Interpretation of Structural Preservation in Low-Dimensional Embeddings

Aindrila Ghosh, Mona Nashaat, James Miller, and Shaikh Quader

Abstract—Despite being commonly used in big-data analytics; the outcome of dimensionality reduction remains a black-box to most of its users. Understanding the quality of a low-dimensional embedding is important as not only it enables trust in the transformed data, but it can also help to select the most appropriate dimensionality reduction algorithm in a given scenario. As existing research primarily focuses on the visual exploration of embeddings, there is still a need for enhancing interpretability of such algorithms. To bridge this gap, we propose two novel interactive explanation techniques for low-dimensional embeddings obtained from *any* dimensionality reduction algorithm. The first technique LAPS produces a local approximation of the neighborhood structure to generate interpretable explanations on the preserved locality for a single instance. The second method GAPS explains the retained global structure of a high-dimensional dataset in its embedding, by combining non-redundant local-approximations from a coarse discretization of the projection space. We demonstrate the applicability of the proposed techniques using 16 real-life tabular, text, image, and audio datasets. Our extensive experimental evaluation shows the utility of the proposed techniques in interpreting the quality of low-dimensional embeddings, as well as with selecting the most suitable dimensionality reduction algorithm for any given dataset.

Index Terms—Interactive data exploration and discovery, algorithms for data and knowledge management, data and knowledge visualization

1 Introduction

Dimensionality reduction algorithms transform high-dimensional datasets into low-dimensional embeddings while attempting to retain most of the original structural relationships (i.e., relative distances) among the data points. On a high level, all dimensionality reduction algorithms perform complex mathematical optimizations to obtain the low-dimensional projection of a dataset that is often hard to interpret. The primary reason behind this is, the dimensions derived by such algorithms do not have any directly interpretable mappings to the original attributes of the high-dimensional data [1]. Hence, dimensionality reduction being one of the first steps in big-data analytics, a vital concern remains [2]: *if the users do not understand the quality of the low dimensional embedding, they will not make efficient decisions during subsequent analysis.* Moreover, the lack of interpretability in dimensionality reduction algorithms also leads to the challenge of selecting the most appropriate algorithm in a given scenario. In their work, Maaten *et al.* [3] and Becht *et al.* [4] have shown that different dimensionality reduction methods perform differently on the same dataset. Also for every such algorithm, there exists a perfectly reasonable metric [4] for which it is superior to its competitors. For example, in case of the maximum amount of preserved variance in an embedding, Principal Component Analysis (PCA) [3] could perform better than others. Or, for the maximum retention of overall distances among data points, Multidimensional Scaling (MDS) [3] could be the best choice. However, given the fact that there is no established way [5] to evaluate the performance of dimensionality reduction methods, data-scientists often follow their intuitions to use any one of these algorithms, without really understanding their behavior.

The quality [3], [6], [7] of a low-dimensional embedding depends on the extent to which an algorithm can preserve the local structural relationships (i.e., the structural similarities in individual neighborhoods) as well as the global structural associations (i.e., the relative differences in overall neighborhoods) from the original dataset. Hence, an interactive assessment of the preserved structure [1] can not only help users to *trust* the relative positioning of individual data points in a projection but also to have confidence in the overall embedding. In recent years, interactive exploration of low-dimensional embeddings has become an increasingly popular [1], [8], [9], [10] mechanism for evaluating the quality dimensionality reduction. However, our investigation shows, the existing research [1], [8], [9], [10], [11] primarily enables visual exploration of embeddings and rarely compare the embedding to the original data [12]. Also, the majority of the existing techniques do not allow simultaneous comparisons of multiple algorithms to evaluate their outcome on a specific dataset. Most importantly, the research of Adadi *et al.* [2] and Guidotti *et al.* [13] confirm that there is still a need for a well-defined mechanism for explaining the structural preservation after dimensionality reduction.

To bridge these gaps, firstly, we present *LAPS - **L**ocal **A**pproximation of **P**reserved **S**tructure*, a method & data-type agnostic technique that provides explanations on the preserved local structure of a low-dimensional embedding. The explanations presented by LAPS justify the fidelity of the relative positioning of any individual data-point in an

- *Aindrila Ghosh, Mona Nashaat, and James Miller are with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton T6G 2R3, Canada. E-mail: {aindrila, nashaata, jimm}@ualberta.ca.*
- *Shaikh Quader is with the IBM Toronto Software Lab, Toronto, Canada. E-mail: shaikhq@ca.ibm.com.*

Manuscript received 5 Sept. 2019; revised 19 May 2020; accepted 25 June 2020. Date of publication 29 June 2020; date of current version 1 Apr. 2022. Corresponding author: Aindrila Ghosh.)
Recommended for acceptance by A. Poulovassilis.
Digital Object Identifier no. 10.1109/TKDE.2020.3005878

embedding by approximating a neighborhood locally around that point. Secondly, we present *GAPS - **G**lobal **A**pproximation of **P**rojection **S**pace*, an interactive technique that presents explanations on the preserved global structure in a low dimensional embedding, by combining non-redundant local-approximations from a coarse discretization of the projection space. As a part of an extensive and comprehensive evaluation, we assess both of the proposed techniques for their flexibility (with 10 different dimensionality reduction algorithms on 16 real-life datasets), applicability (i.e., with tabular, text, image, and audio data), utility (i.e., with a user-study that examines their ability to explain the quality [7] of a projection), and reliability (i.e., to assist with the selection of the most appropriate dimensionality reduction algorithm). Our experiments also reveal the roles of different user-defined parameters in the outcome of the proposed techniques. Moreover, they uncover the ability of the techniques in discovering feature correlations in high-dimensional data.

Our primary contributions in this work are as follows:

1. LAPS, a novel algorithm that can provide interpretable and faithful explanations on the retained local structures in any low-dimensional embedding, by locally approximating the neighborhoods.
2. GAPS, a novel technique that provides explanations on the preserved global structure of a manifold in its low-dimensional embedding, by combining local approximations of discrete non-redundant neighborhoods into a global approximation.
3. An extensive 5-phase experimental evaluation of the proposed methods LAPS and GAPS.

The rest of the paper is organized as follows: Section 2 provides an overview of related work as Section 3 introduces the necessary background information and design requirements for the proposed techniques. Next, whilst Section 4 presents the proposed algorithms, Section 5 describes our experimental evaluations of the presented techniques in detail. Finally, Section 6 concludes the paper with a brief discussion on future work.

2 Related Work

When it comes to interpretability, visual interaction with low-dimensional embeddings [1], [14], [15] has been the most commonly proposed approach by researchers. In the past few years, several tools [8], [11], techniques [9], and frameworks [1], and essays [15] have been presented that aim at making the complex procedure of dimensionality reduction more understandable to its users. Whilst detailed surveys of different interaction paradigms for low-dimensional embeddings can be found in [14] and [16], in this section we highlight the most closely related work to our proposed algorithms.

Covering different aspects of interaction with dimensionality reduction, some existing techniques (e.g., Embedding Projector [8]) allow users to visually explore the neighborhood structures in embeddings. Some other techniques (e.g., Probing Projections [11], CheckViz [6]) visualize the amount of approximation errors in relative distances between the data points in a projection. Among these, whilst Probing Projections [11] assists users to perform distance corrections within neighborhoods, CheckViz [6] enables visualization of false neighborhoods in the projection. Taking the scope for interactivity one step further, some techniques (e.g., Praxis [1], DimStiller [17], LAMP [18]) allow users to interact with the dimensionality reduction process itself. For example, Praxis [1] lets users interactively modify the input feature values for a data-point to see the change in its projection, as well as to alter the position of a point in an embedding to see the changes in original feature values. DimStiller [17] represents the transformation performed during dimensionality reduction as a series of events in a pipeline. The technique allows users to interactively add or remove dimensions in the input and visualize any step in the pipeline at any point in time. The interactive multidimensional projection technique LAMP [18] allows users to interactively steer a projection by enabling them to select the control points that build a family of affine mappings.

To facilitate an efficient selection of hyper-parameters for dimensionality reduction, some techniques such as LDSScanner [19] enable the exploration of the neighborhood structures in the high-dimensional datasets. On the other hand, some tools like SIRIUS [9] enable interactive symmetric dual exploration of the most correlated attributes and neighborhoods in data. At the same time, to explain the quality of embeddings, techniques such as DimReader [10] enable visual exploration of the newly generated axis lines in the projections. Identifying the need to quantify the structural preservation in embeddings, researchers such as Martins *et al.* [20] propose mechanisms to both visually and quantitatively assess low-dimensional embeddings using false and missing neighbors. In an attempt to explain the relative positioning of data points in an embedding, researchers such as Pagliosa *et al.* [21], Silva *et al.* [22], and Self *et al.* [23] present techniques that identify the influences of the original attributes in the formation of neighborhood structures.

Nevertheless, our investigation of related research showed that very few researchers (e.g., Kodali *et al.* [12]) have considered both the aspects of neighborhood preservation and the retention of attribute influences when quantifying the structural quality of an embedding. Even so, a majority of these approaches are designed for only a specific set of dimensionality reduction algorithms (e.g., the approach proposed by Kodali *et al.* [12] is designed for Weighted Multidimensional Scaling). As a result, these techniques rarely provide an opportunity for a side-by-side comparison among embeddings obtained from different dimensionality reduction algorithms or to perform an interactive selection of the most appropriate algorithm for any given dataset. Also, very few approaches [12] enable any interactive comparisons between the original high-dimensional data and their low-dimensional embeddings to explain the quality of the obtained projections. Hence, there is still a need for a well-defined technique that would visually and quantitatively explain [2], [13] the extent of the preserved local and global structures in reduced dimensions, and consider the impacts of both neighborhoods and attribute influence preservations in embeddings.

3 Problem Characterization

The overall procedure of dimensionality reduction can be formally defined as: assuming a matrix X of size $n \times D$, that

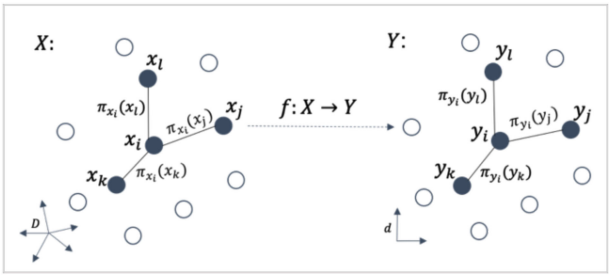


Fig. 1. An overview of the dimensionality reduction process.

represents a high-dimensional dataset with n records and D attributes so that $X = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^{D \times n}$. That is, each x_i represents a data-vector for an individual record in X and cardinality of $x_i = |x_i| = D$. Dimensionality reduction can be defined as a mapping function:

$$f : X \rightarrow Y, \tag{1}$$

where, f transforms X into a low-dimensional embedding Y of size $n \times d$, where $Y = \{y_1, y_2, \ldots, y_n\} \in \mathbb{R}^{d \times n}$ and $|y_i| = d$. On a high level, any f can also be formulated [3] as an optimization problem as:

$$\underset{Y \in \mathbb{R}^{d \times n}}{\arg\min} f(Y; X, \theta), \tag{2}$$

where f represents the objective function that attempts to minimize the structural differences between X and Y as θ represents the hyper-parameters of the dimensionality reduction algorithm. Ideally, the dimension d for Y is the intrinsic dimensionality [3] of the dataset X. The intrinsic dimensionality d represents an estimation of the minimum number of dimensions that can be used to represent X with minimum information loss.

