Measuring Quality and Interpretability of Dimensionality Reduction Visualizations

Bibal, Adrien; Frénay, Benoît

Published in:
SafeML ICLR Workshop

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

MEASURING QUALITY AND INTERPRETABILITY OF DIMENSIONALITY REDUCTION VISUALIZATIONS

Adrien Bibal & Benoît Frénay
PReCISE - Faculty of Computer Science - NADI
University of Namur
Namur 5000, Belgium
{adrien.bibal,benoit.frenay}@unamur.be

ABSTRACT

One first step to get insights about a dataset can be its visualization using dimensionality reduction (DR). However, DR processes induce a loss of information that needs to be quantified in order to evaluate the quality of their results. Furthermore, two DR visualizations with a similar loss value can be really different in the eyes of the user. This paper presents DR quality measures developed in the machine learning community, as well as visual quality measures considered in the information visualization community, which can be used to assess interpretability. We propose to combine several measures from these two categories in order to be able to predict and study users' understanding of DR visualizations.

1 INTRODUCTION

Given the high amount of data generated today, many techniques are developed and used to get insights about these data. Visualization is an important method for understanding hidden patterns in data and is often used as a first explanatory step before any processing or analysis. Indeed, when the studied dataset is high dimensional, the structures and patterns are hard to comprehend.

Dimensionality reduction (DR) is one of the different ways to transform high-dimensional (HD) data so as to allow a visualization (Lee & Verleysen, 2007). The objective of visualization through DR techniques is to find a low dimensional space, typically two or three dimensions, for representing high-dimensional data. Among all DR techniques, one can cite principal component analysis (PCA) (Hotelling, 1933), multidimensional scaling (MDS) (Kruskal & Wish, 1978) and t-distributed stochastic neighborhood embedding (t-SNE) (van der Maaten & Hinton, 2008).

In order to evaluate embeddings of HD data obtained with DR, two goals must be taken into account. On the one hand, it is necessary to define a measure of information preservation for the dimensionality reduction process. On the other hand, the user still needs to interpret the new space where data are projected, as it may serve as a basis for analyses. These two goals, ensuring information preservation and interpretability, should be considered together for measuring the overall quality of an embedding (Liu et al., 2017; Vellido et al., 2012; Frénay & Dumas, 2016; Dumas et al., 2018).

This paper proposes to bridge the gap between DR visualization quality metrics in machine learning and information visualization to measure the two facets of DR visualization quality. The paper is organized as follow. The background on dimensionality reduction is presented in Section 2. Then, Section 3 presents information preservation and interpretability measures in the literature. Propositions on how to bridge the gap between measures of the two categories, information preservation and interpretability, are presented in Section 4. Finally, Section 5 concludes the paper.

2 DIMENSIONALITY REDUCTION VISUALIZATION AND INTERPRETABILITY

Dimensionality reduction (DR) is the process of reducing the large number of dimensions d of a dataset to a lower number $m \ll d$. There are many reasons behind such a process, like the need to escape the curse of dimensionality (Bellman, 1961; Hastie et al., 2009). For instance, when the number of dimensions is too high, each pair of instances tend to have the same distance with respect

to all other pairs. This is a major difficulty when using algorithms with an objective function based on distances.

Another use of dimensionality reduction is to visually analyze the data at hand (Lee & Verleysen, 2007). When the number of reduced dimensions m is equal to 2, high-dimensional patterns can be seen and analyzed, as long as the information loss in the DR process is reasonable. The measures assessing this preservation of information and therefore characterizing the "accuracy" of the DR process are called *DR accuracy* measures in the remainder of this paper.

Among the possible DR techniques, linear DRs, such as PCA, are often considered to be methods providing interpretable embeddings because the way in which their parameters are combined can be easily understood. However, nonlinear DR (NLDR) embeddings, such as the ones computed by MDS or t-SNE, are hard to understand (Liu et al., 2017). Interpretability, in the context of DRs, is therefore understood as how easy it is to understand the mapping between the high and low dimensions. The measures assessing the presence of comprehensible visual patterns are called *DR interpretability* measures in this paper, even though the information visualization literature refers to them as visual quality measures. Indeed, we argue that the main way of assessing the interpretability of an NLDR mapping is through measuring meaningful visual patterns. Measures representing the two categories, DR accuracy and interpretability, are presented in the next section. Then, Section 4 presents a way to combine them in order to assess DR qualities globally.

