An Interactive Tool for Analyzing High-Dimensional Clusterings

Justin Lin[*]and Julia Fukuyama[†]

Abstract

Technological advances have spurred an increase in data complexity and dimensionality. We are now in an era in which data sets containing thousands of features are commonplace. To digest and analyze such high-dimensional data, dimension reduction techniques have been developed and advanced along with computational power. Of these techniques, nonlinear methods are most commonly employed because of their ability to construct visually interpretable embeddings. Unlike linear methods, these methods non-uniformly stretch and shrink space to create a visual impression of the high-dimensional data. Since capturing high-dimensional structures in a significantly lower number of dimensions requires drastic manipulation of space, nonlinear dimension reduction methods are known to occasionally produce false structures, especially in noisy settings. In efforts to deal with this phenomenon, we developed an interactive tool that enables analysts to better understand and diagnose their dimension reduction results. It uses various analytical plots to provide a multi-faceted perspective on results to determine legitimacy. The tool is available via an R package named DRtool.

1 Introduction

The potency of nonlinear dimension reduction methods lies in their flexibility, allowing them to model complex data structures. That same flexibility, however, makes them difficult to use and interpret. Each method requires a slew of hyperparameters that need to be calibrated, and even when adequately calibrated, these methods require a trained eye to interpret. For example, the two most popular nonlinear dimension reduction methods, t-SNE and UMAP, are known to generate unintuitive results [6, 1]. The results often cluster, even when no clusters exist in the data, and cluster sizes/locations can be unreliable. We've developed an interactive tool that analysts may use to conduct a post-hoc analysis of their high-dimensional clustering. The tool uses the minimum spanning tree (MST) to model the global structure of clusters and provide an additional perspective on inter-cluster relationships. This allows analysts to extract more information from their dimension reduction results by making it easier to differentiate the signal and the noise.

In this paper, we describe the analytical plots provided by the tool (Section 2). We present a MST stability experiment, demonstrating the MST's ability to approximate high-dimensional structure (Section 3). And we walk through use of the tool on two separate data sets (Section 4).

2 Methods

2.1 The Minimum Spanning Tree

Graphs have been applied to many multivariate statistical problems. The authors of [15] introduced the minimal ascending path spanning tree as a way to test for multimodality. The Friedman-Rafsky test [9], along with its modern variations [3, 5, 4], use the MST to construct a multivariate two-sample test. Single-linkage clustering [10] and runt pruning [16] are both intimately related to the MST. In the context of dimension reduction, IsoMap [17] makes use of neighborhood graphs, [11] introduces the maximum information spanning tree, and [13] uses the MST. These methods, which fall under the category of manifold learning, use graphs to model high-dimensional data assumed to be drawn uniformly from a high-dimensional manifold. An accurate low-dimensional embedding can then be constructed from these graphs. It's apparent that graphs are useful for modeling high-dimensional data, especially when it

[*] Department of Mathematics, Indiana University
[†] Department of Statistics, Indiana University

Algorithm 1 Simplified Medoid Subtree

```
Require: MST $T = (V, E)$ with cluster medoids $m_1, \ldots, m_k \in V$
$T' = (V', E') \Leftarrow$ minimal subtree of $T$ containing all $m_i$
repeat
    Let $v \in V' \setminus \{m_1, \ldots, m_k\}$ with $deg(v) = 2$ and neighbors $a, b \in V'$. Let $d(v, a)$ and $d(v, b)$ be the weights of the edges incident to $v$.
    Replace $v$ and its two incident edges with an edge connecting $a$ and $b$ with weight $d(v, a) + d(v, b)$.
until $T'$ no longer contains non-medoid vertices with degree two.
output T'
```

comes to dimension reduction and cluster analysis. Our tool uses the MST to analyze the reliability of visualizations produced by nonlinear dimension reduction methods.

We've opted for the MST for a couple of key properties. Firstly, the MST and shortest paths along it are quick to compute. Secondly, the MST contains a unique path between any two vertices, providing a well-defined metric on the data. Lastly, it provides a good summary of the data's structure. It contains as a subgraph the nearest-neighbor graph, and any edge deletion in the MST partitions the vertices into two sets for which the deleted edge is the shortest distance between them [9].

