A Psychophysical Investigation of Dimensionality Reduction

Joshua M. Lewis  
Department of Cognitive Science  
University of California, San Diego  
San Diego, CA 92093  
josh@cogsci.ucsd.edu

Laurens van der Maaten  
Department of Computer Science and Engineering  
University of California, San Diego  
San Diego, CA 92093  
lvdmaaten@gmail.com

Virginia R. de Sa  
Department of Cognitive Science  
University of California, San Diego  
San Diego, CA 92093  
desa@cogsci.ucsd.edu

Abstract

A cornucopia of dimensionality reduction techniques have emerged over the past decade, leaving data analysts with a wide variety of choices for reducing their data. Means of evaluating and comparing low-dimensional embeddings useful for visualization, however, are very limited. When proposing a new technique it is common to simply show rival embeddings side-by-side and let human judgment determine which embedding is superior. This study investigates whether such human embedding evaluations are reliable, i.e., whether humans tend to agree on the quality of an embedding. Our results reveal that, although experts are reasonably consistent in their evaluation of embeddings, novices generally disagree on the quality of an embedding. We discuss the impact of this result on the way dimensionality reduction researchers should present their results, and on applicability of dimensionality reduction outside of machine learning.

1 Introduction

There is an evaluative vacuum in the dimensionality reduction literature. In many other unsupervised machine learning fields, such as density modeling, evaluation may be performed by measuring likelihoods of held-out test data. Alternatively, in domains such as topic modeling, human computation [1] resources such as Amazon’s Mechanical Turk may be employed to exploit the fact that humans are phenoms in evaluating semantic structure [2]. Human evaluations have also been used to assess image segmentation techniques [3]. The field of dimensionality reduction, however, lacks a standard evaluation measure [4], and is not as obvious a target for human intuition. On what intuitive basis can we judge a 200 to 20-dimensional reduction to be good, or even a 20 to 2-dimensional reduction? Nevertheless, with no broadly agreed upon embedding quality measure, human judgment is often explicitly and implicitly solicited in the literature. The most common form of this solicitation consists of placing a scatterplot of the preferred embedding next to those of rival embeddings and inviting the reader to conclude that the preferred embedding is superior (e.g., [5]). If one is interested in applying a dimensionality reduction algorithm to visualize a dataset, is this a valid way to select a technique from the wide range of dimensionality techniques? To answer this question,

1Moreover, one should note that dimensionality reduction comprises only a small part of the “visualization zoo” [6].
we need to evaluate whether humans are good at evaluating embeddings. As there is no external authority we can appeal to, this is a daunting task. However, it is relatively easy to find out whether human data analysts are at least consistent in their evaluations, which is the main aim of this study. Consistency, across individuals and across a wide range of inputs, is a reasonable prerequisite for evaluation.

Motivated by this aim, we solicit embedding quality judgments from both novice and expert subjects in an effort to determine whether they are consistent in their ratings. The embeddings are derived from nine distinct dimensionality reduction algorithms. For the novice subjects, we manipulate dataset knowledge—half read a description and see samples from each dataset, and half do not. We hypothesize that providing dataset information will increase consistency, as it should if the evaluative process is principled.

## 2 Dimensionality reduction techniques

Dimensionality reduction techniques can be subdivided into several categories: linear or non-linear, convex or non-convex, parametric or non-parametric, etc. [7]. Whilst many new techniques have been proposed over the last decade, data analysts still often resort to a linear, convex, parametric techniques such as PCA to visualize their data. Non-linear, convex, non-parametric manifold learners such as Isomap [8], LLE [9], and MVU [10] are also frequently used for visualization purposes [11, 12, 13], even though it is unclear whether these techniques produce superior results [5].

As described in the introduction, one of the key problems of dimensionality reduction is that it lacks a widely agreed upon evaluation measure [4]. In fact, it is very unlikely that there will ever be such an evaluation measure, as it would imply the existence of a free lunch [14]: the “optimal” dimensionality reduction technique would be the technique that optimizes the measure. Also, there is a lot of debate within the field on what a good objective for dimensionality reduction is: for instance, a latent variable model approach to dimensionality reduction suggests one should focus on preserving global data structure [15], whereas a manifold learning viewpoint deems preservation of local data structure more important [9]. The lack of an evaluation measure and the ongoing debate within the field motivate the use of a more anthropocentric approach.

In our study, we investigated nine dimension reduction techniques: (1) PCA, (2) projection pursuit, (3) random projection, (4) Sammon mapping, (5) Isomap, (6) MVU, (7) LLE, (8) Laplacian Eigenmaps, and (9) t-SNE. The nine techniques were selected based on their importance in the literature. All nine techniques are briefly described below (we refer to points in the embedding as map points).