3 Accuracy and Interpretability of DR Visualizations

As an introduction to this section, let us consider an analogy with regression analysis. Regression is a problem in which a relation must be found between a set of features $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_d$ and a target $\mathbf{t}$. In linear regression, a linear combination of the features $w_1\mathbf{x}_1 + w_2\mathbf{x}_2 + ... + w_d\mathbf{x}_d$ is used for predicting $\mathbf{t}$. The *mean squared error* (MSE) is an error measure often considered for evaluating the quality of the feature weights $w_1, w_2, ..., w_d$ found for predicting $\mathbf{t}$. However, the reduction of error may not be the sole objective to optimize. For instance, in Lasso (Tibshirani, 1996), another objective is to set as many weights $w_1, w_2, ..., w_d$ as possible to 0. In addition to overcoming overfitting problems, setting some weights to 0 makes the model more interpretable, as fewer features are used in the prediction.

Overall quality of DR visualizations can be considered in the same terms. The DR information preservation measures quantify how "accurate" the DR model is. DR "accuracy" corresponds to how well the patterns in the high-dimensional space, such as distances between instances or neighborhoods, are reproduced in the new low-dimensional space. The interpretability objective focuses on helping users to understand the model. In Lasso, this is performed by setting some feature weights w_i to 0. In DR visualizations, this second objective is related to how easily users visually understand the 2D or 3D embedding.

As Bertini et al. (2011) mention, quality metrics can evaluate any stage of the Card et al. (1999)'s information visualization pipeline (see Figure 1). In our case, the DR "accuracy" quantifies the information preservation of the DR transformation (first process of the pipeline: data transformation), while the DR interpretability metrics focus on the transformed data (second stage of the pipeline: transformed data). Because the two types of measure are grounded in different stages of the DR visualization process, the accuracy is measured with high-dimensional and low-dimensional data, while interpretability is only assessed using low-dimensional data. Note that further stages, such as the way 2D data are displayed, can influence the interpretability of the DR visualization result.

This section presents these two kinds of quality metrics. The measures quantifying the error made while reducing the dimensions are presented in Section 3.1. Measures assessing the presence of visual patterns in 2D representations of data are presented in Section 3.2.

3.1 On the Accuracy of Dimensionality Reduction

In machine learning, the quality of a DR embedding is defined by its faithful reproducibility of the projected high-dimensional structures and patterns. This quality needs to be objectively quantified, as it is hard for users to visually assess it. Indeed, by definition of the problem, users cannot visualize the high-dimensional patterns, which is why DR visualizations are needed (Mokbel et al., 2013).

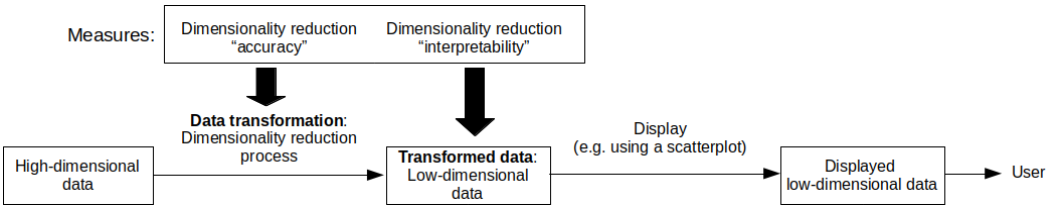

Figure 1: Figure adapted from Bertini et al. (2011) representing the Card's InfoVis pipeline.