2.1.1 MST Stability

The MST is meant to provide a robust estimation of the data's global structure, and more specifically, inter-cluster relationships. As such, it should be stable in the presence of noise and unaffected by local perturbations of the data. To demonstrate MST stability, we study the effect of random noise on the inter-cluster relationships explained by the MST.

To derive the inter-cluster relationships from the MST, we first take the medoid subtree, i.e. the minimal subtree containing the medoid of each cluster, then apply a simplification procedure (Algorithm 1). The algorithm collapses paths of non-medoid vertices into one edge of equal length. We refer to the output as the simplified medoid subtree. It encodes global the inter-cluster relationships within the data.

2.1.2 Robinson-Foulds Metric

To compare simplified medoid subtrees, we used the Robinson-Foulds metric [14]. The R-F metric was originally introduced to quantify the dissimilarity of phylogenetic trees, but the algorithm generalizes to arbitrary weighted trees. It looks at partitions of each tree created by removing individual edges, then counts the number of partitions present in one tree but not in the other. We modified the algorithm (Algorithm 2) to specifically measure the dissimilarity in medoid vertices.

Algorithm 2 Robinson-Foulds Distance

```
Require: Trees $T_1 = (V_1, E_1)$ and $T_2 = (V_2, E_2)$ with medoids $m_1, \ldots, m_k \in V_1$ and $n_1, \ldots, n_k \in V_2$
$P_1 \Leftarrow \{\}$
for $e \in E_1$ do
    $G \Leftarrow (V_1, E_1 \setminus \{e\})$ with connected components $G_1$ and $G_2$
    $M_1 \Leftarrow \{m_1, \ldots, m_k\} \cap V(G_1)$
    $M_2 \Leftarrow \{m_1, \ldots, m_k\} \cap V(G_2)$
    $P_1 \Leftarrow \textsc{Add}(P_1, \{M_1, M_2\})$
$P_2 \Leftarrow \{\}$
for $e \in E_2$ do
    $G \Leftarrow (V_2, E_2 \setminus \{e\})$ with connected components $G_1$ and $G_2$
    $M_1 \Leftarrow \{n_1, \ldots, n_k\} \cap V(G_1)$
    $M_2 \Leftarrow \{n_1, \ldots, n_k\} \cap V(G_2)$
    $P_2 \Leftarrow \textsc{Add}(P_2, \{M_1, M_2\})$
output $\frac{|P_1 \Delta P_2|}{2|P_1 \cap P_2|}$
```

2.2 The Tool

The main objective is to analyze and leverage the structural data embedded in the MST. For example, paths between clusters are used to study inter-cluster relationships in the context of the underlying manifold from which the data is drawn.

To start, the user must provide a data matrix, a low-dimensional embedding, and a clustering. From there, the MST is calculated and various analytical plots are provided. The primary plot is the low-dimensional embedding colored according to the provided clustering. There is an option to overlay the medoid MST to understand the global structure of the clusters.

The remaining plots require the user to select two groups of interest, which can be done interactively in one of two ways. One way is to select two endpoints. The MST path is calculated and projected onto the low-dimensional embedding. The two groups are then the classes each endpoint belongs to. The second way is to select custom groups. The user may interact with the low-dimensional embedding by drawing boundaries for each group. The projected path then connects the medoid of each group. Once the groups and path are specified, the user is provided additional plots used to investigate the relationship between the two selected groups of points.

2.3 Path Projection Plot

To better understand the path of interest, a local projection method is applied to visualize the path and nearby points in two dimensions. The goal of the projection is to "unwind" the path, so that it can be used to study the relationship between the two selected groups. We apply Principal Component Analysis followed by regularized Canonical Correlation Analysis in a method we've dubbed the PCA – rCCA method.

2.3.1 The PCA – rCCA Method

Let $P \in \mathbb{R}^{k \times p}$ be the matrix of high-dimensional path points with endpoints $p_1, p_k \in \mathbb{R}^p$. Let $X \in \mathbb{R}^{n \times p}$ be the matrix of points of interest, i.e. the points belonging to the selected groups and the points along the path.

