A Behavioral Investigation of Dimensionality Reduction

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

Laurens van der Maaten
lvdmaaten@gmail.com
Pattern Recognition & Bio-informatics Lab
Delft University of Technology

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

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. We also investigate what types of embedding structures humans appreciate *a priori*. 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.

Keywords: dimensionality reduction; unsupervised machine learning; psychophysics

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 (Ahn, Maurer, McMillen, Abraham, & Blum, 2008) resources such as Amazon's Mechanical Turk may be employed to exploit the fact that humans are phenoms in evaluating semantic structure (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009). Human evaluations have also been used to assess image segmentation techniques (Martin, Fowlkes, Tal, & Malik, 2001). The field of dimensionality reduction, however, lacks a standard evaluation measure (Venna, Peltonen, Nybo, Aidos, & Kaski, 2010), and is not as obvious a target for human intuition. Two or three dimensional embeddings can be visualized as scatter plots, but on what intuitive basis can we judge a 200 to 2-dimensional reduction to be good? In addition, Gestalt effects or simple rotations may bias human evaluations of scatter plots. Nevertheless, with no broadly agreed upon embedding quality measure (though a few have been proposed, see below), human judgment is often explicitly and implicitly solicited in the literature. The most common form of this solicitation consists of placing a scatter plot of the preferred embedding next to those of rival embeddings and inviting the reader to conclude that the preferred embedding is superior (e.g., (Maaten & Hinton, 2008)). If one is interested in applying a dimensionality reduction algorithm to visualize a dataset, is this a valid way to select from the wide range of techniques?[1] To answer this question, 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 first aim of this study. Consistency, across individuals and across a wide range of inputs, is a reasonable prerequisite for evaluation.

Beyond investigating whether human data analysts are consistent when they evaluate embeddings, the second aim of this study is to investigate what humans are doing when they evaluate embeddings. Such information could be useful for determining whether humans are appropriate for an evaluation task with a known structure (e.g. if they naturally prefer embedding characteristics appropriate to the structure), or for developing techniques that are tailored towards producing results that humans will find helpful (e.g. algorithms that selectively emphasize informative data structure). We can to some extent infer human strategies from the algorithms humans prefer, but we can also investigate those strategies by correlating embedding characteristics with human evaluations.

Motivated by the two aims described above, we solicit embedding quality judgments from both novice and expert subjects in an effort to determine whether they are consistent in their ratings, and which embedding characteristics they find appealing. 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. The study consists of two experiments. The first presents subjects with a selection of embeddings derived from nine distinct dimensionality reduction algorithms; the second uses embeddings from a single algorithm with several different parameter settings for a more controlled comparison between "clustered" and "gradual" embeddings.

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. (Lee & Verleysen, 2007). Whilst many new techniques have been proposed over the last decade, data analysts still often resort to linear, convex, parametric techniques such as PCA to visualize their

[1] Moreover, one should note that dimensionality reduction comprises only a small part of the "visualization zoo" (Heer, Bostock, & Ogievetsky, 2010).

data. Non-linear, convex, non-parametric manifold learners such as Isomap (Tenenbaum, Silva, & Langford, 2000), LLE (Roweis & Saul, 2000), and MVU (Weinberger, Sha, Zhu, & Saul, 2007) are also frequently used for visualization purposes (Jain & Saul, 2004; Lewis, Hull, Weinberger, & Saul, 2008; Mahecha, Martínez, Lischeid, & Beck, 2007), even though it is unclear whether these techniques produce superior results (Maaten & Hinton, 2008).

