Tree Preserving Embedding

Albert D. Shieh SHIEH@FAS.HARVARD.EDU
Tatsunori B. Hashimoto THASHIM@FAS.HARVARD.EDU
Edoardo M. Airoldi AIROLDI@FAS.HARVARD.EDU
Department of Statistics & FAS Center for Systems Biology, Harvard University, Cambridge, MA 02138

Abstract

Visualization techniques for complex data are a workhorse of modern scientific pursuits. The goal of visualization is to embed high-dimensional data in a low-dimensional space while preserving structure in the data relevant to exploratory data analysis such as clusters. However, existing visualization methods often either fail to separate clusters due to the crowding problem or can only separate clusters at a single resolution. Here, we develop a new approach to visualization, tree preserving embedding (TPE). Our approach uses the topological notion of connectedness to separate clusters at all resolutions. We provide a formal guarantee of cluster separation for our approach that holds for finite samples. Our approach requires no parameters and can handle general types of data, making it easy to use in practice.

1. Introduction

Visualization is an important first step in the analysis of high dimensional data. High-dimensional data often has low intrinsic dimensionality, making it possible to embed the data in a low-dimensional space while preserving much of its structure. However, it is rarely possible to preserve all types of structure in the embedding. Therefore, dimensionality reduction methods can only aim to preserve particular types of structure. Linear methods such as principal component analysis (PCA) and multidimensional scaling (MDS) (Mardia et al., 1979) preserve global distances, while nonlinear methods such as manifold learning (Tenenbaum et al., 2000; Roweis & Saul, 2000; Belkin & Niyogi, 2003) preserve local distances defined by kernels or neighborhood graphs. However, most dimensionality reduction methods fail to preserve clusters (Venna et al., 2010), which are often of greatest interest.

Clusters are difficult to preserve in embeddings due to the so-called crowding problem (van der Maaten & Hinton, 2008). When the intrinsic dimensionality of the data exceeds the embedding dimensionality, there is not enough space in the embedding to allow clusters to separate. Therefore, clusters are forced to collapse on top of each other in the embedding. As the embedding dimensionality increases, there is more space in the embedding for clusters to separate and the crowding problem disappears, making it possible to preserve clusters exactly (Roth et al., 2003). However, since the embedding dimensionality is at most two or three for visualization purposes, the crowding problem is prevalent in practice. When the clusters are known, they can be used to guide the embedding to avoid the crowding problem (Xing et al., 2002). However, the clusters are often difficult to find. In fact, the embedding is often used to help find the clusters in the first place. Therefore, it is important to solve the crowding problem without knowledge of the clusters.

Force-based methods such as stochastic neighbor embedding (SNE) (Hinton & Roweis, 2003), variants of SNE (Cook et al., 2007; van der Maaten & Hinton, 2008; Carreira-Perpinán, 2010; Venna et al., 2010), and local MDS (Chen & Buja, 2009), have been proposed to overcome the crowding problem. Force-based methods use attractive forces to pull together similar points and repulsive forces to push apart dissimilar points. SNE and its variants use forces based on kernels, while local MDS uses forces based on neighborhood graphs. Force-based methods have long been used in graph drawing to separate clusters (Di Battista et al., 1998; Kaufmann & Wagner, 2001). Although force-based methods are effective, it is difficult to balance the relative strength of attractive and repulsive forces. When repulsive forces are too weak, they will fail to separate clusters, but when repulsive forces are too strong,

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

they will artificially create clusters. Therefore, force-based methods are sensitive to intrinsic resolution parameters such as kernel bandwidths and neighborhood graph sizes that control the amount of separation between points in the embedding.

We introduce tree preserving embedding (TPE) in order to overcome the limitations of force-based methods. TPE aims to preserve both distances and clusters by preserving the single linkage (SL) dendrogram in the embedding. SL is a hierarchical clustering method that merges clusters with minimum nearest neighbor distance. The SL dendrogram is the associated tree with clusters as vertices and merge distances as vertex heights. TPE preserves the SL dendrogram in the sense that both the data and the embedding have the same SL dendrogram. Embeddings and dendrograms have long been used as representations for dissimilarities (Shepard, 1980). However, there is no guarantee that embeddings and dendrograms will be consistent when used separately. In particular, clusters found by dendrograms may not be found in embeddings due to the crowding problem. TPE combines embeddings and dendrograms in a common representation.