PCA and projection pursuit find a subspace of the original data space that has some desired characteristic. For PCA, this subspace is the one that maximizes the variance of the projected data. For projection pursuit [16], the subspace maximizes the non-Gaussianity of the projected data. Random projections are linear mappings that preserve pairwise distances in the data by exploiting the Johnson-Lindenstrauss lemma [17]. Sammon mapping constructs an embedding that minimizes a weighted sum of squared pairwise distance errors, in which the weights are inversely proportional to the original distances between the data points [18]. Isomap constructs an embedding by performing classical scaling on a geodesic distance matrix that is obtained by running a shortest-path algorithm on the nearest neighbor graph of the data [8]. MVU learns an embedding that maximizes data variance, while preserving the pairwise distances between each data point and its $k$ nearest neighbors, by solving a semidefinite program [10]. LLE constructs an embedding that preserves local data structure by minimizing a sum of squared errors between each map point and its reconstruction from its $k$ nearest neighbors in the original data [9]. Laplacian Eigenmaps try to minimize the squared Euclidean distances between each map point and the map points corresponding to its $k$ nearest neighbors in the original data, while enforcing a covariance constraint on the solution [19]. t-SNE embeds points by minimizing the divergence between two distributions over pairs of points, in which the probability of picking a pair of points decreases with their pairwise distance [5].

## 3 Experimental setup

Our experiment uses stimuli generated from the dimensionality reduction algorithms described above and attempts to determine whether humans are consistent in their embedding evaluations.
when the embeddings are fairly distinct. We divided subjects into (1) an expert group with detailed knowledge of dimensionality reduction and information on the underlying datasets presented in written and pictorial form, (2) a novice group with no knowledge of dimensionality reduction and no information on the visualized data, and (3) a group of similar novices but with the same information on the underlying datasets given to the experts. The dataset information we presented to groups 1 and 3 constituted an intuitive description of the data, as well as images revealing the underlying data (e.g., the Swiss roll, or handwritten character images).

Thirty one undergraduate human subjects were recruited for this study as the novice group, 16 female and 15 male, with an average age of 19.1 years. None of the novice subjects had any specific knowledge of dimensionality reduction techniques. Our expert group consisted of five subjects—three graduate students and two faculty members. The informed novice group had 15 subjects and the novice group 16. We generated two-dimensional point-light stimuli (see Figure 1 for a visualization of all the stimuli) by running the nine dimensionality reduction techniques discussed in Section 2 on seven different high-dimensional datasets, comprising a variety of domains. All techniques were run for a reasonable range of parameter settings, and we selected the embedding that was best in terms of the trustworthiness for presentation to the subjects.

Each trial consisted of nine different embeddings of the same dataset randomly arranged in a 3 × 3 grid. The datasets were shown as scatter plots with white points on a black background. For novice subjects, trials were organized into three blocks of seven, where each dataset appeared once per block and the order of the datasets within each block was randomized. Expert subjects were tested on one block. We instructed subjects to choose the two most useful displays and the one least useful display from the nine available on every trial. From the subject instructions: For each trial, please examine all the scatter plots and choose the two that you find most useful and the one that you find least useful. The task in the second part of this experiment will be much faster and easier if you choose useful scatter plots. Do the best you can to choose useful plots based on whatever criteria you deem appropriate. We intentionally left the task ambiguous so as not to bias subjects towards particular evaluation criteria and the fictitious “second part” of the experiment existed solely for increasing subject motivation.

We analyzed our novice subjects for internal consistency of their positive and negative ratings across blocks and found that even our least consistent subject was more consistent than expected by chance. Hence, we did not exclude any subjects due to internal inconsistency. To analyze consistency across subjects we use Fleiss’ κ and include neutral responses. Fleiss’ κ measures the deviation between observed agreement and the agreement attributable to chance given the relative frequency of ratings. Neutral ratings are twice as frequent as non-neutral, and positive ratings are twice as frequent as negative ratings, so the compensation for relative frequency in Fleiss’ κ makes it well-suited to our data.

We also measured the following six characteristics of our embedding stimuli: (1) variance, (2) skewness, (3) kurtosis, (4) clusteredness, (5) visual span, and (6) Gaussianity. The variance, skewness, and kurtosis were measured per dimension in a map that was scale-normalized such that \( y_i \in [0, 1]^d \) (preserving the aspect ratio of the maps), and averaged over the \( d \) dimensions of the map. We measured clusteredness by constructing \( k \)-nearest neighbor graphs in the map with \( k = 3, \ldots, 12 \), and measuring the maximum clustering coefficient of the resulting graphs \([22]\). The clustering coefficient computes the ratio of connections between the adjacent vertices of map point \( i \), averaged over all map points. The visual span of each map was measured by fitting a Parzen kernel density estimator with Gaussian kernels on the map (the variance \( \sigma \) of the Gaussians was optimized on a small validation set). Subsequently, we measure the ratio of the map surface that has a density of at least 10% of the maximum density of the density estimate. The Gaussianity of the maps was determined by averaging the results of Lilliefors tests \([23]\) performed on 5,000 one-dimensional random projections of the map. We analyze the relationships between novice informed ratings, novice uninformed ratings, expert ratings, and the six quantitative measures by calculating the Pearson’s correlation coefficient \( \rho \) between ratings and measures (after normalization within trial).