For most real-world datasets, $d \ll D$. This means, as D is reduced to d, the points in the dataset are relocated to a much smaller space than the original high-dimensional manifold. Fig. 1 shows such a transformation, where the preservation of the structure of X in Y refers to the fact that the points that were close to each other in X should remain close in Y as well. Also, the points that are far from each other in X should remain the same in Y. The notion of closeness among the data points lying on a manifold is defined using proximity measures [5], [6]. Considering, for any data-point x_i lying on a manifold represented by X, the neighborhood [19], [24] of x_i is a subset Z of X containing x_i, so that $x_i \in Z \subseteq X$. In this case, Z contains the data points that are closest to x_i. In Fig. 1, we define the proximity between the point x_i and its nearest neighbors, i.e., $\forall\, x' \in Z$ and $x_i \neq x'$ as $\pi_{x_i}(x')$. Also, in Fig. 1 as f (cf. (2)) attempts to minimize the overall divergence between X and Y and preserve the neighborhood structure, in an ideal case [24] the following inequalities should hold:

$$\pi_{y_i}(y_j) \begin{cases} < \pi_{y_i}(y_k) & if\ \pi_{x_i}(x_j) < \pi_{x_i}(x_k) \\ > \pi_{y_i}(y_k) & otherwise \end{cases}, \tag{3}$$

where, $y_i = f(x_i)$, $y_j = f(x_j)$, and $y_k = f(x_k)$ and $i \neq j \neq k$. Also, the points x_j and x_k belong to the neighborhood of x_i as y_j and y_k belong to the neighborhood of y_i. Nevertheless, dimensionality reduction being an optimization problem, research [6], [11] shows, its outcome is often likely to converge to a local-optima leading to the inequalities between relative distances not being retained for every data-point in X after the transformation.

A large number of non-linear dimensionality reduction algorithms (e.g., MDS, Isomap [3]) rely on the neighborhood geometry of a manifold to recognize its overall structure. Among them, some algorithms (e.g., MDS [3]) use the euclidean distance [9] as their proximity measure for the data points, considering the manifold to be locally isometric to a euclidean space [6]. The euclidean distance $dist_\varepsilon$ between two points x_i and x_j can be defined as:

$$dist_\varepsilon(x_i, x_j) = \sqrt{\sum_{a=1}^{D} (u_a - v_a)^2}, \tag{4}$$

where, u_a and v_a represent individual features in x_i and x_j respectively. On the other hand, some algorithms (e.g., Isomap [25]) use pairwise Geodesic distance [3] among points to measure the global intrinsic feature of the manifold. The Geodesic distance $dist_\gamma$ among the points x_i and x_j in X, can be defined using the infimum over the lengths of all the smooth paths connecting the two points as:

$$dist_\gamma(x_i, x_j) = \inf\{L(\sigma)\}, \tag{5}$$

where σ is a smooth path from x_i and x_j. The smoothness of σ is measured by the number of continuous derivatives along σ. Formally, assuming S_σ as a set of all points along σ and every $s_\sigma \in S_\sigma \in \mathbb{R}$, then σ is considered to be smooth if it has derivatives of all orders for every $s_\sigma \in S_\sigma$.

Due to the use of different proximity measures and optimization functions, different dimensionality reduction algorithms perform differently [5] on the same dataset. For example, as shown in Fig. 2, the artificial Swiss-roll dataset (cf. Fig. 2a) is transformed using four different dimensionality reduction algorithms. The Fig. 2b, 2c, 2d, and 2e, clearly show the differences in preservation of the local and global structures of the original dataset. Hence, it can be noted that the structural preservation plays an important role in selecting the most suitable dimensionality reduction algorithm for a given dataset. Moreover, as dimensionality reduction is performed prior to any deeper analysis with a high-dimensional dataset (e.g., training a predictive model with Y), the lack of preserved structure can result in poor subsequent analysis. In this paper, we propose interpretable explanations about Y as a solution to the above-mentioned problems.

3.1 Requirements Analysis for Explanations

In this article, we define *explanations of preserved structure* as a *set* of meaningful textual and visual artifacts that describes the ability of an algorithm to retain the original relative distances among the data points in the low-dimensional space. In this section, we define a set of requirements for the explanations of reduced dimensions.

First of all, we need the explanations to be *interpretable* [26], [27] for both expert and novice users. According to Ribeiro *et al.* [28] and Yang *et al.* [29], as humans, we relate to meaningful names much faster than numeric values or complex graphical representations. Hence, along with displaying the relative distances among data points [8], [9],

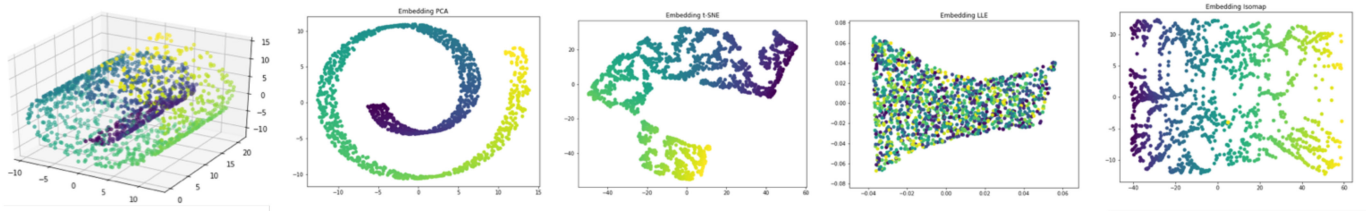


Fig. 2. Structural preservation with different dimensionality reduction algorithms on the artificial Swiss Roll dataset.

[11], there is a need to present the explanations in terms of the attributes of the original dataset. Moreover, to prevent the explanations from overwhelming users in cases of very high-dimensional (e.g., $D > 100$) datasets, there needs to be a way for the users to regulate the amount of information they would want to see.

Secondly, we expect *local fidelity* and *global legitimacy* from the explanations. For example, whilst it is important for LAPS to be locally faithful (i.e., for the data point being explained), for GAPS it is essential to be globally legitimate (i.e., accurate for the entire dataset). We note that it is often impossible to achieve complete global legitimacy of the explanations unless all the data points in a dataset are considered. The explanations need to incorporate this fact.

Thirdly, the explanations should be *algorithm & data-type agnostic*. That is, the selection of a suitable dimensionality reduction algorithm requires the explanations to be applicable for a variety of such algorithms. Moreover, to explore the full potential of the explanations, they should be flexible enough to incorporate any type of data.

Finally, the explanations for local and global structures should be *consistent*. That is, not only the look and feel of the explanations but also the user-interactions with them should be made consistent for both LAPS and GAPS.

4 Explaining Reduced Dimensions

In this section, we present the overall ideas of our proposed methods *LAPS (**L**ocal **A**pproximation of **P**reserved **S**tructure)* and *GAPS (**G**lobal **A**pproximation of **P**rojection **S**pace)*.

Prior to formally defining the methods, we introduce some notations that would be used later in the article. Considering a random point $x_i \in X$, that is represented using a feature vector $U = [u_1, u_2, \ldots, u_D] \in \mathbb{R}^D$, where $\forall\ u_a \in U$, a represents an individual feature. The locality around x_i is defined using a set $Z \subseteq X$, containing k-nearest neighbors of x_i. We define the local explanation for x_i as a *set* containing feature influence explanations $fie(x_i)$ and the local-divergence λ_{x_i} for x_i. Whilst the feature influence explanation $fie(x_i)$ represents an interpretable function that approximates the contribution of each feature in the relative proximity between x_i and its k-nearest neighbors, the local-divergence λ_{x_i} represents the disagreements in the feature influence explanations and neighborhood structures of x_i and y_i . In this case, $y_i \in Y$ represents the low-dimensional counterpart of $x_i \in X$, i.e., $y_i = f(x_i)$. In par with our requirements for *consistency*, we compose the global explanations using $fie(X_S)$ and λ_{X_S}. Here, $fie(X_S)$ represents the unification of local feature explanations of a user-defined subset X_S of X. λ_{X_S} presents the global structural divergence between the original data points in X_S and their counterparts in the embedding Y_S. Formally, we define the local and global explanations as:

$$\begin{aligned} loc\_expl_{x_i} &= \{\ fie(x_i),\ \lambda_{x_i} \mid\ \exists x_i \in X\} \\ glob\_expl_{X_S} &= \{\ fie(X_S),\ \lambda_{X_S} \mid\ \exists X_S \subseteq X\}, \end{aligned} \tag{6}$$

where both $loc\_expl_{x_i}$ and $glob\_expl_{X_S}$ are *sets* of textual and visual artifacts used for interpreting the embeddings.

4.1 Motivating Example

To facilitate a better understanding of the concept of *explanation* defined above, in this section, we demonstrate the idea with a toy-example. In this example, we have our analyst Alice analyze the Animals[1] dataset [9] that contains 30,475 images and distinguishes 50 animal classes using 85 numeric attributes. In this case, X represents the dataset, where $n = 30475$ and $\mathrm{D} = 85$. After transforming the data into a 2D embedding Y using *any* dimensionality reduction algorithm, Alice wants to interpret the preserved local structure in the embedding. To obtain a local explanation using LAPS, Alice selects a single point-of-interest x_i (say, $x_i = rabbit$) from X. The $loc\_expl_{x_i}$ of preserved structure for *rabbit* contains the following: (1) $fie(x_i)$ and $fie(y_i)$: the positive and negative influence scores for all the 85 attributes in the construction of the neighborhood of *rabbit* in X as well as in Y (where, $y_i = f(x_i)$). (2) λ_{x_i}: the local-divergence score for the data-point *rabbit*. Here, λ_{x_i} is computed as a weighted sum of the disagreements between $fie(x_i)$ and $fie(y_i)$ and the disparities in the neighborhood structures for the point *rabbit* in X and Y. Similarly, to obtain an explanation on the preserved global structure in the embedding, Alice interactively selects a subset X_S from X that contains the data points *rabbit*, *mouse*, *hamster*, *mole*, and *squirrel*. The $glob\_expl_{X_S}$ obtained using GAPS consists of (1) $fie(X_S)$ and $fie(Y_S)$: the overall influences of the original attributes in the relative positioning of the neighborhoods of the points in X_S and Y_S. (2) λ_{X_S}: the global-divergence computed by adding the scaled local divergences of the points $\lambda_{x_i} \in \lambda_{X_S}$.

4.2 Local Approximation of Preserved Structure

In order to generate *data-type agnostic* local explanations using LAPS, we avoid making any assumptions about X.