Preserving the SL dendrogram is a natural choice for several reasons. First, the SL dendrogram is the only dendrogram consistent with the topology of the minimum spanning tree (MST) (Gower & Ross, 1969; Zadeh & Ben-David, 2009). Preserving the topologies of neighborhood graphs has been shown to help overcome the crowding problem (Shaw & Jebara, 2009). However, while the topologies of neighborhood graphs such as the MST can only be preserved approximately in general (Eades, 1996), we show that the SL dendrogram can be preserved exactly. Second, the SL dendrogram represents both global and local structure due to its hierarchical nature. Preserving global structure allows TPE to separate clusters, while preserving local structure prevents TPE from artificially creating clusters. Finally, TPE can separate clusters even when the SL dendrogram cannot. Although SL is often criticized as a clustering method for finding poor clusters in practice (Hartigan, 1975), SL finds poor clusters due to the instability of cutting the SL dendrogram at a particular height (Stuetzle, 2003). Since TPE preserves the SL dendrogram at all heights, TPE is not sensitive to the instabilities of the SL dendrogram at any particular height.

We make cluster separation in TPE precise using the topological notion of connectedness. A natural and commonly used notion of a cluster is a set of points that are connected at a particular resolution. It is well known that the SL dendrogram finds clusters of connected points at different resolutions for different heights (Hartigan, 1975). We show that TPE preserves connectedness in the sense that points in the embedding are connected if and only if they are connected in the data. Preserving connectedness guarantees that clusters separated in the data remain separated in the embedding. Thus, TPE is guaranteed to separate clusters at all resolutions, rather than a single resolution.

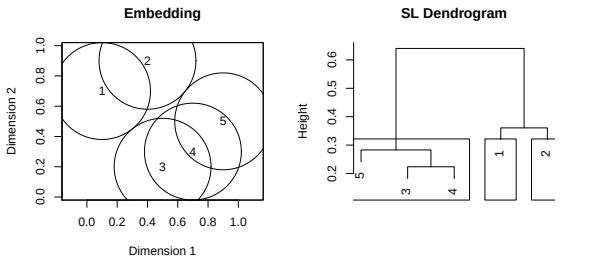

Figure 1. Relation between the SL dendrogram and connectedness in the embedding. Cutting the SL dendrogram at a height of $\varepsilon = 0.3$ produces three clusters of ε-connected points. Points 3 and 5 are ε-connected by point 4 because the ε-ball centered at point 4 contains points 3 and 5, while points 1 and 2 are not ε-connected to any other points because they are not contained by any ε-balls centered at other points.

Besides its theoretical basis, TPE has several attractive features in practice. First, in contrast to many nonlinear dimensionality reduction methods that require careful parameter selection to perform well, TPE requires no parameters. Second, TPE inherits the generality of MDS, which can handle data represented only in terms of dissimilarities rather than as vectors. Dissimilarities are found in many applications where vector representations are either poor or unavailable. Finally, since the merge order of the SL dendrogram is preserved under monotonic transformations of the dissimilarities, TPE is more sensitive to the rank order than the values of the dissimilarities. Therefore, TPE can handle non-metric dissimilarities.

2. Tree Preserving Embedding

In this section, we introduce TPE as a set of constraints in MDS that preserve the SL dendrogram. The constraints arise from a characterization of the SL dendrogram using a notion of connectedness. We introduce an algorithm similar to hierarchical clustering to implement TPE. In order to make the algorithm practical, we propose a variant based on a greedy approximation that maintains the constraints. Finally, we show that TPE preserves connectedness in a precise

sense that corresponds well with separating clusters.

2.1. Algorithm

TPE preserves the SL dendrogram in MDS. Given an $n \times n$ dissimilarity matrix D for a set of n objects $S = \{1, \dots, n\}$, MDS (Buja et al., 2008) finds an embedding $X = \{x_1, \dots, x_n\} \subseteq \mathbb{R}^p$ of the objects in p dimensions that minimizes the stress function

$$\sigma(X) = \sum_{x_i, x_j \in X} (d_{i,j}(X) - D_{i,j})^2,$$

the sum of squared errors between the Euclidean distances $d_{i,j}(X) = \|x_i - x_j\|$ in the embedding and the dissimilarities $D_{i,j}$. In general, stress is a natural objective that aims to preserve the underlying metric of the dissimilarities. However, since stress emphasizes approximating large dissimilarities well, minimizing stress without constraints on the embedding leads to the crowding problem. TPE preserves the SL dendrogram in order to overcome the crowding problem.

SL is a hierarchical clustering method that iteratively merges pairs of clusters $A, B \subseteq S$ with minimum nearest neighbor distance

$$d(A, B) = \min_{i \in A, j \in B} D_{i,j}$$

starting with singleton clusters and ending with the trivial cluster. The SL dendrogram is the associated binary tree of depth $n-1$ with singleton clusters as leaf vertices, the trivial cluster as the root vertex, merged clusters as internal vertices, and merge distances as vertex heights. There are many equivalent characterizations of the SL dendrogram (Hartigan, 1985). We use the following notion of connectedness to express the SL dendrogram as a set of constraints on pairs of objects.

