Computers & Graphics 116 (2023) 287–297

Contents lists available at ScienceDirect

Computers & Graphics

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

Special Section on SIBGRAPI 2023

Measuring the quality of projections of high-dimensional labeled data

Bárbara C. Benato [a,*], Alexandre X. Falcão [a], Alexandru C. Telea [b]

[a] Institute of Computing, University of Campinas, Campinas, Brazil
[b] Department of Information and Computing Sciences, Faculty of Science, Utrecht University, Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 9 May 2023
Received in revised form 16 August 2023
Accepted 16 August 2023
Available online 19 August 2023

Keywords:
Quality of projections
Labeled data
Pseudo labeling

ABSTRACT

Dimensionality reduction techniques, also called projections, are one of the main tools for visualizing high-dimensional data. To compare such techniques, several quality metrics have been proposed. However, such metrics may not capture the *visual separation* among groups/classes of samples in a projection, *i.e.*, having groups of similar (same label) points far from other (distinct label) groups of points. For this, we propose a pseudo-labeling mechanism to assess visual separation using the performance of a semi-supervised optimum-path forest classifier (OPFSemi), measured by Cohen's Kappa. We argue that lower label propagation errors by OPFSemi in projections are related to higher data/visual separation. OPFSemi explores local and global information of data distribution when computing optimum connectivity between samples in a projection for label propagation. It is parameter-free, fast to compute, easy to implement, and generically handles any high-dimensional quantitative labeled dataset and projection technique. We compare our approach with four commonly used scalar metrics in the literature for 18 datasets and 39 projection techniques. Our results consistently show that our proposed metric consistently scores values in line with the perceived visual separation, surpassing existing projection-quality metrics in this respect.

1. Introduction

Multidimensional data occurs in many fields as science, engineering, medicine, and machine learning (ML). Visually inspecting such data to find relevant patterns is challenging when the number of dimensions is high. Dimensionality reduction (DR) algorithms, also called *projections*, are methods of choice for this task. Projections aim to map the high-dimensional data to low-dimensional spaces (typically 2D or 3D) so that the main data patterns are preserved and thus directly explorable.

Projection techniques have been used in ML to explore high dimensional data [1], comprehend and explain models [2,3], design better classifiers [4], and label data [4]. In most such tasks, data features and class labels are supposed to be correlated, i.e., close data points typically have the same labels, so that the features of the former can be used to predict the latter.

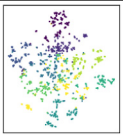
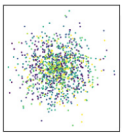
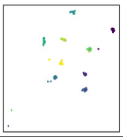
The success of these explorative tasks depends on the *visual separation* (VS) of the projection used to depict it. If a dataset exhibits clear *data* separation (into samples of different classes), then analysts should be able to gauge this by seeing a corresponding *visual* separation in the projection, in terms of densely-packed, ideally non-overlapping, groups of points with the same label (within a given group). Conversely, if a dataset exhibits poor separation, its projection should also show poor visual separation. Assessing VS is useful, for example, to judge the ease of classifying a dataset (or parts thereof) [2].

Many projection methods have been proposed, using different underlying techniques as graphs, linear algebra, optimization, and neural networks [5]. Such techniques generate a wide variety of scatterplots for the same give dataset, especially when one changes their various hyperparameters. Several *metrics* have been proposed to quantify a projection's quality. However, the most used metrics in the DR literature – Trustworthiness (T) [6], Continuity (C) [6], Normalized stress (S) [7], and Neighborhood hit (N) [8] do not directly measure visual separation at a global projection level but rather more local properties (as discussed further in Section 4.3). Table 1 shows this by a simple example of three DR techniques (t-distributed stochastic neighbor embedding [t-SNE] [9], stochastic proximity embedding [SPE] [10], and uniform manifold approximation and projection [UMAP] [11]) and their T, C, S, and N metrics, all ranging between 0 (worst quality) and 1 (best quality). The SPE plot has high metric values but arguably much poorer visual separation (of the 9 color-coded classes) into distinct, same-color, point groups than the t-SNE plot which has much lower metric values. UMAP and SPE have similar (high) metric values but, we argue, visual separation is much stronger in the UMAP than the SPE plot. All in all, this shows that these four metrics do not capture visual separation well.

* Correspondence to: Institute of Computing, University of Campinas, Av. Albert Einstein, 1251, Zip code: 13083-852 Campinas, SP, Brazil.
E-mail addresses: barbarabenato@gmail.com (B.C. Benato), afalcao@ic.unicamp.br (A.X. Falcão), a.c.telea@uu.nl (A.C. Telea).

https://doi.org/10.1016/j.cag.2023.08.023
0097-8493/

Table 1
Values of T, C, S, and N (Section 4.3) and scatterplots of a dataset (cnae9, Section 4.1) for three projection techniques (t-SNE, SPE, and UMAP, Section 4.2). S values are set as $1 - \textit{normalized stress}$ for easy interpretation. We see that the metrics do not correlate well with the perceived visual separation of the label-colored points in the projection.

Projection	T	C	S	N	Scatterplot
t-SNE	0.516584	0.649824	0.809278	0.256746	
SPE	0.833158	0.937247	0.777273	0.839947	
UMAP	0.798405	0.871780	0.762065	0.978571	

Recent ML studies have explored the VS information of 2D projection spaces to assess data separation in high dimensions [1]; understand deep learning classifiers [2]; find misclassified samples [5]; investigate decision boundaries of classifiers [12]; build better classifiers [13–15]; and investigate the correlation among high-dimensional separability, VS, and classifier performance [16]. Even though some studies have investigated 1-near-neighbor classifiers [9,17] and Gaussian mixture models [18] to estimate the quality of clustering, they did not aim to specifically measure the correlation between classes *and* clusters. To our knowledge, ML approaches were not directly used to measure this VS relation in projections.

In this paper, we propose a new VS quality assessment approach based on ML techniques. We exploit earlier findings that studied VS in t-SNE projections to propagate labels, also called pseudo labeling. Projections with high VS (as assessed qualitatively by users) led to good label propagation results [16]. Our hypothesis is that the converse is also true: If we measure a good label propagation score, then the projection will have a high VS. For label propagation, we use the semi-supervised optimum path forest algorithm (OPFSemi) [19] in the 2D projection space provided by DR methods. OPFSemi was shown to lead to very good label propagation accuracies in both high-dimensional and low-dimensional spaces [15,19] and as such is a good candidate for this task. We evaluate the label propagation by computing the coefficient of agreement of Cohen's Kappa (κ) [20] between true and pseudo labels, a simple but fast and effective way to perform this task which works well also for unbalanced labeled datasets. We assess our proposal on 39 projection algorithms for 18 labeled datasets and show that our method correlates with perceived VS better than well-known metrics for projection quality used in the DR literature. As such, we argue that our metric is an additional useful way to characterize the quality of a projection, atop of existing projection quality metrics.

Summarizing, we propose a method to quantify VS separation in projections which

(a) yields better global and local quantification of VS when compared to four popular metrics in DR;
(b) generically handles any high-dimensional quantitative labeled dataset and any projection technique;
(c) is easy to use as it is parameter-free;
(d) is fast to compute and simple to implement.

2. Related work

Let $D = \{\mathbf{x}_i, l_i\}$ with $\mathbf{x}_i \in \mathbb{R}^n$ be a labeled dataset with labels $l_i \in \{1, 2, \ldots, g\}$. Each sample $\mathbf{x}_i$ has a label l_i. A projection P takes a dataset D and produces a scatterplot $P(D) = \{\mathbf{y}_i = P(\mathbf{x}_i) | \mathbf{y}_i \in \mathbb{R}^q\}$, where typically $q \in \{2, 3\}$. In this work, we consider $q = 2$, *i.e.*, 2D scatterplots.

2.1. Classical quality metrics

Classical quality metrics include scalar metrics, point-pair metrics, and local metrics [21]. Four scalar metrics frequently used in DR literature [5] are described below. All these metrics range between 0 (worst case) and 1 (best case).

