PHASE TRANSITION IN TSNE : CRITICAL SAMPLE SIZE FOR LEARNING

Laura Bonde Holst (s173953), Pranjal Garg (s210242) and Virgile Ulrik I. Blanchet-Møhl (s163927)

Supervisor: Lars Kai Hansen
Advanced Machine Learning 02460

ABSTRACT

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a dimensionality reduction technique that is frequently used for visualization of high-dimensional data sets. Our analytic result suggests that learning signal structure using t-SNE is dependent on the signal strength - the sample size and dimensionality and that a critical number of observations is needed before any learning starts. A phase transition is observed in the learning curves - from almost perfect learning to no learning at all in both simulated data and real datasets.

***Index Terms*—** Visualization, dimension reduction, embedding algorithms, Phase transition

1. INTRODUCTION

Visual exploration is an essential tool of data analysis and machine learning, which allows for the development of intuitions and hypotheses for the data and the surrounding processes that generates them.[1] Here, techniques doing automatic analysis of the data in order to project a high dimensional object into a low dimensional space become really useful in the understanding of real world data sets.

A large number of such linear and non-linear embedding techniques have been proposed over the last decades.[2] Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Truncated Singular Value Decomposition (SVD) are examples of linear dimensionality reduction methods. Kernel PCA, t-distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP) and Trimap are examples of non-linear dimensionality reduction methods.

These different methods yield different results under various circumstances, and the desired outcome sometimes relies on variables unknown to the user. As explained by Ipsen and Hansen[3], the popular PCA show poor results under specific circumstances. Even further, PCA also display a phase transition in performance, which is influenced by eg. signal-to-noise-ratio (SNR).

This paper seeks to expand on the conclusions drawn by Ipsen and Hansen [3], by exploring other dimensionality reduction (DR) techniques, and the impact of various parameters on their performance. We aim to show how noise, data set size and class imbalance impacts the performance of these methods, as well as investigate whether the phase transition is also present in other DR techniques than PCA.

2. DIMENSION REDUCTION MODELS

For this paper four dimensionality reduction algorithms were analysed: PCA, t-SNE, UMAP and TriMap.

2.1. Principal Component Analysis

Principal Component Analysis (PCA) is a linear dimension reduction technique that finds the directions of maximum variances. Using this method to visualize the first initial components (describing most of the variance) is inherently a method that captures primarily the global structure of the data. [4]

Even though the method is widely used, it has flaws. As [3] describes, the performance of PCA is correlated to several notable factors, such as noise ratio and sample size, meaning that the technique fails for small sample noises or noisy data. Furthermore, PCA is depending on the data being uniformly scaled or standardized, as the technique seeks to maximizs variance.[4]

2.2. T-distributed Stochastic Neighbor Embedding

T-distributed Stochastic Neighbor Embedding (or t-SNE) is build upon Stochastic Neighbor Embedding (SNE), which computes the likelihood of two points being similar using a Gaussian distribution. [5]. However, SNE creates crowded clusters as data points are grouped too closely together at low sample sizes.

t-SNE successfully solved the crowding problem by substituting the Gaussian distribution, used in the low-dimensional space, with long-tailed t-distributions.[6] It uses gradient descent to find a low dimensional embedding in which a relative distance match the original high-dimensional space, with respect to the perplexity parameter which determines the standard deviation of the conditional distributions

used for the relative distance calculations in high-dimensional space.

This method primarily focuses on local structure, in that it puts focus on nearest neighbors in the relative distance calculations.