Definition 1. *Objects $i, j \in S$ are* ε-connected *if there exists a path $\alpha_1 = i, \dots, \alpha_m = j \in S$ such that $D_{\alpha_l, \alpha_{l+1}} \leq \varepsilon$ for $l = 1, \dots, m-1$.*

Definition 2. *Points $x_i, x_j \in X$ are* ε-connected *if there exists a path $x_{\alpha_1} = x_i, \dots, x_{\alpha_m} = x_j \in X$ such that $d_{\alpha_l, \alpha_{l+1}}(X) \leq \varepsilon$ for $l = 1, \dots, m-1$.*

Intuitively, objects are connected if there exists a path with short hops between them. The SL dendrogram contains the paths with short hops between objects. Cutting the SL dendrogram at a height of ε produces a partition of the objects into clusters with heights at most ε. It is well known that cutting the SL dendrogram at a height of ε produces clusters of ε-connected objects (Hartigan, 1975). Therefore, objects are ε-connected if there exists a path of vertices with heights at most ε between their singleton clusters in the SL dendrogram. The relation between the SL dendrogram and connectedness in an embedding is illustrated in Figure 1.

Algorithm 1 Tree Preserving Embedding

Input: dissimilarity matrix D, embedding dimensionality p
Output: embedding X
// Initialize clusters
Set $S_1 = \{1\}, \dots, S_n = \{n\}$
Set $I_1 = \{1, \dots, n\}$
Set $X_1 = \{x_1 = 0\}, \dots, X_n = \{x_n = 0\}$
for $k = 1$ **to** $n - 1$ **do**
  // Find the next cluster merge
  Set $a_k, b_k = \arg\min_{a,b \in I_k} d(S_a, S_b)$ s.t. $a \neq b$
  Set $d_k = d(S_{a_k}, S_{b_k})$
  // Merge the clusters
  Set $S_{n+k} = S_{a_k} \cup S_{b_k}$
  Set $I_{k+1} = \{i \in I_k : i \neq a_k, b_k\} \cup \{n + k\}$
  // Find the ultrametric distances for the merged cluster
  for $i, j \in S_{n+k}$ **do**

$$\text{Set } U_{n+k}, i, j = \begin{cases} U_{a_k,i,j} & \text{if } i, j \in S_{a_k} \\ U_{b_k,i,j} & \text{if } i, j \in S_{b_k} \\ d_k & \text{otherwise} \end{cases}$$

  end for
  // Embed the merged cluster
  Set $X_{n+k} =$

$$\arg\min_{X = \{x_i \in \mathbb{R}^p : i \in S_{n+k}\}} \sigma(X)$$

  s.t. x_i, x_j are $U_{n+k,i,j}$-connected $\forall i, j \in S_{n+k}$
  $d_{i,j}(X) \geq U_{n+k,i,j}$ $\forall i, j \in S_{n+k}$

end for
Return X_{2n-1}

Cluster merges connect objects in the SL dendrogram. The ultrametric distance between objects is the distance at which they are merged into the same cluster in the SL dendrogram (Johnson, 1967). It is well known that the ultrametric distance in the SL dendrogram is equivalent to the sub-dominant ultrametric distance

$$U_{i,j} = \min_{\alpha_1 = i, \dots, \alpha_m = j} \max_{l=1}^{m-1} D_{\alpha_l, \alpha_{l+1}},$$

the maximum hop in a minimum path between the objects (Carlsson & Mémoli, 2010). The SL dendrogram can be characterized by two constraints on each pair of objects. First, each pair of objects must be connected by its ultrametric distance. Second, each pair of objects must not be connected by any distance less than

its ultrametric distance. The first constraint guarantees that clusters are merged at the same distances as the SL dendrogram, while the second constraint guarantees that clusters are merged in the same order as the SL dendrogram. TPE uses the constraints on pairs of objects as constraints on associated pairs of points in the embedding.

Algorithm 1 implements TPE. The algorithm proceeds similarly to hierarchical clustering. There are $n-1$ iterations, one for each depth of the SL dendrogram. At each iteration, a pair of clusters is merged and the merged cluster is embedded by minimizing stress subject to the connectedness constraints. At the last iteration, the trivial cluster is embedded and returned. The number of objects being embedded changes at each iteration depending on the size of the merged cluster. Since the embeddings at each iteration are independent, only the embedding at the last iteration is needed. However, earlier embeddings can be used to help initialize later embeddings in practice.