As described in the introduction, one of the key problems of dimensionality reduction is that it lacks a widely agreed upon evaluation measure (Venna et al., 2010). 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* (Wolpert, 1996): 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 (Lawrence, 2005), whereas a manifold learning viewpoint deems preservation of *local* data structure more important (Roweis & Saul, 2000). 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 investigate nine dimensionality reduction techniques, selected for their importance in the literature: (1) PCA, (2) projection pursuit, (3) random projection, (4) Sammon mapping, (5) Isomap, (6) MVU, (7) LLE, (8) Laplacian Eigenmaps, and (9) t-SNE. 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 (Friedman & Tukey, 1974), the subspace maximizes the non-Gaussianity of the projected data. Random projections are linear mappings that mostly preserve pairwise distances in the data due to the Johnson-Lindenstrauss lemma (Bingham & Mannila, 2001). Sammon mapping constructs an embedding that minimizes a weighted sum of squared pairwise distance errors, where distances are weighted in inverse proportion to their magnitude (Sammon, 1969). 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 (Tenenbaum et al., 2000). 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 (Weinberger et al., 2007). 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 (Roweis & Saul, 2000). Laplacian Eigenmaps try to minimize the squared Euclidean distances between each points corresponding to its k nearest neighbors in the original data, while enforcing a covariance constraint on the solution (Belkin & Niyogi, 2002). 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 (Maaten & Hinton, 2008).

Experimental setup

We perform two experiments with our human subjects. The first experiment uses stimuli generated from the dimensionality reduction algorithms described above to determine whether humans are consistent in their evaluations when the embeddings are fairly distinct (the first aim of the study). The second experiment uses stimuli that are all generated by t-SNE, but with different parameter settings that affect how clustered the resulting embedding appears. This helps us determine what type of structure humans generally prefer in embeddings (the second aim of our study).

Experiment 1

In the first experiment, 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 of an intuitive description of the data, as well as images representing 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 expert subjects were drawn from the same institution and represent two different departments. Amongst the five expert subjects there are four distinct academic backgrounds at the graduate level. The informed novice group had 15 subjects and the uninformed 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 on seven different high-dimensional datasets, comprising a variety of domains. We ran each technique for a reasonable range of parameter settings, and we selected the embedding that was best in terms of the trustworthiness[2] (Venna & Kaski, 2006) for presentation to the subjects.

Each trial consisted of nine different embeddings of the same dataset arranged randomly per trial in a 3×3 grid. The datasets were shown as scatter plots with white points on a black background to reduce brightness-related eye fatigue. For novice subjects, trials were organized into three blocks

[2]The trustworthiness measures the ratio of k nearest neighbors in the data that is still among the k nearest neighbors in the maps.