Trustworthiness (T) [6]: measures the fraction of points in D that are also close in $P(D)$ or how local visual patterns in a projection truly represent actual data patterns. This is related to the so-called false neighbors of a projected point [22]. In the definition of T (Eq. (1)), $U_i^{(K)}$ is the set of points that are among the K nearest neighbors of point i in 2D but not among the K nearest neighbors of point i in D; and $r(i, j)$ is the rank of the 2D point j in the ordered-set of nearest neighbors of i in 2D.

$$T(\mathbf{x}, \mathbf{y}) = 1 - \frac{2}{NK(2n - 3K - 1)} \sum_{i=1}^{N} \sum_{j \in U_i^{(K)}} (r(i, j), -K) \tag{1}$$

Continuity (C) [6]: measures the fraction of points in $P(D)$ that are also close in D. This is related to the missing neighbors of a projected point [22]. In the definition of C (Eq. (2)), $V_i^{(K)}$ is the set of points that are among the K nearest neighbors of point i in D but not among the K nearest neighbors in 2D; and $\hat{r}(i, j)$ is the rank of the $\mathbb{R}^n$ point j in the ordered-set of nearest neighbors of i in D.

$$C(\mathbf{x}, \mathbf{y}) = 1 - \frac{2}{NK(2n - 3K - 1)} \sum_{i=1}^{N} \sum_{j \in V_i^{(K)}} (\hat{r}(i, j), -K) \tag{2}$$

Normalized stress (S) [7]: measures the preservation of point-pairwise distances from D to $P(D)$ (see Eq. (3)). Euclidean distance is commonly the most used.

$$S(\mathbf{x}, \mathbf{y}) = \frac{\sum_{ij} (\Delta^n(\mathbf{x_i}, \mathbf{x_j}) - \Delta^q(P(\mathbf{x_i}), P(\mathbf{x_j})))^2}{\sum_{ij} \Delta^n(\mathbf{x_i}, \mathbf{x_j})^2} \tag{3}$$

Ideally, a projection should have $S = 0$. To ease comparison with the other metrics, we next use instead $1 - S$ in our work.

Neighborhood hit (N) [8]: measures the fraction of the K neighbors $N_i^{(K)}$ of a point i in $P(D)$ that have the same label l as point i, averaged over all points in $P(D)$ (see Eq. (4)). This is related to the labeled separation in a projection $P(D)$.

$$N(\mathbf{y}) = \sum_{i=1}^{N} \frac{|j \in N_i^{(K)} : l_j = l_i|}{NK} \tag{4}$$

Additional scalar metrics exist for measuring label separation in projections — we discuss them in Section 2.2.

Scalar metrics characterize the quality of an entire projection $P(D)$ by a single value, so they are simple to interpret. However, this inherently averages quality over different parts of $P(D)$ and/or D. Point-pair metrics, *e.g.* the Shepard diagram of pairwise point distances [7], and local metrics, *e.g.* missing and false neighbor plots [22] offer finer-grained quality characterizations. These metrics are typically used to create visualizations of the quality distribution over a projection and cannot be (easily) used to numerically compare several projection algorithms. As such, we do not consider them in our work.

2.2. Visual perception metrics

Several metrics have been used to assess the visual perception of different patterns present in projections. Among them, approaches based on clustering, such as the Silhouette score, explore centroids and labels to assess group separation. Other clustering-based approaches combine information from nD and 2D spaces with labels to gauge visual perception [23]. Class consistency [24] and distance consistency [25] measures assess class separation via distances from defined centroids. Both combine density functions and local neighborhoods to identify class overlap. Although pseudo labels can be used as a strategy with such metrics, they still rely on suitably chosen and parameterized clustering techniques and probability density models and can have difficulties detecting (and characterizing) clusters of complex shapes — the Swiss roll dataset is a famous example. In [18], the authors also explored Gaussian mixture models to measure clustering in monochrome scatterplots, but without taking into account labels. Sedlmair et al. [26] compared cluster separability measures and human observations and concluded that grouping measures might fail to capture multiple sub-groups and groups of different sizes, shapes, and densities. In [27], fifteen metrics and user judgment were used to analyze visual separability in 2D scatterplots. The authors found that the distance consistency measure [25] led to the best agreement with human judgement, but can vary across synthetic and real-world data scenarios. They evaluated their results using only the AUC metric, which can be affected by class unbalances. To circumvent problems related to clustering-based approaches, solutions based on graphs and minimum-spanning trees have been proposed. In [28,29], methods were proposed to find patterns in large scatter plot matrices. Separately, [30] evaluated original and projected spaces for the same purpose. The benefits of such methods include covering global and non-trivial shapes, being parameter-free, and fast to compute. However, these studies did not explore graph-based approaches in a pseudo labeling task to evaluate projections — which is our proposal. Also, they did not compare their methods in a wide experimental setup with well-known projections and datasets as we will be doing.

Human judgment has also been explored in user studies to evaluate the relation between the above-mentioned metrics and visual perception. An important contribution of these studies is designing a method to conduct the experiments and avoid hundreds of scatterplots that have to be inspected by users [27,31,32]. For this, scatterplots are ranked from the best to the worst, and only the top three to five are offered for user inspection [31, 33]. We also use this ranking in our experiments (Section 5.2). The above-mentioned studies do not use many combinations of datasets and projections. Rather, many (dozens) of metrics are compared (or a new one is proposed) for a single [27,32] or a couple of datasets [31,33]. Additionally, the analyzed metrics still have the main issues that we outlined before (see also [33]). In contrast to the above, we aim to evaluate *many* (hundreds of) dataset-projection technique combinations, both quantitatively and by a user study.

2.3. Pseudo labeling in ML

In ML, pseudo labeling refers to assigning labels to data samples to build accurate and large training sets. To do this, supervised samples propagate so-called pseudo-labels to the unsupervised samples. Next, the ML model is trained with the dataset containing both true (supervised) and pseudo-labeled samples.

For this pseudo labeling task, the OPFSemi method [19] was proven to surpass many other semi-supervised methods. Autoencoders [13], convolutional neural networks [4,34], and contrastive models [16] are used to support the label propagation ability of OPFSemi in different learned feature spaces of different dimensions. In our work, we used OPFSemi in the 2D space. As the algorithm explores a complete graph with all samples in a given dataset (Section 3.2), we argue that OPFSemi can capture local and global information of data distribution instead of local information only — as the neighbor-based metrics T, C, and N do. Other advantages of OPFSemi are that the method is free of parameters and does not make assumptions about the shapes of the classes [19].

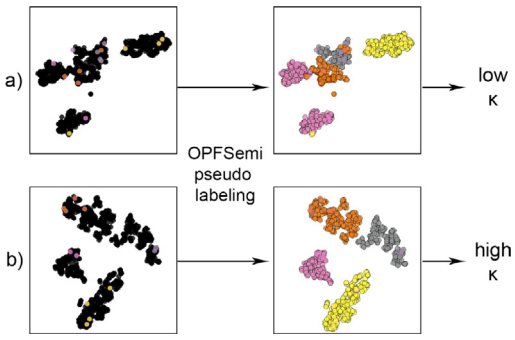


Fig. 1. Pseudo labeling as a measure of visual separation (see Section 3).

2.4. Pseudo labeling, data separation, and visual separation

The success of pseudo labeling depends on how well the data is separated into different groups of similar points, which we next denote as *data separability* (DS). If the data consists of groups of same-label points which are far away from other such groups (high DS), then propagating labels works well. This leads to good training sets which allow constructing good models, *i.e.*, models with a high *classifier performance* (CP). This is related to the well-known fact in ML that a dataset with high DS allows training models with a high CP.

Visual separability (VS) in a projection is related in many ways to DS and CP. Projections can be used to assess DS (VS→DS, [9]). They can also be used to find misclassified samples and assess the difficulty of classifying a dataset (VS→CP, [2,5]). Increasing DS in a dataset can be used to create projections with a higher VS (DS→VS, [35]). Closer to our work, Benato et al. found that OPFSemi's pseudo labeling in 2D t-SNE projected spaces is superior to that in the original high-dimensional space (VS→CP, [4, 14,16,36]).

3. Measuring visual separation by pseudo labeling

Our work builds atop of the observations from related work (Section 2.4) by hypothesizing that high performance in label propagation indicates a high separability of same-label groups in the projection space. Fig. 1 illustrates this. Labels are propagated from supervised samples (colored) to unsupervised samples (black). When there is poor VS in a given projection (a), pseudo labels are wrongly assigned, something which we can measure as described next in Section 3.3. When there is good VS in the projection (b), pseudo labels are accurately assigned.

Fig. 2 shows our VS measurement pipeline which is detailed next.

3.1. Sample selection

We start with a 2D projection $P(D)$ of some labeled dataset D, computed by any desired projection algorithm P. We next randomly split $P(D)$ into a ground-truth dataset A and test dataset B. In our experiments, we take 50% of the samples in $P(D)$ in each of A and B (different fractions can be considered).