The connectedness constraints in the optimization problem may appear to be too rigid to allow TPE to find a low stress embedding since each pair of objects can be connected by an arbitrary sequence of objects. However, the connectedness constraints do not specify that the paths connecting pairs of points must be the same as the paths connecting associated pairs of objects. Preserving the paths that fulfill the connectedness constraints would preserve the MST, which is not possible in general (Eades, 1996). Moreover, the flexibility in choosing the paths to fulfill the connectedness constraints allows points to move freely in the embedding to lower stress. However, this flexibility comes at a cost. Due to the combinatorial nature of choosing paths, the optimization problem is intractable. Nevertheless, we can obtain a tractable approximation by restricting the types of paths that can be chosen.

2.2. Greedy Approximation

The optimization problem allows all points in a merged cluster to be rearranged in the embedding at each iteration. However, since each cluster being merged has already been embedded in prior iterations, it is wasteful to allow each cluster to be rearranged internally. The connectedness constraints within the clusters are already fulfilled in the prior embeddings. Since connectedness is preserved under rigid transformations, if we restrict the placement of the clusters to rigid transformations, then we only need to fulfill the connectedness constraints between the clusters. The paths that fulfill the connectedness constraints between the clusters must pass through their nearest neighbors. Therefore, placing the clusters exactly their merge distance apart fulfills the connectedness constraints between them. The remaining flexibility in placing the clusters can be used to minimize the stress between them.

In place of the optimization problem at each iteration k, the greedy approximation proceeds as follows. We find a rigid transformation that aligns the prior embeddings of the clusters while keeping them separated by their merge distance

$$T^* = \arg\min_{T \in E(p)} \sum_{x_i \in X_{a_k}, x_j \in X_{b_k}} (d_{i,j}(T) - D_{i,j})^2$$
$$\text{s.t.} \quad \min_{x_i \in X_{a_k}, x_j \in X_{b_k}} d_{i,j}(T) = d_k$$

where $d_{i,j}(T) = \|T(x_i) - x_j\|$ is the Euclidean distance in the embedding after alignment $E(p)$ is the set of rigid transformations in p dimensions. We align the clusters by setting $x_i = T^*(x_i)$ for all $x_i \in X_{a_k}$ and and return the merged cluster $X_{n+k} = X_{a_k} \cup X_{b_k}$.

The greedy approximation is reminiscent of Procrustes analysis (Gower & Dijksterhuis, 2004), which can be used to merge different embeddings (Quist & Yona, 2004). However, Procrustes analysis aligns embeddings without constraints, making it sensitive to the crowding problem. In contrast, the greedy approximation has a constraint that keeps the clusters separated in order to preserve the SL dendrogram. The constraint makes the greedy approximation more difficult to solve than Procrustes analysis. Nevertheless, the greedy approximation can be solved efficiently in practice using constrained optimization methods.

The efficiency of the greedy approximation comes at the cost of sensitivity to the merge order of the SL dendrogram. Since the greedy optimization cannot change the shapes of the clusters from prior embeddings, it cannot always align the clusters well. For small cluster merges, the prior embeddings will have little effect on the alignment. However, for large cluster merges, there may not be enough empty space in the prior embeddings to allow the clusters to be aligned well. Nevertheless, we found that a good alignment of the clusters can usually be found in practice.

The greedy approximation has a time complexity of $O(n^3)$ since there are $O(n)$ iterations, each of which requires minimizing the $O(n^2)$ stress between the clusters in order to find the alignment of the clusters. The greedy approximation has the same time complexity as dimensionality reduction methods based on a spectral decomposition of a dissimilarity matrix. While the cubic time complexity of the greedy approximation may be prohibitive for some applications, methods developed to improve the scalability of MDS such as land-

