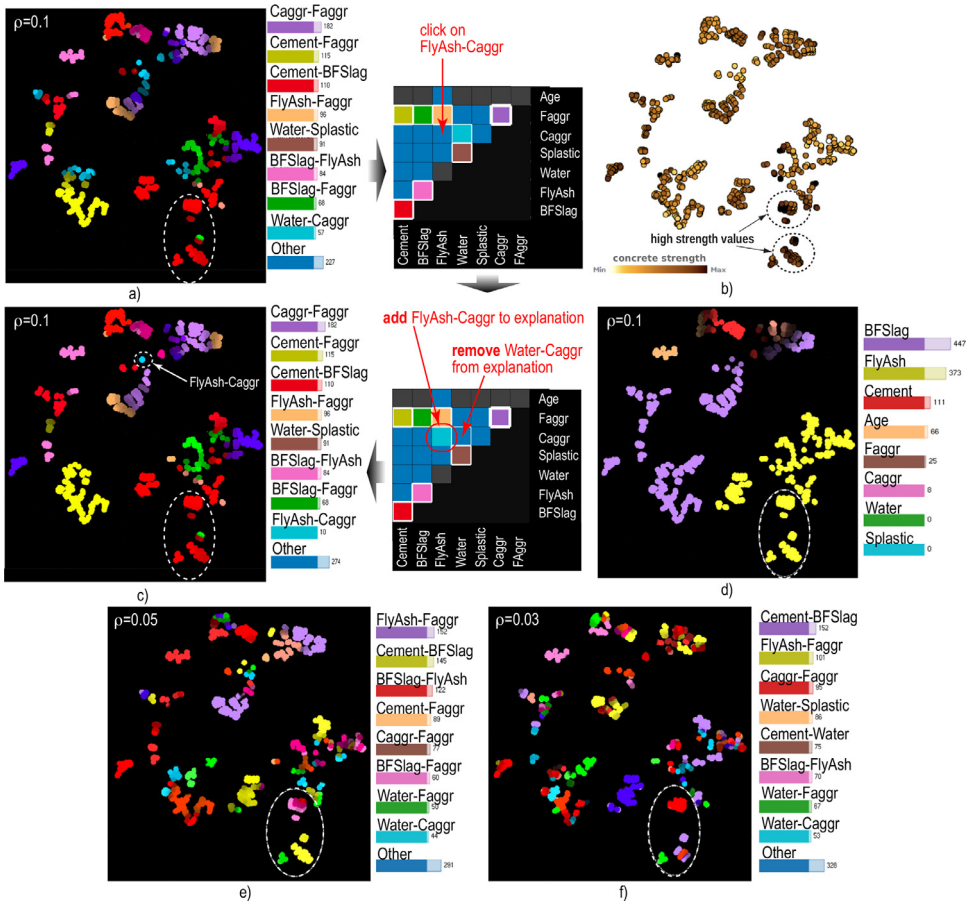

Fig. 2. Matrix view, *concrete* dataset. Clicking on the *FlyAsh-Caggr* cell (a) allocates a color to it, showing where in the t-SNE projection these two variables are strongly correlated. To make room for this, the weakest-correlated pair *Water-Caggr* is removed from the explanation (c) Additional insight is obtained by color-coding the dependent dimension (c), the variance explanation (d), and correlation views using smaller neighborhood sizes ρ (e,f). See Section 3.3.

comparing the values of all independent dimensions over these clusters, Da Silva found that *BFSlag* also had high values in these areas. However, this manual comparison of color-coded dimensions is quite tedious.

We next show how our explanatory views help refining the above insights. In Fig. 2a,c, we see a correlation between *cement* and *BFSlag* attributes in the selected region. Now, if *cement* and *BFSlag* correlate with each other, and *BFSlag* correlates with high concrete strength, *cement* likely correlates to concrete strength as well. To search for additional correlations over subsets of points in the selected region (smaller neighborhoods), we next decrease the radius ρ used to compute the correlation view. In Fig. 2e, computed with $\rho = 0.05$, we see a *BFSlag-Faggr* correlation (pink upper cluster), and also a *water-Faggr* correlation (green lower cluster). Also, the *cement-BFSlag* correlation stays strong in the middle (yellow) cluster. In Fig. 2f, computed with $\rho = 0.03$, we see the *cement-BFSlag* and *water-Faggr* correlations in the purple, respectively green, clusters; the red upper cluster shows an additional *Caggr-Faggr* correlation. Now, because *BFSlag* was found to correlate with *Faggr* in this region, *Faggr* might be related to high concrete strength (especially in combination with large *BFSlag* values). And because *Faggr* might be correlated, and we found a *water-Faggr* correlation and a *Caggr-Faggr* correlation, both *water* and *Caggr* might explain high concrete strength.

We now use the variance view (Fig. 2d) to get extra insights in the selected region. The entire region is yellow, *i.e.*, points there have a small *FlyAsh* variance. Also, *FlyAsh* varies little also *far beyond* the region borders. Putting it all together: *BFSlag*, *cement*, *Faggr*, *water*, and *Caggr* (but not *FlyAsh*) might together help shaping a regressive model for high concrete strength. Wu *et al.* [42] independently studied this dataset for of predictive modeling, showing the Pearson correlation coefficients between the data attributes (Table II in [42]). They found a relatively strong positive *cement-BFSlag* correlation (0.29), inverse correlations of *BFSlag-Caggr* (-0.31) and *BFSlag-Faggr* (-0.31), and an inverse *Faggr-water* correlation (-0.44). Our findings, obtained via our correlation views, are consistent with these results – except that we do not visualize the *sign* of the correlation.

3.4. Parameters

Our explanations depend on the following user parameters:

Neighborhood size: Given as a fraction of the projection size (so $\rho \in [0, 1]$), ρ tells the *scale* of the visual structures we want to explain. Fig. 4 illustrates this for the variance explanation of the *wine* dataset. Smaller ρ values explain finer-grained structures, but can create noisy visualizations, since, in the limit, every (small) neighborhood can be potentially best explained by a different dimension; since we usually do not have as many categorical colors as the dataset's number of dimensions n, many such neighborhoods will not receive an explanation (see Section 3). Large ρ values will attempt to explain large visual structures by a single dimension, which, in the limit, when ρ equals the projection's size, amounts to showing the dimension having globally least variance, which is not insightful. Good values for ρ range around 0.1 of the projection's size. This is the default value used in all the views in this paper unless otherwise specified. Indeed, for a dataset having a

few thousand samples, this ρ value yields a few tens of samples per neighborhood v_i, which is sufficient, as a lower bound, to reliably compute all the proposed explanations.