Next, as a pre-processing before transforming X to Y, we estimate the intrinsic dimensionality d of X using the maximum likelihood intrinsic dimensionality estimator [30] defined as:

1. https://cvml.ist.ac.at/AwA2/

$$\hat{d} = \frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} \hat{d}_k \quad where, \hat{d}_k = \frac{1}{n} \sum_{i=1}^{n} d_k(X), \tag{7}$$

where, $\hat{d}$ represents a unit vector with an estimation for d and $(k_2 - k_1)$ signifies the range of nearest neighbors to consider while estimating d. This pre-processing is necessary [3], [19], [30] as the estimation of d prior to obtaining Y not only ensures noise reduction [19], [24] in Y, such an estimation also enhances the stability [30] of Y. Next, with d as a parameter to f, we obtain Y as $f(X)$. In order for the explanations to be *algorithm-agnostic*, we also avoid making any assumptions about f.

Algorithm 1. The *LAPS* Procedure

Input: dataset X, embedding Y, instance x_i, neighbor count k
Output: $fie(x_i)$, $fie(y_i)$ and local divergence λ_{x_i}
Step 1: Obtain nearest neighbors for x_i and y_i
for all $j \in \{0, 1, \ldots\ldots, k\}$ **do**:
$Z_{x_i} \leftarrow nn_{x_{ij}}, Z_{y_i} \leftarrow nn_{y_{ij}}$ (8)
Step 2: Approximate the local neighborhoods for x_i and y_i
for all $x' \in Z_{x_i}$ and $y' \in Z_{y_i}$ **do**:
$\overline{Z_{x_i}} \leftarrow$ sample_around(x'), $\overline{Z_{y_i}} \leftarrow$ sample_around(y') (9)
Step 3: Compute relative proximities among x_i, y_i and their respective neighbors
for all $x' \in \overline{Z_{x_i}}$ and $y' \in \overline{Z_{y_i}}$ **do**:
$\overline{\pi_{x_i}} \leftarrow \pi_{x_i}(x'), \overline{\pi_{y_i}} \leftarrow \pi_{y_i}(y')$
Step 4: Order data-vectors in terms of ascending proximity
$\overline{\overline{Z_{x_i}}} \leftarrow$ sort$(\overline{Z_{x_i}})$, $\overline{\overline{Z_{y_i}}} \leftarrow$ sort $(\overline{Z_{y_i}})$
Step 5: Compute feature distance contributions for x_i and y_i
Compute $FC_{Z_{x_i}} \leftarrow$ feature_contribution$(\overline{\overline{Z_{x_i}}})$ (11)
Compute $FC_{Z_{y_i}} \leftarrow$ feature_contribution$(\overline{\overline{Z_{y_i}}})$ (11)
Step 6: Compute feature influence explanations for x_i and y_i
$fie(x_i) \leftarrow$ corr$(FC_{Z_{x_i}}, \overline{\pi_{x_i}})$, $fie(y_i) \leftarrow$, $corr(FC_{Z_{y_i}}, \overline{\pi_{y_i}})$ (12)
Step 7: Compute the local divergence score for x_i
Compute λ_{x_i}. (13)

Once Y is obtained, the LAPS process is initiated (cf. Algorithm 1). As the first step, the user interactively selects a single data-point $x_i \in X$. Considering, $y_i \in Y$ being the low-dimensional counterpart of x_i i.e., $f(x_i) = y_i$, LAPS begins with the identification of the localities (i.e., neighborhood structure) around x_i and y_i by performing an unsupervised k-nearest neighbor search using the *ball-tree* [31] algorithm. Where the nearest neighbors nn_{x_i} for x_i and nn_{y_i} for y_i can be defined as:

$$\begin{aligned} nn_{x_i} &= \{\forall x' \in X | \forall x'' \in X, x' \neq x'' : \pi_{x_i}(x') \leq \pi_{x_i}(x'')\} \\ nn_{y_i} &= \{\forall y' \in Y | \forall y'' \in Y, y' \neq y'' : \pi_{y_i}(y') \leq \pi_{y_i}(y'')\}. \end{aligned} \tag{8}$$

After identification of the indexes of the k-nearest neighbors for both x_i and y_i, the original features vectors from X for the closest neighbors of x_i and y_i are selected and combined into feature vector matrices Z_{x_i} and Z_{y_i} respectively, where, $x_i \in Z_{x_i}$ and $y_i \in Z_{y_i}$. To enhance user-interactions with the LAPS process, we allow the value of k to be user-defined. The primary reasons behind using the *ball-tree* algorithm for an unsupervised k-nearest neighbor search are: firstly, the algorithm is well-known [31] for its efficiency with the fast discovery of nearest neighbors in high-dimensional manifolds. Secondly, f being a data-transformation technique, the neighborhood structure after using f has no direct impact from the training labels associated with the data points.

Next, to approximate the local neighborhoods for x_i and y_i, LAPS samples instances around each $x' \in Z_{x_i}$ and $y' \in Z_{y_i}$. During this step, a constant number of data-point samples are drawn uniformly at random having a normal distribution centered around each $x' \in Z_{x_i}$ and $y' \in Z_{y_i}$. Formally, the sampling of each perturbed neighbor for any data point $x' \in Z_{x_i}$ and $y' \in Z_{y_i}$ can be defined as:

$$\begin{aligned} \bar{x}_i &= \forall x' \in Z_{x_i}, \forall {u'}_a \in x', \delta \in [0,1] : \{\delta \times \sigma_{{u'}_a} + \mu_{{u'}_a}\} \\ \bar{y}_i &= \forall y' \in Z_{y_i}, \forall {v'}_a \in y', \delta \in [0,1] : \{\delta \times \sigma_{{v'}_a} + \mu_{{v'}_a}\}. \end{aligned} \tag{9}$$

where δ represents a random perturbation, as $\sigma_{u'_a}$, $\sigma_{v'_a}$ represent the standard deviation[2] of an individual feature value in Z_{x_i} and Z_{y_i} respectively. At the same time, $\mu_{u'_a}$ and $\mu_{v'_a}$ signify the means of individual feature values in Z_{x_i} and Z_{y_i} respectively. To ensure *local fidelity*, such an approximation of local structure of data points is a commonly-practiced [28], [32] approach among researchers. The perturbed neighborhood for each point in both Z_{x_i} and Z_{y_i} are combined into feature vector matrices $\overline{Z_{x_i}}$ and $\overline{Z_{y_i}}$ respectively.

In the next step, the relative proximities: $\pi_{x_i}(x')$ is calculated between the points x_i and $\forall x' \in \overline{Z_{x_i}}$ and $\pi_{y_i}(y')$ is computed between y_i and $\forall y' \in \overline{Z_{y_i}}$. In case of the feature vectors for x_i and y_i containing only continuous values, the euclidean distance (cf. Eq. (4)) is used as the proximity measures $\pi_{x_i}(x')$ and $\pi_{y_i}(y')$. In contrast, in the case of feature vectors with a mixture of both continuous and categorical values, the Gower dissimilarity [9], [11], [33] is used as the proximity measures $\pi_{x_i}(x')$ and $\pi_{y_i}(y')$ between instances. The Gower dissimilarity [33] $dist_\omega$ between any pair of data points x_i, x' can be defined as:

$$dist_\omega(x_i, x') = \sum_{u=1}^{D} \delta_{x_i x' u} \times dist_{\omega_{x_i x' u}} / \sum_{u=1}^{D} \delta_{x_i x' u}, \tag{10}$$

where u represents an individual attribute in $\overline{Z_{x_i}}$ as $dist_\omega$ signifies the distance between x_i and x' for attribute u. In case of continuous variables $dist_\omega$ is calculated as $abs|x_{iu} - x'_u| / range(u)$. For categorical variables, $dist_\omega$ is 0 if $x_{iu} = x'_u$, otherwise 1. In Eq. (10), $\delta_{x_i x' u}$ is 1 if x_{iu} and x'_u are comparable, otherwise 0. In our experiments, we do not consider weighted proximity measures [9] in par with our *algorithm-agnostic* design goal because not every dimensionality reduction method considers additional feature weights in their process [3].

Next, every data-vector $x' \in \overline{Z_{x_i}}$ and $y' \in \overline{Z_{y_i}}$ are ordered in terms of their (ascending) proximity with the original (data) points x_i and y_i respectively. We represent these ordered feature-vectors matrices as $\overline{\overline{Z_{x_i}}}$ and $\overline{\overline{Z_{y_i}}}$ respectively. Also, as the ascending proximity values between x_i and the data points $x' \in \overline{Z_{x_i}}$ are stored in a set $\overline{\pi_{x_i}}$, the same for y_i and the data points $y' \in \overline{Z_{y_i}}$ are stored as $\overline{\pi_{y_i}}$. The feature-vector matrices $\overline{Z_{x_i}}$ and $\overline{Z_{y_i}}$, are then used compose two feature distance

2. The number of data points in the puturbed neighborhood of every data point is fixed to 5000, Hence, due to the effect of central limit theorem, the distribution of feature values were noticed to be very close to a Gaussian distribution.

contribution [34] matrices namely, $FC_{Z_{x_i}}$ and $FC_{Z_{y_i}}$ respectively. Each element in a feature distance contribution matrix holds the impact of each attribute in the overall distances between a pair of consecutive points. We define the elements in $FC_{Z_{x_i}}$ and $FC_{Z_{y_i}}$ as:

$$\forall x'_i, x'_{i+1} \in \bar{\bar{Z}}_{x_i}, \forall u \in x'_i, x'_{i+1} : \pi_{x'_{i_u}}(x'_{i+1_u})/\pi_{x'_i}(x'_{i+1})$$
$$\forall y'_i, y'_{i+1} \in \bar{\bar{Z}}_{y_i}, \forall v \in y'_i, y'_{i+1} : \pi_{y'_{i_u}}(y'_{i+1_u})/\pi_{y'_i}(y'_{i+1}). \tag{11}$$

where u represents an individual attribute in $\overline{Z_{x_i}}$ and v represents the same in $\overline{Z_{y_i}}$. The points x'_i, x'_{i+1} and y'_i, y'_{i+1} signify two consecutive data-vectors in $\overline{Z_{x_i}}$ and $\overline{Z_{y_i}}$ respectively. We build the two matrices based on the concept of feature distance contribution [34], which represents a ratio showing the importance (i.e., contribution) of an individual feature in the overall distance between two points.