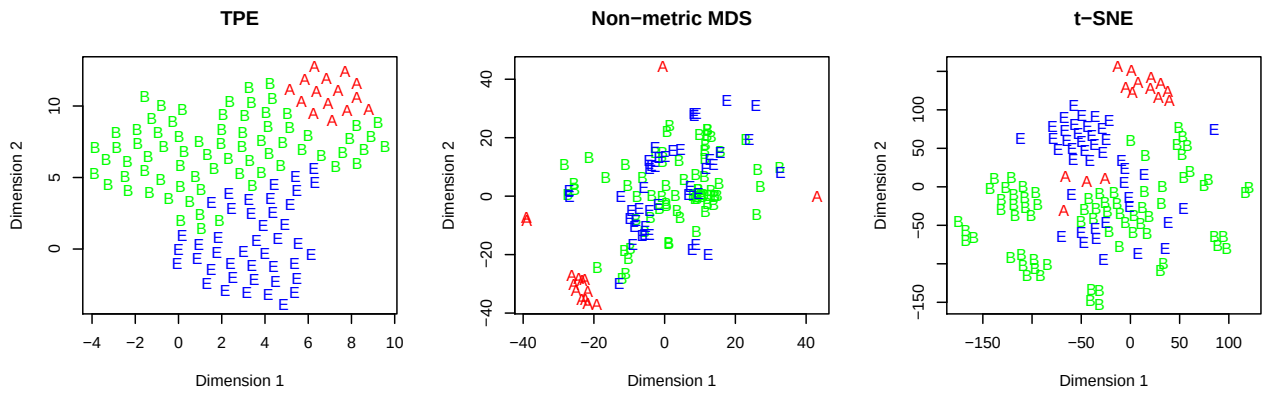

Figure 2. Embeddings of protein sequences by TPE, non-metric MDS, and t-SNE. Each point is a protein sequence labeled by the domain it belongs to where 'A' denotes Archaea, 'B' denotes Bacteria, and 'E' denotes Eukaryota.

mark points (de Silva & Tenenbaum, 2003) can also be applied to TPE in principle.

2.3. Connectedness

TPE preserves clusters at all resolutions rather than just a single resolution due to the hierarchical nature of the SL dendrogram. Since the clusters found by cutting the SL dendrogram at different heights can be characterized by connectedness at different resolutions, TPE preserves connectedness in the following sense.

Theorem 1. *For any $\varepsilon > 0$, points $x_i, x_j \in X$ are ε-connected if and only if objects $i, j \in S$ are ε-connected.*

Proof. First, we show that if objects $i, j \in S$ are ε-connected, then points $x_i, x_j \in X$ are ε-connected. We know that x_i, x_j are δ-connected by the distance $\delta = d_k$ at which x_i, x_j were merged into the same cluster at some iteration k. Let $\alpha_1, \ldots, \alpha_m \in S$ be the path that ε-connects i, j. Since $i \in S_{a_k}$ and $j \in S_{b_k}$, there exists some l such that $\alpha_l \in S_{a_k}$ and $\alpha_{l+1} \in S_{b_k}$. Since α_l and α_{l+1} are in different clusters, $\delta \leq D_{\alpha_l, \alpha_{l+1}} \leq \varepsilon$. Therefore, x_i, x_j are ε-connected.

Now, we show that if points $x_i, x_j \in X$ are ε-connected, then objects $i, j \in S$ are ε-connected. Without loss of generality, let x_i, x_j be points merged into the same cluster at iteration k such that x_i, x_j are ε-connected by the distance $\varepsilon = d_k$. Let $x_{\alpha_1}, \ldots, x_{\alpha_m} \in X$ be the path that ε-connects x_i, x_j. We know that there exists some l such that $d_{\alpha_l, \alpha_{l+1}}(X) = d_k$. Let $k_l \leq k$ be the minimum number of iterations such that $\alpha_1, \ldots, \alpha_l$ are in the same cluster. Then $\alpha_1, \ldots, \alpha_l$ are δ-connected by $\delta = d_{k_l}$. Since d_k is monotonically increasing in k, $\delta \leq \varepsilon$. Therefore, $\alpha_1, \ldots, \alpha_l$ are ε-connected. Similarly, $\alpha_{l+1}, \ldots, \alpha_m$ are ε-connected. Since α_l and α_{l+1} are in different clusters, $D_{\alpha_l, \alpha_{l+1}} \leq \varepsilon$. Therefore, i, j are ε-connected. □

TPE preserves connectedness in the sense that points are connected in the embedding if and only if they are connected in the data. Clusters at any resolution can be neither too close together nor too far apart in the embedding without violating connectedness at some resolution. Therefore, preserving connectedness guarantees that clusters separated in the data remain separated in the embedding. Preserving connectedness is of more than just theoretical interest. Since connectedness applies to finite samples, preserving connectedness provides a formal guarantee of cluster separation for TPE in practice. To our knowledge, TPE is the first method with a formal guarantee of this kind.

Preserving connectedness can be used to obtain distance bounds on points in the embedding. Points in the embedding can be no closer than the merge distance between their clusters and no further apart than the total merge distances between and within their clusters. Therefore, points that are merged into the same cluster earlier in the SL dendrogram will tend to be closer in the embedding.

Theorem 2. *We have*

$$\max_{l=1}^{m-1} D_{\alpha_l, \alpha_{l+1}} \leq d_{i,j}(X) \leq \sum_{l=1}^{m-1} D_{\alpha_l, \alpha_{l+1}}$$

where $\alpha_1 = i, \ldots, \alpha_m = j$ is the path between the objects in the MST.