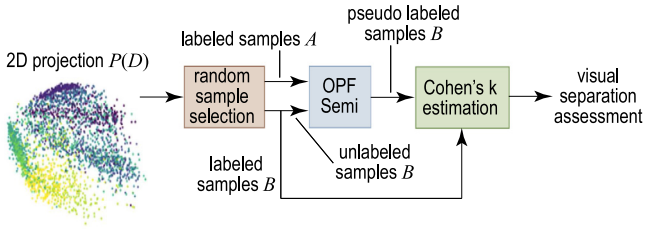


Fig. 2. Pipeline of our approach to assess VS in projections.

3.2. Using OPFSemi for pseudo labeling

We pseudo label the samples in B by propagating the true labels from A using OPFSemi. The OPFSemi algorithm [19] maps both labeled and unlabeled samples to nodes of a complete graph, with edges weighted by the Euclidean distance between samples in a given feature space (the projection space in our case). The labeled samples are taken as prototypes to compete among themselves for the unlabeled ones. Each prototype conquers its most closely connected unlabeled samples by offering minimum-cost paths and assigning its label to them. As a path-cost function, OPFSemi uses the maximal edge-weight along the path. By that, OPFSemi computes a minimum-cost path forest rooted at the prototype set. Its time complexity is $O(m^2)$ for m nodes, since the graph is complete, but it is possible to precompute a minimum-spanning tree in $O(m^2)$ and perform label propagation (optimum-path forest computation) on this tree in $O(m \log m)$ for any randomly chosen set of prototypes in the case of our application. As the process is calculated over a complete graph with all samples in D, we argue that OPFSemi can capture local and global information of data distribution, instead of local information only.

3.3. Pseudo labeling effectiveness measurement

To assess the quality of pseudo labeling, we measure the agreement between the true labels (original labels of samples in B) and the pseudo labels assigned to B by OPFSemi. This agreement could be measured by accuracy, f1 score, or AUC, for example. However, such metrics do not take into account the number of false positives and false negatives, which can highly impact the results for datasets having significant class imbalance. Earlier studies showed the advantage of using Cohen's kappa coefficient (κ) [20] over accuracy to measure the agreement in pseudo labeling [16]. To account for unbalanced classes in $P(D)$, we use κ defined as

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{5}$$

where p_o is the simple accuracy, i.e., the number of correctly classified samples (true positives) over N samples, and

$$p_e = \frac{1}{N^2} \sum_g n_g^\alpha n_g^\beta, \tag{6}$$

where g is the number of classes, N is the number of samples, and n_g^α and n_g^β are the predicted class g given by the true label α and given label β, respectively. The κ coefficient is in a $[-1, 1]$ range, where $\kappa \leq 0$ means no agreement and $\kappa = 1$ means complete agreement between two classifiers α and β.

4. Experimental set-up

To evaluate our usage of OPFSemi to gauge visual separation, we designed several experiments based on the projection-quality benchmark proposed in [21] to our knowledge, the largest public such benchmark for DR. All our results and code are openly available [37].

Table 2
The 18 datasets used in our evaluation and their characteristics.

Dataset	Type	Samples	Dimensions	Labels
bank [38]	tables	2059	63	ordinal
cifar10 [39]	images	3250	1024	categorical (10)
cnae9 [40]	text	1080	856	categorical (9)
coil20 [41]	images	1440	400	categorical (20)
epileptic [42]	tables	5750	178	ordinal
fashion_mnist [43]	images	3000	784	categorical (10)
fmd [44]	images	997	1536	ordinal
har [45]	tables	735	561	categorical (30)
hatespeech [46]	text	3222	100	ordinal
hiva [47]	tables	3076	1617	ordinal
imdb [48]	text	3250	700	ordinal
orl [49]	images	400	396	categorical (40)
secom [50]	tables	1567	590	ordinal
seismic [51]	tables	646	24	ordinal
sentiment [52]	text	2748	200	ordinal
sms [53]	text	836	500	ordinal
spambase [54]	text	4601	57	ordinal
svhn [55]	images	733	1024	categorical (9)

4.1. Datasets

From the benchmark, we chose 18 datasets which are often used in many ML and DR evaluations. Importantly, they are all labeled and, since they are used in ML benchmarks, we know that labels and features are correlated. These datasets come from different application domains and have different sample and dimension counts. Table 2 shows the type of data, sample count, dimension count, and the type and number of labels for each dataset (for more details, see [21]).

4.2. Projection algorithms

From the 44 projection techniques evaluated in [21], we used 39 techniques (see Table 3). The remaining 5 techniques were excluded since their code, as provided in [21], was hard to understand and run. All these techniques are well known in the DR literature and practice. Among them are examples of linear and non-linear and global and local, projections. Also, we consider projections that input the high-dimensional samples themselves and projections which only require a similarity (distance) matrix of the samples. We fixed the parameters of all projection techniques to the default values proposed by each author. More details about the chosen projection techniques and their default parameter values can be found in [21].

4.3. Metrics

As we are proposing a new metric to evaluate visual separation (Section 3), an immediate question is how this metric compares to well established metrics for measuring projection quality. To assess this, we consider, for the latter, the four scalar metrics T, C, N, and S described in Section 2. These are also among the metrics considered by the projection benchmark in [21]. For brevity, we next refer to these four metrics as 'standard' metrics. We compute T, C, and N using $K = 7$ nearest neighbors (see Eqs. (1), (2), and (4)), in line with [21].

4.4. Experimental design

We executed two types of evaluations, as follows:

(a) Quantitative analysis (Section 5.1):

(i) *Correlation plots:* We plot the correlation between our proposed assessment of VS (κ) approach and each standard metric. This yields one scatterplot for each of the four

Table 3
The 39 projection techniques used in our evaluation. We list the linearity, input type, and whether the technique is local or global.

Projection	Linearity	Input	Local or global
DM [56]	nonlinear	samples	local
FA [57]	linear	samples	global
FMAP [58]	nonlinear	samples	global
GPLVM [59]	nonlinear	samples	global
F-ICA [60]	linear	distances	global
IDMAP [61]	nonlinear	distances	local
ISO [62]	nonlinear	distances	local
L-ISO [63]	nonlinear	samples	local
LAMP [7]	nonlinear	samples	local
LE [64]	nonlinear	samples	local
LLC [65]	nonlinear	samples	local
LLE [66]	nonlinear	samples	local
H-LLE [67]	nonlinear	distances	local
M-LLE [68]	nonlinear	samples	local
LPP [69]	linear	samples	global
LSP [8]	nonlinear	samples	local
LTSA [70]	nonlinear	samples	local
L-LTSA [71]	linear	samples	local
MC [72]	nonlinear	samples	local
MDS [73]	nonlinear	samples	global
L-MDS [74]	nonlinear	samples	global
N-MDS [75]	nonlinear	samples	global
L-MVU [76]	nonlinear	distances	global
NMF [77]	linear	distances	global
PBC [78]	nonlinear	samples	local
PCA [57]	linear	samples	global
I-PCA [79]	linear	samples	global
K-PCA-P [80]	nonlinear	samples	global
K-PCA-R [80]	nonlinear	samples	global
K-PCA-S [80]	nonlinear	samples	global
P-PCA [81]	linear	samples	global
S-PCA [82]	linear	samples	global
PLSP [83]	nonlinear	samples	global
G-RP [84]	nonlinear	samples	global
S-RP [84]	nonlinear	samples	global
t-SNE [9]	nonlinear	samples	local
SPE [10]	nonlinear	samples	global
T-SVD [85]	linear	samples	global
UMAP [11]	nonlinear	distances	local

standard metrics. In such a plot, each point is a dataset projected by a projection technique, with all dataset-technique combinations considered. The aim of this analysis is to see whether our new metric correlates or not with existing metrics. If so (which we will show it is not the case), then our new metric does not bring any added value. If not (which is the case), then they cannot *both* gauge visual separation equally well — either our new metric or the standard ones are better for this measurement, but not both of them. We analyze this aspect further via our qualitative analysis described below.

(ii) *Statistical analysis:* We present the main statistical information for our new metric and the standard metrics (minimum, maximum, mean, standard deviation, median, and mode).