\[\text{The trustworthiness measures the ratio of } k \text{ nearest neighbors in the data that is still among the } k \text{ nearest neighbors in the maps.}\]

\[^2\text{For instance, defining a classification task would bias subjects to embeddings that show separated clusters.}\]

\[^3\text{Note that if the distribution of points in the embedding is Gaussian, the point distribution in each of the random projections has to be Gaussian as well.}\]
4 Results

The Fleiss’ kappa consistency measure $\kappa$ for experts is 0.39, for uninformed novices is $-0.28$, and for informed novices is $-0.40$. Fleiss’ kappa $\kappa$ ranges from $-1$ to $+1$, with $-1$ representing complete disagreement, $+1$ representing complete agreement and $0$ representing the amount of agreement expected by chance. Hence, the consistency measures reveal that, whereas experts tend to agree with each other on the quality of an embedding, novices strongly disagree with each other in their evaluations (they disagree more than if the evaluation was done randomly). Surprisingly, novices who received information on the underlying data disagree more strongly with each other than novices who had no information about the underlying data (counter to our hypothesis but interpretable, see below).

In Figure 2, we depict the raw ratings (averaged over each group) as a collection of Hinton diagrams. The diagrams reveal that informed novices exploit dataset knowledge in specific instances to differ significantly from uninformed novices. For example, the t-SNE embedding of the Swiss roll dataset (a relatively clustered embedding) is rated much more negatively by novices when they know that the data are gradual. Experts tend to rate t-SNE positively or negatively depending on the dataset and show a relatively consistent rating for Isomap. Informed novices consistently rate Sammon mapping and projection pursuit positively while generally rating manifold learners such as Isomap and LLE negatively. Uninformed novices are all over the map with the exception of (like all other subjects) rating MVU as not notable in either a positive or negative sense.

Table 1 shows correlation coefficients between the six embedding characteristics and the evaluations by the three human groups. We also present the correlation between the evaluations and the trustworthiness, which gives an indication of the quality of the embedding (in terms of local neighborhood preservation). Notably, expert ratings are the only ratings that correlate in the correct direction with trustworthiness. Another correlation that stands out is visual span: it appears to play a substantial role in informed novice ratings (they apparently surmise an embedding should fill up the available space), whereas it plays little role in expert ratings.

5 Discussion

Experts show themselves to be more consistent than chance and much more consistent than novices. This is reassuring, and indicates that the idea of having experts evaluating embeddings is not flawed to begin with. Novice subjects actually get less consistent with each other if they are informed. While this seems troubling at first, it actually makes some sense after closer consideration. Comparing the Hinton diagrams between novices and informed novices, one can plainly see that informed novices are converging on a smaller selection of techniques for both positive and negative ratings. The issue for the informed novices, however, is that they are not sure whether these stimuli should be rated as positive or negative. As a result, there is often energy for the same cell in both diagrams. Since the base rate of positive and negative ratings is low compared to the neutral ratings, the $\kappa$ consistency measure interprets this as substantial disagreement and thus the negative numbers. Importantly, the informed novice $\kappa$ is further from chance level than the novice $\kappa$.

Expert ratings are laudable in that they correlate in the correct direction with trustworthiness and have a context-dependent appreciation of clusteredness. Both novice groups rate clusteredness negatively regardless of context and are more influenced by elementary characteristics such as visual span. The substantial difference in strategy between novices and experts indicates that one can really benefit from training on the task of evaluating embeddings (unlike in common object recognition).

Overall, our results discourage free-form solicitation of human computation approaches à la [2] and [3] to the evaluation of dimensionality reduction techniques. Moreover, the novices’ lack of consistency lends worry to the prospect of naïve dimensionality reduction-based analysis. Most data analysts seeking to apply dimensionality reduction are not very familiar with the field. As a result, they are likely to be susceptible to the favorable visualizations presented in many dimensionality reduction papers. To ensure that dimensionality reduction techniques are applied wisely, authors should strive to explicate the situations that favor their algorithms (e.g., t-SNE is useful if the data is expected to have cluster structure, Isomap if the data lie on a convex manifold). In addition, data analysts should be encouraged to use sanity checks such as the trustworthiness measure in order to prevent them from basing analysis on interesting, but flawed, embeddings.
<table>
<thead>
<tr>
<th>Method</th>
<th>COIL</th>
<th>Faces</th>
<th>Helix</th>
<th>MNIST</th>
<th>ORL</th>
<th>Words</th>
<th>Swiss roll</th>
</tr>
</thead>
<tbody>
<tr>
<td>MVU</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isomap</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lapl. Eq.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sammon</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t-SNE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: All stimuli from experiment 1. Methods are in rows; datasets are in columns.
Figure 2: Human responses to the embeddings in experiment 1. Positive responses in the first row, negative in the second row. Experts (left), novices (center) and informed novices (right) by column.

Table 1: Correlation coefficients between human responses and dataset characteristics. Text in bold if $|\rho| \geq .30$
References