Proof. Let x_i, x_j be points merged into the same cluster at iteration k. Since the merge distances of the

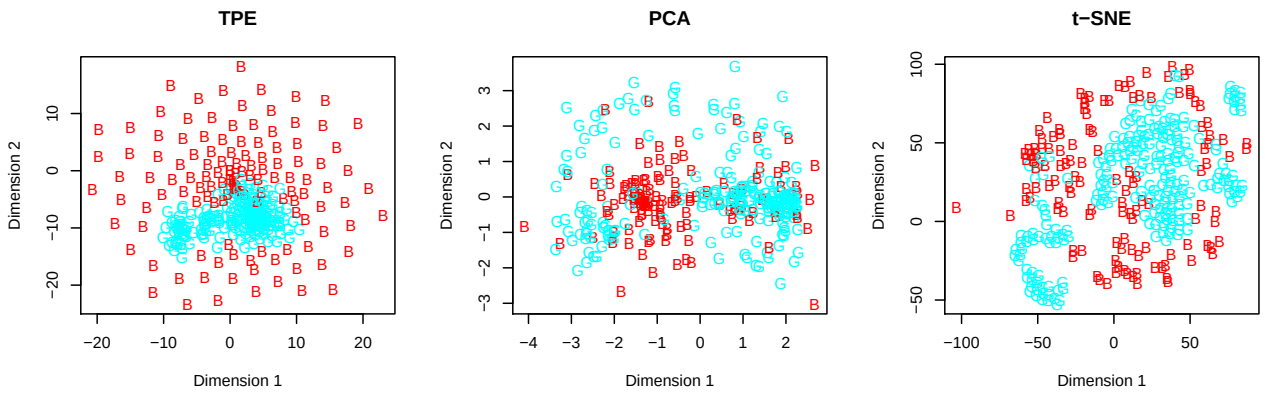

Figure 3. Embeddings of radar signals by TPE, PCA, and t-SNE. Each point is a radar signal labeled by its quality where 'G' denotes a good radar signal and 'B' denotes a bad radar signal.

SL dendrogram are equivalent to the edge lengths of the MST (Gower & Ross, 1969), the lower bound is equivalent to the merge distance between the clusters a_k, b_k and the upper bound is equivalent to the total merge distances between and within the clusters a_k, b_k. Therefore, the lower and upper bounds follow from the connectedness constraints. □

The distance bounds can be used to obtain an upper bound on the stress of the embedding. Therefore, preserving connectedness prevents the embedding from having arbitrarily high stress.

Corollary 1. *We have*

$$\sigma(X) \leq \sum_{x_i, x_j \in X} \max\{(D_{i,j} - U_{i,j})^2, (D_{i,j} - L_{i,j})^2\}.$$

3. Results

In this section, we demonstrate the applicability of TPE by analyzing simulated and real examples drawn from molecular biology, signal processing, and computer vision. Rather than being exhaustive, our goal is to highlight some of the properties of TPE through each example. We compare TPE to both traditional methods, PCA, non-metric MDS, and Isomap (Tenenbaum et al., 2000), and a force-based method, t-SNE (van der Maaten & Hinton, 2008), that recent studies have found separates clusters well (Venna et al., 2010). We demonstrate that TPE most robustly preserves different types of structure.

3.1. Protein Sequences

As a first example, we analyzed 124 protein sequences of 3-phosphoglycerate kinases (3-PGKs) belonging to the domains Archaea, Bacteria, and Eukaryota collected from public databases by Pollack et al. (2005). Since protein sequences cannot be represented as vectors, methods such as PCA that require such a representation cannot be used. Finding a good metric for protein sequences is a difficult and longstanding problem (Atchley et al., 2005). We used sequence alignment scores from the basic local alignment search tool (BLAST) (Althscul et al., 1990) as dissimilarities. Since BLAST scores can be highly non-metric (Roth et al., 2003), they are notoriously difficult to embed without collapsing points on top of each other. Figure 2 shows the embeddings by TPE, non-metric MDS, and t-SNE. TPE clearly separates all three domains, while non-metric MDS and t-SNE mix members of different domains together.

3.2. Radar Signals

As a second example, we analyzed 351 radar signals targeting free electrons in the ionosphere collected by (Sigillito et al., 1989). Each radar signal consisted of 34 integer and real measurements. We treated each radar signal as a 34 dimensional vector and used Euclidean distances as dissimilarities. Good radar signals were defined as those that returned evidence of free electrons in the ionosphere, while bad radar signals were defined as those that passed through the ionosphere and returned background noise. Therefore, good radar signals are highly similar, while bad radar signals can be highly dissimilar. Figure 3 shows the embeddings by TPE, PCA, and t-SNE. TPE clearly separates good and bad radar signals, while PCA and t-SNE mix them together.

