Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices

Michael Sedlmair, *Member, IEEE*, Tamara Munzner, *Member, IEEE*, and Melanie Tory

Abstract—To verify cluster separation in high-dimensional data, analysts often reduce the data with a dimension reduction (DR) technique, and then visualize it with 2D Scatterplots, interactive 3D Scatterplots, or Scatterplot Matrices (SPLOMs). With the goal of providing guidance between these visual encoding choices, we conducted an empirical data study in which two human coders manually inspected a broad set of 816 scatterplots derived from 75 datasets, 4 DR techniques, and the 3 previously mentioned scatterplot techniques. Each coder scored all color-coded classes in each scatterplot in terms of their separability from other classes. We analyze the resulting quantitative data with a heatmap approach, and qualitatively discuss interesting scatterplot examples. Our findings reveal that 2D scatterplots are often 'good enough', that is, neither SPLOM nor interactive 3D adds notably more cluster separability with the chosen DR technique. If 2D is not good enough, the most promising approach is to use an alternative DR technique in 2D. Beyond that, SPLOM occasionally adds additional value, and interactive 3D rarely helps but often hurts in terms of poorer class separation and usability. We summarize these results as a workflow model and implications for design. Our results offer guidance to analysts during the DR exploration process.

Index Terms—Dimensionality reduction, scatterplots, quantitative study

1 Introduction

High-dimensional data analysis is a common challenge amongst experts from many application domains such as science, engineering or finance. When conducting visual analysis of high-dimensional data, one typical approach is to transform the original dataset using a dimensionality reduction (DR) technique to create a lower-dimensional version that preserves as much information as possible from the original, and then visually encode only the reduced data [34]. Many DR techniques exist [45]; the most commonly used for visual data analysis include Principal Component Analysis (PCA) [22] and many variants of Multidimensional Scaling (MDS) [5, 16]. The most common visual encoding (VE) technique for showing the dimensionally reduced data is scatterplots. The three major variants are static 2D scatterplots (abbreviated here as 2D), interactive 3D scatterplots (i3D for short), and static 2D scatterplot matrices (SPLOMs) showing axis-aligned views for every possible pair of reduced dimensions.

A significant amount of previous research has focused on providing broad guidance for high-dimensional data analysis [1, 36, 38, 53], and some has focused more narrowly on guidance for DR in particular [20]. However, there is insufficient empirical guidance on how to visually encode dimensionally reduced data. Although the use of scatterplots for non-reduced data has been extensively studied [29, 31], these findings focus on their use for judging correlation and thus do not generalize to their use with dimensionally reduced data because the new, synthetic dimensions are typically not correlated [34]. While 2D scatterplots have been shown to be more effective than landscapes for both visual search [40] and visual memory [41] tasks with dimensionally reduced data, the different scatterplot variants have not been compared to each other for different types of datasets.

We conducted an empirical study to investigate the interplay between visual encoding and dimensionality reduction techniques. We compared the three scatterplot VE variants (2D, i3D, and SPLOM) over 75 datasets reduced with four different DR techniques: PCA [22], robust PCA [39], Glimmer MDS [21], and t-SNE [44]. In contrast to a typical user study collecting the judgements of a large number of people over a small number of datasets, we conducted a *data study* to collect judgements over a very broad set of data from a small number of trained coders [35]. Two coders judged the class separation of 5460 color-coded classes across 816 scatterplot visualizations.

We then engaged in generating a workflow model that can guide scatterplot choices in the DR exploration process. The workflow model reflects the main findings and implications of our study that 2D is often 'good enough'; that is, i3D and SPLOM do not notably improve visual class separability. If 2D is not good enough, the most promising approach is to keep the same visual encoding but to try another DR technique. Switching to a SPLOM as a next step does occasionally help. Switching to i3D, however, rarely helps and often hurts; that is, it has higher time costs and often provides less class separability, even for artificial datasets specifically designed for 3D.

This work is part of a larger project investigating questions at the intersection of DR and visualization. Our understanding of this intersection is informed by a previous field study of DR and visualization usage across multiple application domains, leading to a better understanding of DR-related visual analysis tasks [34]. Here, we focus on the task of *visual cluster verification*, one of several core tasks identified in that work. The most direct precursor to this work was a taxonomy of factors that contribute to visual cluster separation in scatterplots with DR data [35]. The study presented here was conducted in parallel with the previous data study. It is based on the same set of 816 scatterplots and was conducted by the same two coders. Despite these commonalities, however, the two studies are fundamentally different: we gathered different data, used other analysis techniques, and pursued different research goals. For the cluster separation taxonomy, we collected and analyzed *qualitative* characteristics, with the goal of identifying how interactions between classes occurred *within* a single scatterplot. For this work, we collected and analyzed *quantitative* judgements of class separation to compare *between* the scatterplot variants of 2D, i3D, and SPLOM.

The primary contribution of this paper is the results of a data study across 816 scatterplots, featuring the comparison of three scatterplot visual encoding techniques and four dimension reduction techniques. The secondary contribution is implications for design and usage of scatterplots in visual data analysis with dimensionally reduced data, and an iterative workflow model to guide this use.

- *Michael Sedlmair is with the University of Vienna. E-mail: michael.sedlmair@univie.ac.at.*
- *Tamara Munzner is with the University of British Columbia. E-mail: tmm@cs.ubc.ca.*
- *Melanie Tory is with the University of Victoria. E-mail: mtory@cs.uvic.ca.*

Manuscript received 31 March 2013; accepted 1 August 2013; posted online 13 October 2013; mailed on 4 October 2013.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

1077-2626/13/$31.00 © 2013 IEEE Published by the IEEE Computer Society

Accepted for publication by IEEE. ©2013 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

2 2D, 3D, SPLOM: JUSTIFICATIONS AND ASSUMPTIONS

Our study focuses on three scatterplot techniques: 2D, interactive 3D, and scatterplot matrices. Reducing to 2 dimensions and using a static 2D scatterplot is a very common practice [34]. However, another surprisingly frequent practice is reducing to three dimensions and visually encoding with 3D scatterplots that are interactively inspected [12, 13, 23], despite the known perceptual disadvantages of visualizing non-spatial, abstract data in 3D [9]. Our own previous work has advocated the use of SPLOMs in conjunction with reducing to a moderate number of dimensions rather than assuming that either two or three dimensions suffices for the reduced dimensionality [20].

A typical justification for the use of 3D scatterplots is that the *intrinsic* dimensionality of a dataset is likely to be greater than 2; that is, that it would take more than just 2 dimensions to closely approximate the information in the dataset. The scree plot analysis technique [20], where the fidelity of the reduced data to the original is plotted against the number of dimensions used, is a visual representation of exactly this property: the intrinsic dimensionality of a dataset is estimated by finding a knee in this curve.

Although estimating intrinsic dimensionality is a useful step in the DR analysis process, we focus on a somewhat different task: examining the visual separation of clusters after visually encoding the reduced data. We have observed that reducing to far fewer dimensions than the dataset's intrinsic dimensionality often suffices to make clusters clearly visible after visually encoding. For example, a dataset with an intrinsic dimensionality of 10 might have all of its clusters visually separated even if it is reduced to 2 dimensions and shown with a single 2D scatterplot. We thus reaffirm Kruskal's long-ago suggestion that the idea of *appropriate* dimensionality might be more useful than *intrinsic* dimensionality [25]. Our data study thus focuses on the question of the interplay between reducing to an appropriate dimensionality, and the VE technique used to show that reduced data.

A different rationale for the use of 3D scatterplots is that a choice of reducing the original dataset to three dimensions dictates the choice of visually encoding that data using 3D. However, SPLOMs are an alternative VE technique that can encode three dimensions of data into two spatial dimensions by trading off space for time; that is, they require more screen real estate than a single scatterplot, rather than requiring that the user spend time interacting with them. Moreover, they can be used for more than three dimensions of data.

The idea that interaction imposes significant time cost has been noted by many previous authors [26, 46]. We similarly argue that there are different time costs to using the three scatterplot variants of 2D, i3D and SPLOMs. A single static 2D scatterplot has a very low time cost: all of the information is directly visible in one region without the need to interact with the representation or mentally relate the views. A SPLOM has medium cost: there is no interaction, but a user must switch visual attention between regions and mentally relate the information in different views [26]. To our knowledge, the costs of using a SPLOM have not been studied empirically, but we conjecture that this cost increases as the number of views within the SPLOM increase. The time cost of an interactive 3D scatterplot is high because the user must spend significant time rotating the view to see the structure from different angles in order to see relationships hidden by occlusion in any single viewpoint. We conjecture that these interaction costs in 3D are substantially higher than view-change costs in SPLOMs.