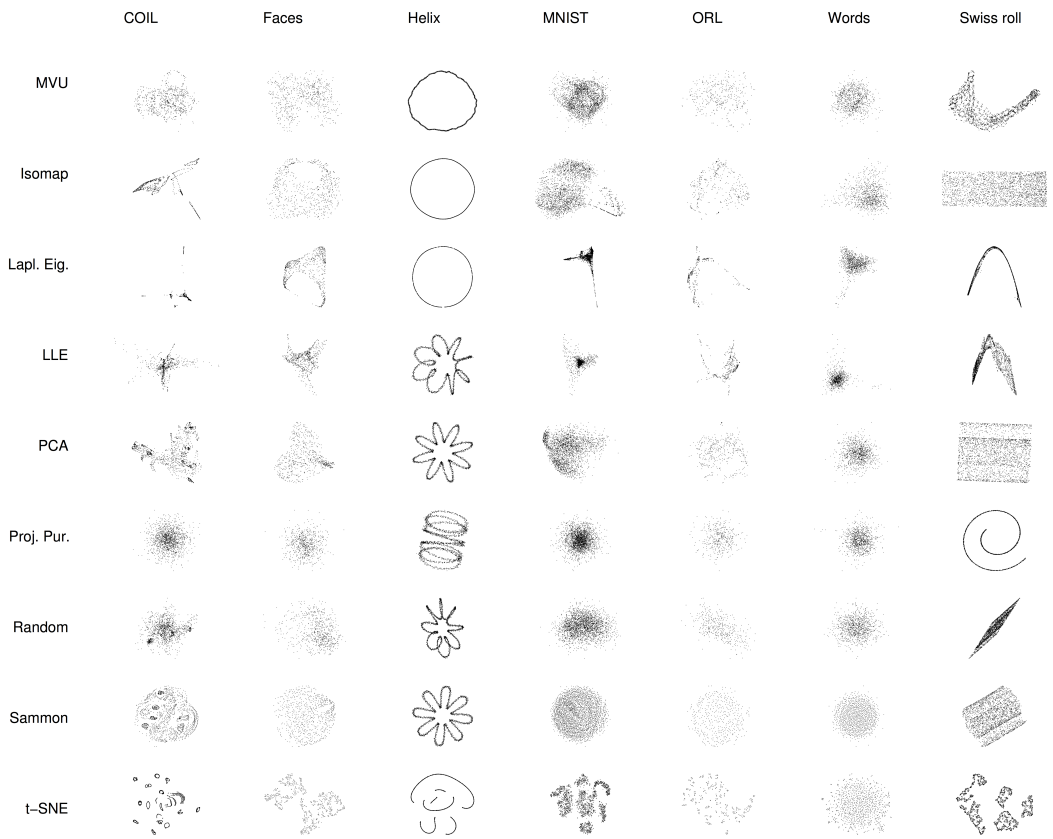

Figure 1: All stimuli from experiment 1. Methods are in rows; datasets are in columns.

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. After describing what a scatter plot is and emphasizing that each set of nine plots is a different perspective on the same dataset, we gave subjects the following 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[3], 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 (the first aim of this study) we use Fleiss' κ (Fleiss, 1971) and include neutral responses. Fleiss' κ measures the deviation between observed agreement and the agreement attributable to chance given the relative frequency of ratings, and normalizes for the number of raters. 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 $\mathbf{y}_i \in [0,1]^d$ (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 (Watts & Strogatz, 1998). 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 σ 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 max-

[3]For instance, defining a classification task would bias subjects to embeddings that show separated clusters.

imum density of the density estimate. The Gaussianity of the maps was determined by averaging the results of Lilliefors tests (Lilliefors, 1967) performed on 5,000 one-dimensional random projections of the map[4]. 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 ρ between ratings and measures (after normalization within trial).

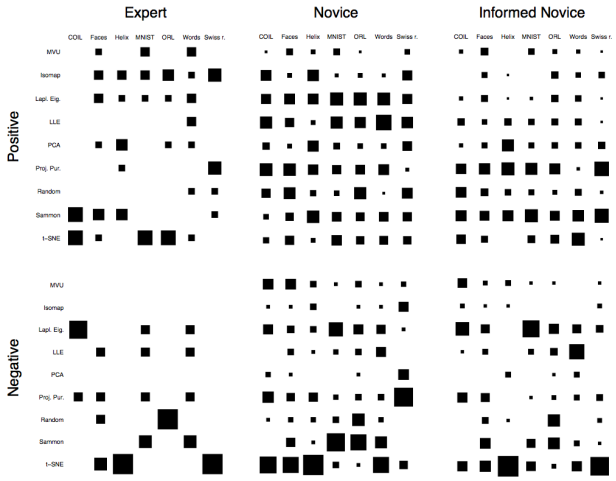

Figure 2: Human responses to the embeddings in experiment 1. Positive responses in the first row, negative in the second row. Experts (left), novices (center) and informed novices (right) by column. Algorithm and dataset ordering are the same as in Figure 1 within each block.

Experiment 2

The second experiment was run directly following experiment 1 on the same subject pool using the same methods, save stimulus design. In experiment 2, the nine stimuli in each trial are obtained by running t-SNE with nine different degrees of freedom v (viz. $v = 0.5, 0.8, 1.0, 1.2, 1.5, 2.0, 2.5, 3.0, 4.0$) on the seven datasets. The degrees of freedom in t-SNE determine to what extent the visualizations are "clustered" (Maaten, 2009). Sample stimuli are shown in Figure 3.

Results

Experiment 1

For the first experiment, the Fleiss' kappa consistency measure κ for experts is 0.39, for uninformed novices is -0.28, and for informed novices is -0.40. Fleiss' kappa κ ranges from -1 to $+1$, with -1 representing complete disagreement, $+1$ representing complete agreement and 0 representing the amount of agreement expected by chance. Though there is no standard significance test for Fleiss' kappa, based on the Landis and Koch scale (Landis & Koch, 1977), experts exhibited fair to moderate agreement, while both groups of novices exhibited poor agreement. 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).