Dimensionality threshold: The value $\theta \in [0, 1]$ (Table 1) specifies how much of the data's local dimensionality we want to explain. For TV and VR, a high θ value explains more of the local dimensionality, but can lead to projections where most points are marked as high-dimensional, which is not very useful. A too low θ value can generate false confidence that the 2D projection captures all the intrinsic dimensionality of the data. For MV, θ behaves oppositely – low values explain more of the intrinsic data dimensionality. We empirically found that $\theta \in [0.6, 0.9]$ (for TV and VR), respectively $\theta \in [0.05, 0.1]$ (for MV) yield an informative, but not too strict, visualization.

Splat radius: The value R gives the size, in pixels, of the splats that render the explanation and its confidence (Section 3). Small R values create discrete-looking scatterplots, where the colors of neighbor points do not visually merge, thereby breaking the color-and-luminance gradients which are key to explaining *regions* in the scatterplot. High R values create too much overlap between neighbor points, so regions smaller than R cannot be visually distinguished. R and the neighborhood radius ρ act as dual scale parameters – ρ controls the scale at which we *compute* explanations, and R controls the scale at which we *render* them. We studied several options of setting R automatically, *e.g.*, based on the average local density of scatterplot points, following similar work in [10]. We found such automatic methods risky, as they tend to indiscriminately 'fill in' gaps of all sizes in a projection, including those which separate faraway point clusters. Hence, we leave R as a parameter for the user to set. A good preset for R is the average distance-to-the-closest-neighbor in the projection, which amounts to $\rho \in [0.03, 0.05]$ of the image size for the figures in this paper.

4. Applications

We show next how the six explanatory views – distance contribution, variance, correlation, and local dimensionality computed by total variance, minimal variance, and variance ratio – can be combined to extract insights from four non-synthetic datasets. We also correlate these insights with ground truth extracted by independent research that studied the same datasets.

4.1. Wine quality dataset

We first consider the *wine* dataset, which has 6497 samples of Portuguese *vinho verde* [43], each with $n = 12$ physicochemical attributes such as acidity, residual sugar, and alcohol rate. Fig. 3a shows the raw projection of this dataset using LAMP [6]. Besides a dense-point cluster bottom-right, there is not much else this image tells us. While other projection methods, *e.g.* t-SNE, may show better separated clusters, the question still remains how to explain these.

Fig. 3b-c show the contribution and variance explanations respectively. These are quite similar and split the projection roughly into four areas, explained by small variations of alcohol (purple), chlorides (yellow), sugar (red), and acidity (beige), respectively. The correlation view (Fig. 3d) brings additional insights: We see a large purple area bottom-right that matches well the area earlier explained by small variations of chlorides, alcohol, and acidity. Over this purple area, the legend of image (d) tells that sugar and density strongly correlate. Also, we see that the red area in Figs. 3b-c, where sugar has a low variation, is now roughly split in Fig. 3d into smaller areas – red (fixed acidity-citric acid correlation), yellow (fixed acidity-pH correlation), beige (fixed acidity-density correlation), and brown (chlorides-density correlation). Note that the contribution-variance and correlation explanations are *complementary*: They cannot, when taken separately, split the projection into fine-grained local explanations, but do so when *combined*. Indeed, the red area in Figs. 3b-c is further split (explained) by using correlation, as explained above; conversely, the purple area in Fig. 3d is further split (explained) by using contribution or variance.

At this point, the analyst may wonder which projection areas are *sufficiently* explained by the above views. The dimensionality view helps here. Fig. 3e shows the local dimensionality of the projected data, computed by total variance (Section 3.1). We see how increasingly more dimensions are needed to capture increasing fractions $\theta \in [0.3, 0.9]$ of the total variance – in the limit, we need all $n = 12$ dimensions to explain $\theta = 100\%$ of the variance. More interestingly, we see in Fig. 3e a gradient of local dimensionality, from highest in the bottom-right area (red-purple colors for $\theta \geq 0.85$) to blue in the top-left area (blue for $\theta \leq 0.75$). Besides color hue, the local dimensionality gradient is also visible in the brightness, which tells the confidence κ that the color-coded number of dimensions locally explain θ percent of the variance. The effect is very similar to the enridged contour maps used to visualize scalar fields [44]: The visual nesting of the 'cushions' created by varying brightness conveys the absolute value of the encoded signal, *i.e.*, the local dimensionality. The way we compute these cushions (Section 3.1) is, however, completely different to [44].

The local dimensionality view helps interpreting the contribution-variance and correlation views as follows: As we have seen, local dimensionality is high in the bottom-right (red-purple) area, where we need 7 to 9 dimensions to explain $\theta = 0.85$ of the data variance. In this area, the contribution-variance and correlation views jointly give us information about only *five* variables – alcohol, chlorides, acidity, sugar, and density. Hence, these two views do not *fully* explain this area, so we need to search for more explanations here. In contrast, the local dimensionality is low in the top-left (blue) area, where we can explain $\theta = 0.75$ of the data variance by a single dimension. From the contribution-variance views, we see that this area is well explained by a small variance of sugar. Hence, in this area, sugar's low variance is sufficient to explain the data.

Fig. 3f shows the local dimensionality computed by VR as opposed to TV (Fig. 3e, for the three largest θ values. While the exact borders of the explained regions differ, we see overall the same pattern, *i.e.*, low dimensionality to the left, respectively high dimensionality to the right, of the projection. The insights described above – obtained with TV dimensionality – stay the same. The actual dimensions assigned to comparable regions in the two explanations are similar – for instance, the blue areas in Fig. 3f ($\theta = 0.75$), of local dimensionality 1 and 2, match well the blue-and-green areas in Fig. 3e ($\theta = 0.75$) which are also of dimensionality 1 and 2.