Even though t-SNE was solving earlier problems, as newer algorithms emerged, t-SNE has been proved to be both slower and less sensitive to global structure than the two following algorithms. [7]

2.3. Uniform Manifold Approximation and Projection

UMAP is constructed from a theoretical framework based in Riemannian geometry and algebraic topology. UMAP is competitive with t-SNE for visualization quality, and preserves more of the global structure with superior run time performance. It constructs a weighted graph of nearest neighbors with vertices being data points, and uses stochastic gradient descent to optimize a lower-dimensional graph to be as structurally similar to the high-dimensional one.[8]

2.4. TriMap

TriMap is a dimensionality reduction method based on triplet constraints, which focuses on preserving the global structure of the data in the embedding.[9] It performs better than the other methods such as t-SNE, LargeVis, and UMAP in preserving global spread. It creates sets of three observations predominantly from nearest neighbors, but with a fraction of the triplets also containing one or two randomly sampled points. Then it uses batch gradient descent, which preserves the ordering of distances of the triplets. Theoretically this method manages to include both local and global structure, but in practice it can be prone to struggle with local structure.

3. EXPERIMENTAL SETUP

Different experiments were run with the DR techniques using the mentioned data sets, using the following methodology

1. Scaling the data set to values within the interval of [0:1]
2. Add noise with $\sigma^2 = [0, 0.5, 1]$
3. Scaling the noisy data set to values within the interval of [0:1]
4. Find low dimensional embeddings for PCA, t-SNE, UMAP and Trimap
5. Find clusters in the low dimensional embedding using K-Means clustering for all DR methods
6. Calculate percentage of correctly classified data points for all the DR methods.
7. Repeat the above three steps by sampling a subset of 4 to 350 points to create learning curves

The notable variables and the experiments are described in details in the following sections.

3.1. Data

For the experiments we use two data sets: the handwritten digits "MNIST" data set and the "Fashion-MNIST" data set, which is a collection of clothing items. Both data sets are grey scale, and the dimensions 8x8 and 28x28 respectively. The small MNIST data set was chosen over the regular 28x28 to quickly perform tests. Both data sets were scaled to a [0:1] interval. Only two classes from each data set were used: for MNIST only 0 and 1 were considered, and for Fashion-MNIST only trousers and shoes were considered.

3.2. Noise signal

In order to test the robustness of the algorithms on noisy data sets, a normal distribution of noise with mean 0 and varying σ^2 was added to each image. After adding noise, the signal was scaled to the interval [0:1] again. A demonstration of the strength of the noise signal may be seen on Figure 1

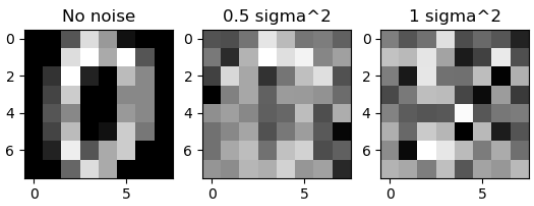


Fig. 1. Noise demonstration on a 0-digit from MNIST with varying noise at $\sigma^2 = [0, 0.5, 1]$ (no, medium and high noise)

3.3. Stratification factor

Another parameter we wanted to control was the distribution of each class in the test data set. In order to test with an equal distribution of classes, under-sampling the original data was done to achieve a 25/75% distribution of classes.

3.4. Code Setup

In the experiment each DR technique was run on a specific data set with the class imbalance stratification factor being toggled on or off. The experimental setup has an outer loop going through the selected noise levels and an inner loop running the methods on an increasing percentile of the data set, retaining the same train-test partition of the data for each method for the sake of comparison. Additionally, the experiment was repeated more times on lower sample sizes to minimise noise on the accuracy. You can find the code here.

3.5. Accuracy

In order to calculate accuracy, an independent K-Means model with two clusters was ran on the low dimensional embedding of the DR models. Half of the data was used as training set, and the accuracy was then calculated as the percentage of correct classifications in the test set based on classes identified from training set. To correctly identify the label of each cluster from K-Means, we introduce an additional bias by assuming that the clusters' true class must be that where the highest accuracy is obtained.

4. RESULTS

The learning curves obtained from our experiments showed some promising results. The results for the different experiment are as follows-

4.1. No Noise and Equal Classes

In the experiment where we used original data sets without adding any noise and keeping equal distribution of classes, we observed a phase transition between 30 and 40 data points for t-SNE where the algorithm jumped from random accuracy(50%) to approximately 95 percent. This sudden jump in learning was only observed with t-SNE whereas the other models started with a higher accuracy of 75% and improved gradually. We obtained similar learning curves for both MNIST and Fashion-MNIST where used the original data sets with equal distribution of both classes and no addition of extra noise.