One obvious question is whether a SPLOM might also incur a time cost if a viewer needs time to mentally reconstruct overall shape from axis-aligned projections; in this case, the interaction support of i3D would probably reduce the cognitive load, and thus the total time cost, compared to a non-interactive SPLOM. We argue that shape reconstruction is similar to correlation: important for the general case of non-reduced data but not for the special case of reduced data. That is, it is not an important aspect for analysts engaged in the task of visual cluster separation with dimensionally reduced data.

These cost assumptions are based on an extensive body of work that suggests the superiority of 2D visualizations over 3D ones for non-spatial data [6, 8, 10, 30, 40, 41, 51]. We summarize this related work in Section 3.1.

Based on this reasoning, we were skeptical that i3D would be a valuable visual encoding technique for DR data. We hypothesized that the less costly 2D and SPLOM VE techniques would be *good enough* for the majority of cases; that is, they would show the class structure in a sufficient way. Our analysis approach is based on the idea that when multiple VE techniques are good enough, the one with least time cost is the best alternative.

3 RELATED WORK

We review related work on empirical evidence about 3D vs. 2D visualizations, and empirical studies and guidelines for high-dimensional data analysis involving DR.

3.1 3D vs. 2D: Empirical Evidence

Our work relies on the assertion that 3D scatterplots incur much higher interaction costs than 2D ones. We do not test this assertion directly; it is based on a substantial body of empirical evidence about the appropriateness of 2D and 3D representations for different tasks, of which we describe some below:

Three-dimensional visualizations, particularly 3D scatterplots, suffer from several limitations: occlusion of objects by other objects in the scene [6, 37], scene complexity [6], depth ambiguity [30], perspective distortion of distances and angles [37], and difficulty of interacting and navigating in 3D [30, 6, 51]. Clouds of disconnected points are perhaps one of the worst possible cases for 3D, since depth cues such as shadows and shape from shading [49] cannot be used.

2D and 3D visualizations have been experimentally compared for a wide variety of tasks and data types. For 3D spatial data and tasks, most empirical evidence suggests that interactive 3D visualizations often outperform 2D projections [32, 37]. In terms of non-spatial data, however, there exists an extensive body of previous work that suggests that 2D visualizations are superior to interactive 3D [8, 10, 40, 41]. DR data falls in the latter category of non-spatial data.

Our work specifically focuses on scatterplots. For scatterplots of non-DR data, there is some limited evidence that 3D may be helpful for questions requiring integrated knowledge of 3 dimensions [52]. (The data set used in this study was tiny by modern standards, with only six data points.) However, we are interested only in scatterplots of DR data, where there is substantial empirical evidence that 3D scatterplots are ineffective. In a usability study of 3D DR scatterplots (with no comparison to 2D), Newby [30] reported that people could manage to use the 3D scatterplot representation to navigate an information space, but had difficulty judging distances between items and became disoriented when navigating through 3D space. Chalmers [6] similarly reported that a 3D point cloud representation of a document space suffered from occlusion, and users found it difficult to orient themselves and navigate. In a direct comparison between 2D and 3D scatterplots of DR data, Westerman and Cribbin [51] found that 2D outperformed 3D for a search task. In fact, they reported that 2D was as good or better than 3D even when the data variance accounted for by the 2D representation was only 50-70% of that of the 3D representation. Similarly, Fabrikant [14] demonstrated that 3D DR scatterplots were ineffective compared to 2D DR scatterplots and 2D or 3D information landscapes, for distance judgment and spatial arrangement tasks. A later study by Westerman et al. [50] on a browsing task was less conclusive, but showed that people needed to do significantly more navigation in 3D compared to 2D, and were able to complete a more exhaustive search in 2D within approximately the same amount of time as 3D.

This paper provides complementary evidence that 3D scatterplots are not suitable for cluster verification with DR data, from the perspective of a data study rather than a user study.

3.2 DR: Studies and Guidance

There is very little related work regarding empirical studies comparing human judgement of projections produced by different DR techniques. One notable exception is Lewis et al., who compared how experts and novices judged the quality of 2D scatterplot projections from different DR [28]. In a study with 36 participants, seven datasets and nine

DR techniques, they found that experts agreed in their DR judgements but novice users did not. While the focus of this study was on DR techniques, user differences and a generic quality judging task, we specifically focus on visual encodings of the data, on the breadth of datasets, and the interaction between VE and DR choices for visual cluster separation tasks.

On the other hand, there is a substantial body of previous work providing guidelines for how to explore high-dimensional data using scatterplots. Recent advances in the visualization literature specifically have focused on finding interesting 2D scatterplots by computing and comparing a score for each 2D projection [36, 38, 53]. Wilkinson et. al., for instance, specified nine measures to judge 2D projections, such as stringy, outlying or clumpy. More recently, similar approaches have been proposed that were specifically designed for cluster verification tasks, that is, to find 2D projections that nicely separate points of given classes [1, 36, 38]. While these efforts focus on designing measures, our work aims to develop a workflow model to help users choose among DR and VE techniques. In that sense, our work is similar in spirit to DimStiller [20], a system that provides workflows to guide steps in the high-dimensional data analysis process including choices on selecting and parameterizing DR techniques.

4 METHODS

To empirically evaluate visual encodings for DR data, we conducted a *data study*, where two trained coders inspected 816 scatterplot visualizations and judged the visual separability of color-coded classes of these datasets. The study was conducted together with a previous data study which was based on the same 816 scatterplots inspected by the same two coders [35]. In our previous work, we reported on data that we collected for qualitatively assessing class separation factors and for evaluating automatic separation measures. Here, we analyze a different set of data we gathered, for which the two coders judged and quantified visual class separability.

We first explain how our research interest led to the methodological choice of a data study. We then describe the data study in terms of our guiding hypothesis, the 816 scatterplots, the data we gathered based on these scatterplots, and how we analyzed this data.

4.1 Method Rationale

Our methodological choice to conduct a data study was informed by a long and thorough exploration of other evaluation methods.

Initially, we planned to conduct a user study in order to compare 2D, i3D and SPLOMs for DR data. However, we found that a user study was not the right methodological approach for this research question; a pilot user study with five participants, six datasets and one DR technique, revealed that the results strongly depended on the characteristics of the data as viewed in the scatterplots and not on differences between participants. This finding suggested that it is imperative to include a broad set of dataset characteristics to make generalizable claims; subjective differences and timing costs, as mainly tested in traditional user studies with many users and few datasets, are only of marginal interest for studying class separability across DR and VE choices.

Judging class separability on a very broad set of datasets, DR techniques, and VE choices requires a significant amount of work. We therefore sought automatic class separation measures to conduct these class separability judgements. Such quality measures have gained recent attention [4] and researchers have proposed using them for evaluation purposes [3]. We used two state-of-the-art measures to judge class separability for our study [36, 38], however, we found that they produced unreliable results. By comparing these automatic measures to our human judgment, we identified strong discrepancies in class separation judgments for half of our 816 scatterplots [35]. These failure rates are not acceptable for our purposes.

Since reliable automatic separation measures were not available, we decided to conduct a data study in which class separation is judged by a small number of trained human *coders*. This decision is supported by recent empirical evidence that humans are consistent in their visual cluster evaluation tasks, especially if they are trained experts as in our case [27]. Consistency among expert coders, or *inter-coder reliability*, is a crucial precondition for the methodological approach we took.

While conducting such studies with one coder is not uncommon, we followed the recommended practice of using two coders and assessing objectivity through inter-coder reliability. Given the significant workload of such studies, more than two coders would be unusual.

4.2 Guiding Hypothesis

Our data collection and analysis was informed by a guiding hypothesis. We call it *guiding* to mean that it expressed our intuitions at the beginning of this project, and not to reflect unambiguous and testable cause-effect relations. Our guiding hypothesis was that:

- 2D is **often** good enough for showing visible class structure;
- SPLOM **sometimes** adds more information;
- i3D **rarely** provides additional benefits in real-world datasets, but **sometimes** does for specifically designed synthetic datasets;
- **sometimes** none of these visual encodings reveals visible class structure.

The major goals of our study were to get a better understanding of how often these four situations occur, that is, the quantification of "often", "sometimes", and "rarely", and of how they change under different circumstances, such as choosing between different DR techniques or for datasets with different characteristics.

4.3 Scatterplot samples

The basis for the quantitative judgements of the coders was the same set of 816 *scatterplot samples* that we generated for previous work in which coders made qualitative judgements about them [35]. These samples resulted from the combination of 75 datasets reduced with 4 DR techniques (PCA, Robust PCA, Glimmer MDS, and t-SNE) and visualized with 3 scatterplot VE techniques (2D, i3D, and SPLOM). We call them scatterplot samples to emphasize that while the SPLOM contains many individual scatterplots as subcomponents, we count the entire SPLOM as a single scatterplot sample rather than separately adding each subcomponent to the total. We summarize the generation process briefly here and provide further details in the supplemental material.