(b) Qualitative analysis (Section 5.2): We qualitatively study a subset of such combinations, aiming to find out which metrics – our new one and/or the standard ones – agrees with the perceived visual separation in the projection scatterplots. For this, we perform four qualitative analyses, as follows.

(i) *Random analysis:* We select 8 datasets randomly from the 18 studied ones. For each dataset, we analyze 3 scatterplots of distinct projections and the respective correlation plots.

(ii) *Ranked analysis:* For each dataset, we rank the projections by each quality metric. For the three best and worst projections in terms of this ranking, we study their visual separation *vs* the computed metric values.

(iii) *Correlation plot and ranked analysis:* We plot the same as in (a,i), highlighting the best and worse cases in (b,ii) in terms of good and poor visual separation, respectively.

(iv) *User study:* We ask 108 participants to rank a total of 2916 projections in terms of visual separation and compute the correlation of their rankings with κ.

To use any quality metric in practice, one needs to interpret its values. In our concrete case, all metrics range between 0 (worst case) and 1 (best case). Assuming that a given metric encodes some quality aspect, it is clear that values very close to 1 will indicate 'good' projections in that respect, whereas values close to 0 will indicate 'poor' projections. To simplify the analysis, we next proceed by binning the $[0, 1]$ range in three bins, as follows. Metric values above a superior boundary sb, *i.e.*, in the range $[sb, 1]$, are considered to indicate good projections. Metric values below an inferior boundary (ib), *i.e.*, in the range $[0, ib]$, are considered to indicate poor projections. Metric values in the range $[ib, sb]$ will indicate projections with average quality. Setting these thresholds, thus, allows us to split the study of projection quality in three categories. For T, C, S, and N, we set $ib = 0.4$ and $sb = 0.8$, following earlier studies on how these metrics capture a projection's quality from the respective four viewpoints [21]. For κ, which measures our proposed visual separation, we set $ib = 0.4$ and $sb = 0.7$ based on our empirical observation of visual separation in projections discussed next in Section 5.

Practically put, this leads to the following automated workflow for usage of κ in practice: For a given projection, a computational pipeline measures κ following Section 3.3). If $\kappa > sb$, the projection has good visual separation — so, it can be further shown to its intended users. If $\kappa < ib$, the projection has poor visual separation, so it should not be offered for visual exploration to the users. If $\kappa \in [ib, sb]$, we cannot automatically determine if the projection is 'fit for visual consumption' from the perspective of visual separation, so other metrics or factors should be considered in its assessment. This workflow can be used *e.g.* by a system that computes many projections of a given dataset, *e.g.*, using several algorithms or hyperparameter grid-search, and uses κ to find the best one to serve to its users.

5. Results

We next detail the results of our experiments and our observations in terms of how κ surpasses the standard metrics for VS assessment in projections.

5.1. Quantitative analysis

Correlation plots: As outlined in Section 4.4, we have $18 \times 39 = 702$ dataset-projection combinations, each assessed by five metrics (T, C, S, N, κ). Analyzing this information in table form is not very insightful. As such, we aggregate it in terms of four correlation plots. Each plot compares the values of one of the standard metrics with κ (Fig. 3a). In each plot, we set κ on the x-axis and the standard metric on the y-axis. The plotted points are the 702 dataset-projection combinations. Each red line represents the trend of same-dataset points — there are thus 18 such lines. We see that, for all the four correlation plots in Fig. 3a, blue points are concentrated in the middle-right regions of the plots, indicating that all quality metrics score mostly average or high values. More interestingly, we do not see any positive (or negative) correlation between κ and the standard metrics. Also, for T, C, and S, there are more horizontal red lines than increasing or decreasing trend lines, while, for N, we see more increasing trend lines. This suggests no clear correlation (strictly

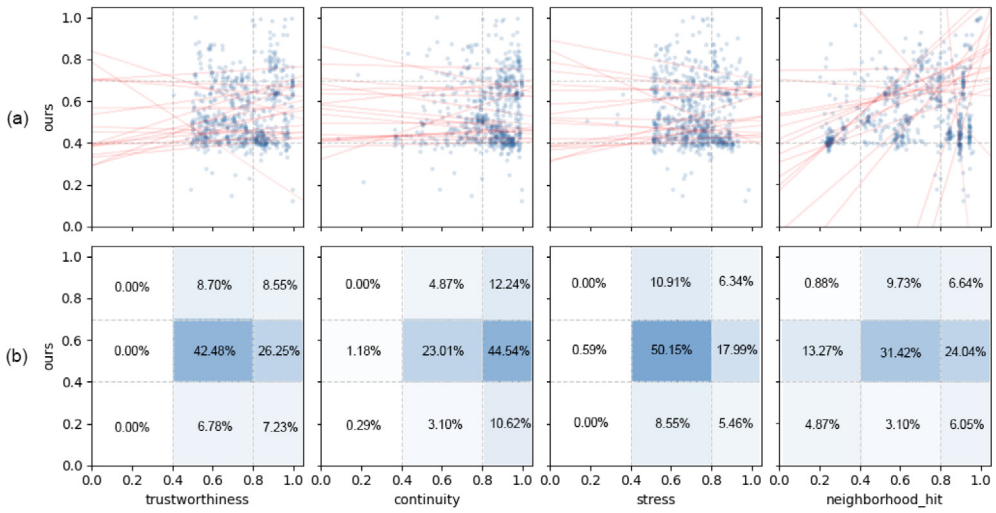


Fig. 3. Correlation plots of κ (ours) and the standard metrics (trustworthiness T, continuity C, stress S, and neighborhood hit N). Stress values are set as $1 - S$ for easy interpretation. In the plots in (a), each blue point is one of the 702 dataset-projection algorithm combinations. Red lines trend all points for the *same* dataset. In (b), we show the fraction of dataset-technique combinations that fall within the nine categories created by the *ib* and *sb* thresholds.

positive or negative) between κ and the standard metrics. For example, if there was a strong pattern of increasing lines for all datasets in one of the four plots, then κ and the standard metric for that plot would agree, *i.e.*, they would gauge visual separation similarly. Conversely, if there was a strong pattern of decreasing lines, then κ and the standard metric for that plot would still capture the same information — high values for one metric would tell the same as low values for the other metric. In these cases, κ would not bring clear added value atop of the respective standard metric. However, our plots do not show this. Since κ and the standard metrics do not show a clear correlation, two situations can happen:

- if the considered projections all have *similar* VS, then all metrics are equally poor predictors, since their values vary a lot while the actual VS is not;
- if the considered projections have *different* VS, then either κ is correlated with it and the standard metrics are not (thus, κ measures VS better than the standard metrics), or the standard metrics are correlated with it and κ is not (thus, the standard metrics measure VS better than κ).

Section 5.2 studies the above hypotheses in further detail by considering the actual perceived VS in the projections.

Fig. 3(b) shows the binning of the metric ranges using the *ib* and *sb* thresholds introduced in Section 4.4). The nine cells are colored to indicate the fraction of the total amount of projection-dataset combinations that fall within each range. T, S, and N present medium-right regions with higher percentages of points (densely populated), while C presents right-medium regions with higher percentages of points. This shows, in short, that interpreting κ is easier than the standard metrics because it has a higher variance. The fact that the standard metrics have a low variance – thus small changes in their values – means that it is harder to use their values in practice to determine the quality of a projection. This is also visible in the selected examples in Table 1. More interestingly, earlier work has observed that projections with very different visual separation yield standard metrics of quite similar values [21,86]. We show next in Section 5.2 that κ's variance is connected to the perceived visual separation in the projections.

Statistical analysis: Table 4 refines these insights on the variance of the compared metrics. We show here the minimum, maximum, mean, standard deviation (std), median, and mode values for all metrics computed for the 702 dataset-projection combinations. Minimum and maximum values are similar for all metrics, except for T, which shows a minimum value of 0.4. Mean and median values are higher than 0.72 for T, C, and S. For N, mean and median values are higher than 0.65 with the highest standard deviation of all considered metrics. κ presents values between 0.49 and 0.55 for mean and median. Mode values are higher than 0.8 for T, C, S, and N, while around of 0.7 for κ. Summarizing, T, C, S, and N mostly assign high values for projections, while κ suggests a wider range of values. This supports our earlier observation that κ may be easier to interpret than the standard metrics. Interestingly, mean, median, and mode are in the same range of the most densely populated regions highlighted in the scatterplots of Fig. 3.

Table 4
Minimum, maximum, mean, standard deviation (std), median and mode values for each metric. Values are calculated over all datasets and projection techniques.