[4]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.

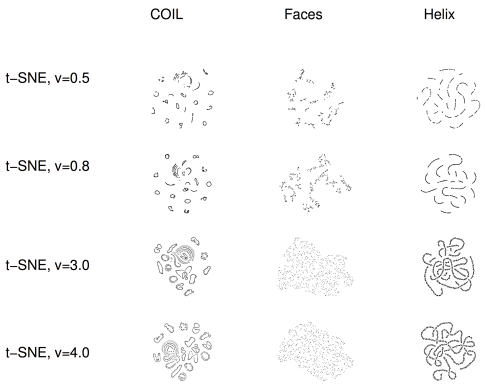

Figure 3: Sample stimuli from experiment 2. Parameter values are in rows; datasets are in columns.

Table 1: Correlation coefficients between human responses and dataset characteristics. Text in bold if $p < .0036$ after Bonferroni correction for $n = 14$ comparisons per subject group and $\alpha = .05$.

ρ	Lilliefors	Skewness	Kurtosis	Variance	Visual Span	Clusteredness	Trustworthiness
Ex. Pos.	.26	-.01	-.19	.34	.17	.22	**.41**
Ex. Neg.	-.08	.17	.19	-.14	-.17	.08	-.08
Nov. Pos.	.07	-.03	**.50**	-.18	-.29	.01	-.08
Nov. Neg.	.00	.17	-.10	.22	.10	-.03	.24
Inf. Pos.	-.02	-.16	-.10	-.11	**.44**	**-.45**	-.09
Inf. Neg.	.03	.31	.19	.10	-.19	.20	.19

In Figure 2, we depict the raw ratings (averaged over each group) as a collection of Hinton diagrams. In the figure, a large square indicates that a relatively large number of subjects gave a positive or negative evaluation of the embedding of the corresponding dataset, constructed by the corresponding technique. The top row of diagrams represent positive responses and the bottom negative, so if subjects are in disagreement about a stimulus, there will be a large box in its corresponding location in both rows. 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 rated 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 preservations). The significant correlations are in bold type, after a Bonferroni correction for multiple comparisons (14 comparisons per subject group, $\alpha = .05$). Notably, expert positive ratings are the only ratings that correlate significantly and 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.

Experiment 2

For the second experiment, the consistency measure κ for experts is 0.35, for uninformed novices is -0.32, and for informed novices is -0.26. The results of the second experiment thus reveal a similar trend: experts have fair agreement on the quality of embeddings, whereas novices give ratings have poor agreement. The ratings reveal that, whereas experts selectively rate more clustered or more continuous embeddings positively depending on the dataset, novices overwhelmingly rate the more clustered embeddings as negative. On the other hand, for positive ratings the novices tend to choose embeddings at either end of the spectrum while avoiding the moderate values of v. Moderate values of v might be avoided since subjects want to classify displays closest to the prototypical clustered or graded display (Rosch, 1975). Using the same set of correlations from Experiment 1 we find that experts ratings do not strongly correlate with any of the characteristics (including trustworthiness), but both groups of novices show a correlation between negative ratings and those stimuli with low kurtosis and high clusteredness.

Discussion

In both experiments, experts show themselves to be more consistent than chance and much more consistent than novices in either condition. This is reassuring, and indicates that the idea of having experts evaluating embeddings is not flawed to begin with. In the first experiment, 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 κ consistency measure interprets this as substantial disagreement and thus the negative numbers. Importantly, the informed novice κ is further from chance level than the novice κ. In Experiment 2, uninformed novices actually differ more from chance but the effect is about half the size, and experts remain consistent.

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 novices could really benefit from training on the task of evaluating embeddings (unlike evaluating results from topic modeling, image segmentation, or object recognition).

Conclusion

With respect to the first aim of our study (determining whether humans are consistent in rating embeddings), we conclude that dimensionality reduction experts are indeed reasonably consistent judges of embedding quality. This supports the practice of soliciting expert judgment for embedding evaluations, as nowadays is common in the literature on dimensionality reduction. However, we also conclude that novices are very inconsistent with one another in terms of their rating of an embedding, even when they have detailed information on the dataset the embedding is visualizing. In fact, novices even correlate negatively with a measure of embedding quality.