Finally, from the feature distance contribution matrices, LAPS generates the first component of the local explanations: feature influence explanations $fie(x_i)$ for x_i using the Pearson's correlation[3] [4] between each column (representing the distance contribution for each feature) in $FC_{Z_{x_i}}$ with ordered overall distances $\overline{\pi_{x_i}}$ between the data points in $\overline{Z_{x_i}}$. Similarly, feature influence explanations $fie(y_i)$ are calculated for y_i, the embedding counterpart of x_i from $FC_{Z_{y_i}}$ and $\overline{\pi_{y_i}}$. Formally, the feature influence explanations $fie(x_i)$ and $fie(y_i)$ can be defined as:

$$\begin{aligned} fie(x_i) &= \left\{ \forall fc_{xa} \in FC_{Z_{x_i}} : \frac{cov(fc_{xa}, \overline{\pi}_{x_i})}{\sigma_{fc_{xa}} \sigma_{\overline{\pi}_{x_i}}} \right\} \\ fie(y_i) &= \left\{ \forall fc_{ya} \in FC_{Z_{y_i}} : \frac{cov(fc_{ya}, \overline{\pi}_{y_i})}{\sigma_{fc_{ya}} \sigma_{\overline{\pi}_{y_i}}} \right\} \end{aligned}, \tag{12}$$

Where, fc_{xa} and $\bar{\pi}_{x_i}$ as well as fc_{ya} and $\bar{\pi}_{y_i}$ are the pairs of random variables under consideration. In Eq. (12), $cov\ (fc_{xa}, \overline{\pi}_{x_i}) = \sum_{j=1}^{n} (fc_{xa_j} - \mu_{fc_{xa}})(\bar{\pi}_{x_i}(x_j) - \mu_{\bar{\pi}_{x_i}})$ and $\sigma_{fc_{xa}} = \sqrt{\sum_{j=1}^{n} (fc_{xa_j} - \mu_{fc_{xa}})^2}$ as $\sigma_{\overline{\pi}_{x_i}} = \sqrt{\sum_{j=1}^{n} (\bar{\pi}_{x_i} - \mu_{\bar{\pi}_{x_i}})^2}$. The overall $fie(x_i)$ and $fie(y_i)$ are represented using a *set* of key-value pairs, where, the keys fc_{xa} and fc_{ya} represent individual attributes in matrices $FC_{Z_{x_i}}$ and $FC_{Z_{y_i}}$ respectively. Finally, we compute the local divergence λ_{x_i} for x_i as:

$$\lambda_{x_i} = w_1\ \pi_{fie(x_i)}(fie(y_i)) + w_2 \frac{nn_{x_i} \cap nn_{y_i}}{|nn_{x_i}|} + w_3 d_{r_{nn_{x_i}, nn_{y_i}}}, \tag{13}$$

where, $\pi_{fie(x_i)}\ (fie(y_i))$ signifies the *cosine* distance between $fie(x_i)$ and $fie(y_i)$. As $d_{r_{nn_{x_i}, nn_{y_i}}}$ represents the difference between the relative orders of neighborhoods of x_i and y_i. In Eq. (13), $w_1,\ w_2,$ and w_3 signify user-defined scalar weights of the three components of λ_{x_i}. By default, $w_1,\ w_2,$ and w_3 are equal, i.e., 0.33.

Since Algorithm 1 produces explanations for a single datapoint in X, its complexity does not depend on the size of X, but on the user-defined size of the sampled neighborhood $\overline{Z_{x_i}}$ (Eq. (8)) for the selected instance. As per our analysis, the run-time complexity of Algorithm 1 is $O(n^2)$, n being the number of samples in $\overline{Z_{x_i}}$. In practice, on a personal computer (i.e., with 4 cores and 8 GB main memory) LAPS executes in less than 15 seconds for $n = 5000$ data points.

3. Experiments with Spearman's correlation (non-parametric) returned the same values with Pearson's (parametric) until the second decimal point

4.3 Global Approximation of Projection Space

To ensure the *global legitimacy* of explanations, we now propose the algorithm *GAPS* that generates an estimation for the retained global structure of the projection space. The preserved global structure is explained as a unification of preserved local structures [28], [35] for a subset X_S of non-redundant data points in X. Acknowledging the importance of a judicious selection of X_S for an accurate global approximation, *GAPS* enables two different ways for formulating X_S from X. Here, either the users interactively pick data points around the manifold or GAPS selects a fixed number of instances uniformly at random belonging to each training label around X. In either case, GAPS lets the users determine the number of instances that they are willing to investigate, and represents it with a budget B. Potentially, the formation of X_S can also be represented using an Exhaustive Subset Enumeration [36] problem. We envision the maximization of diversified sample selection for X_S as future work.

Algorithm 2. The *GAPS* Procedure

Input: dataset X, data subset X_S, budget B, neighbor count k
Output: global divergence λ_X
Step 1: Generate local feature influence explanations
  for all $x_i \in X_S$ **and** $y_i \in Y_S$ **do**:
    $LFI_{X_S} \leftarrow fie(x_i),\ LFI_{Y_S} \leftarrow fie(y_i), \lambda_{X_S} \leftarrow \lambda_{x_i}$ Algo. 1
Step 2: Approximate the local neighborhoods for the nearest neighbors of all $x_i \in X_S$ and $y_i \in Y_S$
  for all $x_i \in X_S$ **and** $y_i \in Y_S$ **do**:
    for all $j \in \{0, 1, \ldots\ldots, k\}$ **do**:
      $Z_{X_S} \leftarrow$ sample_around($nn_{x_{ij}}$) (8),(9)
      $Z_{Y_S} \leftarrow$ sample_around($nn_{y_{ij}}$)
Step 3: Compute pairwise proximities between each pair of points in feature vectors in Z_{X_S} and Z_{Y_S} followed by an estimation of the overall feature distance contribution and global feature influence explanations
  for all $x_i,\ x_j \in Z_{X_S}$ **and** $y_i,\ y_j \in Z_{Y_S}$ **do**: (10),(11)
    Compute $\pi_{x_i}(x_j), \pi_{y_i}(y_j)$, Compute $fie(X_S),\ fie(Y_S)$
Step 4: Obtain an approximation of the global divergence $\widehat{\lambda_{X_S}}$ for the selected subset X_S
  Compute $\widehat{\lambda_{X_S}}$ (12)
Step 5: Calculate the overall global divergence for X using the Global-Local Approximation (GLA) approach
  Compute λ_X (13)

Once the data points in X_S are selected, as shown in Algorithm 2, GAPS obtains a set of local explanations for $x_i \in X_S$ and $y_i \in Y_S$, where, $y_i = f(x_i)$. Next, as the local feature influence explanations for the instances in X_S and Y_S are used to compose two $B \times D$ dimensional matrices LFI_{X_S} and LFI_{Y_S} respectively, the local divergences for each point in X_S and Y_S are represented using the sets λ_{X_S} and λ_{Y_S} respectively. In parallel, GAPS obtains a global estimate of the structural relations among the data points in X_S. As the first step towards obtaining this, the approximated local neighborhoods (cf. Eq. (9)) for each $x_i \in X_S$ and $y_i \in Y_S$ are

combined into two feature vector matrices namely Z_{X_S} and Z_{Y_S} respectively. Next, pairwise proximities between each pair of points in feature vectors in Z_{X_S} and Z_{Y_S} are calculated. Considering the proximities among data points around a high-dimensional manifold, in GAPS, we use the Geodesic distances (cf. Eq. (5)) among the pairs of points in Z_{X_S} and Z_{Y_S}. After ordering the data point pairs in ascending order of proximity, using equations (11) and (12) an estimation of the overall feature distance contribution and global feature influence explanations are obtained. Finally, similarly as LAPS, an approximation of the global divergence $\lambda_{\widehat{X_S}}$ for the selected subset X_S is obtained as a weighted sum of the disagreements in the overall estimation of the feature influences, and the disagreements in the neighborhood structures for X_S and Y_S.

Boyd *et al.* [35] and Haftka *et al.* [37] show, on the one hand, the local approximation of divergence for each data-point in X_S is the most effective near the point where it was calculated. However, the accuracy of such local approximations can deteriorate [37] as it moves away from the point where it was constructed. In contrast, a global approximation may not be accurate for every data-point in the manifold, its quality does not deteriorate with distance. Hence, in GAPS we follow the Global-Local Approximation (GLA) [37] approach. GLA allows an additive blending of local approximations to form a globally-valid approximation. Here, before the unification of the local-approximations, the ratio of the global estimate to each of the local approximation is used as a scaling factor [35] to multiply the local-approximations. Hence, we define the overall divergence in the preserved global structure λ_X as:

$$\lambda_X = \sum_{j=1}^{B} \frac{\lambda_{X_{S_j}}}{\lambda_{\hat{X}_S}} \lambda_{X_{S_j}}, \tag{14}$$

where, λ_X represents an additive blending of scaled local divergence scores $\lambda_{x_i} \in \lambda_{X_S}$.

Although Algorithm 2 presents a unified approximation for B instances in X, it has a run-time complexity of $O(n^2)$, n being the number of row vectors in the unified perturbed neighborhood matrix Z_{X_S}.

5 Experimental Evaluation

In this section, we present the results of our experimental evaluations of the two proposed techniques. This section aims at answering the following questions:

1. Do LAPS and GAPS fulfill their design requirements discussed in Section 3.1?
2. Can the proposed methods instill confidence in the projection and enable the selection of a suitable dimensionality reduction algorithm?
3. Are the explanations able to effectively explain the structural preservations in embeddings?
4. Can the explanations be considered as an improvement over the most closely related work?
5. Do the explanations remain consistent for different user-selected parameter combinations?

Based on the above-mentioned questions, our evaluation of the proposed methods was performed in five phases. Firstly, we applied the techniques on 16 real-world datasets to assess the applicability of the methods. Secondly, we investigated the role of the local and global divergence scores in the selection of the most suitable dimensionality reduction algorithms. Thirdly, we executed a user-study to assess the utility of the two techniques. Fourthly, we compared the proposed techniques to the most closely related research that aims at interpreting embeddings. Finally, we analyzed the impact of different user-defined parameter combinations of the proposed techniques. Overall, this section is divided into two sub-sections; first, we define the algorithms and datasets that were used in our experiments, followed by a detailed analysis of the answers to the above-mentioned questions.

5.1 Experimental Setup

To ensure the *model & data-type agnostic* nature of the proposed algorithms, we compare the structural retention of 10 state-of-the-art linear and non-linear dimensionality reduction methods for 16 real-world datasets. The algorithms include popular techniques such as PCA [38], t-distributed Stochastic Neighbor Embedding (t-SNE) [39], Uniform Manifold Approximation and Projection (UMAP) [40], openTSNE [41], MDS [42], Isomap [25], Locally Linear Embedding (LLE) [43], Variational Autoencoder (VAE) [44], Local tangent space analysis (LTSA) [45], and KernelPCA [46]. The 16 high-dimensional datasets[4] used in our experiments belong to four different data-types namely tabular, text, audio, and images. As 14 out of the 16 datasets were selected from the Kaggle[5] data repository, the Animals image dataset [9] and the UrbanSound8k[6] dataset were selected from related literature. For our experiments, we decided to consider labeled datasets only. Moreover, since dimensionality reduction only works with tabular numeric data [3], we pre-processed the non-tabular datasets before the experiments. For example, as the audio datasets were converted to time-series data of sound amplitudes using the *librosa*[7] library, the text datasets were converted into word embeddings using *Word2Vec*[8] models.