Metric	Minimum	Maximum	Mean	Std	Median	Mode
κ	0.120832	1.000000	0.541826	0.150	0.499891	0.696969
T	0.407621	0.998752	0.753015	0.145	0.762368	0.820627
C	0.087105	0.999063	0.833071	0.143	0.876587	0.940296
S	0.185023	1.000000	0.723541	0.128	0.729920	0.817838
N	0.203571	1.000000	0.653787	0.237	0.686142	0.914286

5.2. Qualitative analysis

Our quantitative analysis showed that κ is not correlated with the standard metrics (Section 5.1). We further study if κ better reflects visual separation by several qualitative analyses.

Random analysis: Since it is not practical to study all 702 projection-dataset scatterplots, we first randomly select several such scatterplots to show the diversity of VS among different projection techniques (Fig. 4). For this, we first randomly choose eight of the 18 datasets. For each standard metric T, C, S, and N, we next randomly select three different projection techniques per dataset, yielding a total of 24 projection scatterplots considered for each standard metric. We next show the correlation plot between κ and each metric. In this correlation plot, the points associated to the three selected projection techniques are highlighted (blue). In each row of Fig. 4, projections are sorted left-to-right by decreasing κ.

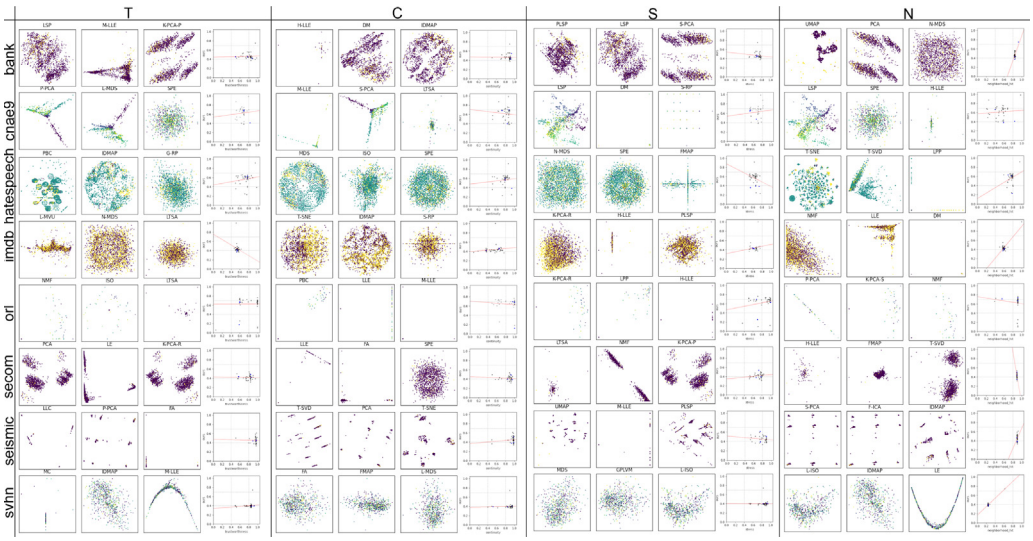


Fig. 4. Randomly chosen projections. Rows contain eight randomly selected datasets. For each of the T, C, S, and N metrics and for each dataset, we show three randomly chosen projections sorted left-to-right on decreasing κ. The fourth column (for each of the four metrics) shows the correlation plot between κ and the metric. Blue points in this plot indicate the three selected projections.

Several things are visible in this figure, as follows. We first notice that dense correlation plots, *i.e.*, datasets for which the projections have small ranges of both κ and the compared standard metric, presented projections with *poor* VS — see, for example, the projections for the *imdb*, *secom*, *seismic*, and *svhn* datasets. The selected projections for these datasets have values of T, C, S, and N higher than 0.8. For these projections, κ ranges from 0.3 to 0.5. Thus, here, the low κ agrees with the perceived poor visual separation, while the standard metrics do not. In the leftmost columns for each metric, we notice *average* visual separation — see *e.g.* the datasets *bank*, *cnae9*, and *hatespeech*. For these cases, κ values exceed 0.6. The standard metrics for these cases range from very low (0.4) to very high (0.9). This is a second indication that κ reflects perceived visual separation better than the standard metrics. Finally, we cannot see any projection with *average* or *good* visual separation in the third (rightmost) column of each row. These are the scatterplots with the lowest κ among the selected ones. This also shows a good agreement between κ and the perceived visual separation.

Ranked analysis: Fig. 5 shows projections ranked by each metric for all 18 studied datasets. In each row (metric) per dataset, we show the best three projections (three left columns in each dataset, surrounded in green) and the worst three projections (three right columns in each dataset, surrounded in red). Within each group of three such projections, the projections are sorted left-to-right on decreasing values of the respective metric. A larger version of Fig. 5 showing more details is given in the supplementary material.

An immediate observation is that projections having the highest (respectively lowest) κ values are also the best, respectively worst, in terms of perceived visual separation. We see many projections having similar VS that are ranked either best or worst by the standard metrics, see *e.g.* the *bank*, *cifar10*, *hiva*, and *imdb* datasets. So, standard metrics are not good predictors of VS. Projections with *average* VS are ranked as worst by at least one of the standard metrics — see *e.g.* the *coil20*, *fashion_mnist*, *fmd*, and *har* datasets. An interesting point concerns N: When both N and κ agree in the (first) best projection, the second-best N value actually has poor VS — see *e.g.* the *cnae9*, *coil20*, *fashion_mnist*, *fmd*, *hatespeech*, and *imdb* datasets. This matches the fact that N shows more increasing trend lines for some datasets in the correlation plots (Section 5.2) compared to the other standard metrics. Hence, N is also not a good predictor of VS. Also, we see that one of the best three projections in terms of κ is seen as the worst projection by the standard metrics for the *cnae9*, *coil20*, *fmd*, *har*, *hiva*, and *sentiment* datasets. All in all, we consistently see that κ has high values for high perceived VS and low values for poor perceived VS, while the standard metrics do not correlate with VS.

Correlation plot and ranked analysis: Fig. 3 showed that there is no correlation of κ with the standard quality metrics (Section 5.1). However, this figure did not show whether there is a correlation between κ and the perceived visual separation. To do this, we select, from the best three ranked projections by κ in Fig. 5, those with convincingly good visual separation as perceived by ourselves. These are UMAP (*bank, cnae9, coil20, fashion_mnist, fmd, har, hatespeech, imdb, seismic, sentiment, spambase*); t-SNE (*cnae9, coil20, fashion_mnist*); Projection by Clustering (PBC) [78] (*coil20, fashion_mnist*); and Interactive Document Maps (IDMAP) [61] (*sms*). Note that, for some datasets, we did not find any projection with a convincingly good visual separation. Separately, we take all the worst-three-ranked projections by κ in Fig. 5 which we visually confirm that have a very poor visual separation.

Fig. 6 shows the correlation plots between κ and the standard metrics — same as Fig. 3, but with the projections selected as best, respectively worst, marked in green, respectively red. We see that the green and red points are far apart from each other along the vertical (κ) axis. The green points clearly at the top, above $\kappa = 0.7$. The red points are nearly all below $\kappa = 0.4$, with and all below $\kappa = 0.5$. Hence, κ correlates very well with our perception of visual separation. However, we see that both green and red points spread quite uniformly along the entire range of the co-plotted standard metric (horizontal axis). For example, there are many red points with $\kappa < 0.4$ which have standard metric values above 0.8 and even close to 1; and there are also many green points with standard metric values below 0.7. All in all, this shows that the standard metrics do not correlate in any significant way with the perceived visual separation.