The concept is to use Canonical Correlation Analysis to determine the two-dimensional linear projection that best unwinds the path. Given two matrices, CCA iteratively calculates linear combinations of the matrix variates for each matrix, known as canonical variate pairs, that maximize covariance. These pairs are chosen to be orthogonal, so they give rise to a projection subspace. To unwind P, we use CCA to compare P against a polynomial design matrix P_d,

$$P_d = \begin{bmatrix} 1 & 1^2 & \cdots & 1^d \\ 2 & 2^2 & \cdots & 2^d \\ \vdots & \vdots & & \vdots \\ n & n^2 & \cdots & n^d \end{bmatrix}.$$

The first two canonical variate pairs are used to construct a two-dimensional projection that maximizes the covariance between the projections of P and P_d. This process generates a two-dimensional subspace on which we can project all of X. Regularization is required to avoid singularity because p is often much greater than k. The regularization constant for P is chosen using cross-validation. No regularization constant is needed for P_d. See [18] for details.

One issue with this method is the projected path often travels along the outskirts of the plot. This is due to the near-orthogonality of high-dimensional data [8]. Because the non-path points are often nearly orthogonal with the projection subspace, they are overly shrunk in the projection. The path points are less affected because the projection subspace is selected to retain the path's shape. While this phenomenon doesn't discredit the entire plot, it leads to misrepresentation of the path's location relative to the rest of the points.

To alleviate this issue, we apply PCA on the entirety of X to prior to applying rCCA. Removing extraneous dimensions containing mostly noise limits the confusion of excess noise for independence. When rCCA is applied post-PCA, the projected path's relative position to the rest of the points is more credible.

Algorithm 3 Simplify Subtree

```
Require: Tree $T = (V, E)$, group one vertices $V_1 \subset V$, and group two vertices $V_2 \subset V$
  $T' = (V', E') \Leftarrow$ minimal subtree of $T$ containing $V_1 \cup V_2$
  repeat
      Let $v \in V' \setminus (V_1 \cup V_2)$ with $deg(v) = 2$ and neighbors $a, b \in V'$. Let $d(v, a)$ and $d(v, b)$ be the weights of the edges incident to $v$.
      Replace $v$ and its two incident edges with an edge connecting $a$ and $b$ with weight $d(v, a) + d(v, b)$.
  until $T'$ no longer contains non-group vertices with degree two.
  repeat
      Let $v_1, v_2 \in V' \setminus (V_1 \cup V_2)$ be adjacent.
      Collapse the edge connecting $v_1$ and $v_2$. The combined vertex is adjacent to all neighbors of $v_1$ and $v_2$.
  until $T'$ no longer adjacent non-group vertices.
```

2.3.2 Calibrating Hyperparameters

The user is responsible for calibrating the dimensionality of the PCA step and the degree d of the reference polynomial design matrix. To pick a number of dimensions, the user is recommended to start with a moderately large number, relative to the dimensionality of the original data. The proportion of variance retained in the selected number of dimensions is conveniently displayed in the upper righthand corner of the plot. A larger number of dimensions retains more information, but may misrepresent the location of the path relative to the rest of the points, while a smaller number of dimensions may diminish some of the variation in the data. As such, the user is encourages to try different numbers of dimensions. To calibrate d, it is recommended to start with $d = 2$ then increment d until the the shape of the path stabilizes.

The user is also given the option to overlay a kernel density estimate. In order to do so, the bandwidth must be calibrated. The recommend procedure is to begin with a large bandwidth that estimates one mode, then gradually decrease the bandwidth until two modes appear. If the two modes correspond with the two groups of interest, and more modes do not immediately appear when continuing to decrease the bandwidth, then a bimodal distribution is a reasonable way to describe the data.

2.4 MST Test

Another perspective on the relationship between the two selected groups can be gained by studying their connectivity in the MST. In particular, the number of MST crossings between the two groups serves as a measure of separation. A large number of crossings indicates lesser separation, while a small number of crossings indicates more separation. This idea motivates a hypothesis test.

2.4.1 The Test Statistic

The test statistic is the number of crossings between the two groups, which is counted according to the following procedure. The minimal subtree containing both groups is isolated. Because the two groups may not be adjacent in the MST, this subtree may points belonging to other clusters as well. To extract the structural relationship between the two groups of interest, the subtree must be simplified. The simplification process collapses paths between the two groups of interest into edges that can be counted (Algorithm 3).