We used a set of 75 datasets divided into four different categories: 31 *real* datasets from our colleagues and collaborators [19, 36, 38] or online data repositories [17, 33, 42, 47, 48]; 16 *synthetic-gaussian* datasets with 3 to 5 randomly distributed gaussian clusters; 24 *synthetic-entangled* datasets with higher-dimensionally entangled classes; and 4 *synthetic-grid* datasets that are simply regular high-dimensional grids, to provide a baseline of known and highly regular dataset structure. Figure 5 shows examples of entangled datasets; these were specifically designed so that they could not be untangled with linear DR techniques, and were intended to be the best possible case for 3D scatterplots.

These four categories are ordered from *very realistic* to *highly artificial*, and are presented in this order throughout the paper. The datasets ranged from 77 to 43,500 points (median=500), from 3 to 159 original dimensions (median=7), and between 2 and 53 classes (median=5). All datasets were pre-classified, either in our generation process, by using clustering algorithms, or as provided with the data.

We carefully selected a set of four representative DR techniques to reduce these 75 datasets: The venerable *PCA* [22] technique, which finds linear projections based on variance, is the first choice of many analysts in the real world [34]. As PCA is known to be vulnerable to outliers, we included another linear technique, *Robust PCA* [39] that has been found to be tolerant of outliers. Given the limitations of linear DR techniques, we additionally included two non-linear techniques: *Glimmer MDS* [21] is a representative of the well-known family of multidimensional scaling techniques that seek to optimally map point distances from the high-dimensional space into a low-dimensional projection. The recent approach of *t-Distributed Stochastic Neighbor Embedding (t-SNE)* [44] is a non-linear DR approach specifically designed to separate clusters well—the task we are interested in.

The result of running these 4 DR techniques on the 75 datasets was *272 dimension reduction operations* rather than 300; in the remaining

28 cases, the computations did not complete because the technique assumptions were not met by the dataset characteristics. All computations were done in R.

The dimensionally reduced data was then visually encoded with three scatterplot techniques: 2D, i3D and SPLOMs. Points in all scatterplot samples were color-coded based on the given class structure. For 2D scatterplots we reduced the data to two dimensions, and for i3D to three dimensions. For SPLOMs, we generated a set of *n-way* SPLOMs per dataset, where n is the number of dimensions shown in the SPLOM. The value of n ranged from 3 to d_{max}, a maximum value determined by a combination of the original dimensionality of the dataset and the values of class separation measures [36], with a cap of 15 dimensions as the upper limit. We introduced this cap to keep the study manageable in terms of time costs for the coders. From this set of SPLOMs, each coder individually selected one SPLOM for the data collection process, by making a judgement about when adding more dimensions to the SPLOM stopped providing any benefits in terms of class separation. The outcome of this process was a set of 816 scatterplot samples.

4.4 Data Collection

The data collection was conducted by the same two trained coders[1] and in parallel with the study in our previous work [35]. For the study reported here, the coders used their in-depth manual inspection of 816 scatterplot samples to rate how visually separable each of the color-coded classes were within these scatterplots. Data gathering took two months, with a total of approximately 80 hours of coding time for each person. In the first six weeks, the coders met two to four times a week to discuss many scatterplot samples and iteratively adapt the coding schemes and strategies, re-coding samples as necessary as the schemes changed. This iterative development of the coding scheme is a lengthly process but is crucial for scientific rigor [7, 18].

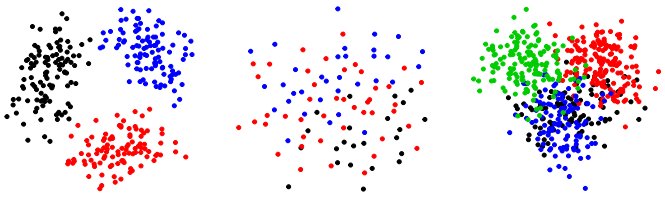
With the number of classes being variable across datasets, each coder made a total of 5460 *classwise ratings*. Each of these classes was given a score between 1 and 5, where 1 means the class is not separated at all and 5 means the class is nicely separated. Initially, we had only three categories, but in the iterative coding process we found that a 3-point scale did not provide enough depth for representing the separability of clusters.

Figure 1 shows examples of different classes and how they were judged by the coders. The coders' definition of visual class separability followed our previously proposed taxonomy of visual cluster separation factors [35]. The most important separability factors we took into account were the amount of spatial overlap between classes and factors related to connectedness between points of a class. A class of points that is connected to each other and that has no overlap with any other class, for instance, gets a "5". Larger spatial overlaps and disconnectedness of points reduce the score. Our manual coding was also robust with respect to a variety of factors that are not reliably testable with state-of-the-art separation measures, such as the shape of a class, the variance of size, point count or point density between classes. That is, as long as a class is connected and without overlay, its shape, size, point count, or density had no negative impact on its rating score. Classes with only 1 point were not judged because the separation of a one-point class is not meaningful. The supplemental material contains a full set of all examples and ratings.

To asses the objectivity of these class-wise ratings, we computed the inter-coder reliability using Krippendorff's alpha. In contrast to other inter-coder measures, Krippendorff's alpha can be used for any number of coders, is robust to missing values, and works with different data types such as nominal, ordinal or ratio; for our classwise ratings Krippendorff's alpha was 0.858 for ordinal data; a score of .8 or greater is considered acceptable in most situations [24].

We also recorded other data that we intended to use to address our research questions. In particular, the coders additionally recorded their *subjective preference* of (a) which combinations of DR and VE they considered the best for each dataset, and (b) which VE they found most helpful for a *dataset* $\times$ *DR* combination[2]. However, this data showed effects of personal preference, with Krippendorff's alpha for (a) being 0.32 and for (b) 0.461. In particular, we found that one coder tended to subjectively pick 3D scatterplots, and the other one tended to prefer 2D and SPLOM. We therefore excluded this data from our analysis. However, as shown above, the low-level classwise ratings were resistant to such subjective biases and provide therefore reliable and un-biased data for our methodological approach.

[1] One of the coders is the first author of this paper, as is standard practice with the methodological approach we took.



(a) max sep: (5,5,5) (b) no sep: (1,1,1) (c) mixed: (4,3,2,1)

Fig. 1. 2D scatterplot samples and class codings from the data study. (a) `entangled3-m-3d-smallOverlap` reduced with PCA representing three classes that were each coded as 5, nicely separated; (b) `gauss-n100-5d-3largeCl` reduced with Glimmer MDS representing three classes that were each coded as 1, not separated at all. (c) `entangled2-4d-overlap` reduced with robust PCA showing an example in between these: the green class got a 4, the red class a 3, the blue class a 2, and the black class, which is completely mixed with the other classes, a 1. Both coders agreed on all these ratings. As with all following scatterplot figures, point sizes have been increased for easier readability in the paper.

4.5 Data Analysis

We undertook extensive exploratory analysis of the collected data. We report only the most interesting results here.

The analysis in Sections 5.1 through 5.3 is primarily presented using heatmaps that either directly show the classwise ratings or show derived data about differences between VE and DR techniques. Heatmaps in the paper show averaged ratings of both coders; the supplemental material contains larger and labeled versions of these averaged heatmaps, as well as separate heatmaps for each coder. We used these representations for our own exploratory analysis and found them very useful. They make visible as many of the class-wise ratings as possible, and allow readers to make their own judgements about the data at both overview and detail levels. Moreover, we also did not want to impose our own opinion of what constitutes 'better'. For instance, is a 'better' VE one where at least one class is more separable, or one where more classes improve than decline? To allow for such differing interpretations, we decided to show the rating data that we gathered in as much detail as possible.

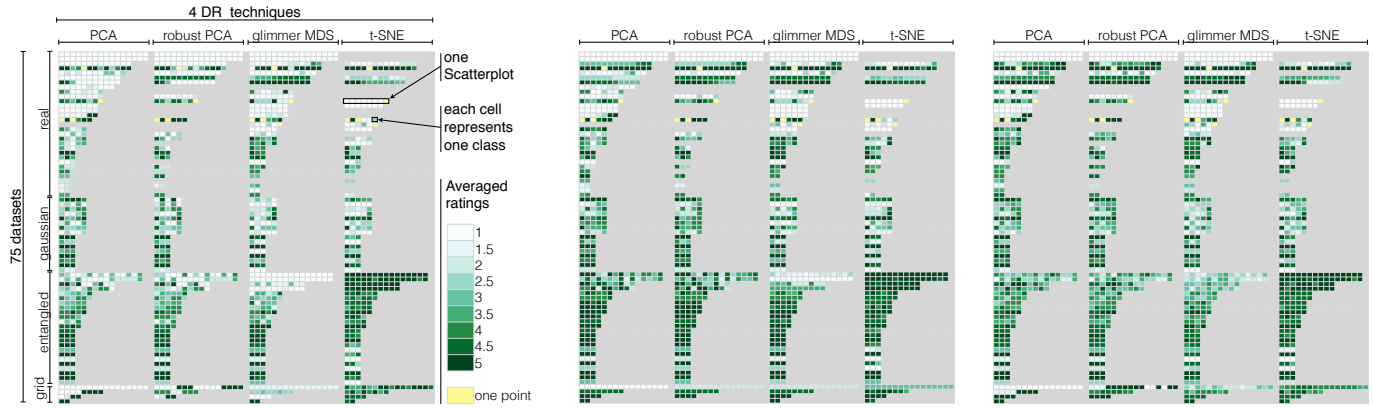
We complemented the heatmap analysis with inferential statistics. Because our rating data are ordinal and not normally distributed, we used the non-parametric Wilcoxon signed rank test (two tailed). Bonferroni correction was applied within each group of tests.