Beh and Holdsworth [45] studied this dataset by correspondence analysis, multiple regression analysis, classification, and visual evaluations. Using the classification technique of Cortez *et al.*,[43], they examined the mean value of each attribute for the classification as scored by assessors. They found a relationship between low sugar, density, fixed acidity and volatile acidity, and higher-quality white wine. Also, stronger values of alcohol, pH and sulfur are implied to lead to higher-quality wine. For red wine, high levels of alcohol and sulfur are also found to be a strong quality indicator, while low chloride levels can lead to higher quality red wine. Residual sugar and density are found to be statistically irrelevant in predicting red wine quality. If we compare Fig. 3 to these findings, checking for value ranges by brushing the projection, we find several matches: The high-quality wines (brown area, Fig. 3b) have indeed high sulfur (brown area, Fig. 3c) and are in a region of high sugar-density correlation (both these attributes having low values, confirmed by brushing – purple area,

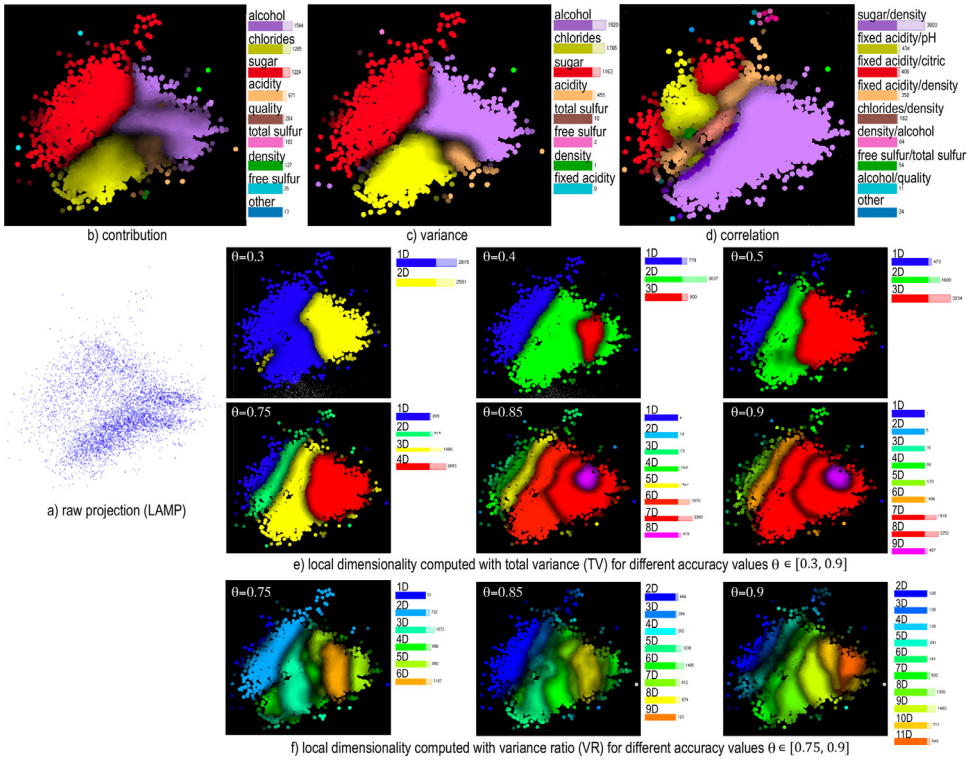

Fig. 3. Explanation of *wine* dataset. The contribution and variance views (b,c) split the projection in four main clusters, characterized by similar values of sugar (red), chlorides (yellow), and alcohol (purple). The correlation view (d) further explains the yellow and red clusters by the correlation of sugar with density (similar interpretations exist for the red cluster). The dimensionality views (e,f) tells that the blue area, which falls inside the red zone in (b,c), can be explained by a single dimension, which is thus the earlier-identified sugar dimension. See Section 4.1.

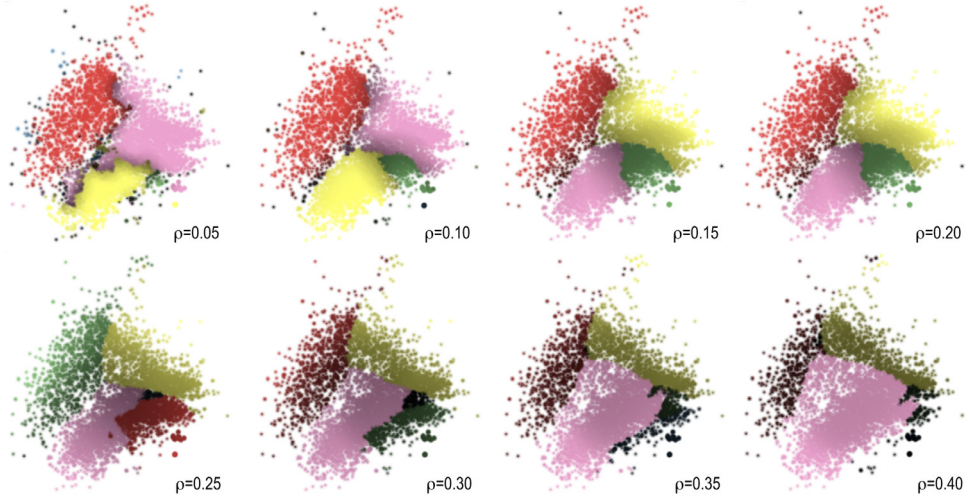

Fig. 4. Variance explanation for the *wine* dataset, projected by LAMP, for eight values of ρ (as fraction of the projection size). *rho* functions as a scale parameter: As it increases, the computed explanation becomes coarser, and small-scale details are removed. See Section 4.1.

Fig. 3c). We confirm the additional layer behind sugar-density correlation (purple area, Fig. 3c), specifically in regions where similarity is explained by chlorides and alcohol (purple and yellow areas, Figs. 3b,c), as all these attributes add to predicting wine quality. In the purple area in Fig. 3c, the sugar-density correlation is roughly of 0.9. This is in line with the sugar-density correlation of 0.83 reported for *all* the samples of this dataset by earlier studies [46].