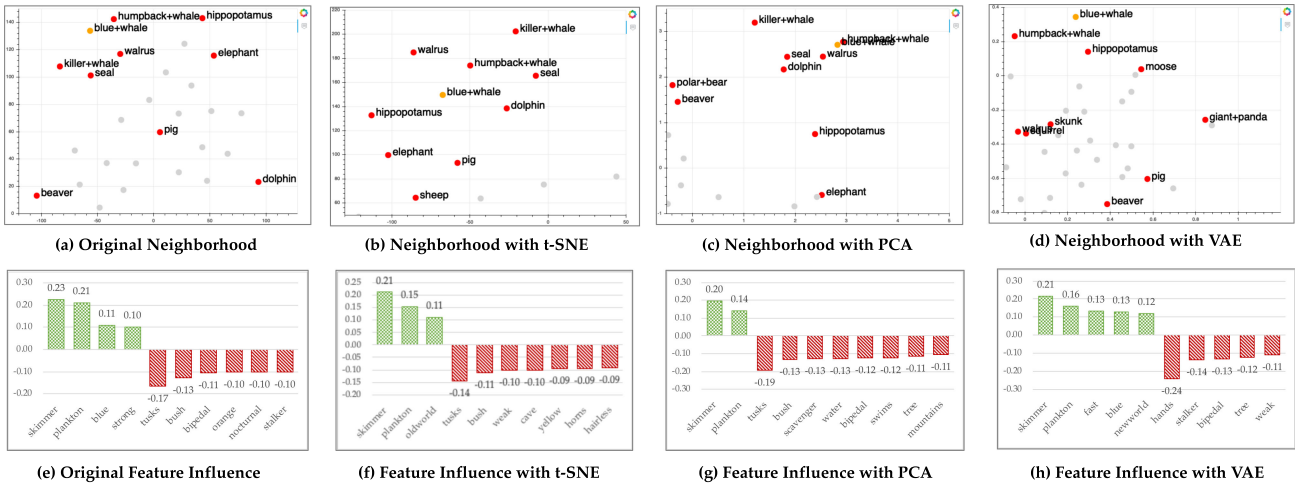
5.2 Experimental Results

This section presents the experimental results for LAPS and GAPS. We divide the section into four parts based on the questions defined at the beginning of Section 5.

5.2.1 Applicability Analysis of LAPS and GAPS

To assess whether LAPS and GAPS fulfill their design requirements discussed in Section 3.1, now we present two case-studies applying the techniques on the image, and tabular datasets. Two more case-studies with audio and text datasets are presented as *supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2020.3005878*.

4. Breast Cancer, Adult, Wine Quality, Credit Card, Animals, MNIST, Flower17, Fashion-MNIST, UrbanSound8K, ESC50, GTZAN, Free-Spoken-Digits, Sentiment140, BBC-Text, SMS Spam Collection, Quora Question Pairs
5. https://www.kaggle.com/datasets
6. https://urbansounddataset.weebly.com/urbansound8k.html
7. https://librosa.github.io/librosa/
8. https://pypi.org/project/gensim/



(a) Original Neighborhood (b) Neighborhood with t-SNE (c) Neighborhood with PCA (d) Neighborhood with VAE

(e) Original Feature Influence (f) Feature Influence with t-SNE (g) Feature Influence with PCA (h) Feature Influence with VAE

Original	blue whale	humpback whale	walrus	seal	dolphin	hippopotamus	killer whale	elephant	pig	beaver
t-SNE	blue whale	humpback whale	walrus	dolphin	hippopotamus	pig	elephant	seal	killer whale	sheep
PCA	blue whale	humpback whale	walrus	seal	dolphin	killer whale	hippopotamus	elephant	polar bear	beaver
VAE	blue whale	humpback whale	hippopotamus	walrus	giant panda	moose	squirrel	pig	beaver	skunk

(i) Order to data-points in Neighborhood

	t-SNE	PCA	VAE
Discrepancy in Feature Influence	0.78	0.79	0.73
Discrepancy in Neighborhood Contents	0.10	0.10	0.20
Discrepancy in Neighborhood Order	0.70	0.30	0.70

Overall Local Divergences	
t-SNE	0.5268
PCA	0.3983
VAE	0.5434

Fig. 3. Application of LAPS on the Animals dataset. Scatter plot in (a) shows the original neighborhood structure of data-point blue-whale, as the plots in (b to d) represent the same after running different dimensionality reduction algorithms. The bar-graphs in (e to h) show the feature influence explanations for the relative distances in the neighborhood of the point. In the above figure, each of the three components of the local divergence has a weight of 0.33 (i.e., the default weight).

Case Study 1: Image data - Animal Dataset

Fig. 3 shows an application of LAPS on the Animal dataset [9]. The Animal dataset contains 30,475 images and is composed of 85 numeric attributes and 50 animal classes. In Fig. 3, we explain the use of LAPS with a subset of the original dataset to enhance visual clarity. Here we select the data-point with the label *blue-whale* as our point of interest. Fig. 3a shows the 10 nearest-neighbors of *blue-whale* in the original 85-dimensional dataset on a two-dimensional projection. As shown in Fig. 3a, in the original dataset some of the most closely related data points to *blue-whale* are the *humpback-whale*, *walrus*, *seal*, *dolphin*, and *killer-whale*, whilst points such as *pig* and *elephant* are also considered neighbors of *blue-whales* for resemblances in their values for the attribute *strong* (cf. Fig. 3e). Showing our analysis with LAPS, Fig. 3e explains the most influential attributes for the neighborhood structure shown in Fig. 3a. As the green bars in Fig. 3e represent a positive correlation of an attribute's contribution to the relative distances in the neighborhood, the red bars show the negatively influencing attributes for the same. From Fig. 3e it can be seen that for the neighborhood of *blue-whale* the most positively influential attributes are *skimmer*, *plankton*, *blue*, and *strong*. On the other hand, the most negatively influential attributes include *tusks* and *bush* that separate *blue-whale* from some of its closest neighbors such as the *elephant*. Fig. 3e also shows that the attributes that are positively or negatively influential with similar magnitude. Due to this similarity, we consider them to be *highly correlated* with each other.

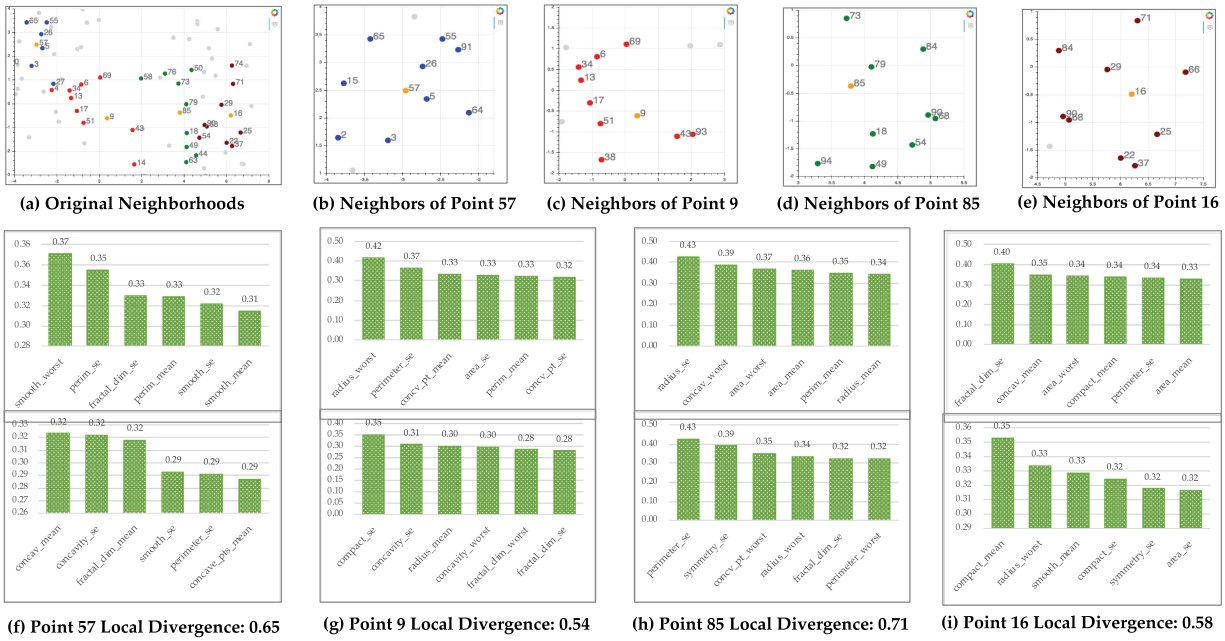
Figs. 3b and 3c show the neighborhoods for *blue-whale* after the application of t-SNE and PCA on the dataset respectively. As shown in the Figs. 3b and 3c as well in Fig. 3i, both t-SNE and PCA preserved 9 out of 10 neighbors for *blue-whale* in the embedding, while replacing the original neighbor *beaver* with *sheep* and *pig* with *polar bear* respectively. Nevertheless, the attribute influencing in the neighborhood structures are significantly changed for both t-SNE and PCA (cf. Figs. 3f and 3g). Looking into Fig. 3.i it can be seen that VAE preserved 6 out of 10 neighbors in the projection and considered points such as *giant panda*, *moose*, *squirrel*, and *skunk* as the neighbors of *blue whale*. As a result, with VAE additional attributes such as *hands*, *tree*, and *weak* show significant negative contribution on the neighborhood structure.

Overall, Fig. 3 shows that all algorithms have performed poorly in terms of preserving the attribute influences in the embeddings. In terms of preserving the neighborhood components and neighborhood orders, t-SNE and PCA have performed relatively better than VAE. Hence, the neighborhood order contributed the most in the comparison of their local-divergences. Here, PCA has performed much better than both the other algorithms, resulting in the lowest local divergence score among the three.

Case Study 2: Tabular data - Breast Cancer Dataset

Our second case study focuses on the Breast Cancer dataset [9] that classifies tumors into malignant and benign. This tabular numeric dataset is composed of 32 attributes and 569 data points. Fig. 4 primarily shows the utility of GAPS with the breast cancer data using *t-SNE* as the used dimensionality reduction technique. The analysis begins with an interactive selection of four[9] non-redundant data points (with indexes: 9,

9. Here we select only four data points only to enhance visual clarity. A detailed discussion on appropriate budget size is presented in Section 5.2.5



(a) Original Neighborhoods (b) Neighbors of Point 57 (c) Neighbors of Point 9 (d) Neighbors of Point 85 (e) Neighbors of Point 16

(f) Point 57 Local Divergence: 0.65 (g) Point 9 Local Divergence: 0.54 (h) Point 85 Local Divergence: 0.71 (i) Point 16 Local Divergence: 0.58

Original	57	5	3	65	55	26	4	27	0	6
Embedding	57	5	26	15	64	3	91	65	55	2
Original	9	17	43	14	6	13	51	69	4	34
Embedding	9	51	43	17	38	93	69	6	13	34
Original	85	49	79	44	18	73	50	58	63	76
Embedding	85	79	18	73	84	90	68	54	94	49
Original	16	22	29	25	37	68	54	71	74	90
Embedding	16	29	25	66	22	68	37	90	71	84

Original Distance Ordering: [[16, 57], [16, 9], [9, 57], [85, 9], [16, 85], [85, 57]]
Embedding Distance Ordering: [[16, 57], [85, 57], [9, 57], [85, 9], [16, 85], [16, 9]]

(j) Original Vs. Embedding Neighborhoods

Global Divergence with default weights (i.e., $w_1, w_2, w_3 = 0.33$) = 0.32074

Fig. 4. Application of GAPS on the Breast Cancer dataset. As the scatter plot in (a) shows the original neighborhoods of the four data points, the plots in (b to e) show their neighborhoods after transforming the dataset using t-SNE. In the scatter plots, the points selected for analysis are colored in 'orange' and the neighborhoods of the four points 57, 9, 85, and 16 are colored in 'blue', 'red', 'green', and 'maroon' respectively. The bar-graph pairs (f to i) show the feature influence explanations of the points in the original dataset on top and the same in the embedding at the bottom. The tabular representation in (j) shows the original versus the projected neighborhoods of the selected points. Here the four selected data points are highlighted in yellow as their original and embedding neighborhoods are highlighted in grey and white respectively.