DR accuracy metrics can be categorized by the aspect of information loss on which they focus. The two main categories for assessing DR accuracy are *distance preserving* and *neighborhood preserving* measures (Lee & Verleysen, 2009). Distance preserving measures have long been used as objective functions in algorithms such as *multidimensional scaling* (MDS). Under the name *stress function*, we find measures such as the famous *Kruskal's stress* (Kruskal & Wish, 1978), which measures how well pairwise distances in high dimension (HD) are preserved in the low-dimensional embedding (LD). The non-metric version of Kruskal's Stress (NMS) (Kruskal, 1964) measures how well the pairwise distance ranks are preserved, instead of the pairwise distances themselves. The *Sammon's non-linear mapping stress* (NLM) (Sammon, 1969) is another stress function, similar to the Kruskal's stress. The *Curvilinear component analysis* stress (CCA) (Demartines & Hérault, 1997) stands out from the other stress metrics by gradually focusing on small distances. Finally, the *correlation coefficient* (CC) (Geng et al., 2005) is a measure of correlation between the vector of pairwise distances in HD and the vector of pairwise distances in LD.

The second DR "accuracy" metric category focuses on neighborhood preservation. The *stochastic neighbor embedding* (SNE) (Hinton & Roweis, 2003), *t-distributed stochastic neighbor embedding* (t-SNE) (van der Maaten & Hinton, 2008) and *Jensen-Shannon embedding* (JSE) (Lee et al., 2013) algorithms include similar objective functions that focus on the neighborhood preservation of each instance, which is formalized by probability distributions. The size of the neighborhood to consider is controlled by a meta-parameter called the *perplexity*. *Neighbor retrieval visualizer* (NeRV) (Venna et al., 2010) is an algorithm that takes its inspiration from information retrieval, with a DR accuracy metric based on the precision/recall balance. *AUClogRNX* (Lee et al., 2015) is a widely used accuracy metric for DR. AUClogRNX is defined by a sum of the neighborhood preservation over all neighborhood sizes in logarithmic scale:

$$Q_{NX}(K) = \frac{1}{KN} \sum_{i=1}^{N} |v_i^K \cap n_i^K| \tag{1}$$

$$R_{NX}(K) = \frac{(N-1)Q_{NX}(K) - K}{N - 1 - K} \tag{2}$$

$$AUC_{lnK}(R_{NX}(K)) = \frac{\sum_{K=1}^{N-2} R_{NX}(K)/K}{\sum_{K=1}^{N-2} 1/K}, \tag{3}$$

where K is the number of neighbors, N is the number of instances, v_i^K is the set of K nearest neighbors of instance i in HD and n_i^K is the set of K nearest neighbors of instance i in LD (Lee et al., 2015). Other neighborhood preservation measures include the *local continuity meta criterion* (LCMC) (Chen & Buja, 2009), *trustworthiness & continuity* (T&C) (Venna & Kaski, 2006) and Q_Y (Meng et al., 2011). LCMC is a penalized stress that increases the loss for close instances in LD that are not neighbors in HD. T&C compares the difference of neighborhood for each instance in HD and in LD. While LCMC and T&C are local measures, as they focus on neighborhoods, Meng et al. (2011) proposes to mix these local measures (called Q_{LC}) with a global measure

$$Q_{GB} = 1 - \frac{6 \sum_{i=1}^{k} d_i^2}{F}, \tag{4}$$

where d_i is a global comparison of ranks in LD and HD for instance i and F is used for normalization. They obtain the measure

$$Q_Y = \mu Q_{GB} + (1 - \mu) Q_{LC}. \tag{5}$$

Among the above DR accuracy metrics, some are intended to be used as objective functions of DR algorithms (e.g. the Kruskal's stress), and some others can only be used for measuring the DR "accuracy" (e.g. AUClogRNX). The advantage of the latter is that they can be mathematically defined without the constraint of being easy to optimize, such as being differentiable (Lee & Verleysen, 2010; Mokbel et al., 2013).

All the above metrics quantify the information preserved, with respect to distances or neighborhoods, by the DR embedding. However, if the DR is used for visualization, these metrics are not sufficient. Indeed, as with the Lasso analogy of Section 3, involving users implies adapting the result to them. This means that the way 2D data is presented in a scatterplot is crucial, even if it requires distorting the patterns present in HD a little bit more in LD. As Behrisch et al. (2018) write: "the essence of effectiveness resides in the identification of *interpretable visual patterns* that contribute to the overarching goal." The visualization interpretability metrics, considered here as the metrics assessing the presence of these interpretable visual patterns, are presented in the next section.

