Report 1 Follow-up

Bipasha Garg (2022111006)

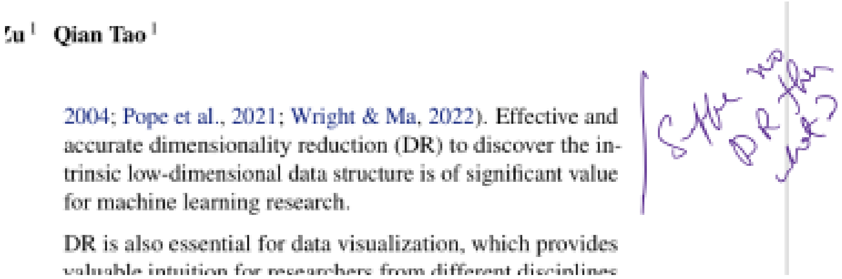
January 2025

1 Paper Comment Review

1.1 What is the manifold property?

- It is basically how data points are organized and related to each other within a lower-dimensional space that is embedded within a high-dimensional space.
- example: a 2d paper crumpled and put into 3d space. Paper is still 2d but put into (example a room) 3d space.
- Important because helps understand true dimensionality of the data and subsequently helps in accurate dimensionality reduction.

1.2 Suppose no DR then what?



:u[1] Qian Tao[1]

2004; Pope et al., 2021; Wright & Ma, 2022). Effective and accurate dimensionality reduction (DR) to discover the intrinsic low-dimensional data structure is of significant value for machine learning research.

DR is also essential for data visualization, which provides valuable intuition for researchers from different disciplines

Suppose no DR then what?

Figure 1: Snippet from paper

- Might lead to various issues like
  - Curse of dimensionality
  - Inaccurate Visualizations
  - Complexity and computational cost increases.

1.3 Is 15 to 20 is a good merge?

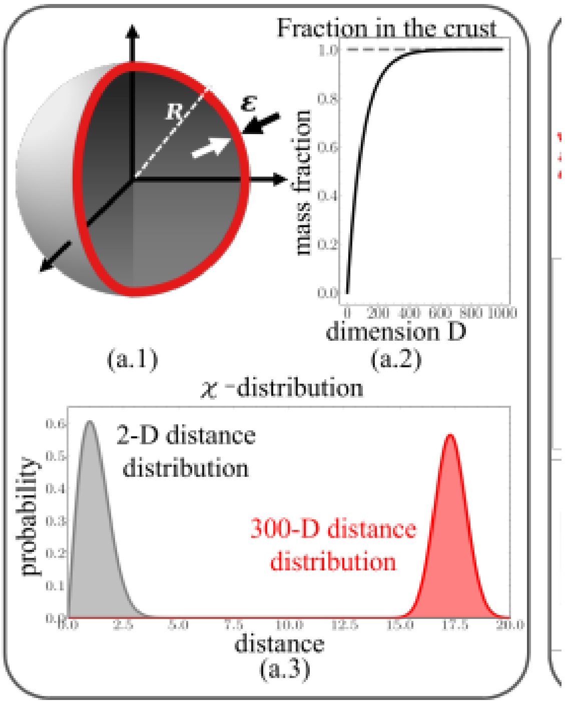


Figure 2: Enter Caption

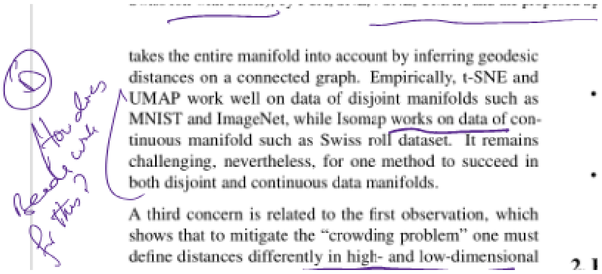
- Image Description: Distributions of the pairwise distances in a 2-dimensional (gray) and 300-dimensional (red) spaces
- **2D Space (Grey Line):** Most of the distances are concentrated towards the lower end of the distance range. This means that in a 2D space, data points have a relatively broad range of distance separations, but most pairs are not extremely far from each other.
- **300D Space (Red Line):** The key takeaways are:
  - **Concentration of Distances:** In the 300D space, the distribution is highly concentrated towards the higher end of the distance range (around 12.5-15). This means that most pairwise distances in the high-dimensional space are similar to each other and *very large*.
  - **Few Close Pairs:** The probability of finding very close pairs is much lower in the 300D space compared to the 2D space.