The results for MNIST data set without any noise or stratification can be seen in Figure 2. The results for Fashion-MNIST which were similar to MNIST can be in Appendix Figure 5.

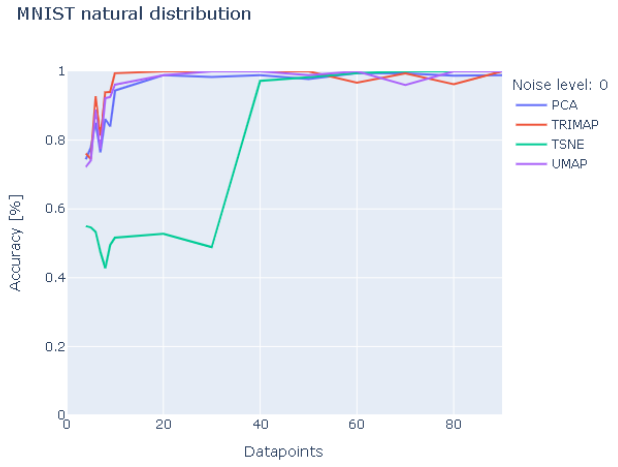
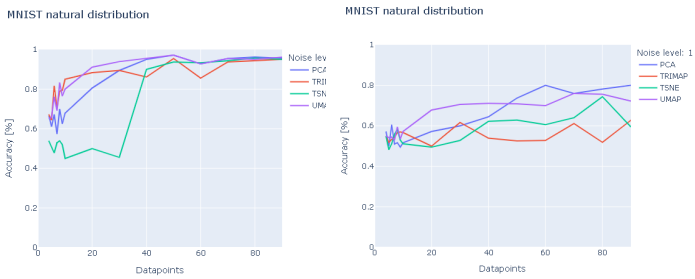


Fig. 2. MNIST with no noise

4.2. Addition of noise

When different intensities of noise was added to the datasets and the learning curves were generated for the different DR techniques, we observed a phase transition in TSNE again for moderate noise levels. But the phase transition stayed between 30 and 40 and didn't change. So it seems like the critical sample size needed for learning in TSNE was not affected by noise. Although the slope of the transition decreased so noise did reduce the rate of learning for TSNE. The other algorithm deteriorated too so it could be just simply mean that noise in signal leads to worser accuracy and might not reveal anything about the behavior of the phase transition.

The results with varying levels of noise is shown on Figure 3 for MNIST data set, but similar results for Fashion-MNIST can be seen in Appendix Figure 6.



(a) MNIST (Medium noise) (b) MNIST (High noise)

Fig. 3. MNIST dataset accuracy with several noise levels

4.3. Unequal classes in dataset

When we created an imbalance between the two classes being tested by making a new dataset with 75% data points from one class and 25% from the other, we saw a similar effect to addition of noise. The phase transition was observed between 30 and 40 data points for TSNE and not the other algorithms and the accuracy of all the algorithm was decreased. This was potentially due to the decrease in the signal strength of the weakest class in the mix. The stratified results can be seen on Figure 4 for the MNIST-data set, and in Appendix Figure 9 for the Fashion-MNIST data set.

5. DISCUSSION

From the results presented in Figure 2 and 3 it is evident that the strength of the noise signal has an influence on both the overall accuracy of the various DR techniques, but also on the phase transition of t-SNE. Furthermore, the class imbalance created, also has a significant effect on both accuracy and phase transition. This could potentially mean that a class