4.2. Software quality dataset

This dataset contains 6773 software projects from SourceForge written in C [47]. Each project has 10 independent dimensions, these being metrics used in software engineering to gauge software quality: coupling between modules, complexity, lack of cohesion, number of source files, number of lines of code, number of function parameters, number of public variables, number of methods, number of data members, and structural complexity. Two additional dimensions measure the number of downloads and number of developers of a given software project.

Fig. 5 a shows the dataset projected with LAMP. As for the wine dataset (Section 4.1), the raw projection is not very informative. Fig. 5b,c show the projection explained by contribution, respectively variance. As for the wine dataset, these two explanations are very similar: The purple and yellow regions in both Fig. 5b,c

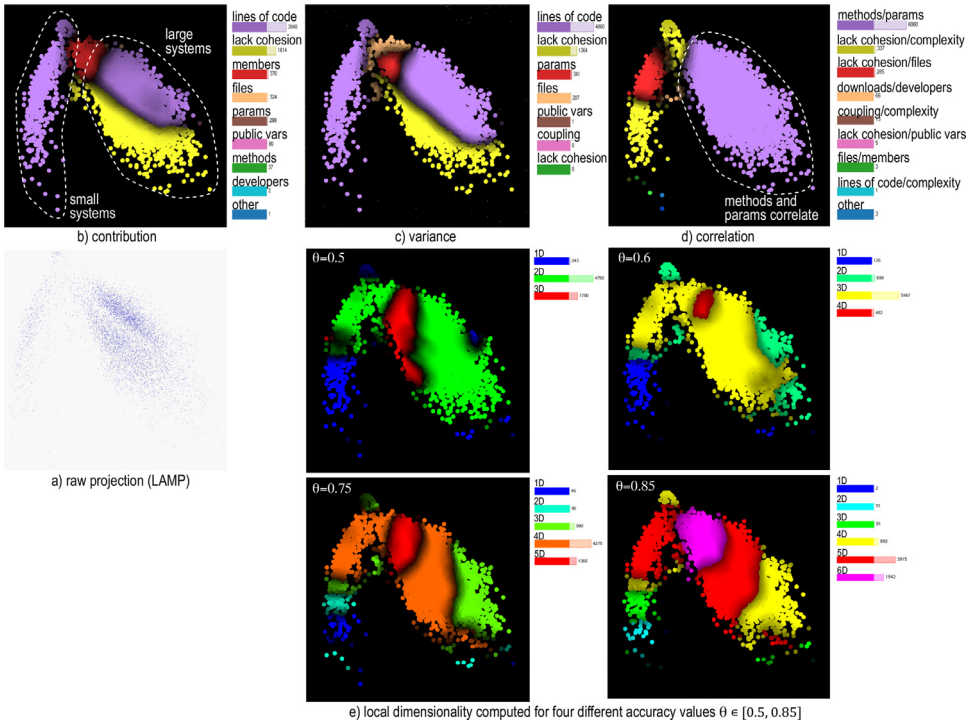

Fig. 5. Explanation of *software* dataset. The contribution (b) and variance (c) views show two purple lobe-like clusters corresponding to small, respectively large, systems. The correlation view (d) shows that large systems also have their method and parameter counts correlated. The local dimensionality views (e) shows that the two lobes can be explained by about three dimensions, while the area connecting them requires more effort to explain. See Section 4.2.

show software systems which are mostly similar due to size (lines of code), respectively complexity. The two *disjoint* purple regions indicate two groups of systems which are similar due to two different *value ranges* of lines of code. Brushing the image shows that projection is roughly split into a left lobe consisting of small software systems, and a right lobe containing large systems. However, the contribution and variance explanations are not *identical*: The red region in Fig. 5b shows systems which are similar in number of members. This region matches very well the union of the red and beige regions in the variance explanation (Fig. 5c), *i.e.*, systems with similar number of parameters or files. Hence, the number of members, parameters, and files appear to be correlated in this region.

The correlation view (Fig. 5d) adds more insights: The large purple area indicates systems which have correlated numbers of methods and parameters. From the earlier correlation/variance analysis, we know that these are large systems. Upon further study of the names of these systems in the original data [47], we find that these are mainly software libraries – for which, indeed, the total number of methods and total parameter count are correlated, since, in libraries (APIs), methods have typically similar parameter counts. The left lobe of the projection, *i.e.*, the small software systems, are yellow and red, indicating correlated lack-of-cohesion and complexity, respectively correlated lack-of-cohesion and number of files. Like for the wine dataset, such findings are only possible when joining the three different explanatory views. The correlated lack-of-cohesion with complexity is also a known signal in software quality analysis: Poor quality software is very often incohesive *and* complex [48].

We now examine the dimensionality of the projected data. Fig. 5e shows this for four different values of θ. Overall, these views tell us that the extremities of the two projection lobes are quite low-dimensional, being well explained by about three dimensions. In contrast, the area connecting the lobes requires five to six dimensions to explain. This area roughly corresponds to the red, respectively red-and-beige, regions in the contribution, respectively variance, views. The dimensionality view tells us that more explanations are needed in this central area since the projection is there not sufficiently well explained by the number of members, respectively lack-of-cohesion and number of parameters dimensions.

We next compare our findings with those of Meirelles *et al.* [47]. They found high correlations of complexity *vs* lack of cohesion (Pearson: 0.786, the highest correlation of all dataset dimension-pairs; Spearman: 0.773; Kendall tau: 0.597); and number of methods *vs* parameters (Pearson: 0.762; Spearman: 0.765; Kendall tau: 0.596). They also found a strong correlation between complexity and lines of code (Pearson: 0.666; Spearman: 0.685; Kendall tau: 0.497), the third strongest correlation for complexity, and a correlation between lack of cohesion and lines of code (Pearson: 0.472; Spearman: 0.490; Kendall tau: 0.341), the second strongest for the lack-of-cohesion attribute. These two correlations combined match our finding of complexity and lack of cohesion correlated (Fig. 5d, yellow areas) over a region of similar lines-of-code values (Fig. 5b, left purple lobe). Their strong-reported correlation of number of methods *vs* number of parameters noted above matches the purple lobe in Fig. 5d, on which we found a correlation of roughly 0.92. Note that the findings of Meirelles *et al.* are *averages* over the entire dataset. Our correlation view refines such insights by showing *local* correlations over subsets of the data.