User study: We further check the correlation of κ with perceived visual separation by a user study. We recruited $S = 108$ participants (37 female, 64 male, 7 other/undisclosed) via an online questionnaire. The participants are first shown a few examples of projections with good, average, and poor separation (gauged

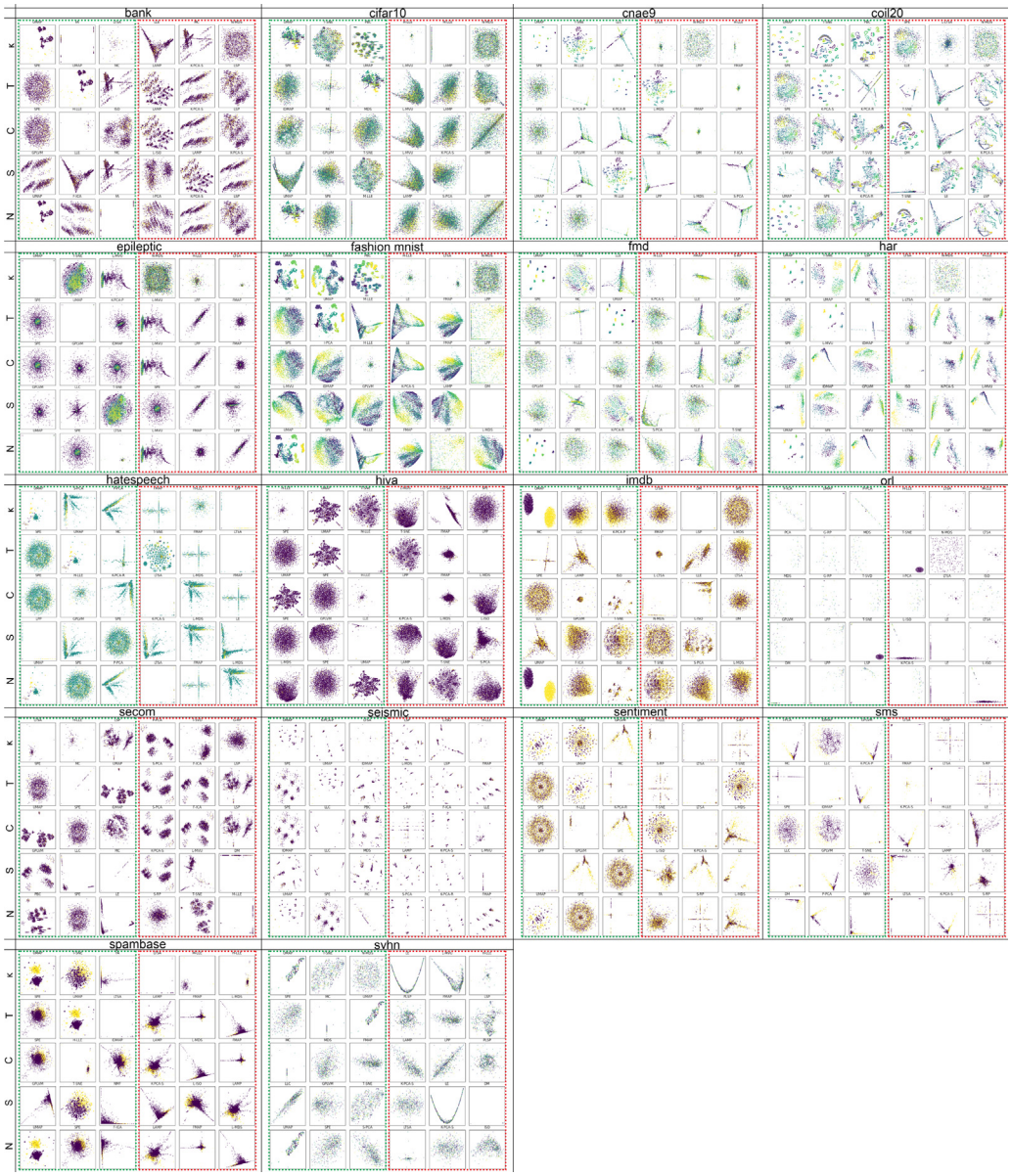


Fig. 5. Projections ranked by each metric (κ, T, C, S, and N) for all 18 studied datasets. In each row per dataset, the first three columns show the best three (dashed green) projections; the last three columns (dashed red) show the worst three projections according to each metric.

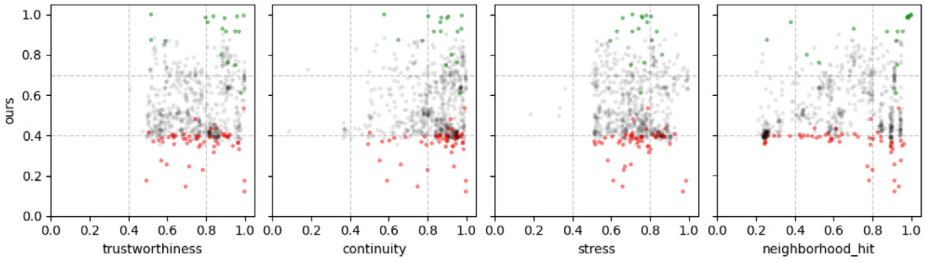


Fig. 6. Correlation plots of κ with the standard metrics with points denoting projections with good perceived VS in green and poor perceived VS in red, respectively. We see that perceived VS correlates very well with κ but not with any of the standard metrics.

by us) so as to understand the idea of scoring visual separation. Next, they are asked each to rank $T = 27$ projection scatterplots on a 5-point Likert scale ranging from very poor to very good, without being given a time limit. The T projections are computed, for each user, by random sampling from the distribution of κ values over all 702 projection-dataset scatterplots we computed previously (Section 5.1). This ensures that we (a) show to users projections with all obtained κ values; (b) show relatively more projections for the more frequent κ values; and (c) users get different projections to score.

Fig. 7 plots the users' recorded scores *vs* the κ values for the $P \cdot T = 2916$ evaluated projection scatterplots. To reduce visual

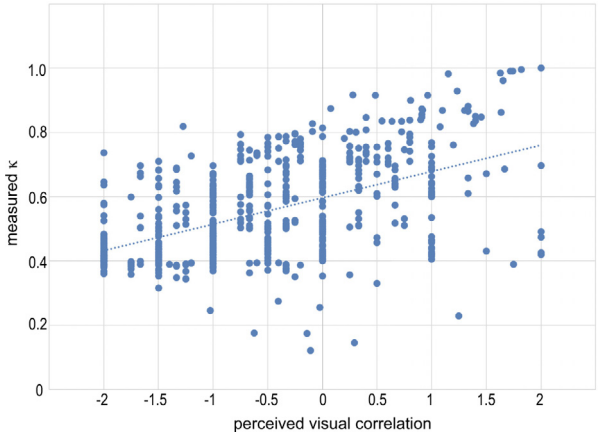


Fig. 7. Correlation plot of measured κ with perceived visual separation scores (measured by 108 users on 2916 projections).

clutter, we averaged scores computed over the same scatterplot by multiple users. The Pearson correlation of visual separation scores with κ is 0.55. Moreover, if we leave out the projections with $\kappa \leq 0.2$ – that is projections with an extremely poor estimated visual separation (which are also very hard to assess by users), we get a correlation score of 0.64. While, as expected, we see some spread of the user scores for the same κ value (and conversely), Fig. 7 and the computed correlation factors mentioned above tell us that κ is in good agreement with the perceived visual separation.

6. Discussion

We next discuss our main findings.

Assessing VS by existing metrics: Our experiments show that the T, C, S, and N projection-quality metrics cannot be (easily) used to predict visual separation of same-label clusters in projections. Our statistical analysis indicates that these metrics have high mean, median, and mode values — they tend to assign values above 0.8 to many projections of many datasets. Hence, even high values of these metrics can lead to poor or indistinct VS. Our qualitative analysis shows that projections which have narrow ranges of these metrics have poor or indistinct VS. Also, for a given dataset projected by several methods, the one having the best VS does not necessarily have the highest value of all (or some) of the standard metrics. Conversely, we see cases in which the highest metric value leads to one of the worst-VS projections for a given dataset projected by several methods.

Our approach to assess VS: Using κ to gauge OPFSemi's performance in label propagation on projected spaces — was consistently shown to better capture VS of projections than the aforementioned four metrics. Our statistical analysis indicates that κ shows reasonable values for mean and median when considering all compared datasets. A mean and mode around 0.5 and 0.7, respectively, suggests that our approach evaluated a large number of projections with values around 0.7, but also a significant amount of low values to compensate the mean. Our qualitative analysis shows that κ values can better capture the extreme cases: Values of κ roughly above 0.7 all have good perceived VS; values of κ below 0.4 correspond to projections where no discernible VS is present. Values of κ in the range [0.4, 0.7] indicate projections with an average amount of VS.

In our analysis, UMAP, t-SNE, and PBC consistently score high VS values for all datasets. These were also the best techniques found by the independent study of Espadoto et al. [21] which used the average of T, C, S, and N. Importantly, this does *not* mean that the said average can be used to measure visual separation. As shown in Table 1 and Fig. 6 projections can have high T, C, S, and N values and *still* poor VS. The said four metrics measure how well a projection captures data patterns (neighborhoods and distances); our κ measures how well a projection is visually separated into different same-label groups. Hence, a good projection should have ideally high values of *all* T, C, S, N, and κ. Our κ is an additional quality factor that complements, but does not replace, existing quality metrics.