We used the results of the quantitative analysis to select a set of interesting example scatterplots for further qualitative discussion; the supplemental material again contains more of these examples. Our data analysis is thus in the spirit of mixed methods approaches [11].

5 Results and Discussion

We first present the *base data* from the averaged classwise ratings. We then present the *within-DR* analysis of the data for each of the four DR techniques separately, followed by the *cross-DR* analysis across all of them together. After a quantification of results into the four bins given by our guiding hypothesis, we briefly discuss secondary results on the usage of SPLOMs that support our usability assumption.

[2] (a) was collected as nominal data, and (b) as ordinal rankings.



(a) 2D (b) i3D (c) SPLOM

Fig. 2. Base data showing classwise ratings for each scatterplot variant, shown as heatmaps. (a) 2D scatterplots, annotated to show the meaning of the visual representation. (b) Interactive 3D scatterplots. (c) SPLOM. Ratings range from 1 (not separated at all) to 5 (nicely separated).

5.1 Base Data

Figure 2(a) shows an annotated heatmap with 75 rows, representing the average score of both coders on 2D Scatterplots. Each row represents one dataset. The columns are organized into four sections, one per DR technique, ordered as PCA, then Robust PCA, followed by Glimmer MDS, and finally t-SNE. Within these sections one row reflects one scatterplot, in this case 2D, with k cells representing the number of classes of this dataset. The length of these within-section rows vary, as not all datasets have the same number of classes. This encoding enables the visibility of (almost) all classes, and makes datasets with many classes more visually salient. As the task of cluster verification indeed becomes more difficult as the number of classes increases, we decided to normalize the amount of screen space used to classes rather than datasets. These encoding choices should be taken into account when interpreting the heatmaps.

Within a cell, the average rating of a class is encoded with a white-to-green color ramp, with 1 (not separable) in light green, to 5 (nicely separable) in dark green. Yellow cells indicate that this class consisted of only 1 point. The heatmaps cap the number of classes to 16 to maintain legibility. The two topmost rows are the only two datasets that have more than 16 classes; we deemed that the additional information was not crucial for the analysis, as both have consistently low classwise scores for all VEs and DRs.

The rows are sorted by dataset classification, starting with `real` at the top, then `synthetic-gaussian`, `synthetic-entangled` and finally `synthetic-grid` at the bottom, reflecting our ordering from very realistic to highly artificial. Within each category, rows are further sorted by the number of classes.

Figures 2(b) and 2(c) show the base data for 3D scatterplots and SPLOMs encoded in the same way. All subsequent heatmaps also use the same spatial encoding. Intuitively, this visual representation is a 4×4 matrix, with the four different dataset categories on the vertical axis, and the four DR techniques on the horizontal axis.

We will use the following naming convention as notation for describing heatmaps and the sections within them:

- ve, with $ve \in \{2d, 3d, splom\}$, refers to the three visual encoding heatmaps in Figure 2(a), 2(b), and 2(c) respectively;
- $ve : dr$, with $dr \in \{pca, rob, mds, tsne\}$, refers to the vertical DR-technique sections within these heatmaps, e.g. $2d : pca$;
- $ve : type$, with $type \in \{real, gauss, entangled, grid\}$, refers to the horizontal dataset-category sections in the heatmaps, e.g. $2d : real$;
- $ve : dr \times type$ thus refers to one of the 16 sections of the matrix, e.g. $2d : pca \times real$.

Visually comparing the base data of 2D, i3D, and SPLOM reveals some interesting insights. We can observe situations where all classes of a dataset are nicely separable, reflected by only dark green cells (e.g., $2d : tsne \times entangled$); situations with only un-separable classes, only light green cells; and situations where some classes are separable and some classes are not, a mix of dark and light green. The $2d : pca \times entangled$ section contains an example of the latter. There is also a noticable difference between the three VE techniques. For instance, comparing the different $ve : pca \times entangled$ across the three heatmaps reveals that in general the cells in this section are darker for i3D and SPLOM than for 2D. Finally, there are also notable differences between DR choices that we can see by comparing the four DR sections along a particular row. Consider again the $2D : entangled$ datasets in Figure 2(a): while in the PCA, robust PCA and Glimmer MDS sections we can see many light cells, in the rightmost t-SNE section there are mainly dark green cells indicating that this DR technique performed well on nearly all of these datasets.

We continue by deriving Δ-heatmaps that directly show the differences between heatmaps, revealing further results and implications in a way that is perceptually easier to interpret than simply visually comparing the base maps side by side.

5.2 Scatterplot Choices: Within-DR

We first compare the class separation performance of the three visual encoding techniques for each DR technique separately. Our analysis is based on the cost assumptions discussed in Section 2, that 2D is less costly than SPLOM, which is in turn less costly than i3D. We thus analyze our results as two comparisons of the base heatmaps: the difference between SPLOM and 2D, and the difference between i3D and the best of those two.

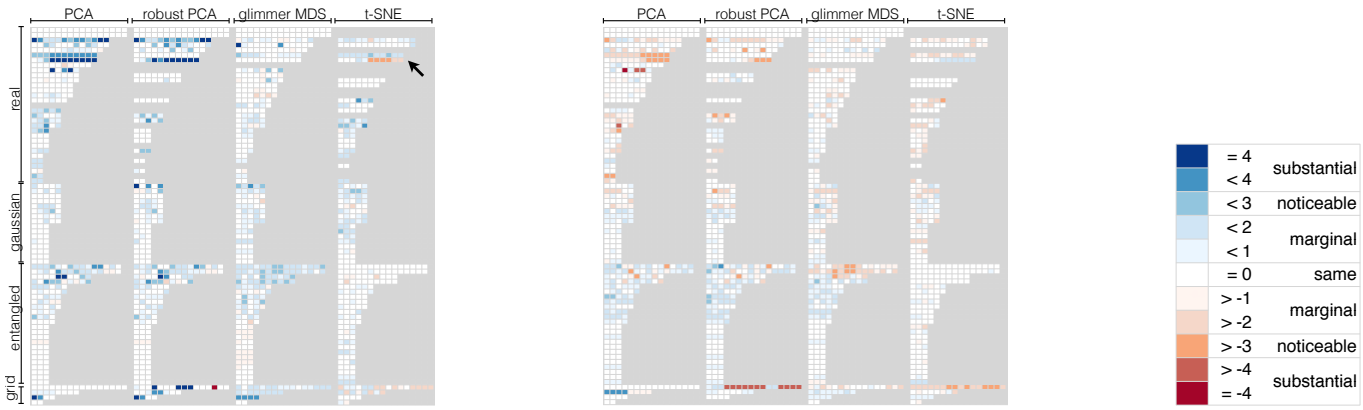
5.2.1 Within-DR: SPLOM

We first show the difference between the SPLOM and 2D classwise ratings for each DR technique separately. Visually speaking, we show a cell-wise subtraction of Figure 2(c) (SPLOM) and Figure 2(a) (2D). This comparison shows both how many classes scored better in the SPLOM compared to 2D within a specific DR technique, and how much better they scored. Using the notation specified above, we formally specify this first data derivation $\Delta^{(1)}$ as:

$$\Delta^{(1)} = splom - 2d$$

The mathematical operator "–" refers to a cell-wise subtraction. We omit line and column indices for notational simplicity and clarity, in this and all following formulae.

The resulting range of cell values val from this Δ-computation is $val \in \{-4, -3.5, ..., 4\}$. The fractional values arise because these scores are averaged between the two coders. In our discussion, we further classify this scale into three bins: when $val \geq 3$ we call the



(a) SPLOM – 2D (b) 3D – max(2D, SPLOM) (c) Color ramp

Fig. 3. Within-DR differences. Heatmap layout is identical to Figure 2(a). (a) $\Delta^{(1)} = splom - 2d$. (b) $\Delta^{(2)} = 3d - \max(2d, splom)$. (c) Diverging blue/red color ramp for all Δ heatmaps. Differences are binned into categories of marginal, noticeable, and substantial. Blue = better, Red = worse (e.g. in (a): a certain class in the SPLOM is better/worse than in 2D).

difference in separability *substantial*; if $3 > val \geq 2$ we call the difference *notable*; if $val < 2$ we call the difference only *marginal*. We call a VE technique *good enough* when the difference between it and the next one is marginal; we consider that upgrading to a more costly scatterplot technique, such as from 2D to a SPLOM, is only worthwhile for notable or substantial differences.