1. Between 15-20
   (a) The range between 15 and 20 in the red line represents the distances where the probability of observing that distance is low, but still greater than zero.
   (b) The problem that most points have very large and similar distances is not localized to 12-15, but rather persists in the range 15-20.
   (c) Even those smaller probabilities for large distances between 15-20 represent a large number of points. When mapping to lower dimensions, where the capacity to fit large distances is limited, the methods would have to squish these points together. Therefore shows the crowding problem.

(d) And the fact that the red line only extends out to 20 means there is not much diversity in the distances. The lack of diversity of the data structure is a key concern as this hinders the ability to differentiate between points.

(e) The smaller range of distances in 300D means that any representation in 2D or 3D will have to make points similar even if they are far away in 300D space. That is why other methods gives distorted global structure.

1.4 How does Beads work for this?



D

How does Beads work for this?

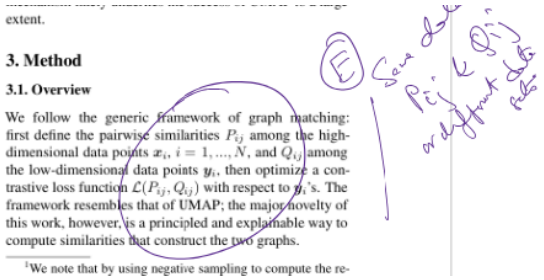
takes the entire manifold into account by inferring geodesic distances on a connected graph. Empirically, t-SNE and UMAP work well on data of disjoint manifolds such as MNIST and ImageNet, while Isomap works on data of continuous manifold such as Swiss roll dataset. It remains challenging, nevertheless, for one method to succeed in both disjoint and continuous data manifolds.

A third concern is related to the first observation, which shows that to mitigate the "crowding problem" one must define distances differently in high- and low-dimensional

Figure 3: Enter Caption

- t-SNE: Can be if we want to separate beads and show internal clustering within beads, but spatial positioning of beads can be misinterpreted.
- Isomap: Only use when beads form a continuous manifold, but will not do well in disjoint manifolds. But here another task adds up which is interpreting the manifolds in the dataset (beads).
- UMAP: Is a good general-purpose visualization tool for beads, providing a balance between cluster separation, spatial positioning and overall shape of beads. Since it works reasonably well for both disjoint and continuous manifolds.
- SpaceMAP: Might help preserve both local and global structure of beads, providing better visualization of the spatial location of beads, especially for datasets with complex hierarchies. **Will try to check and implement finding and working with manifolds in beads.**

1.5 Same data has Pij, Qij or different datasets?



extent.

3. Method

3.1. Overview

We follow the generic framework of graph matching: first define the pairwise similarities P_{ij} among the high-dimensional data points $\boldsymbol{x}_i$, $i = 1, ..., N$, and Q_{ij} among the low-dimensional data points $\boldsymbol{y}_i$, then optimize a contrastive loss function $\mathcal{L}(P_{ij}, Q_{ij})$ with respect to $\boldsymbol{y}_i$'s. The framework resembles that of UMAP; the major novelty of this work, however, is a principled and explainable way to compute similarities that construct the two graphs.