3.2 On the Interpretability of DR Visualizations

In the case of nonlinear dimensionality reduction (NLDR), the link between the new dimensions and the original ones is hard to understand. Liu et al. (2017) propose to see this as a trade-off between interpretability and the *intrinsic structure* of the reduction. Linear dimensionality reductions are often considered as easy to interpret because the new dimensions are linear combinations of the original ones. For NLDR, the intrinsic structure of the embedding is much more complex, resulting in a much less interpretable embedding. The difficulty of identifying the link between the high and the low dimensions is all the more important since many NLDR are non-parametric.

There are two main ways to solve the interpretability problem. First, techniques can be developed to interpret the new LD axes. For instance, regression analysis can be used to interpret the new axes with external variables. In psychology, some data obtained in an experiment A can be used to understand the dimensionality reduction performed by multidimensional scaling on data obtained from an experiment B (Koch et al., 2016; Bibal et al., 2018; Marion et al., 2019). Second, another way to get a better understanding of the embedding is to analyze the position of the instances in the scatterplot. If the instances are positioned such that users can understand these positions with the original dimensions in mind, then the embedding can be considered interpretable. Indeed, if a DR algorithm projects clusters of instances that users understand based on HD features, then the projection can be said to be interpretable, even for a non-parametric DR.

Metrics that measure the position of instances in the 2D space are called visual quality metrics in the information visualization literature. These measures, which help in interpreting the embedding, have different aspects. Among all possible measures, Bertini et al. (2011) present typical categories such as grouping/clustering, correlation, outliers and "complex patterns." Some measures that consider clusters in the 2D space use the instance labels (if present in the dataset) to measure if the 2D visual clusters correspond to those labels (see e.g., Sedlmair & Aupetit (2015); Aupetit & Sedlmair (2016)). For instance, one state-of-the-art supervised cluster measure is the distance consistency (DSC):

$$\text{DSC} = \frac{|x' \in v(X) : \text{CD}(x', centr'(c_{clabel(x)})) \neq true|}{N}, \tag{6}$$

where N is the number of instances, $v(X)$ is the 2D visualization of the dataset X, $centr'(c_i)$ is the 2D centroid of the class c_i, $clabel(x)$ is the provided label of the instance x and $\text{CD}(x, centr'(c_i))$ is true if the closest centroid to x is the one corresponding to the class c_i of x (Sips et al., 2009). This measure computes the proportion of instances for which the closest 2D centroid does not correspond to their label in the original dataset. In addition, measures based on graphs (graph-theoretic scagnostics), such as measures of density (by computing statistics on edge lengths of a minimal spanning tree) or the presence of outliers (detected by comparing edge lengths), can be found in Wilkinson et al. (2005). For more information on these measures, the reader is referred to the recent survey on visual quality measures by Behrisch et al. (2018).

All measures considered in this section assess the presence of patterns in the low-dimensional space. While the end goal is to measure if interesting visual patterns are present in the 2D space, it is not useful to produce a DR visualization with meaningful patterns if these patterns are not present in the high-dimensional space. In other words, a visualization must be both interpretable and accurate. This is why the gap between measures presented in the sections 3.1 and 3.2 should be filled.

4 BRIDGING THE GAP BETWEEN DR QUALITY MEASURE CATEGORIES

In order to globally assess DR visualization quality, accuracy and interpretability measures can be combined. Bibal & Frénay (2016) linearly combined visual clustering measures with AUClogRNX and found that AUClogRNX (the measure of DR "accuracy") outperforms the measures of visual patterns when predicting user preferences of embedding understandability. Johansson & Johansson (2009) also linearly combined some visual quality measures (presence of outliers, correlations and clusters) and estimated the weights through user interaction.