Computational cost to assess VS: Measuring VS by our method is fast and requires no parameter settings. For example, for the *hiva* dataset – $N = 3076$ samples, $n = 1617$ dimensions, the largest among the evaluated datasets (Section 4.1) – computing κ took only 0.1216 s on a consumer-grade laptop on average for all the considered 39 projections (0.1149 s to run OPFSemi; 0.0067 s to compute Cohen's Kappa). In contrast, assessing the four standard metrics requires an expensive grid-search procedure to factor out their hyperparameter values and is quite slow to compute (minutes per dataset [21]).

Limitations: We measure OPFSemi's performance by propagating labels from 50% of the samples in a dataset to the remaining ones. It is not currently clear how our results – and the ability of κ to measure visual separation – depends on this data split. Yet, earlier work has shown that OPFSemi has consistent performance even when using far fewer labels [13–15]. Using this fraction as a parameter is interesting to consider as this would define a *multiscale* visual separation metric. We aim to study this aspect in future work, together with a comparison of our kappa score with Silhouette coefficient based metrics computed for a wide range of clustering methods and clustering hyperparameter settings.

A separate aspect relates to the interpretation of visual separation. A projection having poor visual separation is not necessarily a 'bad' one – the labels may be intrinsically mixed up in the high-dimensional space, in which case it is hard to assume that a projection can separate them well. Conversely, if we know that a projection is poor in terms of its T, C, S, and N quality metrics, the fact that it has (or not) a good visual separation is of little relevance to its actual usefulness for data exploration tasks – in general, one should not further use such a projection since it does not represent well the data structure. However, for datasets where we *know* that labels are well separated in the data *and* we know that the projection has high data-structure-preserving quality, we expect the projection to keep this aspect. In these cases, we can use our approach to gauge the quality of the projection. As such, visual separation should be used *in conjunction* with other information to judge the suitability of a projection for visual exploration tasks — a conclusion drawn from different viewpoints also by earlier authors [5].

Lastly, while our user study (Section 5.2) shows that κ correlates with perceived visual separation, extra analysis is needed to show how this depends on projection techniques, datasets, and user experience. We aim to cover this in future work.

7. Conclusion

We proposed a novel approach to assess the visual separation quality of 2D projections. Our approach is based on assessing the performance of a graph-based semi-supervised classifier in propagating labels in the projection (2D) space. If high label propagation performance is achieved, *i.e.*, few wrongly labels are assigned then the projected space is well separated into distinct groups of same-label samples. To evaluate our proposal, we executed both quantitative and qualitative analyses using 18 datasets and 39 projection techniques in line with the benchmark

of [21]. We showed that our proposed approach can better gauge visual separation in projections than common projection-quality metrics. Up to our knowledge, this is the first time that the visual separation quality of 2D projections is assessed through label propagation task for many projection techniques.

We next aim to evaluate the impact of different amounts of labels in the classifier to assess visual separation in projections. Also, we aim to explore the OPFSemi classifier to evaluate projection quality in reducing the high-dimensional space while preserving patterns of the original data, by combining optimum path forests computed in both high and low dimensional spaces.

CRediT authorship contribution statement

Bárbara C. Benato: Conceptualization, Methodology, Software, Data curation, Visualization, Investigation, Writing – original draft, Writing – review & editing. **Alexandre X. Falcão:** Writing – review & editing, Supervision. **Alexandru C. Telea:** Conceptualization, Methodology, Writing – review & editing, Supervision.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

The authors acknowledge CAPES grants with Finance Code 001, FAPESP grants #2014/12236-1, #2019/10705-8, #2022/12668-5, and CNPq grants #303808/2018-7. We also acknowledge Carlijne Govers, Utrecht University, for organizing the user study presented in Section 5.1.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cag.2023.08.023.

References

[1] Maaten LVD, Postma E, den Herik JV. Dimensionality reduction: a comparative review. J Mach Learn Res 2009;10:66–71.
[2] Rauber P, Falcão A, Telea A. Projections as visual aids for classification system design. Inf Vis 2017;17(4):282–305.
[3] Rauber PE, Falcão AX, Telea AC. Visualizing time-dependent data using dynamic t-SNE. In: Proc. eurovis – short papers. 2016, p. 73–7.
[4] Benato BC, Gomes JF, Telea AC, Falcão AX. Semi-supervised deep learning based on label propagation in a 2d embedded space. In: Proc. CIARP. 2021, p. 371–81.
[5] Nonato L, Aupetit M. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. IEEE TVCG 2018.
[6] Venna J, Kaski S. Visualizing gene interaction graphs with local multidimensional scaling. In: Proc. ESANN, vol. 6. 2006, p. 557–62.
[7] Joia P, Coimbra D, Cuminato JA, Paulovich FV, Nonato LG. Local affine multidimensional projection. In: Proc. IEEE TVCG. 2011, p. 2563–71.
[8] Paulovich FV, Nonato LG, Minghim R, Levkowitz H. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. IEEE TVCG 2008;564–75.
[9] van der Maaten L, Hinton G. Visualizing data using t-SNE. JMLR 2008;9:2579–605.
[10] Agrafiotis DK. Stochastic proximity embedding. J Comput Chem 2003;24(10):1215–21.
[11] McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction. 2018, arXiv preprint arXiv:1802.03426.
[12] Rodrigues FCM, Espadoto M, Hirata Jr R, Telea A. Constructing and visualizing high-quality classifier decision boundary maps. Information 2019;10(9):280–97.
[13] Benato BC, Telea AC, Falcão AX. Semi-supervised learning with interactive label propagation guided by feature space projections. In: Proc. SIBGRAPI. 2018, p. 392–9.
[14] Benato BC, Gomes JF, Telea AC, Falcão AX. Semi-automatic data annotation guided by feature space projection. Pattern Recognit 2021;109:107612.
[15] Benato BC, Telea AC, Falcão AX. Deep feature annotation by iterative meta-pseudo-labeling on 2D projections. Pattern Recognit 2023;141:109649.
[16] Benato BC, Falcão AX, Telea A-C. Linking data separation, visual separation, and classifier performance using pseudo-labeling by contrastive learning. In: Proc. VISAPP. 2023.
[17] Zhou Z, Zu X, Wang Y, Lelieveldt BF, Tao Q. Deep recursive embedding for high-dimensional data. IEEE TVCG 2022;1237–48.
[18] Abbas MM, Aupetit M, Sedlmair M, Bensmail H. Clustme: A visual quality measure for ranking monochrome scatterplots based on cluster patterns. Comput Graph Forum 2019.
[19] Amorim W, Falcão A, Papa J, Carvalho M. Improving semi-supervised learning through optimum connectivity. Pattern Recognit 2016;60:72–85.
[20] Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 1973;33(3):613–9.
[21] Espadoto M, Martins R, Kerren A, Hirata N, Telea A. Toward a quantitative survey of dimension reduction techniques. IEEE TVC 2019;27(3):2153–73.
[22] Martins R, Coimbra D, Minghim R, Telea A. Visual analysis of dimensionality reduction quality for parameterized projections. Comput Graph 2014;41:26–42.
[23] Marghescu D. Evaluating the effectiveness of projection techniques in visual data mining. In: Proc. IASTED. 2006, p. 186–93.
[24] Sips M, Neubert B, Lewis JP, Hanrahan P. Selecting good views of high-dimensional data using class consistency. Comput Graph Forum 2009;28:831–8.
[25] Tatu A, Bak P, Bertini E, Keim D, Schneidewind J. Visual quality metrics and human perception: an initial study on 2d projections of large multidimensional data. In: Proc. AVI. 2010, p. 49–56.
[26] Sedlmair M, Tatu A, Munzner T, Tory M. A taxonomy of visual cluster separation factors. Comput Graph Forum 2012;31:1335–44.
[27] Sedlmair M, Aupetit M. Data-driven evaluation of visual quality measures. Comput Graph Forum 2015;34:201–10.
[28] Wilkinson L, Anand A, Grossman R. Graph-theoretic scagnostics. In: Proc. IEEE infoVis. 2005, p. 21–21.
[29] Wilkinson L, Wills G. Scagnostics distributions. J Comput Graph Statist 2008;17(2):473–91.
[30] Motta R, Minghim R, de Andrade Lopes A, Oliveira MCF. Graph-based measures to assist user assessment of multidimensional projections. Neurocomputing 2015;150:583–98.
[31] Tatu A, Albuquerque G, Eisemann M, Schneidewind J, Theisel H, Magnork M, et al. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In: Proc. IEEE VAST. 2009, p. 59–66.
[32] Albuquerque G, Eisemann M, Magnor M. Perception-based visual quality measures. In: Proc. IEEE VAST. 2011, p. 13–20.
[33] Tatu A, Bak P, Bertini E, Keim D, Schneidewind J. Visual quality metrics and human perception: an initial study on 2d projections of large multidimensional data. In: Proc. AVI. 2010, p. 49–56.
[34] Amorim W, Rosa G, Rogério, Castanho J, Dotto F, Rodrigues O, et al. Semi-supervised learning with connectivity-driven convolutional neural networks. Pattern Recognit Lett 2019;128:16–22.
[35] Kim Y, Telea AC, Trager SC, Roerdink JB. Visual cluster separation using high-dimensional sharpened dimensionality reduction. Inf Vis 2022;21(3):197–219.
[36] Benato BC, Telea AC, Falcao AX. Iterative pseudo-labeling with deep feature annotation and confidence-based sampling. In: Proc. SIBGRAPI. IEEE; 2021, p. 192–8.
[37] Benato BC, Falcão AX, Telea AC. Code repository. 2023, https://github.com/barbarabenato/measuring_quality_of_projections.
[38] Moro S, Cortez P, Rita P. A data-driven approach to predict the success of bank telemarketing. Decis Support Syst 2014;62:22–31.
[39] Krizhevsky A, Hinton G, et al. Learning multiple layers of features from tiny images. Tech. rep, Toronto, Canada; 2009.
[40] Ciarelli PM, Oliveira E. Agglomeration and elimination of terms for dimensionality reduction. In: Proc. IEEE ISDA. 2009, p. 547–52.
[41] Nene SA, Nayar SK, Murase H, et al. Columbia object image library (coil-20). Tech. rep., Columbia University; 1996.
[42] Andrzejak RG, Lehnertz K, Mormann F, Rieke C, David P, Elger CE. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. Phys Rev E 2001;64(6):061907.