To count the number of crossings, the number of edges between the two groups in the simplified subtree are counted. It is also possible for a point of non-interest to act as a mediator along a path between the two groups of interest. To account for this scenario, for each point of non-interest adjacent to both groups, we also count its maximal degree to both groups.

2.4.2 The Null Distribution

The null distribution should represent the number of crossings in the case when both groups belong to the same cluster. Therefore, the composite null hypothesis must account for all distributions the cluster may have been drawn from. This family of distributions is constructed as follows. For simplicity, assume the one-dimensional case. We aim to construct a family $\mathcal{F}$ of unimodal distributions on $[-1, 1]$ from which the cluster may have been drawn from. Let n_1 and n_2 be the sample sizes of each group, respectively. The family contains all distribution functions $f : [-1, 1] \to \mathbb{R}_{\geq 0}$ satisfying the following criteria:

- there exists $c \in [-1, 1]$ such that f is increasing on $[-1, c]$ and decreasing on $[c, 1]$,
- $\int_{-1}^{0} f(x)\, dx = \frac{n_1}{n_1+n_2}$, and
- $\int_{0}^{1} f(x)\, dx = \frac{n_2}{n_1+n_2}$.

The hyperplane is located at $x = 0$, and the number of crossings is assumed to scale with $f(0)$, the density at the hyperplane. To ensure the test has the correct size, the null distribution that maximizes the probability of rejection must be used. That way, the null would also be rejected under any other member of the composite null hypothesis, ensuring the probability of Type I error does not exceed the pre-determined significance level under any $f \in \mathcal{F}$.

It is not difficult to see

$$f'(x) = \begin{cases} \frac{n_1}{n_1+n_2} & -1 \leq x < 0 \\ \min(\frac{n_1}{n_1+n_2}, \frac{n_2}{n_1+n_2}) & x = 0 \\ \frac{n_2}{n_1+n_2} & 0 < x \leq 1 \end{cases}$$

minimizes $f'(0)$. f' is uniform on each side of the splitting hyperplane, and $f'(0)$ is equal to the lesser of the densities of each group.

Generalizing to higher dimension, we consider a uniform distribution on a hyperrectangle. The density of the distribution is determined by the lesser of the densities of each individual group. The density of each cluster is approximated by the number of samples divided by the product of singular values after truncating extraneous dimensions,

$$D_i = \frac{n_i}{\prod_j \sigma_j^i} \text{ for } i = 1, 2.$$

In principle, there are ways other than the product of singular values to estimate cluster volume, but high-dimensional clusters often appear rectangular due to under-sampling and near-orthogonality [8].

Now suppose $D_1 \leq D_2$, i.e. group one is less dense than group two. Then n_1 points are sampled from a hyperrectangle with side lengths $\sqrt{12}\sigma_j^1$. The scaling factor of $\sqrt{12}$ ensures the variance in each principal direction of the uniform distribution is equal to the variance contained in each principal component of the group one data. Once the points are sampled, the MST is calculated, and the number of edges crossing a splitting hyperplane is recorded. The simulation process yields an approximate distribution to which the test statistic is compared. The p-value is the percentile of the test statistic within this simulated distribution. A 1-sided test is employed because we are only interested in rejecting the null given sufficiently small values of the test statistic.

2.5 Heatmap

The heatmap is a very useful tool for comparing groups because it provides a feature-by-feature perspective. It pinpoints the exact features in which the two groups differ the most. The interactive heatmap also allows users to select and analyze sub-heatmaps, providing a more focused view on specific features. The features are ordered according to difference in group means.

2.6 Meta Data Plot

Along with the data and clustering, the user may also supply meta data corresponding to the samples in the original data. The meta data for each group is presented via pie charts for categorical data and box plots for numerical data. These plots are useful for discovering trends in the data.

3 Results

3.1 MST Stability Experiment

1,500 samples were randomly chosen from the MNIST data set of handwritten digits [7]. Each 28 × 28-pixel image was flattened into a vector of length $28^2 = 784$, so the data contain 1,500 samples in 784 dimensions. A PCA pre-processing step was employed to reduce the number of dimensions to 300. The simplified medoid subtree T was then calculated.