We propose to combine a large set of measures for each objective, DR "accuracy" and embedding "interpretability." By doing so, the combination would balance the two objectives, while considering different aspects of each objective. The overall quality measure could have the linear form

$$\begin{aligned}\text{overall quality measure} = (\alpha_1 * AM_1) + ... + (\alpha_i * AM_i) + ... + (\alpha_n * AM_n) \\ + (\beta_1 * IM_1) + ... + (\beta_j * IM_j) + ... + (\beta_k * IM_k),\end{aligned} \tag{7}$$

where AM_i is the i^{th} normalized accuracy measure, IM_j is the j^{th} normalized interpretability measure, n is the number of accuracy measures, k is the number of interpretability measures and the α's and β's are parameters to estimate. These parameters, which can be estimated based on a user-based experiment, allow the overall quality measure to be used for assessing other DR visualizations. The importance of each goal would be identified by comparing all α's with all β's. It would also be possible to rank α's (resp. β's) among all α's (resp. β's): the higher the α (resp. β) value, the greater the importance of its corresponding measure among the measures of accuracy (resp. interpretability). If a sparsity penalty is added, α's and β's set to 0 would allow us to know the measures that are not necessary for mimicking users. For instance, if the β_j associated with the measure of visual outliers is set to 0, that would mean that users may not consider visual outliers when assessing the overall quality of DR visualizations. Furthermore, collinearity between measures would highlight redundancies among them.

For estimating α and β values that best represent reality, a user-based experiment should be run. This means that a set of DR visualizations should be assessed by users who would give quality scores to these different visualizations. These scores would make up a vector $\mathbf{t}$ to predict. Optimizing α's and β's in Eq. 7 for predicting $\mathbf{t}$ would make it possible to get insights on the importance of DR accuracy with respect to interpretability for users when assessing visualizations, as well as on the importance of each quality measure for modeling users. Furthermore, multiple regressions can be considered to account for different user profiles. For instance, α's and β's can be estimated for a first profile (e.g. users accustomed to scatterplot analyses), and also for a second profile (e.g. novice users). Finally, it is possible that optimizing the overall quality measure based on user feedback results in bias in favor of the interpretability measures (i.e. users might not consider accuracy when evaluating visualizations). In order to avoid this issue, some information regarding the accuracy should be shown to users during the experiment. For instance, the information loss of the DR visualization or visual signals indicating local DR mistakes can be provided, e.g. (Aupetit, 2007; Lespinats & Aupetit, 2011).

5 CONCLUSION

In this paper, we presented how to approach the problem of interpretability in DR visualizations produced by dimensionality reduction (DR) techniques. Two kinds of measures were discussed. The first kind aims at assessing the quality of the DR process through the idea of information loss. These DR quality measures are mainly developed in the machine learning community, which rarely consider users as part of the evaluation. The other kind of measures, from the information visualization community, characterize the presence of meaningful visual patterns in the low-dimensional space. These measures focus on the visual patterns in 2D, even if these patterns are not present in HD.

We propose to combine these two categories of measures in order to account for the information loss, as well as the interpretability of DR visualizations. This would make it possible to highlight measures that best correspond to user's perception. In future works, we plan to set up a user-based experiment to find the parameters α's and β's from Eq. 7 that best fit user's understandability of DR visualizations. These parameters would allow us to compare state-of-the-art measures with each other and with respect to the real perception of users.

REFERENCES

Michaël Aupetit. Visualizing distortions and recovering topology in continuous projection techniques. *Neurocomputing*, 70(7-9):1304–1330, 2007.

Michaël Aupetit and Michael Sedlmair. Sepme: 2002 new visual separation measures. In *IEEE Pacific Visualization Symposium (PacificVis)*, pp. 1–8, 2016.

M. Behrisch, M. Blumenschein, N. W. Kim, L. Shao, M. El-Assady, J. Fuchs, D. Seebacher, A. Diehl, U. Brandes, H. Pfister, T. Schreck, D. Weiskopf, and D. A. Keim. Quality metrics for information visualization. *Computer Graphics Forum*, 37(3):625–662, 2018.

Richard E Bellman. *Adaptive control processes: a guided tour*. Princeton university press, 1961.

Enrico Bertini, Andrada Tatu, and Daniel Keim. Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2203–2212, 2011.