4.3. City pollution dataset

This dataset, from the UCI Machine Learning repository, contains 420768 measurements of 6 air pollutants (PM2.5, PM10, SO2, NO2, CO, O3) and 6 meteorological variables (temperature, pressure, dew point temperature, rain, wind direction, and wind speed) measured hourly from March 2013 to February 2017 at 12 sites in Beijing [49]. We removed the time dimension (aggregating all measurements together) and projected the resulting dataset using both PCA and t-SNE.

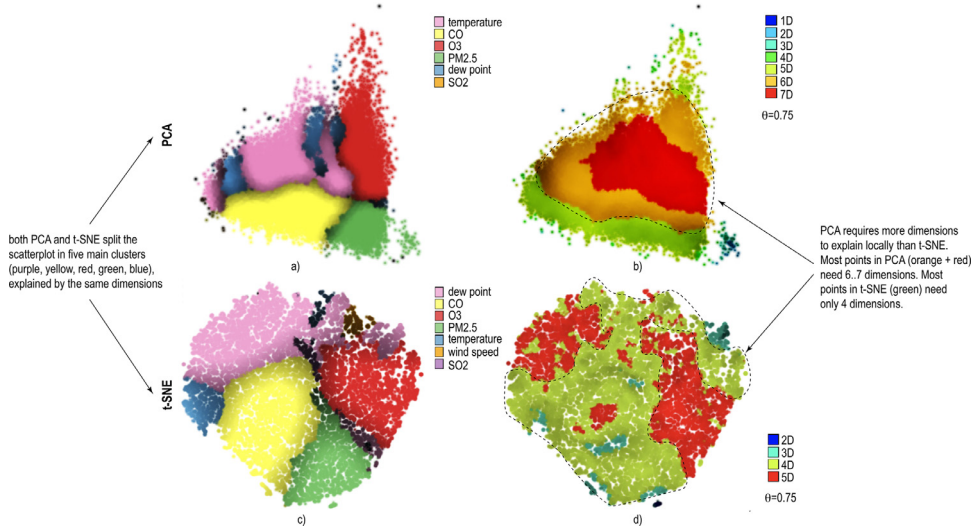

Fig. 6. Explanation of *city pollution* data, PCA and t-SNE projections. The variance views (a,c) show that both projections split the data into clusters with similar explanations. The dimensionality views show that PCA needs more additional dimensions to explain its clusters (b) than t-SNE (d). See Section 4.3.

We use this dataset to contrast how our explanations work for different projection types. Fig. 6a shows the variance explanation for PCA. This projection is split into four similar-size regions explained by the temperature, CO, O3, and PM2.5 dimensions. The dimensionality explanation of the PCA projection (Fig. 6b, $\theta = 0.75$) shows that we need five to seven dimensions to explain the projection, with more dimensions needed in the center thereof. The t-SNE projection is also split into similar-variance zones explained by the same variables (temperature, CO, O3, and PM2.5). Interestingly, these regions are placed relatively to each other quite similarly to their counterparts in the PCA projection. The dimensionality explanation of the t-SNE projection (Fig. 6d, $\theta = 0.75$) is very different from PCA's one: We do not see the low-to-high dimensionality gradient present in Fig. 6b; rather, the projection is locally either 4-dimensional (green) or 5-dimensional (red). Hence, t-SNE achieves a better 'spread' of the high-dimensional dataset in 2D than PCA. More interestingly, the red-green borders in Fig. 6 match relatively well the borders of the red and pink regions in Fig. 6c. This tells us that the dew-point and O3 explained regions in that figure are five-dimensional, whereas the CO, PM2.5, and temperature explained regions are four-dimensional, respectively.

4.4. Air quality dataset

This dataset, also from the UCI repository, has 9358 samples of air quality measurements (CO, NOx, NO2, benzene, and non-metanic hydrocarbons (NMHC)) done by both an experimental sensor and a reference ground-truth (GT) analyzer. Apart from these, temperature, relative humidity (RH) and absolute humidity (AH) are measured. Data were recorded from March 2004 to February 2005 in a highly polluted area of an Italian city [50], and its authors outline significant differences between the experimental sensor and GT values.

As for the city pollution dataset, we use our views to explain the PCA and t-SNE projection of this data (aggregating the time dimension). Fig. 7a shows the variance explanation of the PCA projection. This projection shows five visually separable clusters (dashed outlines A-E). Cluster D is actually an overlap of three clusters explained by the dimensions CO(GT) – pink, AH – yellow, and NMHC (GT) – red. The dimensionality view (Fig. 7b, $\theta = 0.68$) increases the confidence in the variance explanation: Clusters A, B, and C, which showed little overlap of explanations, are intrinsically two-dimensional, so we can trust the PCA projection here. Cluster E, which has a line structure, is intrinsically one-dimensional, so its explanation by the single dimension NOx (GT) in Fig. 7a is complete. In contrast, cluster D is two-to-three dimensional, which is exactly what its explanation by three 'overlapping' dimensions in Fig. 7a tells us. Fig. 7c shows the variance explanation of the t-SNE projection. We see here six visually distinct clusters (A′-F′). Upon closer inspection, by brushing, we found that A′ corresponds roughly to the union of A, B, and the pink part of D; B′ corresponds to the red part of D; D′ and F′ correspond to the yellow part of D; C′ corresponds to C; and E′ corresponds to E. Saliently, the colors in Fig. 7c correspond almost perfectly to visually distinct clusters. We also see no dark points in this figure, meaning that the confidence of the explanation is very high. Hence, the t-SNE projection both *groups* similar-value points better than PCA (see the pink points), and *separates* different-value points better (see the red, yellow, and green points). The dimensionality view (Fig. 7d, $\theta = 0.68$) confirms this: except a tiny red area, all points indicate neighborhoods of intrinsic dimensionality of one (blue) or two (green). Since this is a 2D projection, this tells us that t-SNE did a very good job in preserving the high-dimensional data structure, and in any case, better than PCA.

5. Discussion

We detail several aspects of our method, as follows.