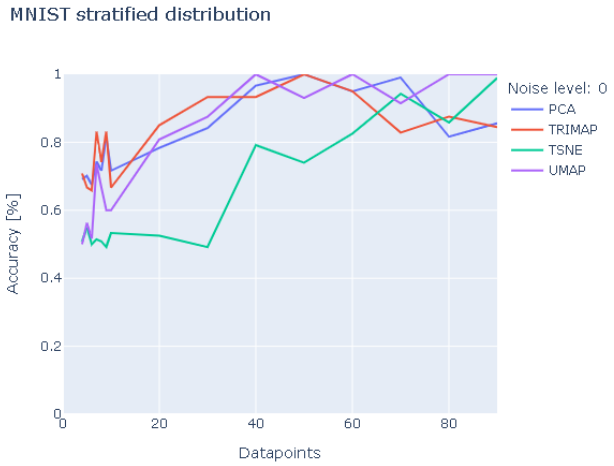


Fig. 4. Stratified MNIST dataset accuracy with no noise

signal needs to be of a certain strength with a minimum number of observations in a data set to be correctly identified and seperated by t-SNE.

These results are similar to the identical experiments with the Fashion-MNIST data set, where both strength of the noise signal as well as class imbalance seemed to skew both accuracy and phase transition. Fashion MNIST being a higher dimensional dataset provides better insight in the differences between the DR methods. The phase transition of t-SNE seems to get smaller and smaller with the addition of noise and class imbalance. but methods like UMAP still managing to get high levels of accuracy when enough samples arre provided.

The affect of the phase transition seem to be amplified by the dimensionality of the datset.

Furthermore, the universal drop in accuracy with stronger noise signal is an expected outcome, as the different classes loose their differentiation as the signal gets lost within the noise and the different DR algorithms can no longer differentiate between the different classes.

However, regardless of noise strength and class distribution one thing remains: the phase transition of t-SNE. Only with the strongest noise signal when the signal is completely nullified and learning is random is when the phase transition of t-SNE disappears, and it is seemingly a matter of merely tens of data points that makes the difference in performance in learning by t-SNE.

Since t-SNE tries to optimize relative distances between points using a t shaped distribution in low dimensional space, its possible that a minimum number of points are needed before the relative distances between different classes s large enough to be distinct in the low dimensional embbedding.

Interestingly, *only* t-SNE shows an obvious phase transition. The remaining models outperform t-SNE from the beginning, showing no transition. Even PCA, which [3] has proven to have a phase transition for both the MNIST and Fashion-MNIST, does not show any such behaviour under our conditions. However, this may be due to different experimental setups.

Even though the phase transition is present in multiple scenarios, it would take further experiments to conclude the characteristics and nature of it which we elaborate in the following section.

6. FUTURE WORK

We would like to do further test these DR techniques with simulated data with a very high number of dimensions. We would also like to implement mutual information(MI) to more accurately measure the accuracy of these DR tech- niques. also we would like to conduct this experiments with more and more DR techniques LargeVis, Kernel PCA , Lapla- cian Eigenfolds. and PAC-MAP, to see if t-SNE is the only one with this phase transition or not. We would further like to implement the metric of signal-to-noise ratio in order to better understand our results over different datasets. We would also like to test the algorithms with missing rate to handle situations with missing data in order to investigate the phase transition.

7. CONCLUSION

Overall, our hypothesis about a critical number of data points needed for learning to start from t-SNE simulation seems to match the two real world data sets, MNIST and Fashion-MNIST. The hypothesis predicts a phase transition in the learning curves which is seen in the case of t-SNE but not any other DR methods. Also, the phase transition doesn't seem to shift with increase in noise or imbalance in classes of the dataset. In our experiments, we observe a phase transition between 30 and 40 data points regardless of the dataset used and with/without moderate levels of noise and class imbalance. This potentially means that there is a beginning of learning at a minimum threshold number of data points, but as the results also show, the learning speed is affected by the noise in the signal and the strength of its weakest class. Some of the deviation could be a result of using KMeans clustering with half the set for classification training to ana- lyze the accuracy of the low dimensional embeddings. This could be improved in the future by using mutual information to analyze the quality of the embbedings. Nonetheless, a constant phase transition in t-SNE- from almost perfect learning to no learning at all should be a sign of caution for the practitioners of this dimensionality reduction technique.