[1] We note that by using negative sampling to compute the re-

E

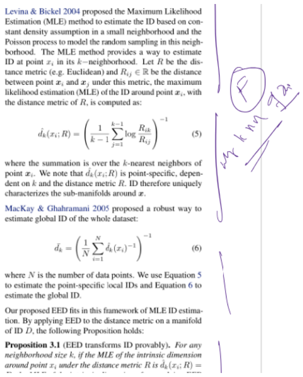
Same data Pij & Qij or different data sets

Figure 4: Enter Caption

- Pij and Qij are calculated using the same data, but in different spaces and with different metrics.
- They measure similarities between data points, but in different spaces (high-dimensional vs. low-dimensional).

- The goal is to adjust the locations of the low-dimensional points (the yi's) so that the similarity structure in the low-dimensional space as expressed by Qij comes as close as possible to the similarity structure in the high-dimensional space as expressed by Pij.
- the lower dimensionality points can be found out using dr methods like PCA.

1.6 Why are k no. of dims used? [section 3.2]



Levina & Bickel 2004 proposed the Maximum Likelihood Estimation (MLE) method to estimate the ID based on constant density assumption in a small neighborhood and the Poisson process to model the random sampling in this neighborhood. The MLE method provides a way to estimate ID at point x_i in its $k-$neighborhood. Let R be the distance metric (e.g. Euclidean) and $R_{ij} \in \mathbb{R}$ be the distance between point $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ under this metric, the maximum likelihood estimation (MLE) of the ID around point $\boldsymbol{x}_i$, with the distance metric of R, is computed as:

$$\hat{d}_k(x_i; R) = \left(\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{R_{ik}}{R_{ij}} \right)^{-1} \quad (5)$$

where the summation is over the k-nearest neighbors of point $\boldsymbol{x}_i$. We note that $\hat{d}_k(x_i; R)$ is point-specific, dependent on k and the distance metric R. ID therefore uniquely characterizes the sub-manifolds around $\boldsymbol{x}$.

MacKay & Ghahramani 2005 proposed a robust way to estimate global ID of the whole dataset:

$$\bar{d}_k = \left(\frac{1}{N} \sum_{i=1}^{N} \hat{d}_k(x_i)^{-1} \right)^{-1} \quad (6)$$

where N is the number of data points. We use Equation 5 to estimate the point-specific local IDs and Equation 6 to estimate the global ID.

Our proposed EED fits in this framework of MLE ID estimation. By applying EED to the distance metric on a manifold of ID D, the following Proposition holds:

Proposition 3.1 (EED transforms ID provably). *For any neighborhood size k, if the MLE of the intrinsic dimension around point x_i under the distance metric R is $\hat{d}_k(x_i; R) =$*

F

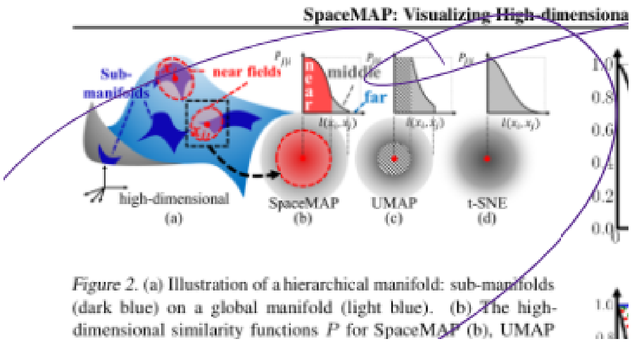
for k nn

for x_i

Figure 5: Enter Caption

- Because:
  - Manifolds can be assumed/approximated to be locally linear in the small neighborhood. Therefore, we consider k nearest neighbours.
  - We can also conclude that then to find the manifold (linear) relatively smaller k would work better since with increasing k the distanced would increase and the manifolds can be highly non-linear.
  - Trade-off with choosing K:
    - Very small k: Can lead to undersampling, the estimated local could be unstable since it will be based on a very few points.
    - Large k: We will lose locality, we will give us less accurate local approximation.