16, 57, 85) from the dataset. In the original neighborhood structure of these points (cf. Fig. 4a) we can see the points are far away from each other on the manifold and have no overlaps in neighbors.

From Figs. 4b, 4c, 4d, 4e, and 4j it can be seen that after t-SNE, as two of the original neighbors of points 9 and 16 are replaced by two different points in the embedding, for the points 57 and 85 four and five points are replaced respectively. In terms of attribute influences, in the original neighborhood of point 57 (Fig. 4f), the most influencing feature *smoothness_worst* is replaced by *concavity-mean* in the embedding. Moreover, some of the highly influencing attributes such as *fractal_dimension_mean*, *perimeter_mean*, and *smoothness_mean* are not included in the group of six most influential attributes in the embedding neighborhood. Similar discrepancies (cf. Figs. 4g, 4h, 4i) in feature influences are noticed for all the points. Nevertheless, from all the original attributes influence bar-graphs, it can be seen that for the chosen points attributes such as *perimeter-mean*, *area-se*, *fractal_dimension_mean*, and *area-mean*, are *highly correlated* attributes in the neighborhoods of all the four selected data points. Along with the divergence in the preserved global structure, in Fig. 4 the analyst can also see the disagreement in the order of preserved relative distances between the four selected data points in the original and their low-dimensional embedding. This disagreement shows that t-SNE has failed to preserve the original relative proximities among the four selected points in the embeddings. We think the reason for this is, t-SNE being a locally focused dimensionality reduction technique has preserved 50% to 80% of the nearest neighbors for each of the selected points whilst disrupting the global distances among the neighborhoods of the four points in the embedding.

5.2.2 *Selection of Appropriate Algorithm Using LAPS*

In this section, we present our assessment of LAPS and GAPS for their ability to instill confidence in a single projection within a group of projections and to assist with the selection of an appropriate dimensionality reduction algorithm for any dataset. Table 1 presents the local divergence scores obtained using LAPS for all the 16 datasets. In Table 1, we compute the mean local divergence scores for 100 random data points from every dataset. To validate the local divergences obtained from LAPS, following the guidelines of Maaten *et al.* [3], we use the generalization error of the 1-nearest neighbor classification algorithm. 1-NN generalization error being a popular metric [3] for the validation of retained local structure in an embedding. In Table 1, we present a side-by-side comparison between the divergence scores obtained from LAPS and the 1-NN generalization errors for the 10 algorithms. Similar to Maaten *et al.* [3], we compute the generalization errors using leave-one-out cross-validation. The results in Table 1 show that in 75% of cases LAPS agree with the 1-NN generalization error scores on the algorithm that preserved most of the local structure.

TABLE 1
Divergence Scores of Dimensionality Reduction Algorithms Using LAPS Vs. 1-NN Generalization Errors

		LAPS Divergence Scores										1-NN Generalization Errors									
Datasets	Type	tSNE	PCA	UMAP	oTSNE	MDS	ISMP	LLE	KPCA	LTSA	VAE	tSNE	PCA	UMAP	oTSNE	MDS	ISMP	LLE	KPCA	LTSA	VAE
Animal	Img	0.472	0.308	0.298	**0.275**	0.334	0.457	0.492	0.308	0.432	0.417	0.418	0.292	**0.263**	0.298	0.353	0.342	0.362	0.290	0.327	0.332
MNIST	Img	**0.357**	0.399	0.458	0.362	0.382	0.362	0.386	0.477	0.362	0.391	0.225	0.388	0.545	0.265	**0.213**	0.622	0.604	0.846	0.323	0.47
FLOWER17	Img	0.560	0.556	0.559	0.560	0.407	**0.375**	0.406	0.407	0.523	0.544	0.632	0.812	0.716	0.628	0.791	**0.577**	0.813	0.818	0.880	0.742
F-MNIST	Img	0.490	0.527	**0.419**	0.489	0.525	0.521	0.528	0.427	0.546	0.443	**0.249**	0.529	0.320	0.255	0.493	0.450	0.529	0.357	0.640	0.387
Brst Cancer	Tabl	**0.469**	0.667	0.535	0.497	0.526	0.593	0.662	0.56	0.66	0.524	**0.046**	0.088	0.056	0.047	0.090	0.056	0.070	0.088	0.903	0.159
Magic	Tabl	**0.313**	0.61	0.446	0.313	0.613	0.646	0.711	0.612	0.541	0.398	**0.186**	0.369	0.285	0.186	0.338	0.372	0.380	0.336	0.350	0.247
Wine Qlty	Tabl	**0.503**	0.537	0.537	0.603	0.604	0.56	0.57	0.537	0.504	0.518	**0.404**	0.415	0.412	0.495	0.470	0.452	0.541	0.415	0.545	0.46
Crdt Card	Tabl	**0.46**	0.587	0.592	0.581	0.527	0.529	0.662	0.587	0.556	0.507	**0.258**	0.296	0.311	0.293	0.305	0.289	0.370	0.290	0.280	0.281
ESC50	Aud	**0.597**	0.664	0.629	0.631	0.696	0.63	0.695	0.664	0.632	0.641	**0.705**	0.898	0.826	0.717	0.907	0.935	0.895	0.898	0.886	0.887
Urban8k	Aud	**0.329**	0.596	0.463	0.329	0.663	0.529	0.629	0.596	0.658	0.415	**0.122**	0.486	0.276	0.128	0.448	0.524	0.594	0.486	0.840	0.413
Spkn Digits	Aud	**0.496**	0.694	0.591	0.502	0.688	0.588	0.69	0.682	0.601	0.577	**0.005**	0.273	0.009	0.005	0.191	0.122	0.089	0.273	0.286	0.122
GTZAN	Aud	0.495	0.487	0.46	0.523	**0.457**	0.523	0.558	0.504	0.558	0.624	0.459	0.326	0.256	0.454	**0.299**	0.311	0.324	0.345	0.425	0.351
SMS Spam	Txt	**0.474**	0.508	0.513	0.499	0.475	0.475	0.508	0.508	0.474	0.499	**0.256**	0.325	0.544	0.260	0.345	0.360	0.311	0.300	0.446	0.379
Quora	Txt	**0.57**	0.57	0.537	0.558	0.603	0.536	0.603	0.57	0.57	0.535	**0.021**	0.211	0.199	0.022	0.054	0.164	0.254	0.337	0.185	0.154
BBC-Text	Text	**0.447**	0.678	0.546	0.571	0.571	0.646	0.621	0.675	0.685	0.688	**0.149**	0.270	0.449	0.149	0.356	0.225	0.297	0.315	0.333	0.27
Sntmt140	Txt	0.575	0.572	**0.502**	0.593	0.569	0.638	0.615	0.567	0.571	0.603	0.658	0.717	0.682	**0.655**	0.745	0.803	0.811	0.752	0.699	0.721

Description of Acronyms in *Table 1: Img: Image, Tabl: Tabular, Aud: Audio, Txt: Text, oTSNE: openTSNE, ISMP: Isomap, F-MNIST: Fashion MNIST, Brst Cancer: Breast Cancer, Wine Qlty: Wine Quality, Crdt Card: Credit Card, Spkn Digits: Spoken Digits, Sntmt140: Sentiment140.*
Note: *In the above table, the left hand side shows an average of LAPS divergence scores for 100 data points from the 16 datasets using 10 different dimensionality reduction algorithms. The right hand side shows the 1-NN generalization errors using the same algorithms on the same datasets. The algorithms with lowest local-divergences and 1-NN gen. errors are highlighted in bold & red. In both the sides, the column representing the algorithm with the lowest local divergence and 1-NN generalization error for more than 50% of the datasets are highlighted in grey.*

For the remaining 25% cases, where the lowest divergences from LAPS do not agree with lowest 1-NN generalization errors, we perform paired t-tests[10] [47] to compare the embeddings obtained from the algorithms suggested by LAPS and 1-NN generalization error. As shown in Table 2, the p-values obtained from statistical significance analysis show no significance differences between the results of the two techniques. Hence, from Table 1 it can be seen that multiple iterations of LAPS can help users to select an algorithm that has preserved most of the local structure of the original dataset in its embedding. A similar analysis with GAPS divergence scores are presented as *supplemental material, available online*.

One could argue on the superiority of LAPS over 1-NN generalization error for evaluating the preserved local structure in an embedding [3]. However, the computation of the 1-NN generalization error is black-box to its users, as users cannot interact with the computation of the metric. Whereas, LAPS allows users to interpret and interact with each component of the metric local-divergence. Moreover, by allowing users to define weights for different components of the divergence calculation, users can decide the importance of the feature influences over the neighborhood structures in divergence calculation, based on their domain expertise.

5.2.3 Evaluation With Human Subjects

To assess the utility of LAPS and GAPS in real-world data analysis, we have performed a user study with 10 human subjects. In this section, we present a brief overview of our study, more detailed information regarding the study questions, gathered insights, and format of discussion are presented as supplemental material, available online.

The participants of our study included both industry professionals and Ph.D. candidates with strong analytical backgrounds. During our study, the subjects were divided into two groups namely novice and expert participants. The novice-group contained 6 individuals with no prior knowledge of dimensionality reduction. Whereas, the expert-group comprised 4 individuals with moderate experience with dimensionality reduction techniques. In this study, we asked the subjects to analyze the Wine Quality dataset (cf. Table 1) that classifies 4898 wine samples into 10 quality categories using 12 attributes. The first stage of the study was conducted in three phases. At first, both the participant groups were briefed about the details of the given dataset (i.e., attributes and labels). Then, both novice and expert participants were presented with embeddings of the dataset obtained using t-SNE, Isomap, and UMAP and were asked

TABLE 2
Statistical comparison of suggested algorithms

Dataset	Suggested algorithms		p-value paired t-test	
	LAPS	1-NN GE	Dim-1	Dim-2
Animal	oTSNE	UMAP	0.404	0.667
MNIST	tSNE	MDS	0.826	0.847
FashionMNIST	UMAP	tSNE	0.186	0.108
Sentiment140	UMAP	oTSNE	0.088	0.095

Note: GE: Generalization Error, Dim-1: first target dimension obtained from dimensionality reduction, Dim-2: second target dimension after dimensionality reduction. For the paired t-tests, the threshold α is considered to be 0.05 (i.e., the most commonly used value for α [27]). The results show, all obtained p-values are more thanα accepting the null hypothesis that embeddings are not statistically significantly different from each other.