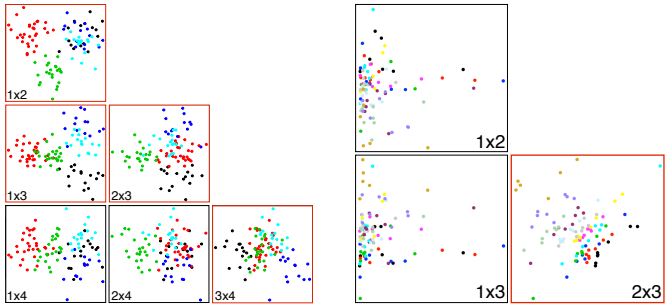
Figure 3(a) shows $\Delta^{(1)}$. All of the Δ heatmaps use the diverging color ramp shown in Figure 3(c). In this case, blue cells show that a class was better in SPLOM, with the saturation of blue encoding how much better. Red cells indicate that 2D was better than SPLOM, and white indicates that they were equally good.

Inspecting Figure 3(a) reveals several scatterplots with saturated blue boxes. In these cases, we argue that changing to the more costly SPLOM would be worthwhile, since some classes are either substantially or notably more separable than in 2D. Two examples in which SPLOM scored better than 2D from the same DR technique are shown in Figure 4. In Figure 4(a), the 2D scatterplot (identical to first 1x2 view) nicely separates the green and the red class, yet the black, blue and cyan are mixed together. All five of these classes can be seen separably by showing more of the principal components in a SPLOM: black can be seen in the 2x3 view, blue in the 3x4 view, and cyan in the 1x3 view. In Figure 4(b), no class is visually separable in the 2D plot (equivalent to the 1x2). But the 2x3 view in the SPLOM shows reasonable separation. The classes appear as adjacent strings that are not visible in the 2D PCA projection. In contrast to Figure 4(a), here a single view in the SPLOM provides the benefit over 2D.

We note that most of the substantial differences appear for the linear DR techniques PCA and robust PCA, and many of them can be found in the topmost section representing the real datasets. We also note that for these linear techniques, SPLOM is never notably or substantially worse than 2D; there are almost no red cells in these columns. This finding is hardly surprising, as in these cases the first 1x2 view in the SPLOM is exactly the 2D plot: both are showing the first principal components. We consider the marginal differences to be uninteresting artifacts of our data collection process.

For the non-linear techniques, switching to SPLOM is not helpful as often, underscoring the conventional wisdom that these techniques have more power to create meaningful embeddings with fewer dimensions that the linear techniques require. In one of the t-SNE instances, 2D is actually noticeably better than SPLOM, shown with a black arrow in Figure 3(a); this situation arises both because t-SNE is a non-deterministic technique that produces different embeddings for every computation and because t-SNE is specifically designed to work well in lower dimensions [43].

We statistically compared the set of classwise $\Delta^{(1)}$ values for each DR technique to a theoretical value of zero, thereby testing the null



(a) (b)

Fig. 4. Within-DR examples where SPLOMs were better than 2D. In both cases, the 1x2 view at the top left is identical to the 2D scatterplot. **(a)** 4-way SPLOM is better than 2D, with the synthetic-gaussian dataset `gauss-n100-10d-5smallCl` reduced by robust PCA. **(b)** 3-way SPLOM is better than 2D, with the real dataset `industryIndices` reduced by PCA. Red boxes mark the views that were judged to best separate specific classes.

hypothesis that there is no overall difference between SPLOM and 2D. Later statistical analyses use this same approach. Wilcoxon tests showed a significant effect in all cases ($V > 9221$, $p < 0.001$ for each DR technique). In other words, SPLOM was overall significantly better than 2D for all DR techniques, though our heatmap analysis shows that 2D was good enough for many individual datasets.

5.2.2 Within-DR: i3D

In the case where neither 2D nor SPLOM are good enough, an analyst might hope that the highest cost interactive 3D Scatterplot could reveal class separability more effectively. To investigate whether 3D might be a fruitful strategy within-DR, we show the difference of 3D classwise ratings compared to the maximum of the 2D classwise ratings and SPLOM classwise ratings. Formally, $\Delta^{(2)}$ can be noted as:

$$\Delta^{(2)} = 3d - \max(2d, splom)$$

Visually speaking, we show a cell-wise subtraction of Figure 2(b) (i3D) and the maximum of Figure 2(c) (SPLOM) and Figure 2(a) (2D). In our exploratory analysis, we found that the *3d – max(2d,splom)* subtraction is only marginally different from *3d – splom*, providing evidence that our cost assumptions hold. The results of this subtraction are shown in Figure 3(b).

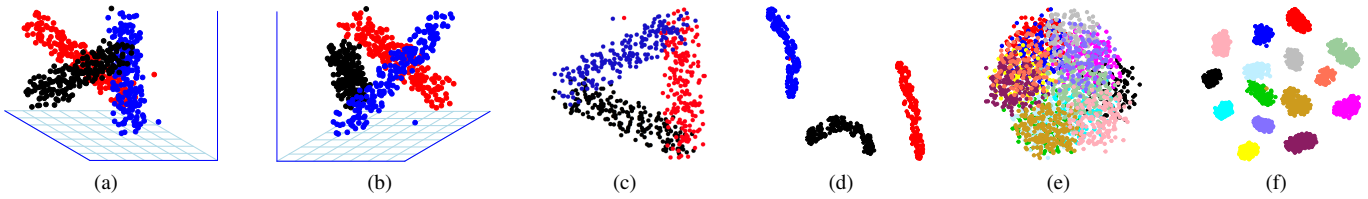
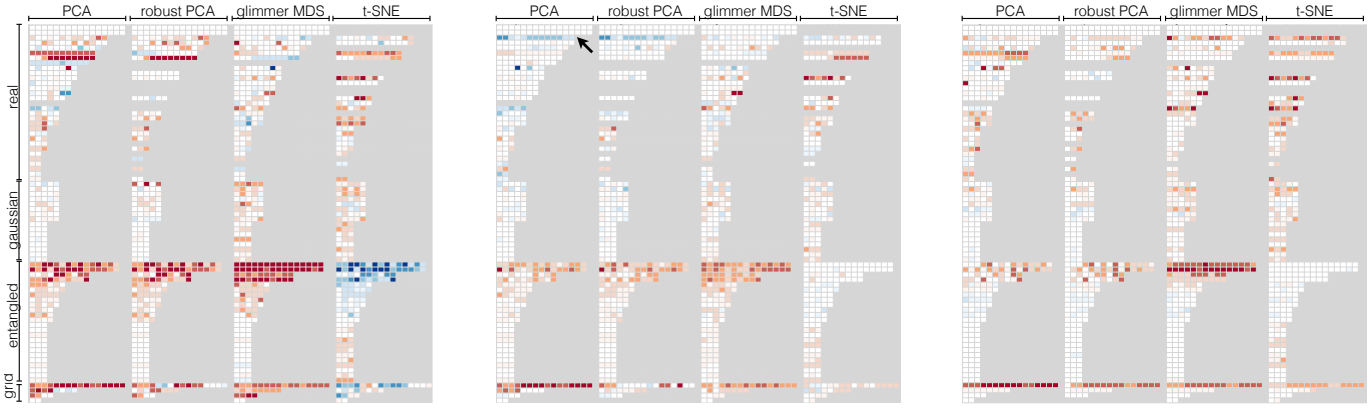


Fig. 5. **(a)-(d):** Screenshots of the entangled dataset `entangled1-3d-3cl-separate` designed to show the most possible benefits for i3D. (a),(b) two viewpoints of the same i3D PCA scatterplot. An accompanying video shows the full 3D rotation. (c) 2D PCA projection. (d) t-SNE untangles this class structure in 2D. **(e)-(f):** 2D scatterplots of the reduced `entangled2-15d-adjacent` dataset which we designed to have a ground truth entangled class structure in 15D. (e) Glimmer MDS cannot untangle the classes, neither can PCA and robPCA (see supplemental material). (f) t-SNE nicely untangles and separates the ground truth classes in 2D.



(a) 2D – 2D from other DRs (b) SPLOM – 2D from all DRs (c) i3D – max (SPLOM, 2D from all DRs)

Fig. 6. Cross-DR differences. Red = another configuration is better, Blue = this configuration is best. (a) 2D compared to all other DRs in 2D. (b) SPLOM compared to all DRs in 2D. (c) i3D compared to SPLOM and all DRs in 2D.