Genericity and scalability: Our method can handle any type of quantitative data projected by any MP technique. Correlations and PCA are computed with Eigen [51]. Since explanations are computed and rendered independently on local point neighborhoods, we parallelized this using multithreading on the CPU. We generated all images in this paper in seconds for datasets up to tens of thousands of points, tens of dimensions, on a modern PC (3.6 Ghz CPU, GeForce 900 GPU). Table 2 shows timing measurements for several datasets having a wide range of dimensions n, samples N, and sizes ρ of the neighborhoods ν_i, sorted ascendingly on the total attribute count $n \cdot N$.

Combining explanations: The examples in Sections. 3 and 4 show that no single explanation suffices. One has to combine the partial insights of different explanations from the total six ones (distance contribution, variance, three local dimensionality variants, and dimensions correlation) to arrive at relevant, stronger, findings. In this process, one can (a) use explanations of the *same* type, *e.g.*

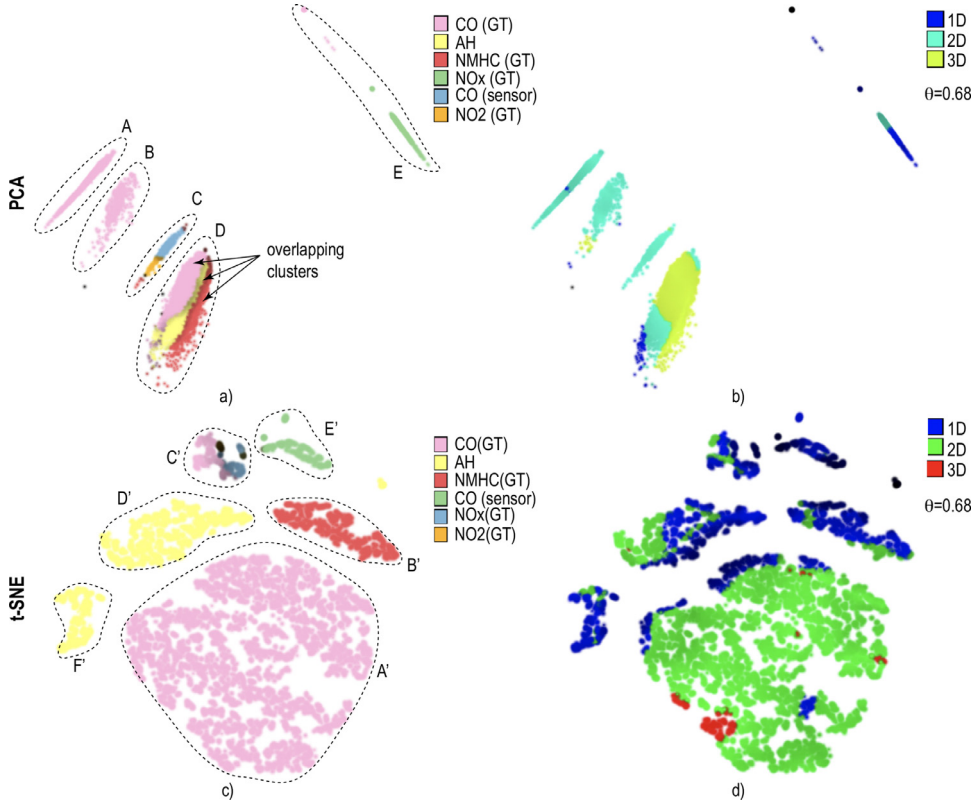

Fig. 7. Explanation of *air quality* dataset, PCA and t-SNE projections. Colors in the variance views (a,c) help finding the main variable explaining what makes points in a cluster similar. The local dimensionality views (b,d) tell us how many extra variables we need to fully explain these clusters. See Section 4.4.

Table 2
Computational performance of explanatory views

Dataset	Dimensions	Samples	Total	Time (secs)		
	n	N	$n \cdot N$	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$
D1	17	143	2431	0.013	0.016	0.016
D2	20	740	14800	0.025	0.028	0.029
D3	32	520	16640	0.016	0.019	0.019
D4	11	4177	45947	00.68	0.069	0.082
D5	25	2584	64600	0.046	0.045	0.047
D6	11	6497	71467	0.133	0.136	0.165
D7	179	11500	2058500	2.611	3.168	5.033
D8	64	41188	2636032	0.845	3.082	13.884

Table 3
Quality metrics for all projections and datasets in this paper.

Dataset	Projection	Continuity	Trustworthiness	Shepard
Concrete (Fig. 2)	t-SNE	0.99810535	0.99517108	0.53527163
Wine (Figs. 3,4)	LAMP	0.84132354	0.92384026	0.79137224
Software (Fig. 5)	LAMP	0.90646675	0.98470294	0.91487058
City pollution (Fig. 6)	PCA	0.93095898	0.99232401	0.95164997
	t-SNE	0.99888747	0.98818749	0.84766134
Air quality (Fig. 7)	PCA	0.94080419	0.99208358	0.97113638
	t-SNE	0.99916219	0.99601412	0.56614243

local dimensionality, which, where matching, strengthen the obtained findings; or (b) explanations of *different* types, *e.g.* correlation and variance, which performs 'logical AND' like operations on their partial insights.

Projection quality: Our explanations rely on the assumption that points close in $P(D)$ correspond to points close in D – that is, that the projection exhibits high values of trustworthiness [20]. In other words, our explanations require that *the neighborhoods shown in a projection are meaningful*. If they are, then we can explain them. If not, then we will produce wrong explanations, but arguably *any* use of such a projection will be flawed, not only our explanations, since the projection contains errors. The extent to which various MP techniques realize this neighborhood preservation varies [18]. One way to address this is to use projection error views [10] to exclude neighborhoods which do not respect this condition [33], or refine their computation by *e.g.* using larger radii ρ. To address this issue, Table 3 shows the continuity, trustworthiness, and Shepard correlation quality metrics computed for all the datasets and all the projections discussed earlier in this paper. For the exact definitions of these metrics, we refer, for brevity, to Table 5 in [18].

Table 3 shows that all the computed projections are of high quality, their values being very close to the maximum value of 1. For t-SNE, the Shepard correlation is relatively lower, but this is expected, as this metric quantifies the preservation of distances and the t-SNE technique does not aim to preserve distances, but neighborhoods. All in all, the projections shown in this paper are of sufficiently high quality to vouch their visual exploration by means of our explanatory techniques, and also to trust their computation which relies on the assumption of high trustworthiness already mentioned above.