10. The paired t-test [47] is the most common parametric statistical test to compare the mean of two sample populations.

TABLE 3
User-Agreement Analysis on LAPS and GAPS

Analytical Aspects	Fleiss' Kappa	
	Novice	Experts
Efficiently Explains	0.66	0.47
Enhances Trust	0.24	0.33
Helps with decision making	-0.17	0.11
Reduces analytical time	1.00	1.00

Note: *The table above summarizes the results of our user study. Here we analyze four different aspects of utility with LAPS and GAPS using 10 human subjects. Among them 6 are novice and 4 were expert participants.*

to manually investigate the embeddings. Finally, the participants were given a brief overview of the expected outcomes of executing LAPS and GAPS on these embeddings. In the next stage of the study, both novice and expert participants were given 10 minutes to execute the techniques and gather insights. In the final stage, that is after the 10 minutes, the participants were asked to provide feedback on four different aspects of LAPS and GAPS. These included: (i) Can the two techniques efficiently explain the structural preservations in the embeddings? (ii) Can executing the two techniques enhance user-trust on the embeddings? (iii) Can the techniques help users with decision making regarding the best performing algorithm? (iv) Can the techniques reduce the analytical time?

To quantitatively summarize the results of this study, following the idea of Lewis *et al.* [7] we computed Fleiss' Kappa consistency measure κ to assess the participant agreements on the feedbacks for LAPS and GAPS. The value of κ ranges from -1 to +1, where -1 represents no observed agreement, +1 signifies perfect agreement and 0 denotes agreement due to random chance. The results of our study are summarized in Table 4. The table shows that in terms of reduction of analytical time both novice and expert users had a perfect agreement about the utility of LAPS and GAPS over manual analysis. Besides, due to their prior experience with dimensionality reduction, the expert users could trust the results of the two techniques more easily (i.e., $\kappa = 0.33$) than the novice users (i.e., $\kappa = 0.24$). Regarding the ability of the techniques to provide an efficient explanation of embedding quality, both novice and expert users had a moderate agreement with κ of 0.66 and 0.47 respectively. In terms of decision making, approximately 50% of both novice and expert users agreed on the utility of LAPS and GAPS hence making κ close to 0.

Overall, the participants agreed on the utility of the two techniques in all four aspects of our analysis with some suggestions for improvements. For example, only 33% of the novice participants altered the relative weights during their analysis of the local and global divergence scores. At the same time, only 50% of the expert subjects could easily understand the discrepancies in all the individual components of local and global divergence scores, whilst the rest needed more assistance. As a proposed solution for both the problems, both novice, and expert users have proposed to integrate the LAPS and GAPS procedures as a part of a visual interactive framework. We envision this integration as future work.

5.2.4 Comparison With Related Research

In this section, we present the results of a behavioral comparison between our proposed techniques and the most closely related research. A more detailed result of this comparative analysis is presented as *supplemental material, available online*. The results of our analysis are summarized in Table 4. The table compares the design requirements of explanations presented in Section 3.1, with related techniques such as SIRIUS [9], Andromeda [23], along with the approaches presented by Pagliosa *et al.* [21], Martins *et al.* [20] and Silva *et al.* [22]. As shown in Table 4, the results of this analysis validate our claim from Section 2 which stated that the related research primarily focuses on individual aspects of interpreting and evaluating embeddings. For example, SIRIUS, Andromeda, and the techniques proposed by Pagliosa *et al.* [21] and Silva *et al.* [22] focus on identifying the most influential attributes in certain points or regions in the embeddings. Nevertheless, they do not attempt to quantify the different aspects of local and global structural preservations. At the same time, Martins *et al.* [20] present a diverse set of visualizations for neighborhood quality analysis they do not look into the aspect of the contributions of the original attributes in the structural preservations of embeddings. Moreover, none of the existing work considers model or data-type agnosticism among their design goals. Overall, our analysis shows that LAPS and GAPS indeed unifies the different aspects of analyzing structural preservation in embeddings. As a result, they provide an more elaborate visual and quantitative interpretations for the low-dimensional embeddings than its closely related research.

5.2.5 User-Defined Parameter Analysis

From our description of the LAPS and GAPS in Section 4, it can be noticed that two user-defined parameters can significantly influence the outcome of the proposed methods. These parameters include: (1) user-defined scalar *weights*

TABLE 4
Behavioral Comparison of LAPS and GAPS With Closely Related Research

Requirements for Explanations	LAPS & GAPS	SIRIUS	Andromeda	Pagliosa *et al.*	Martins *et al.*	Silva *et al.*
Interpretability (Present attribute influences)	●	○	●	●		●
Local Fidelity (Explain preserved local structure)	●	○	○	○	●	○
Global Legitimacy (Preserved global structure)	●			○	●	○
Model Agnostic (Applicable to any algorithm)	●					
Datatype Agnostic (Applicable to any datatype)	●					
Consistency (In local and global explanations)	●			●	●	●

Note: *In the table above '●' represents complete support '○' represents partial support for the requirement.*

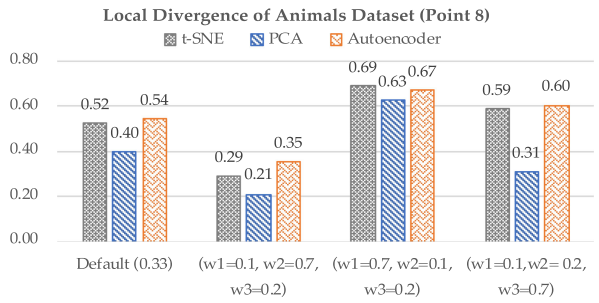


Fig. 5. Analysis of weight-combinations in the computation of local-divergence using LAPS on Animals dataset. *Note: In the graph above, the left-most set of bars represent the local-divergence with a default value of 0.33 for each of the three components of* λ_{x_i} (Eq. (13)). *The second, third, and fourth set of bars from the left represent the weight combinations of (0.1, 0.7, 0.2), (0.7, 0.1,0.2), and (0.1, 0.2, 0.7) on the three components of* λ_{x_i}.

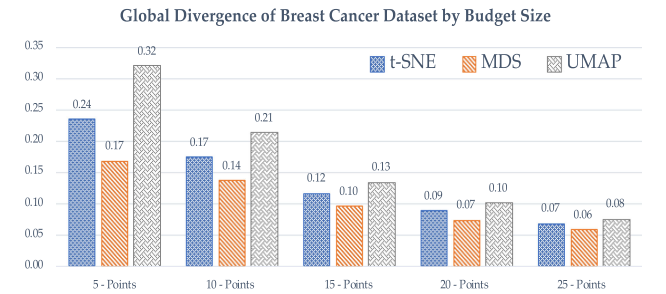


Fig. 6. Analysis of global divergence scores by budget size. *Note: The bars in the above graphs represent the global divergence* λ_X (Eq. (14)) *computed using t-SNE, MDS, and UMAP respectively. The graphs show that with a gradual increase in the selection budget, the global divergence steadily dropped for all three algorithms.*

(cf. Eq. (13)) for the components of local and global divergence scores and (2) the *selection budget* B (cf. Section 4.3) in the computation of global divergence. In this section, we analyze whether the selection of these parameters impacts the consistency of outcome for LAPS and GAPS.

The role of weight in local and global divergences:

Since the scalar weights of each sub-component of local and global divergence can be user-defined, it can be argued whether a strategic selection of these weights can help in manipulating the results or hiding any imperfections in the embeddings. To find answers to this question, we further investigate the local divergence scores presented in our first case study (on the Animals dataset) discussed in Section 5.2.1. As shown in Fig. 5, in this study we analyze the data-point *blue-whale* from the Animals dataset [9] and compare the local divergence for the point in the embeddings obtained using the algorithms t-SNE, PCA, and VAE. Fig. 3 shows that, with the default relative weights (i.e., 0.33) among the three algorithms, PCA has the lowest local-divergence for *blue-whale*. Fig. 3 computes the local divergence scores for the algorithms with the default weights of 0.33 for each of its components. In this section, we analyze the impact of any changes in the weights of the individual components of the local divergences for the point *blue-whale*. Our analysis results are summarized in Fig. 5. The results show that different combinations of weights for the individual components of local-divergence does not allow users to manipulate the results but only shows the differences between the embeddings more clearly. For example, as shown in Fig. 3, in all the embeddings, the discrepancy for attribute influence explanations have been the highest (i.e., $>73\%$) among the three components. Whilst the false and missing neighborhood has the lowest discrepancy (i.e., $\leq 20\%$), the inconsistency in the order of neighbors is high for t-SNE and VAE (i.e., $\sim 70\%$) but low for PCA ($\sim 30\%$). Hence, in Fig. 5, when increasing the weight of attribute influence explanation component (i.e., w1) to 0.70 the local-divergence for all the algorithms is increased by 25% to 58% from their original local divergences obtained using default weight combinations. Similarly, increasing the weight of the neighborhood order component (i.e., w2) to 0.70 reduces the local divergence scores for all the algorithms by 50% to 80%. However, as shown in Fig. 5 in all the cases, PCA still has the lowest local-divergence among the three algorithms in every case.

The role of budget in global-divergence computation:

In this section, we investigate the impact of the budget (i.e., the data subset size) in the computation of global divergence. For this analysis, we extend the results of our second case study presented in Section 5.2.1 and investigate the impact of a budget size 5, 10, 15, 20, and 25 on the global-divergence scores of the Breast Cancer dataset using t-SNE, MDS, and UMAP. We summarize the results of our analysis in Fig. 6. As shown in Fig. 6, with a gradual increase in the selection budget, only the absolute value of the overall global divergence steadily dropped for all the three algorithms with MDS being the best performing algorithm in all the 5 cases. Hence, from Fig. 6, it can be seen that GAPS appropriately shows the best performing algorithm in terms of preservation of global structure for a given dataset. However, the absolute value of the divergence might get more accurate with a higher budget size.

6 Conclusions and Future Work

In this paper, we propose two interactive explanation techniques for low-dimensional embeddings obtained from *any* dimensionality reduction algorithm. The first technique LAPS produces a local approximation of the neighborhood structure to generate interpretable explanations on the preserved locality for a single instance in an embedding. The second method GAPS explains the retained global structure of a high-dimensional dataset in its embedding, by unifying non-redundant local-approximations from a coarse discretization of the projection space. Our experimental evaluation of the techniques with tabular, image, text, and audio data demonstrates the flexibility of these techniques. Moreover, our extensive experiments show the utility of the proposed techniques in demonstrating the preserved structural relationships in lower dimensions, as well as determining the most correlated attributes in a dataset, along with an interactive selection of the most appropriate dimensionality reduction algorithm for any given dataset.