Inspecting Figure 3(b) reveals that i3D rarely helps compared to the best of 2D and SPLOM, and sometimes actually hurts. That is, i3D shows less class structure, visible as red cells with notable or substantial saturation levels, but has higher usage costs. The only notable blue cells are with the highly artificial entangled and grid datasets, reflecting our main hypotheses that i3D would work well for these specifically designed datasets.

Figure 5(a)-(b) shows a simple example of one of our artificial entangled datasets in 3D. The dataset was specifically designed for 3D and has three oblong and thin classes that cannot be fully distinguished with any linear 2D projection. Figure 5(c) shows the 2D PCA Scatterplot of this dataset. We note that even in this example the 2D Scatterplot contains only minimal overlap of the classes, and the clear shape of the clusters makes the classes easy to tell apart.

Wilcoxon tests showed that $\Delta^{(2)}$ was significantly different from zero only for t-SNE ($V = 2514$, $p < 0.001$). In this case, i3D was overall significantly worse than 2D and SPLOM for t-SNE.

5.3 Scatterplot Choices: Cross-DR

The analysis presented thus far has focused on scatterplot choices within a certain DR technique. We now investigate how the three scatterplot techniques perform *cross-DR*; that is, how choosing among different DR techniques influences the VE performance. For that purpose, we first compare how 2D reduced scatterplots compare among the four DR techniques. We then analyze whether SPLOM or i3D add additional benefit in such cross-DR explorations.

5.3.1 Cross-DR: 2D

We are interested in understanding whether one DR technique adds notable or substantial benefits over all others when analyzing a dataset. For this purpose, we compare the 2D classwise ratings of one DR technique to the maximum 2D classwise ratings of the remaining three DR technique. Formally, this $\Delta^{(3)}$ derivation can be written as follows. Let $DR = \{pca, rob, mds, tsne\}$, then

$$\forall dr \in DR : \Delta^{(3)}_{dr} = 2d_{dr} - \max_{\substack{\forall i \in DR \\ i \neq dr}} (2d_i)$$

Visually speaking, we show a cell-wise subtraction between the four vertical DR sections—the set of columns per DR technique—in Figure 2(a). For all four DR techniques, we take their respective DR section and subtract the maximum of the remaining three DR sections. The results are shown in Figure 6(a), using the same color ramp as before. Red cells indicate that for this class there is another DR that better separates it. Blue cells indicate that the specific DR technique separates this class better than all other techniques. White indicates that there are no differences across all or a subset of DR techniques.

Interpreting Figure 6(a) reveals a number of datasets for which picking a different DR technique indeed makes a substantial difference for many of their classes. This finding is reflected by the high number of dark red cells indicating the superiority of another DR technique. In some of these cases, one DR technique is substantially better than all others, which can be seen by the blue boxes. The most noticeable pattern is that t-SNE has superior performance for many of the entangled datasets, seen in the $\Delta^{(3)}$: *entangled* section. Figure 5(e)-(f) gives an example of a dataset with 15 ground truth classes entangled in 15 dimensions. Neither PCA, robust PCA, nor Glimmer MDS, as shown in Figure 5(e)[3], reveals this entangled class structure. In contrast, t-SNE clearly untangels 15 separable classes as shown in Figure 5(f). Figure 5(d) shows how t-SNE performs on the previously discussed example of Figure 5; again, t-SNE clearly untangles the classes.

Statistical analysis supported the superior performance of t-SNE. Wilcoxon tests showed that $\Delta^{(3)}$ was significantly different from zero for PCA, robust PCA, and Glimmer MDS ($V > 1990$, $p < 0.001$ in all

[3] Screenshots of PCA and robust PCA are in the supplemental material.

cases), that is, another DR technique was better overall. There was no significant difference for t-SNE ($V = 15790$, $p < 0.74$).

Although t-SNE provides a benefit for the artificial entangled datasets, and this finding is reflected in the aggregated statistical analysis, for the real datasets seen in the top $\Delta^{(3)}$: *real* section, there is no clear preference for one DR technique. Red boxes can be found across all four DR techniques, indicating differences between DR techniques but also that there is no one-and-only DR solution. While changing from linear techniques to non-linear techniques was fruitful for some datasets, for others the reverse change was also beneficial.

5.3.2 Cross-DR: SPLOM and 3D

One question that remains is how much SPLOMs and i3Ds help in cross-DR explorations. To investigate this question, we provide two comparisons. First, in $\Delta^{(4)}$ we compare, for each DR-technique, how much a SPLOM is better than the 2D scatterplots of all four DR techniques. Second, we investigate the question of how much i3D adds on top of that. Based on our cost arguments, for $\Delta^{(5)}$ we compare i3D of a certain DR technique first to the SPLOM of the same technique. We then compare it to the 2D scatterplots from all four DR techniques. Formally, let $DR = \{pca, rob, mds, tsne\}$, then

$$\forall dr \in DR : \Delta^{(4)}_{dr} = splom_{dr} - \max_{\forall i \in DR}(2d_i)$$

$$\forall dr \in DR : \Delta^{(5)}_{dr} = 3d_{dr} - \max(splom_{dr}, \max_{\forall i \in DR}(2d_i))$$

Figure 6(b) shows the cross-DR performance of SPLOMs $\Delta^{(4)}$, and Figure 6(c) shows it for i3D $\Delta^{(5)}$.

Interestingly, there are very few blue boxes, indicating that changing the visual encoding technique rarely helps on top of cross-DR exploration. Wilcoxon tests supported this claim, with $\Delta^{(4)}$ being significantly less than zero for all DRs ($V > 1676$, $p < 0.001$) and the same for $\Delta^{(5)}$ ($V > 578$, $p < 0.001$). In other words, switching to another DR is a better overall strategy than changing from 2D to SPLOM or i3D.

One of the few examples where a SPLOM added notable class separability is the `industryIndices` real dataset that is marked with a black arrow in 6(b) and shown in Figure 4(b). For this dataset instance, the structure of adjacent clusters was revealed by the 2nd and 3rd principal components, which apparently could not be brought forward by non-linear techniques.

In general, there are also fewer red boxes and more white boxes for SPLOMs as compared to a pure 2D cross-DR comparison 6(a). This change of pattern indicates that in many cases a SPLOM of one DR technique shows the same cluster separability information as a 2D scatterplot created with another DR technique. Nevertheless, there are many notable and substantial red boxes visible, indicating that a 2D plot from another DR technique is better than a SPLOM with this particular one.

For i3D, we note that there is no single class in the tested datasets where i3D would add any notable or substantial difference in a cross-DR exploration scenario. There are also many saturated red cells, showing that i3D often performs worse than the 2D encodings of other DR techniques or the SPLOM of the same DR technique; that is, it 'hurts' in terms of having less class separability.

5.4 Quantification

Following our main guiding hypothesis, we were interested in learning about the quantities on how often 2D is good enough, how often SPLOM adds further class separability, how often i3D adds further class separability, and how often none of these visual encoding techniques reveal visible class structure. To get a better understanding of these quantities, we assigned numbers to the four bins based on the results presented above. We use a conservative measure where, for a given dataset, its visual encoding a is better than b if at least one class in a is notably ($\geq +2$) better than in b. Visually speaking, this measure means that at least one cell is encoded with a saturated blue. Using this conservative measure avoids any potential bias towards 2D.

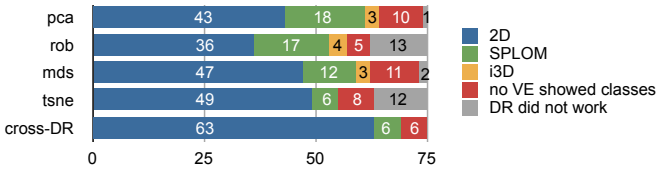


Fig. 7. Blue bars reflect how often 2D was good enough; green show how often SPLOM was notably *better* than 2D for at least for one class; yellow shows how often i3D was better than that; and red indicates that neither 2D, SPLOM, nor i3D revealed any class that scored higher than two. Gray indicates the 28 cases when R could not compute the dimension reduction. Top four rows = within-DR analyses, Bottom row = cross-DR. Note that for t-SNE and for cross-DR i3D never added any notable or substantial benefit for any class in our study.

Figure 7 shows counts for within-DR (top four rows) and cross-DR results (last row). We consider that 'no visual encoding (VE) showed classes' if all average class scores for all VEs were ≤ 2.

Following up on our main hypothesis, we found that our intuitions about 2D, SPLOM and i3D choices were largely reflected in within-DR situations, in particular, those of the linear PCA and robust PCA techniques. These analyses reflect hypothetical situations in which an analyst sticks with one DR technique and does not try any other ones. As we hypothesized, in those cases 2D is "often" good enough, SPLOM "sometimes" adds more visible class structure, i3D does not add on top of that except for some artificial datasets, and "sometimes" none of these visual encoding techniques reveals any class structure.