Adrien Bibal and Benoît Frénay. Learning interpretability for visualizations using adapted Cox models through a user experiment. *NIPS Workshop on Interpretable Machine Learning in Complex Systems*, 2016.

Adrien Bibal, Rebecca Marion, and Benoît Frénay. Finding the most interpretable MDS rotation for sparse linear models based on external features. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, pp. 537–542, 2018.

Stuart Card, Jock D. Mackinlay, and Ben Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

Lisha Chen and Andreas Buja. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association*, 104(485): 209–219, 2009.

Pierre Demartines and Jeanny Hérault. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on Neural Networks*, 8(1):148–154, 1997.

Bruno Dumas, Benoît Frénay, and John A. Lee. Interaction and user integration in machine learning for information visualisation. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, pp. 97–104, 2018.

Benoît Frénay and Bruno Dumas. Information visualisation and machine learning: Characteristics, convergence and perspective. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, pp. 623–628, 2016.

Xin Geng, De-Chuan Zhan, and Zhi-Hua Zhou. Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(6):1098–1107, 2005.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: Data Mining, Inference, and Prediction*, volume 2. Springer-Verlag New York, 2009.

Geoffrey E Hinton and Sam T Roweis. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 857–864, 2003.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Sara Johansson and Jimmy Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics*, 15 (6):993–1000, 2009.

Alex Koch, Roland Imhoff, Ron Dotsch, Christian Unkelbach, and Hans Alves. The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology*, 110(5):675, 2016.

Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

Joseph B Kruskal and Myron Wish. *Multidimensional scaling*. Sage, 1978.

John A Lee and Michel Verleysen. *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

John A Lee and Michel Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7-9):1431–1443, 2009.

John A Lee and Michel Verleysen. Scale-independent quality criteria for dimensionality reduction. *Pattern Recognition Letters*, 31(14):2248–2257, 2010.

John A Lee, Emilie Renard, Guillaume Bernard, Pierre Dupont, and Michel Verleysen. Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing*, 112:92–108, 2013.

John A Lee, Diego H Peluffo-Ordóñez, and Michel Verleysen. Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure. *Neurocomputing*, 169:246–261, 2015.

Sylvain Lespinats and Michaël Aupetit. CheckViz: Sanity check and topological clues for linear and non-linear mappings. *Computer Graphics Forum*, 30(1):113–125, 2011.

Shusen Liu, Dan Maljovec, Bei Wang, Peer-Timo Bremer, and Valerio Pascucci. Visualizing high-dimensional data: Advances in the past decade. *IEEE Transactions on Visualization & Computer Graphics*, 23(3):1249–1268, 2017.

Rebecca Marion, Adrien Bibal, and Benoît Frénay. BIR: A method for selecting the best interpretable multidimensional scaling rotation using external variables. *Neurocomputing*, 342:83–96, 2019.

Deyu Meng, Yee Leung, and Zongben Xu. A new quality assessment criterion for nonlinear dimensionality reduction. *Neurocomputing*, 74(6):941–948, 2011.

Bassam Mokbel, Wouter Lueks, Andrej Gisbrecht, and Barbara Hammer. Visualizing the quality of dimensionality reduction. *Neurocomputing*, 112:109–123, 2013.

John W Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18(5):401–409, 1969.

Michael Sedlmair and Michaël Aupetit. Data-driven evaluation of visual quality measures. In *Computer Graphics Forum*, volume 34, pp. 201–210, 2015.

Mike Sips, Boris Neubert, John P Lewis, and Pat Hanrahan. Selecting good views of high-dimensional data using class consistency. In *Computer Graphics Forum*, volume 28, pp. 831–838, 2009.

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Alfredo Vellido, José David Martín-Guerrero, and Paulo JG Lisboa. Making machine learning models interpretable. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, pp. 163–172, 2012.

Jarkko Venna and Samuel Kaski. Local multidimensional scaling. *Neural Networks*, 19(6-7):889–899, 2006.

Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11:451–490, 2010.

Leland Wilkinson, Anushka Anand, and Robert Grossman. Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization*, pp. 157–164, 2005.