Limitations: While we can *technically* handle datasets of any dimensionality n, we need more variables for the explanation as local dimensionality grows. Also, the correlation is $O(n^2)$ in computation and space needed for the dimension matrix (see Fig. 2 and related text). Our method works well up to 20 dimensions in practice; it does not target datasets with hundreds of dimensions such

as from deep learning. Yet, such datasets have *abstract* dimensions which do not have a meaning for users, so using them to explain projections is likely not desirable. Our method scales *visually* well even for many dimensions, since it uses only the *top ranked* ones which contribute to explaining most of the projected points (Section 3).

One can ask whether using nD point neighborhoods $\xi_i = \{\mathbf{x} \in D | \ \|\mathbf{x} - \mathbf{x}_i\| \leq \rho\}$, $P(\mathbf{x}_i) = \mathbf{y}_i$, instead of 2D neighborhoods ν_i (and their correspondents μ_i in nD), is a valid option. Doing this is technically trivial, but we argue against it: We aim to explain the point-groups one sees in a *projection* (2D scatterplot) and not the point-clusters that exist in nD, but may *not* be visible in 2D due to *e.g.* projection continuity issues [20]. Also, setting the neighborhood size ρ would be tricky for ξ_i, as one has to assess what is the 'natural' scale of patterns in nD. This motivates our choice to use 2D neighborhoods as a basis for our explanations.

A separate limitation involves color coding, which is used to create categorical color maps (contribution, variance, and correlation plots) and also ordered color maps (dimensionality plot). As explained in Section 4, several such plots are to be used together to arrive at a good understanding of a projection. This may potentially confuse users since the respective colormaps contain similar colors. The problem can be partly alleviated by designing colormaps with a smaller overlap in terms of such colors. However, as we next aim to extend our approach with additional explanatory views, this alleviation strategy is not a full solution. For now, we prominently display the respective color legends next to each explanatory plot, aiming thereby to attract the attention of the user of the particular meaning of colors in that plot.

User perception: As our techniques aim to explain the patterns one sees in a projection, they should be tested in experiments where subjects use them to to perform some explanatory tasks. Earlier studies [52] provide good guidelines of perceptual cues and visual tasks that users address with projections. We aim to extend this work by making such tasks more specific to include explanations that refer to the names of involved dimensions. With this set of tasks, we can next present various combinations of datasets D and projections $P(D)$, computed by several projection techniques P to the users, to find which are the dataset and/or projection-technique aspects that best suit our explanatory techniques. A similar study can be used to find optimal parameters for our explanatory techniques.

6. Conclusions

We have presented a set of visualizations for explaining the visual patterns present in 2D projections of high-dimensional data in terms of the underlying data dimensions. We extended the explanations proposed in earlier work [11] by three ways to evaluate the local data dimensionality and a technique to detect and inspect local dimension correlations. We show that the combined visual analysis of all these explanatory techniques can lead to non-trivial insights in the data that correlate well with independent findings obtained using other methods. We illustrate our approach on five experimental datasets. Our methods are simple to use, have a few parameters with good presets and clear effects, and scale well computationally to datasets of hundreds of thousands of samples and 10..20 dimensions.

Several extensions to our work are possible. Adding more explanation types, such as inverse correlation, correlation of more than two dimensions, or the presence of specific nD data patterns, is a low hanging fruit. We aim to compute, in parallel, a wide range of local explanations based on a pattern library, and next show the most salient ones in the final view, thereby enriching the current contribution, variance, correlation, and dimensionality views. This would perform a scagnostics-like [53] local analysis of the projection, but using patterns described by the high-dimensional data rather than by the scatterplot. Computing a hierarchical explanation, where projection regions are recursively split by additional explanations, is another direction we aim to pursue.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Zonglin Tian: Methodology, Software, Formal analysis, Validation, Investigation, Writing - original draft. **Xiaorui Zhai:** Methodology, Formal analysis, Data curation, Validation, Investigation, Writing - original draft. **Daan van Driel:** Conceptualization, Methodology. **Mateus Espadoto:** Methodology, Software, Visualization, Writing - original draft. **Alexandru Telea:** Methodology, Conceptualization, Supervision, Writing - original draft.

Acknowledgments

Z. Tian was supported by the China Scholarship Council under grant 201906080046.

References

[1] Greenacre M. Biplots in practice. Fundacion BBVA, Bilbao; 2010.
[2] Gower J, Lubbe S, Roux N. Understanding biplots. Wiley; 2011.
[3] Broeksema B, Baudel T, Telea A. Visual analysis of multidimensional categorical datasets. Computer Graphics Forum 2013;32(8):158–69.
[4] Coimbra D, Martins R, Neves T, Telea A, Paulovich F. Explaining three--dimensional dimensionality reduction plots. Information Visualization 2016;15(2):154–72.
[5] Pagliosa P, Paulovich F, Minghim R, Levkowitz H, Nonato L. Projection inspector: Assessment and synthesis of multidimensional projections. Neurocomputing 2015;150:599–610.
[6] Joia P, Coimbra D, Cuminato JA, Paulovich FV, Nonato LG. Local affine multidimensional projection. IEEE TVCG 2011;17(12):2563–71.
[7] Rauber P, da Silva R, Feringa S, Celebi M, Falcao A, Telea A. Interactive image feature selection aided by dimensionality reduction. In: Proc. EuroVA; 2015. p. 97–101.
[8] Aupetit M. Visualizing distortions and recovering topology in continuous projection techniques. Neurocomputing 2007;10(7-9):1304–30.
[9] Schreck T, von Landesberger T, Bremm S. Techniques for precision-based visual analysis of projected data. Information Visualization 2010;9(3):181–93.
[10] Martins R, Coimbra D, Minghim R, Telea AC. Visual analysis of dimensionality reduction quality for parameterized projections. Computers & Graphics 2014;41:26–42.
[11] da Silva R, Rauber P, Martins R, Minghim R, Telea A. Attribute-based visual explanation of multidimensional projections. In: Proc. EuroVA; 2015. p. 97–101.
[12] van Driel D, Zhai X, Tian Z, Telea A. Enhanced attribute-based explanations of multidimensional projections. In: Proc. EuroVA. Eurographics; 2020.
[13] Tenenbaum JB, De Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000;290(5500):2319–23.
[14] De Silva V, Tenenbaum JB. Sparse multidimensional scaling using landmark points. Tech. Rep.. Stanford University; 2004.
[15] van der Maaten L, Hinton GE. Visualizing data using t-SNE. JMLR 2008;9:2579–605.
[16] McInnes L., Healy J., Melville J.. UMAP: Uniform manifold approximation and projection for dimension reduction. 2018. ArXiv:1802.03426v2 [stat.ML].
[17] Nonato LG, Aupetit M. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. IEEE TVCG 2018;25(8):2650–73.
[18] Espadoto M, Martins R, Kerren A, Hirata N, Telea A. Towards a quantitative survey of dimension reduction techniques. IEEE TVCG 2019. Doi:10.1109/TVCG.2019.2944182
[19] Geng X, Zhan D, Zhou Z. Supervised nonlinear dimensionality reduction for visualization and classification. IEEE Trans Syst Man Cybern 2005;35(6):1098–107.
[20] Venna J, Kaski S. Visualizing gene interaction graphs with local multidimensional scaling. In: Proc. ESANN; 2006. p. 557–62.
[21] Paulovich FV, Nonato LG, Minghim R, Levkowitz H. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. IEEE TVCG 2008;14(3):564–75.