There are several avenues of future work that we would like to explore. For instance, in any interactive technique, one of the most important aspects is scalability. Although, both the proposed algorithms have a computational complexity of $O(n^2)$, for our current design of LAPS and GAPS, we restrict the user-defined neighborhood size (cf. Eq. (9)) to be as large as 10 and the number of perturbed samples (cf. Eq. (10)) to be a maximum of 5000. These design

constraints are inspired by Ribeiro *et al.* [28] who confirm the adequacy of 10 nearest neighbors and 5000 sampled instances in determining the local properties of a data-point. However, we leave experimenting with different sizes of neighborhoods (i.e., > 10) to future work. Although, improving any inherent open challenges [3], [4] of dimensionality reduction techniques (e.g., computational-complexity [4], optimization of hyperparameters [41]) is beyond the scope of this research.

Apart from scalability, we think there are a few more aspects where the proposed work can be improved. Firstly, although both the proposed algorithms allow for user interactions with the processes, the overall interactivity of the approaches can be improved by integrating them as a part of a unified visual framework. As ongoing work, we are working on creating such a framework. To enhance the overall scalability of the framework, we are currently exploring parallel processing for LAPS and GAPS. Secondly, to enhance the fidelity of GAPS, as discussed in Section 3.2, our ongoing work also includes defining the diversified sample selection for GAPS as an Exhaustive Subset Enumeration [36] problem.

References

[1] M. Cavallo and Ç. Demiralp, "A visual interaction framework for dimensionality reduction based data exploration," 2018, *arXiv181112199 Cs*. [Online]. Available: http://arxiv.org/abs/1811.12199, Accessed: Jul. 12, 2019.

[2] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[3] L. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality reduction : A comparative review," *J Mach. Learn. Res.*, vol. 66, no. 71, 2008, pp. Art. no. 13.

[4] E. Becht *et al.*, "Dimensionality reduction for visualizing single-cell data using UMAP," *Nature Biotechnol.*, vol. 37, no. 1, pp. 38–44, Dec. 2018, doi: 10.1038/nbt.4314.

[5] M. Sedlmair, M. Brehmer, S. Ingram, and T. Munzner, "Dimensionality Reduction in the Wild: Gaps and Guidance," Dept. Comput. Sci., Univ. Brit. Columbia, Vancouver, BC, Canada, Tech. Rep. TR-2012-03, Jun. 2012.

[6] S. Lespinats and M. Aupetit, "CheckViz: Sanity check and topological clues for linear and non-linear mappings," *Comput. Graph. Forum*, vol. 30, no. 1, pp. 113–125, Mar. 2011, doi: 10.1111/j.1467-8659.2010.01835.x.

[7] J. M. Lewis and V. R. de Sa, "A behavioral investigation of dimensionality reduction," in *Proc. Annu. Meet. Cogn. Sci. Soc.*, 2012, vol. 34, no. 34, pp. Art. no. 7.

[8] D. Smilkov, N. Thorat, C. Nicholson, E. Reif, F. B. Viégas, and M. Wattenberg, "Embedding projector: Interactive visualization and interpretation of embeddings," 2016, *arXiv161105469 Cs Stat*. [Online]. Available: http://arxiv.org/abs/1611.05469, Accessed: Jul. 12, 2019.

[9] M. Dowling, J. Wenskovitch, J. T. Fry, S. Leman, L. House, and C. North, "SIRIUS: Dual, symmetric, interactive dimension reductions," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 172–182, Jan. 2019, doi: 10.1109/TVCG.2018.2865047.

[10] R. Faust, D. Glickenstein, and C. Scheidegger, "DimReader: Axis lines that explain non-linear projections," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 481–490, Jan. 2019, doi: 10.1109/TVCG.2018.2865194.

[11] J. Stahnke, M. Dork, B. Muller, and A. Thom, "Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 629–638, Jan. 2016, doi: 10.1109/TVCG.2015.2467717.

[12] L. Kodali, J. Wenskovitch, N. Wycoff, L. House, and C. North, "Uncertainty in Interactive WMDS Visualizations," in *Proc. Symp. Vis. Data Sci. Posters VDS'19*, Vancouver, BC, Canada, 2019.

[13] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, Aug. 2018, doi: 10.1145/3236009.

[14] D. Sacha *et al.*, "Visual interaction with dimensionality reduction: A structured literature analysis," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 241–250, Jan. 2017, doi: 10.1109/TVCG.2016.2598495.

[15] M. Wattenberg, F. Viégas, and I. Johnson, "How to use t-SNE effectively," *Distill*, vol. 1, no. 10, 2016, doi: 10.23915/distill.00002.

[16] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci, "Visualizing high-dimensional data: Advances in the past decade," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 21–30, Mar. 2017.

[17] S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Moller, "DimStiller: Workflows for dimensional analysis and reduction," in *Proc. IEEE Symp. Vis. Anal. Sci. Technol.*, 2010, pp. 3–10, doi: 10.1109/VAST.2010.5652392.

[18] P. Joia, F. V. Paulovich, D. Coimbra, J. A. Cuminato, and L. G. Nonato, "Local affine multidimensional projection," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 12, pp. 2563–2571, Dec. 2011, doi: 10.1109/TVCG.2011.220.

[19] J. Xia *et al.*, "LDSScanner: Exploratory analysis of low-dimensional structures in high-dimensional datasets," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 1, pp. 236–245, Jan. 2018, doi: 10.1109/TVCG.2017.2744098.

[20] R. M. Martins, D. B. Coimbra, R. Minghim, and A. C. Telea, "Visual analysis of dimensionality reduction quality for parameterized projections," *Comput. Graphics*, vol. 41, pp. 26–42, Jun. 2014, doi: 10.1016/j.cag.2014.01.006.

[21] L. Pagliosa, P. Pagliosa, and L. G. Nonato, "Understanding attribute variability in multidimensional projections," in *Proc. 29th SIBGRAPI Conf. Graph. Patterns Images*, 2016, pp. 297–304, doi: 10.1109/SIBGRAPI.2016.048.

[22] R. R. O. Da Silva, P. E. Rauber, R. M. Martins, R. Minghim, and A. C. Telea, "Attribute-based visual explanation of multidimensional projections," in *Proc. EuroVis Workshop Vis. Anal.*, 2015, Art, no. 5, doi: 10.2312/EUROVA.20151100.

[23] J. Z. Self, L. House, S. Leman, and C. North, "Andromeda: Observationlevel and parametric interaction for exploratory data analysis," Tech. Rep., Dept. Comput. Sci., Virginia Tech, Blacksburg, Virginia, 2015.

[24] K. Bunte, M. Biehl, and B. Hammer, "Dimensionality reduction mappings," in *Proc. IEEE Symp. Comput. Intell. Data Mining.*, 2011, pp. 349–356, doi: 10.1109/CIDM.2011.5949443.

[25] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.

[26] A. A. Freitas, "Comprehensible classification models: A position paper," *ACM SIGKDD Explor. Newsl.*, vol. 15, no. 1, pp. 1–10, Mar. 2014, doi: 10.1145/2594473.2594475.

[27] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144, [Online]. Available: http://arxiv.org/abs/1602.04938

[28] F. Yang, L. T. Harrison, R. A. Rensink, S. L. Franconeri, and R. Chang, "Correlation judgment and visualization features: A comparative study," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 3, pp. 1474–1488, Mar. 2019, doi: 10.1109/TVCG.2018.2810918.

[29] E. Levina and P. J. Bickel, "Maximum likelihood estimation of intrinsic dimension," in *Proc. Int. Conf.. Neural Inf. Process. Syst.*, 2005, pp. 777–784.

[30] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2005, pp. 513–520.

[31] G. Plumb, D. Molitor, and A. Talwalkar, "Model agnostic supervised local explanations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 2515–2524.

[32] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, vol. 27, no. 4, pp. 857—871, 1971, doi: 10.2307/2528823.

[33] H. Wang and M. Hong, "Distance variance score: An efficient feature selection method in text classification," *Math. Problem Eng.*, vol. 2015, pp. 1–10, 2015, doi: 10.1155/2015/695720.

[34] J. P. Boyd, "Additive blending of local approximations into a globally-valid approximation with application to the dilogarithm," *Appl. Math. Lett.*, vol. 14, no. 4, pp. 477–481, 2001, doi: 10.1016/S0893-9659(00)00180-4.

[35] F. Pan, A. Roberts, L. McMillan, D. Threadgill, and W. Wang, "Sample selection for maximal diversity," in *Proc. 7th IEEE Int. Conf. Data Mining*, 2007, pp. 262–271, doi: 10.1109/ICDM.2007.16.
[36] R. Haftka, "Combining global and local approximations," *AIAA J.*, vol. 29, no. 9, Sep. 1991, Art. no. 1523.
[37] K. Pearson, "On lines and planes of closest fit to systems of points in space," *London Edinburgh Dublin Philos. Mag.*, vol. 2, no. 11, pp. 559–572, Nov. 1901, doi: 10.1080/14786440109462720.
[38] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.
[39] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*, Accessed: Apr. 08, 2019. [Online]. Available: http://arxiv.org/abs/1802.03426
[40] P. G. Poličar, M. Strazar, and B. Zupan, "openTSNE: A modular python library for t-SNE dimensionality reduction and embedding," Cold Spring Harbor Laboratory, Aug. 2019, doi: 10.1101/731877.
[41] W. S. Torgerson, "Multidimensional scaling: I. Theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, Dec. 1952, doi: 10.1007/BF02288916.
[42] J. B. Tenenbaum, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000, doi: 10.1126/science.290.5500.2319.
[43] S. T. Roweis, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000, doi: 10.1126/science.290.5500.2323.
[44] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. 2nd Int. Conf. Learn. Representations*, Banff, AB, Canada, 2014.
[45] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM J. Sci. Comput.*, vol. 26, no. 1, pp. 313–338, Jan. 2004, doi: 10.1137/S1064827502419154.
[46] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a Kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998, doi: 10.1162/089976698300017467.
[47] J. Demśar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.

Aindrila Ghosh received the master's degree in computer science from the University of Paderborn, she is currently working toward the PhD degree in the University of Alberta. As her graduate research, she is currently working on enhancing the interpretability of machine learning models, along with improving user-engagement in the machine learning process.

Mona Nashaat received the BS and master's degrees in computer engineering from the Faculty of Engineering, Port Said University, she is currently working toward the PhD degree with the Department of Electrical and Computer Engineering, University of Alberta. Her research focuses on machine learning, intelligent systems, and big data.

James Miller received the BSc and PhD degrees in computer science from the University of Strathclyde, Scotland. After working as a principal scientist at the United Kingdom's National Electronic Research Initiative, then a senior-lecturer at the University of Strathclyde, he joined the University of Alberta as a professor. He has published more than one hundred journal and conference articles and sits on the editorial board of the *Journal of Empirical Software Engineering*.

Shaikh Quader received the BSc degree in computer science from the University of New Brunswick, and the master's degree in computer science from the University of Waterloo. He is a lead AI architect with IBM Canada. During his work, he developed and managed complex software projects. He leads research collaboration between IBM and academia and has published at several IEEE conferences.