8. REFERENCES

[1] Syed Mohd Ali, Noopur Gupta, Gopal Krishna Nayak, and Rakesh Kumar Lenka, "Big data visualization: Tools and challenges," in *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, 2016, pp. 656–660.

[2] Shaeela Ayesha, Muhammad Kashif Hanif, and Ramzan Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Information Fusion*, vol. 59, pp. 44–58, 2020.

[3] Niels Bruun Ipsen and Lars Kai Hansen, "Phase transition in pca with missing data: Reduced signal-to-noise ratio, not sample size!," 2019.

[4] Svante Wold, Kim Esbensen, and Paul Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37–52, 1987, Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

[5] Geoffrey Hinton and Sam Roweis, "Stochastic neighbor embedding," vol. 15, 06 2003.

[6] Laurens van der Maaten and Geoffrey Hinton, "Viualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 11 2008.

[7] Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik, "Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization," 12 2020.

[8] Leland McInnes, John Healy, and James Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 02 2018.

[9] Ehsan Amid and Manfred K. Warmuth, "Trimap: Large-scale dimensionality reduction using triplets," 10 2019.

9. APPENDIX

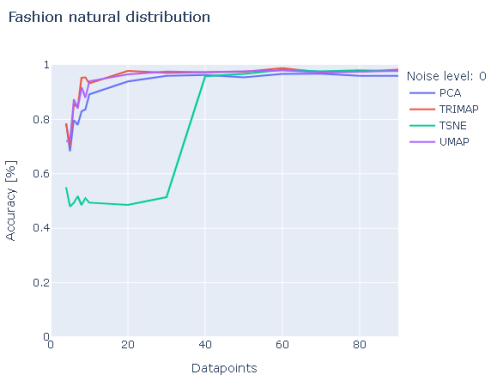
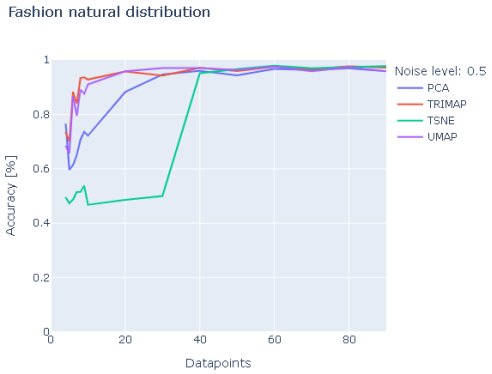
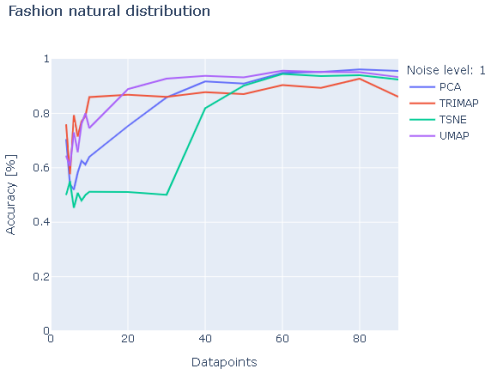


Fig. 5. Fashion-MNIST with no noise

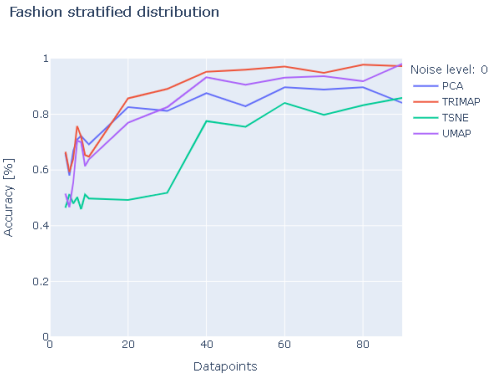


(a) Fashion-MNIST with 0.5 σ^2 noise level



(b) Fashion-MNIST with 1 σ^2 noise level

Fig. 6. Fashion-MNIST at different noise levels



Stratified Fashion-MNIST with no noise