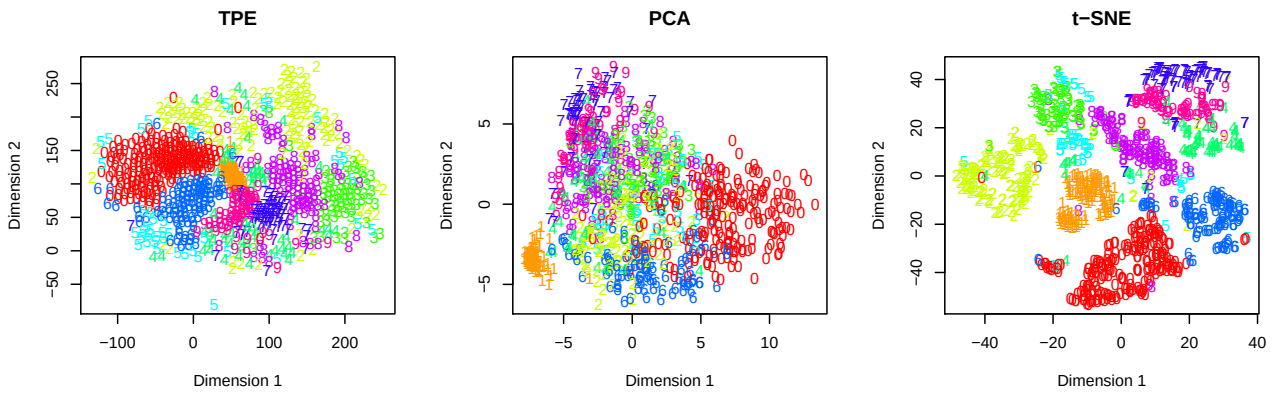

Figure 4. Embeddings of handwritten digits by TPE, PCA, and t-SNE. Each point is an image labeled by the digit it represents.

3.3. Handwritten Digits

As a third example, we analyzed 1000 images of handwritten digits collected by the United States Postal Service in (Hull, 1994). Each image was 16×16 pixels and greyscale color. We treated each image as a 256 dimensional vector and used Euclidean distances as dissimilarities. Since the intrinsic dimensionality of handwritten digits is thought to be much higher than two or three (Saul & Roweis, 2003), it is notoriously difficult to separate all ten digits in an embedding due to the crowding problem. Figure 4 shows the embeddings by TPE, PCA, and t-SNE. TPE and t-SNE separate all ten digits, while PCA only separates some.

3.4. Stability

In order to test the sensitivity of TPE to sampling variability, we generated 100 samples of the barbell, a popular example of a two-dimensional non-convex manifold (Saul & Roweis, 2003). For each sample, we generated 150 points as follows. We sampled 50 points each from multivariate normal distributions with means $(0, 0)$ and $(10, 10)$ and standard deviation 1 and 50 points with coordinates $(u, u) + z$ where u is sampled uniformly from the interval $[0, 10]$ and z is sampled from a multivariate normal distribution with mean $(0, 0)$ and standard deviation 0.05. The barbell is notoriously difficult to embed due to the presence of both clusters in the bells and continuity in the bar (Saul & Roweis, 2003), making it a good example to test for stability.

We compared TPE, Isomap, and t-SNE by using Procrustes analysis to align their embeddings to the exact embedding. The average sample variance of the coordinates of the points in the embeddings was 3.45 for the exact embedding, 2.72 for TPE, 3.40 for Isomap, and 984.54 for t-SNE. TPE and Isomap had comparable stability to the exact embedding, while t-SNE was two orders of magnitude less stable than TPE and Isomap.

4. Discussion

Revealing clusters is one of the main goals of visualization. However, most dimensionality reduction methods have difficulty preserving clusters due to the crowding problem. In three difficult examples, TPE was able to separate clusters of interest well compared to other dimensionality reduction methods. It is important to emphasize that the success of TPE is not a mere consequence of the ability of the SL dendrogram to separate clusters. In all of the examples, the clusters found by cutting the SL dendrogram were no better than random clusters in terms of accuracy with respect to the clusters of interest. TPE succeeds by preserving clusters at all resolutions rather than just a single resolution.