With respect to the second aim of our study (determining what types of structure humans appreciate in embeddings), we conclude that humans do not appear to have overwhelmingly strong *a priori* preferences, although novices appear to appreciate gradual embeddings that span a large portion of the space. Experts can adapt their preference for gradual vs. clustered depending on the dataset.

Overall, our results discourage free-form solicitation of human computation approaches á la (Chang et al., 2009) and (Martin et al., 2001) 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 re-

duction papers. To ensure that dimensionality reduction techniques are applied wisely, authors should strive to explicate the dataset characteristics 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). Authors could also cover usage scenarios appropriate to their algorithm (e.g., if a researcher is interested only in visualizing points that are most different then PCA would suffice and other techniques would be overkill), including guidelines for interpreting the relationship between the high and low dimensional spaces (sometimes this relationship will be very clear, as in PCA; other times, as in MVU, there is not a clear relationship). 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.

Acknowledgments

This work is funded by NSF Grant #SES-0963071, Divvy: Robust and Interactive Cluster Analysis (PI Virginia de Sa).

References

Ahn, L. von, Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*, *321*(5895), 1465-1468.

Belkin, M., & Niyogi, P. (2002). Laplacian Eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems* (Vol. 14, pp. 585–591). Cambridge, MA, USA: The MIT Press.

Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the* 7^{th} *acm sigkdd* (pp. 245–250).

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (Vol. 21).

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382.

Friedman, J., & Tukey, J. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, *23*, 881-890.

Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. *ACM Queue*, *8*(5).

Jain, V., & Saul, L. (2004). Exploratory analysis and visualization of speech and music by locally linear embedding. In *Proceedings of the international conference of speech, acoustics, and signal processing* (Vol. 3, pp. 984–987).

Landis, J. R., & Koch, G. G. (1977, March). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, *6*(Nov), 1783–1816.

Lee, J., & Verleysen, M. (2007). *Nonlinear dimensionality reduction*. New York, NY, USA: Springer.

Lewis, J., Hull, P. M., Weinberger, K., & Saul, L. (2008). Mapping uncharted waters: exploratory analysis, visualization, and clustering of oceanographic data. In *Proceedings of the international conference on machine learning and applications* (pp. 388–395).

Lilliefors, H. (1967). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*, 399–402.

Maaten, L. van der. (2009). Learning a parametric embedding by preserving local structure. In *Proceedings of the* 12^{th} *international conference on artificial intelligence and statistics* (pp. 384–391).

Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2431–2456.

Mahecha, M., Martínez, A., Lischeid, G., & Beck, E. (2007). Nonlinear dimensionality reduction: Alternative ordination approaches for extracting and visualizing biodiversity patterns in tropical montane forest vegetation data. *Ecological Informatics*, *2*, 138–149.

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001, July). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the* 8^{th} *international conference on computer vision* (Vol. 2, pp. 416–423).

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, *7*(4), 532 - 547.

Roweis, S., & Saul, L. (2000). Nonlinear dimensionality reduction by Locally Linear Embedding. *Science*, *290*(5500), 2323–2326.

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

Tenenbaum, J., Silva, V. de, & Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323.

Venna, J., & Kaski, S. (2006). Visualizing gene interaction graphs with local multidimensional scaling. In *Proceedings of the* 14^{th} *european symposium on artificial neural networks* (pp. 557–562).

Venna, J., Peltonen, J., Nybo, K., Aidos, H., & Kaski, S. (2010). Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, *11*(Feb), 451–490.

Watts, D., & Strogatz, S. (1998). Collective dynamics of small-world networks. *Nature*, *393*, 440–442.

Weinberger, K., Sha, F., Zhu, Q., & Saul, L. (2007). Graph Laplacian regularization for large-scale semidefinite programming. In *Advances in neural information processing systems* (Vol. 19).

Wolpert, D. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, *8*, 1341-1390.