[43] Xiao H, Rasul K, Vollgraf R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. 2017, arXiv:1708.07747.
[44] Sharan L, Rosenholtz R, Adelson E. Material perception: What can you see in a brief glance? J Vis 2009;9(8):784.
[45] Anguita D, Ghio A, Oneto L, Parra X, Reyes-Ortiz JL. Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In: Proc. IWAAL. 2012, p. 216–23.
[46] Davidson T, Warmsley D, Macy M, Weber I. Automated hate speech detection and the problem of offensive language. In: Proc. AAAI ICWSM, vol. 11. 2017, p. 512–5.
[47] Guyon I, Saffari A, Dror G, Cawley G. Agnostic learning vs. prior knowledge challenge. In: Proc. IEEE IJCNN. 2007, p. 829–34.
[48] Maas A, Daly RE, Pham PT, Huang D, Ng AY, Potts C. Learning word vectors for sentiment analysis. In: Proc. NAACL-HLT. 2011, p. 142–50.
[49] Samaria FS, Harter AC. Parameterisation of a stochastic model for human face identification. In: Proc. IEEE applications of computer vision. 1994, p. 138–42.
[50] McCann M, Johnston A. SECOM dataset. In: UCI Machine Learning Repository. 2008.
[51] Sikora M, et al. Application of rule induction algorithms for analysis of data collected by seismic hazard monitoring systems in coal mines. Arch Min Sci 2010;55(1):91–114.
[52] Kotzias D, Denil M, De Freitas N, Smyth P. From group to individual labels using deep features. In: Proc. ACM SIGKDD. 2015, p. 597–606.
[53] Almeida TA, Hidalgo JMG, Yamakami A. Contributions to the study of SMS spam filtering: new collection and results. In: Proc. ACM symposium on document engineering. 2011, p. 259–62.
[54] Hopkins M, Reeber E, Forman G, Suermondt J. Spambase dataset. Hewlett-Packard Labs; 1999.
[55] Netzer Y, Wang T, Coates A, Bissacco A, Wu B, Ng AY. Reading digits in natural images with unsupervised feature learning. In: Proc. NIPS. 2011.
[56] Coifman RR, Lafon S. Diffusion maps. Appl Comput Harmon Anal 2006;21(1):5–30.
[57] Jolliffe IT. Principal component analysis and factor analysis. In: Principal component analysis. 1986, p. 115–28.
[58] Faloutsos C, Lin K-I. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In: Proc. ACM SIGMOD. 1995, p. 163–74.
[59] Lawrence N. Gaussian process latent variable models for visualisation of high dimensional data. In: Proc. NIPS, vol. 16. 2003, p. 329–36.
[60] Hyvarinen A. Fast ICA for noisy data using Gaussian moments. In: Proc. IEEE ISCAS, vol. 5. 1999, p. 57–61.
[61] Minghim R, Paulovich FV, de Andrade Lopes A. Content-based text mapping using multi-dimensional projections for exploration of document collections. In: Proc. SPIE, vol. 6060. 2006, p. 259–70.
[62] Tenenbaum JB, De Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000;290(5500):2319–23.
[63] Chen Y, Crawford M, Ghosh J. Improved nonlinear manifold learning for land cover classification via intelligent landmark selection. In: Proc. IEEE IGARSS. 2006, p. 545–8.
[64] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Proc. NIPS, vol. 14. 2001, p. 585–91.
[65] Teh Y, Roweis S. Automatic alignment of local representations. In: Proc. NIPS, vol. 15. 2002, p. 865–72.
[66] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science 2000;290(5500):2323–6.
[67] Donoho DL, Grimes C. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. In: Proc. of the national academy of sciences, no. 10. 2003, p. 5591–6.
[68] Zhang Z, Wang J. MLLE: Modified locally linear embedding using multiple weights. In: Proc. NIPS, vol. 19. 2006, p. 1593–600.
[69] He X, Niyogi P. Locality preserving projections. In: Proc. NIPS, vol. 16. 2003, p. 153–60.
[70] Zhang Z, Zha H. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM J Sci Comput 2004;26(1):313–38.
[71] Zhang T, Yang J, Zhao D, Ge X. Linear local tangent space alignment and application to face recognition. Neurocomput 2007;1547–53.
[72] Brand M. Charting a manifold. In: Proc. NIPS. 2002, p. 985–92.
[73] Torgerson WS. Theory and methods of scaling. Wiley; 1958.
[74] De Silva V, Tenenbaum JB. Sparse multidimensional scaling using landmark points. Tech. rep., Stanford University; 2004.
[75] Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 1964;29(1):1–27.
[76] Weinberger K, Packer B, Saul L. Nonlinear dimensionality reduction by semidefinite programming and kernel matrix factorization. In: AISTATS. 2005, p. 381–8.
[77] Lee D, Seung HS. Algorithms for non-negative matrix factorization. In: Proc. NIPS, vol. 13. 2000, p. 556–62.
[78] Paulovich FV, Minghim R. Text map explorer: a tool to create and explore document maps. In: Tenth international conference on information visualisation. IEEE; 2006, p. 245–51.
[79] Lim J, Ross D, Lin R-s, Yang M-H. Incremental learning for visual tracking. In: Proc. NIPS. 17, 2004, p. 793–800.
[80] Schölkopf B, Smola A, Müller K-R. Kernel principal component analysis. In: Proc. ICANN. 1997, p. 583–8.
[81] Tipping ME, Bishop CM. Probabilistic principal component analysis. J R Statist Soc 1999;61(3):611–22.
[82] Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. J Comput Graph Statist 2006;15(2):265–86.
[83] Paulovich FV, Eler DM, Poco J, Botha CP, Minghim R, Nonato LG. Piece wise laplacian-based projection for interactive data exploration and organization. Comput Graph Forum 2011;30:1091–100.
[84] Dasgupta S. Experiments with random projection. In: Proc. UAI. 2000, p. 143–51.
[85] Halko N, Martinsson P, Tropp J. Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. 2009, arXiv:0909.4061.
[86] Castelein W, Tian Z, Mchedlidze T, Telea A. Viewpoint-based quality for analyzing and exploring 3D multidimensional projections. In: Proc. IVAPP. 2023.