However, this impression changes if we look at non-linear techniques, and especially if we look at cross-DR exploration scenarios. For cross-DR, **2D "almost always"** is good enough, **SPLOM "only occasionally"** adds on top of that, and **i3D "never"** adds.

We note that our dataset collection is not perfectly representative. Some interesting cases could have been missed, and our quantitative results should be interpreted with this in mind. However, we made every endeavor to give i3D the best possible chance, by creating artificial datasets specifically designed for 3D. It is worth noting that i3D still very rarely helped even under these conditions.

5.5 SPLOM Usage

The usability assumption behind our data analysis is that SPLOMs are less costly in terms of usability than interactive 3D scatterplots, as we discuss in Section 2. However, this assumption only holds true if the dimensionality of SPLOM does not increase in an unbounded way.

To verify this assumption, both coders also tracked the dimensionality of the SPLOMs they chose, and the number of views that they used for the class verification tasks. We excluded SPLOMs that just replicated 2D (e.g., all classes visible with the first two principal components) to avoid a bias towards lower dimensional SPLOMs and fewer views. We found that the maximum SPLOM dimensionality was 7, with a mean of 4.1. The average number of used views was only 2.7, and the maximum was 11 views (only for one case). In 88% of the cases, 1-4 views revealed the necessary class structure. These findings indicate that when using SPLOMs for cluster verification with DR data, their size is limited and thus they are indeed usable. In particular, our usability cost assumptions are not violated. Further details can be found in the supplemental material.

6 Design Implications and Workflow Model

Based on our findings, we formulate a workflow model, and derive implications for making scatterplot choices in the DR exploration process. The workflow model, presented in Figure 8, provides sequences of concrete steps in the analysis process, as well as decision support for visual encoding choices.

At the beginning of the analysis process, an analyst *picks a DR technique* and *visualizes its output with a 2D Scatterplot*, the lowest cost visual encoding technique. The choice of DR will usually be informed by the analyst's previous experience with DR and/or the mathematical complexity of the DR approach. As the next step the analyst will

engage in the actual task of *visually verifying class separability*. The outcomes of this task depend on the analyst's prior expectations, and based on them will be either *"good enough" or "not good enough"*. If an analyst considers the visible results "good enough", s/he is already done assuming that what s/he sees is a *true positive*; that is, there is real class structure in the data.

However, the more common case is that classes are not nicely separable. When there is no visible class structure, an analyst is faced with the challenge of differentiating *true negatives* from *false negatives*. True negatives represent situations where there are really no separable classes, while false negatives appear when separable classes exist in higher dimensions but are absent in the lower-dimensional projection. This situation can be caused by artifacts arising from choices of how to reduce the data and/or how to visually encode it. In these situations, an analyst will often engage in an iterative process of investigating class separability from other DR and visual encoding perspectives [34], with the goal of building up more confidence in the real high-dimensional class structure of a dataset step by step. At the end of the process, an analyst will be able to make a more informed decision about true or false positives or negatives with respect to the visible class structure. If the set of tested DRs and visual encodings is large, the confidence that the result is a *true negative* will be higher. Our findings suggest several implications guiding such an iterative exploration process:

Change between DR. Use 2D. Our results indicate that *trying different DR techniques* with 2D scatterplots is a fruitful approach to investigate class separability. If a broad set of DR techniques are considered, other scatterplot techniques only occasionally add value beyond that exploration. Even if a set of classes is entirely mixed together in the embedding of one DR technique, another DR technique might reveal visually separable classes. Trying specific DR techniques with different parameterizations might also be a promising approach.

There is no one-and-only DR. While trying different DR techniques, it can be useful not only to change from linear to non-linear DR, but also in the other direction. There is no one-and-only DR technique that is superior to all others—at least not among the four we tested. t-SNE performed very well with untangling our artificial entangled datasets, but did not reveal certain structures in real world datasets, such as adjacent, stringy classes, which were revealed by the linear PCA techniques. We also noticed that while PCA, Glimmer MDS and t-SNE resulted in different and interesting projections, robust PCA rarely added additional insights on top of those. Note, however, that our focus was on visual encoding techniques; we do not claim a complete analysis of strength and weaknesses of the four different DR techniques.

SPLOM occasionally can help. There are two cases where *looking at SPLOMs* can help in the DR process: first, if the tested set of DR techniques is small. (Our within-DR analysis is a detailed discussion of the case where the set is exactly equal to one.) The second case is when the analyst has a strong hypothesis that existing classes are not visible in 2D plots of various DR techniques. Our study revealed one example, shown in Figure 4(b), for which only a PCA SPLOM revealed the separability of many classes. Our results in Section 5.2.1 suggest that SPLOMS are particularly helpful for linear DR techniques such as PCA or robust PCA. However, note that as the set of analyzed DR techniques grows larger, a SPLOM is less likely to reveal more structure than another DR technique in 2D. Also, trusting in SPLOMs alone is dangerous: Linear SPLOMs, for instance, cannot help in untangling high-dimensional entangled structures.

Do not use 3D for cluster verification with DR data. Especially when considered within the context of a cross-DR exploration process, i3D rarely, if ever, adds value on top of the less costly 2D and SPLOM VE techniques. On the contrary, it often hurts by hiding class structure that can be seen with other techniques, and by adding higher interaction costs. Even for the highly artificial datasets designed to showcase the potential of i3D, its benefits were questionable for within-DR analysis. With cross-DR analysis, t-SNE outperformed all i3Ds for those datasets, and we found no single example where i3D added value; i3D does therefore not appear in our workflow model.

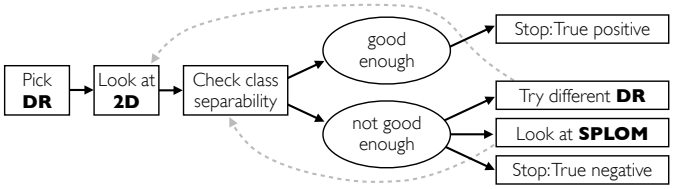


Fig. 8. A workflow model for guiding VE and DR technique choices in the DR exploration process.

7 Limitations and Future Work

As with all empirical work, there are limitations to our work stemming from the specific study design we picked and its implementation. While we intentionally based our data study on a broad set of 75 datasets and four DR techniques, we do not claim all-encompassing coverage of either. On the one hand, there might be datasets with interesting characteristics that do not follow general trends we observed here. On the other hand, there are many other DR techniques that we could have tested such as Linear Discriminate Analysis (LDA) [15] or Laplacian Eigenmaps [2]. However, based on our focus on visual encoding and the deliberate breadth of our dataset and DR collection, we do expect that these trends will hold true for many other situations.

We also specifically tested the task of cluster verification for DR data. We do not claim that our results generalize to other kinds of data; most importantly, they do not apply to intrinsically spatial non-reduced data. In terms of other DR tasks [34], we expect that some of our findings might generalize to cluster identification tasks, which are similar to cluster verification. In contrast, tasks related to dimensions such as naming newly derived dimensions are substantially different; exploring VEs for such tasks is an interesting topic for future work.

In terms of guiding the DR exploration process, previous work has focused on providing pre-specified workflows [20]. Our findings suggest that exploring different DR techniques with 2D scatterplots can be useful. How to provide specific guidance for such cross-DR choices and exploration is an interesting question for future work.

We also argue for more breadth in study approaches, including data studies. If we had followed our original user study plan, we would have only focused on visual encoding techniques, thus missing the crucial influence of dataset characteristics and DR techniques. Considering visual encoding techniques in a vacuum could lead to findings that are at best incomplete and at worst misleading. Ultimately, visualization is often only one step in a larger data analysis chain.

8 Conclusions

We presented a data study comparing class separability of DR data in 2D Scatterplots, interactive 3D Scatterplots and Scatterplot Matrices. The study was based on a broad set of 816 scatterplot samples and led to four implications and a workflow model that can be used to guide explorative high-dimensional data analysis.

As the most promising approach, our results suggest using 2D scatterplots to explore output of different DR algorithms. On top of that, SPLOMs can occasionally help reveal more class structure. In contrast, i3D rarely helps and often hurts, since it often reduces class separability and nearly always comes with higher interaction costs. Based on these findings, we recommend avoiding interactive 3D scatterplots for DR data, especially for cluster verification tasks. Instead, we advocate cross-DR exploration with 2D Scatterplots and the use of SPLOMs when the set of considered DR techniques is small.

Supplemental Material

All supplemental material is available at `http://www.cs.ubc.ca/labs/imager/tr/2013/ScatterplotEval/`.