TPE is a promising approach to visualization because it has a formal guarantee of cluster separation, requires no parameters, and can handle general types of data. However, there are a few issues with TPE that limit its applicability. First, TPE has a cubic time complexity, which can be prohibitively slow for large data sets. Second, since TPE only provides an embedding rather than a mapping, it cannot be applied to out-of-sample data. Finally, although we have found that the greedy approximation works well in practice, better optimization methods may significantly improve the performance of TPE. We hope that these issues will be addressed by future research.

References

Althscul, S F, Gish, W, Miller, W, Myers, E W, and Lipman, D J. Basic local alignment search tool. *J Mol Biol*, 215:403–410, 1990.

Atchley, W R, Zhao, J, Fernandes, A D, and Drüke, T. Solving the protein sequence metric problem. *Proc Natl Acad Sci USA*, 102:6395–6400, 2005.

Belkin, M and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput*, 15:1373–1396, 2003.

Buja, A, Swayne, D F, Littman, M L, Dean, N, Hofmann, H, and Chen, L. Data visualization with multidimensional scaling. *J Comput Graph Stat*, 17:444–472, 2008.

Carlsson, G and Mémoli, F. Characterization, stability and convergence of hierarchical clustering methods. *J Mach Learn Res*, 11:1425–1470, 2010.

Carreira-Perpinán, M A. The elastic embedding algorithm for dimensionality reduction. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

Chen, L and Buja, A. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *J Am Stat Assoc*, 104:209–219, 2009.

Cook, J A, Sutskever, I, Mnih, A, and Hinton, G E. Visualizing similarity data with a mixture of maps. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, 2007.

de Silva, V and Tenenbaum, J B. Global versus local methods for nonlinear dimensionality reduction. In *Advances in Neural Information Processing Systems*, volume 15, 2003.

Di Battista, G, Eades, P, Tamassia, R, and Tollis, I G. *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall, New York, 1998.

Eades, P. The realization problem for euclidean minimum spanning trees is np-hard. *Algorithmica*, 16:60–82, 1996.

Gower, J C and Dijksterhuis, G B. *Procrustes Problems*. Oxford University Press, Oxford, 2004.

Gower, J C and Ross, G J S. Minimum spanning trees and single linkage cluster analysis. *Appl Stat*, 18:54–64, 1969.

Hartigan, J A. *Clustering Algorithms*. Wiley, New York, 1975.

Hartigan, J A. Statistical theory in clustering. *J Classif*, pp. 63–76, 1985.

Hinton, G E and Roweis, S T. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*, volume 15, 2003.

Hull, J J. A database for handwritten text recognition research. *IEEE T Pattern Anal*, 16:550–554, 1994.

Johnson, S C. Hierarchical clustering schemes. *Psychometrika*, 32:241–254, 1967.

Kaufmann, M and Wagner, D. *Drawing Graphs: Methods and Models*. Springer-Verlag, New York, 2001.

Mardia, K V, Kent, J T, and Bibby, J M. *Multivariate Analysis*. Academic Press, London, 1979.

Pollack, J D, Li, Q, and K, Pearl D. Taxonomic utility of a phylogenetic analysis of phosphoglycerate kinase proteins of archaea, bacteria, and eukaryota: Insights by bayesian analyses. *Mol Phylogenet Evol*, 35:420–430, 2005.

Quist, M and Yona, G. Distributional scaling: an algorithm for structure-preserving embedding of metric and non-metric spaces. *J Mach Learn Res*, 5:399–420, 2004.

Roth, V, Laub, J, Kawanabe, M, and Buhmann, J M. Optimal cluster preserving embedding of nonmetric proximity data. *IEEE T Pattern Anal*, 25:1540–1551, 2003.

Roweis, S T and Saul, L K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

Saul, L K and Roweis, S T. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *J Mach Learn Res*, 4:119–155, 2003.

Shaw, B and Jebara, T. Structure preserving embedding. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.

Shepard, R N. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210:390–398, 1980.

Sigillito, V G, Wing, S P, Hutton, L V, and Baker, K B. Classification of radar returns from the ionosphere using neural networks. *J Hopkins Apl Tech D*, 10:262–266, 1989.

Stuetzle, W. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *J Classif*, 20:25–47, 2003.

Tenenbaum, J B, de Silva, V, and Langford, J C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

van der Maaten, L and Hinton, G E. Visualizing data using t-sne. *J Mach Learn Res*, 9:2579–2605, 2008.

Venna, J, Peltonen, J, Nybo, K, Aidos, H, and Samuel, K. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *J Mach Learn Res*, 11:451–490, 2010.

Xing, E P, Ng, A Y, Jordan, M I, and Russell, S. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, volume 14, 2002.

Zadeh, R B and Ben-David, S. A uniqueness theorem for clustering. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence*, 2009.