1.7 How is Figure 2 related to Beads?



SpaceMAP: Visualizing High-dimensiona

Figure 2. (a) Illustration of a hierarchical manifold: sub-manifolds (dark blue) on a global manifold (light blue). (b) The high-dimensional similarity functions P for SpaceMAP (b), UMAP

Figure 6: Enter Caption

FIGURE 2:

- a) Shows sub-manifolds (dark blue circles) which are a part of a larger manifold (light blue). Can be related as - cluster (larger manifold) and beads (sub manifold).
- It can be considered as a hierarchical division.
- Since SpaceMap shows view of the similarities, depending on how far away the data points are It may help in better clustering but I am at this point not sure that how it can help in the spatial arrangement of points.

1.8 For SpaceMAP, we divide the space with respect to each data point i into the near field ($S_{i,near}$), middle field ($S_{i,middle}$), and far field ($S_{i,far}$). How is it done?

- **Near Field ($S_{i,near}$):** Measures how similar a data point is to its very closest neighbors.
- **Middle Field ($S_{i,middle}$):** Measures how similar a data point is to neighbors that are further away (but not the furthest).
- **Far Field ($S_{i,far}$):** Measures how similar a data point is to neighbors that are extremely far away.
- Method:
  - Use k-NN.
  - Determine two sets: knear and kmiddle.
  - knear is the number of nearest neighbors used to define the near field.
  - kmiddle is the number of nearest neighbors *beyond* knear to define the middle field.
  - The knear closest data points to xi are defined to be in the near field. The next kmiddle closest data points (beyond the knear already assigned to the near field) to xi are considered in the middle field. All other data points not in the near or middle fields are considered in the far field.
  - As stated in the paper, they empirically set knear to 20 and kmiddle to 50. Eg: If knear = 20, then the 20 nearest neighbors to xi are assigned to Si,near.
  - Next we compute pairwise distances, and assign corresponding field.

1.9 How is the loss function of SpaceMAP related to Beads?

- The purpose of the lose function is to ensure that similar points in the high dimensional space are closer to each other in low dimensional space, and dissimilar points are further apart.
- This can be correlated in a way that if two points belong to the same "bead," their similarity in the high-dimensional space (Pij) will tend to be high (at least in the near and middle fields, if not the far field).
- By using SpaceMAP's hierarchical approach to computing the high-dimensional similarities Pij, the loss function can preserve the structure of hierarchical data.

1.10 Figure 4

Upper Panel (a.1 - a.6): This panel shows a synthetic "Swiss roll" dataset that is designed to test the ability of dimensionality reduction methods to handle a continuous manifold.

- SpaceMAP unfolds the Swiss roll to 2D, preserving the continuous structure and the nature of the data. Whereas other methods are not able to do so completely.

Lower Panel (b.1 - b.6): This panel shows a modified "Swiss roll" dataset with a hole. This tests the method's ability to handle both a continuous manifold and a discontinuous region (the hole).

- SpaceMAP unfolds the Swiss roll and accurately represents the hole in the data as a blank region whereas other methods are not able to do so completely.

1.11 Last question

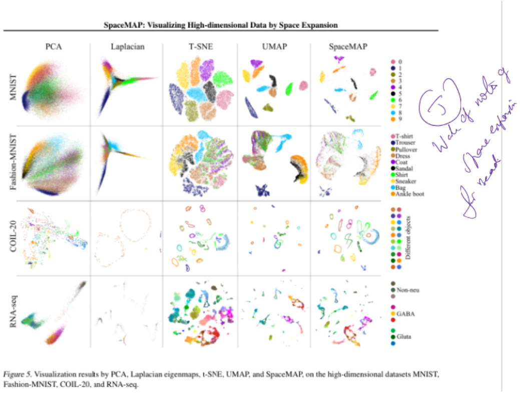

SpaceMAP: Visualizing High-dimensional Data by Space Expansion

Figure 5. Visualization results by PCA, Laplacian eigenmaps, t-SNE, UMAP, and SpaceMAP, on the high-dimensional datasets MNIST, Fashion-MNIST, COIL-20, and RNA-seq.

Ⓙ
Write of nots g
More explori
for read

Figure 7: Enter Caption

Did not really understand what the question is, Sir. Could you tell me that again?