Acknowledgments

The authors thank A. Tatu for participating in the data study as a coder, as well as M. Brehmer, J. Dawson, J. Ferstay, A. Ghane, S. Ingram, T. Möller, M. Phillips, and T. Torsney-Weir for their feedback.

REFERENCES

[1] A. Anand, L. Wilkinson, and T. N. Dang. Visual Pattern Discovery using Random Projections. In *Proc. IEEE Conf. Visual Analytics Science and Technology (VAST)*, pages 43–52, 2012.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems*, 14:585–591, 2001.

[3] E. Bertini and G. Santucci. Visual Quality Metrics. In *Proc. AVI Workshop on BEyond time and errors: novel evaluation methods for information visualization (BELIV)*. ACM, 2006.

[4] E. Bertini, A. Tatu, and D. A. Keim. Quality metrics in high-dimensional data visualization: an overview and systematization. *IEEE Trans. on Visualization and Computer Graphics (InfoVis)*, 17(12):2203–2212, 2011.

[5] I. Borg and P. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

[6] M. Chalmers. Using a landscape metaphor to represent a corpus of documents. In *Proc. European Conf. Spatial Information Theory (COSIT '93)*, pages 377–390. Springer LNCS 716, 1993.

[7] K. Charmaz. *Constructing grounded theory. A practical guide through qualitative analysis*. Sage Publications, Inc, 2006.

[8] A. Cockburn and B. McKenzie. An evaluation of cone trees. In *People and Computers XIV: Usability or Else. British Computer Society Conf. on Human Computer Interaction*, pages 425–436. Springer, 2000.

[9] A. Cockburn and B. McKenzie. 3D or not 3D?: evaluating the effect of the third dimension in a document management system. In *Proc. ACM Conf. Human Factors in Computing Systems (CHI)*, pages 434–441, 2001.

[10] A. Cockburn and B. McKenzie. Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments. In *Proc. ACM Conf. Human Factors in Computing Systems (CHI)*, pages 203–210, 2002.

[11] J. W. Creswell and V. L. P. Clark. *Designing and conducting mixed methods research*. Wiley Online Library, 2007.

[12] I. Doyle, M. Ratcliffe, A. Walding, E. V. Bon, M. Dymond, W. Tomlinson, D. Tilley, P. Shelton, and I. Dougall. Differential gene expression analysis in human monocyte-derived macrophages: impact of cigarette smoke on host defence. *Molecular immunology*, 47(5):1058, 2010.

[13] A. L. Duran, J. Yang, L. Wang, and L. W. Sumner. Metabolomics spectral formatting, alignment and conversion tools (msfacts). *Bioinformatics*, 19(17):2283–2293, 2003.

[14] S. Fabrikant. *Spatial metaphors for browsing large data archives*. PhD thesis, University of Colorado Boulder, 2000.

[15] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, 7(2):179–188, 1936.

[16] S. L. France and J. Carroll. Two-way multidimensional scaling: A review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 41(5):644–661, 2011.

[17] A. Frank and A. Asuncion. University of California Irvine (UCI) Machine Learning Repository, 2010.

[18] D. Furniss, A. Blandford, and P. Curzon. Confessions from a Grounded Theory PhD: Experiences and Lessons Learnt. In *ACM Trans. Computer-Human Interaction (ToCHI)*, pages 113–122, 2011.

[19] C. Holt and M. Bradford. Evaluating benchmarks of population status for Pacific salmon. *North American Journal of Fisheries Management*, 31(2):363–378, 2011.

[20] S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Möller. DimStiller: Workflows for dimensional analysis and reduction. In *Proc. IEEE Conf. Visual Analytics Science and Technology (VAST)*, 2010.

[21] S. Ingram, T. Munzner, and M. Olano. Glimmer: Multilevel MDS on the GPU. *IEEE Trans. Visualization and Computer Graphics (TVCG)*, 15(2):249–261, 2009.

[22] I. T. Jolliffe. *Principal Component Analysis, 2nd ed.* Springer, 2002.

[23] H. S. Kim, J. P. Schulze, A. C. Cone, G. E. Sosinsky, and M. E. Martone. Dimensionality reduction on multi-dimensional transfer functions for multi-channel volume data sets. *Information visualization*, 9(3):167–180, 2010.

[24] K. Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage, 2004.

[25] J. Kruskal and M. Wish. *Multidimensional Scaling*. Sage, 1978.

[26] H. Lam. A framework of interaction costs in information visualization. *IEEE Trans. Visualization and Computer Graphics (InfoVis)*, 14(6):1149–1156, 2008.

[27] J. M. Lewis, M. Ackerman, and V. de Sa. Human cluster evaluation and formal quality measures: A comparative study. In *Proc. 34th Conf. of the Cognitive Science Society (CogSci)*, pages 1870–1875, 2012.

[28] J. M. Lewis, L. van der Maaten, and V. de Sa. A behavioral investigation of dimensionality reduction. In *Proc. 34th Conf. of the Cognitive Science Society (CogSci)*, pages 671–676, 2012.

[29] J. Li, J.-B. Martens, and J. J. van Wijk. Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization*, 9(1):13–30, 2008.

[30] G. Newby. Empirical study of a 3D visualization for information retrieval tasks. *J. Intelligent Information Systems*, 18(1):31–53, 2002.

[31] R. Rensink and G. Baldridge. The perception of correlation in scatterplots. *Computer Graphics Forum (EuroVis)*, 29(3):1203–1210, 2010.

[32] T. Sando, M. Tory, and P. Irani. Effects of animation, user-controlled interactions, and multiple static views in understanding 3d structures. In *Proc. Symp. Applied Perception in Graphics and Visualization (APGV)*, pages 69–76, 2009.

[33] SAP. HANA, 2010. http://www.sap.com/hana/, last accessed 01/10.

[34] M. Sedlmair, M. Brehmer, S. Ingram, and T. Munzner. Dimensionality reduction in the wild: Gaps and guidance. Technical report, Dept. of Computer Science, University of British Columbia, 2012.

[35] M. Sedlmair, A. Tatu, M. Tory, and T. Munzner. A taxonomy of visual cluster separation factors. *Computer Graphics Forum (EuroVis)*, 31(3):1335–1344, 2012.

[36] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum (EuroVis)*, 28(3):831–838, 2009.

[37] M. St. John, M. B. Cowen, H. S. Smallman, and H. M. Oonk. The use of 2-D and 3-D displays for shape understanding versus relative position tasks. *Human Factors*, 43(1):79–98, 2001.

[38] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnor, and D. Keim. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST)*, pages 59–66, 2009.

[39] V. Todorov and P. Filzmoser. An object oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32(3):1–47, 2009.

[40] M. Tory, D. Sprague, F. Wu, W. Y. So, and T. Munzner. Spatialization design: comparing points and landscapes. *IEEE Trans. on Visualization and Computer Graphics (InfoVis)*, 13(6):1262–9, 2007.

[41] M. Tory, C. Swindells, and R. Dreezer. Comparing dot and landscape spatializations for visual memory differences. *IEEE Trans. on Visualization and Computer Graphics (InfoVis)*, 15(6):1033–9, 2009.

[42] University of Massachusetts. Statistical Data and Software Help, 2011. http://www.umass.edu/statdata/statdata/, last accessed 11/11.

[43] L. van der Maaten. Learning a parametric embedding by preserving local structure. *Proc. Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 5:384–391, 2009.

[44] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[45] L. Van der Maaten, E. Postma, and H. Van Den Herik. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10:1–41, 2009.

[46] J. J. van Wijk. Views on visualization. *IEEE Trans. Visualization and Computer Graphics (TVCG)*, 12(2):421–432, 2006.

[47] VisuMap Technologies Inc. VisuMap Data Repository, 2011. http://www.visumap.net/, last accessed 11/11.

[48] M. O. Ward. Xmdv data repository, 2011. http://davis.wpi.edu/xmdv/datasets.html, last accessed 11/11.

[49] C. Ware. Designing with a 2 1/2 D Attitude. *Information Design Journal*, 10(3):255–262, 2001.

[50] S. Westerman. Browsing a document collection represented in two- and three-dimensional virtual information space. *Intl. Journal Human Computer Studies (IJHCS)*, 62(6):713–736, 2005.

[51] S. Westerman and T. Cribbin. Mapping semantic information in virtual space: dimensions, variance and individual differences. *Intl. Journal Human Computer Studies (IJHCS)*, 53(5):765–787, 2000.

[52] C. Wickens and E. Merwin, D.H.and Lin. Implications of graphics enhancements for the visualization of scientific data: Dimensional integrality, stereopsis, motion, and mesh. *Human Factors*, 36(1):44–61, 1994.

[53] L. Wilkinson and A. Anand. Graph-theoretic scagnostics. *Proc. IEEE Symp. Information Visualization (InfoVis)*, pages 157–164, 2005.