[22] Sips M, Neubert B, Lewis J, Hanrahan P. Selecting good views of high-dimensional data using class consistency. Comp Graph Forum 2009;28(3):831–8.
[23] Lee JA, Verleysen M. Quality assessment of dimensionality reduction: Rank-based criteria. Neurocomputing 2009;72(7):1431–43.
[24] Lueks W, Gisbrecht A, Hammer B. Visualizing the quality of dimensionality reduction. Neurocomputing 2013;112:109–23.
[25] Lespinats S, Aupetit M. CheckViz: Sanity check and topological clues for linear and non-linear mappings. Comp Graph Forum 2011;30(1):113–25.
[26] 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.
[27] Oeltze S, Doleisch H, Hauser H. Interactive visual analysis of perfusion data. IEEE TVCG 2007;13(6):1392–9.
[28] Olsen K, Korfhage R, Sochats K. Visualization of a document collection: the VIBE system. Inform Process Manag 1993;29(1):69–81.
[29] Endert A, Flaux P, North C. Semantic interaction for visual text analytics. In: Proc. ACM CHI; 2012. p. 324–33.
[30] Yi J, Melton R, Stasko J. Dust & magnet: multivariate information visualization using a magnet metaphor. Inform Visual 2005;4(4):239–56.
[31] Piringer H, Kosara R, Hauser H. Interactive F + C visualization with linked 2D/3D scatterplots. In: Proc. IEEE CMV; 2004. p. 49–60.
[32] Elmqvist N, Dragicevic P, Fekete J-D. Rolling the dice: multidimensional visual exploration using scatterplot matrix navigation. IEEE TVCG 2008;14(8):1141–8.
[33] Rodrigues FCM, Espadoto M, Hirata R, Telea A. Constructing and visualizing high-quality classifier decision boundary maps. Information 2019;10(9):280–97.
[34] Cliff N. The eigenvalues-greater-than-one rule and the reliability of components. Psychological Bulletin 1988;103(2):276–9.
[35] Jolliffe IT. Principal Component Analysis. Springer; 2002. 2nd edition
[36] O'Donnell LJ, Westin CF. An introduction to diffusion tensor image analysis. Neurosurg Clin N Am 2011;22(2):185–96.
[37] P PB, Falguerolles A. Application of resampling methods to the choice of dimension in principal component analysis. In: Computer Intensive Methods in Statistics. Springer; 1993. p. 167–76.
[38] North GR, Bell TL, Cahalan RF, Moeng FJ. Sampling errors in the estimation of empirical orthogonal functions. Mon Weather Rev 1982;110:699–706.
[39] Yeh I-C. Modeling of strength of high performance concrete using artificial neural networks. Cement and Concrete Research 1998;28(12):1797–808.
[40] Lichman M.. UCI machine learning repository. 2013. http://archive.ics.uci.edu/ml.
[41] da Silva R. Visualizing multidimensional data similarities – improvements and applications. University of Groningen, Netherlands; 2016.
[42] Wu S, Li B, Yang J, Shukla S. Predictive modeling of high-performance concrete with regression analysis. In: Proc. IEEE Intl. Conf. on Industrial Engineering and Engineering Management; 2010.
[43] 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):547–53.
[44] van Wijk JJ, Telea A. Enridged contour maps. In: Proc. IEEE Visualization; 2001. p. 69–74.
[45] Beh EJ, Holdsworth CI. A visual evaluation of a classification method for investigating the psysicochemical properties of Portugese wine. Current Anal Chem 2012;8(2):205–17.
[46] Zeng L.. The wine dataset analysis. 2021. https://rpubs.com/Li2019/Wine.
[47] Meirelles P, Santos C, Miranda J, Kon F, Terceiro A, Chavez C. A study of the relationships between source code metrics and attractiveness in free software projects. In: Proc. Brazilian Symposium on Software Engineering (SBES); 2010. p. 11–20.
[48] Richter C. Designing Flexible Object-Oriented Systems with UML. New Riders Publishing; 1999.
[49] Zhang S, Guo B, Dong A, He J, Xu Z, Chen S. Cautionary tales on air-quality improvement in Beijing. Proc Royal Society A 2017;473(2205):20170457. https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data
[50] Vito SD, Massera E, Piga M, Martinotto L, Francia GD. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. Sensors and Actuators B: Chemical 2008;129(2):750–7. https://archive.ics.uci.edu/ml/datasets/Air+Quality
[51] Eigen numerical library. 2020. http://eigen.tuxfamily.org.
[52] Etemadpour R, Motta R, de Souza Paiva J, Minghim R, Oliveira MD, Linsen L. Perception-based evaluation of projection methods for multidimensional data visualization. IEEE TVCG 2014;21(1):81–94.
[53] Wilkinson L, Arland A, Grossman R. Graph-theoretic scagnostics. In: Proc. InfoVis; 2005. p. 157–64.