Random Gaussian noise was then added to the data and the new simplified medoid subtree T' was calculated. The R-F distance $RF(T, T')$ was recorded. This process was repeated 30 times.

To better interpret the R-F distances, we designed a null distribution of distances as a reference for comparison. These distances should represent R-F distances between trees that do not portray similar global structures and inter-cluster relationships. To generate the null distribution from the data, we randomly permuted the class labels and computed the R-F distances between the resulting simplified medoid subtrees and the original simplified medoid subtree. By randomly re-labelling the clusters, we are simulating examples with distinct global structures. Figure 1 shows the R-F distances produced by adding noise and permuting the class labels. The simplified medoid subtrees generated by adding noise were significantly closer to the original simplified medoid subtree than those generated by randomly permuting the class labels in terms of R-F distance, showing inter-cluster relationships in the MST are robust to noise.

4 Application

4.1 Image Data Example

To demonstrate use of the tool, we explore the MNIST data set in detail. Again, the 784 × 784-pixel images were flattened and 1,500 samples were randomly sampled. A PCA pre-processing step was applied prior to applying UMAP [12] to construct a two-dimensional embedding. To replicate a real use case, we study a k-means clustering instead of the true class labels (Figure 2). The reader may follow along using the `run_example(example="MNIST", cluster="kmeans")` function in our *DRtools* package.

At first glance, there are three major instances of disagreement between the low-dimensional embedding and the k-means clustering. Classes 1 and 2 seem to form one cluster together, class 4 is split into two separate clusters, and class 9 is merged with points from other clusters.

4.1.1 Classes 1 and 2

There seems to be minimal separation between classes 1 and 2, suggesting they may correspond to the same digit. We select a path from point 25,483 in class 1 to point 44,483 in class 2. To get a closer look, we first look at the Path Projection Plot. The chosen number of dimensions is 100, which retains 97% of the variance, and the path stabilizes at a degree of $d = 4$.

The resulting plot depicts overlap between the two classes. Adjusting the bandwidth of the density estimate to 1.5 shows unimodal density, suggesting the two classes may come from the same population. Showing the MST edges also does not provide any evidence of separation. The MST test results, however, may suggest otherwise. Seven crossings are counted when the approximate expectation under the null is 11.03 with a standard error of 3.523. While the bootstrapped p-value of 0.06 is insignificant under a significance level of 5%, the closeness indicates a more careful examination is necessary.

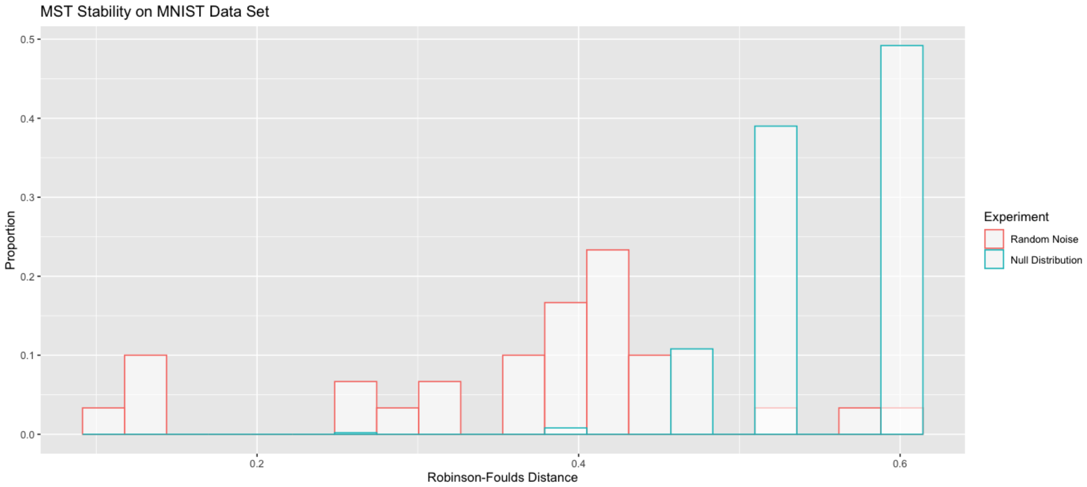

Figure 1: MST Stability on MNIST Data Set

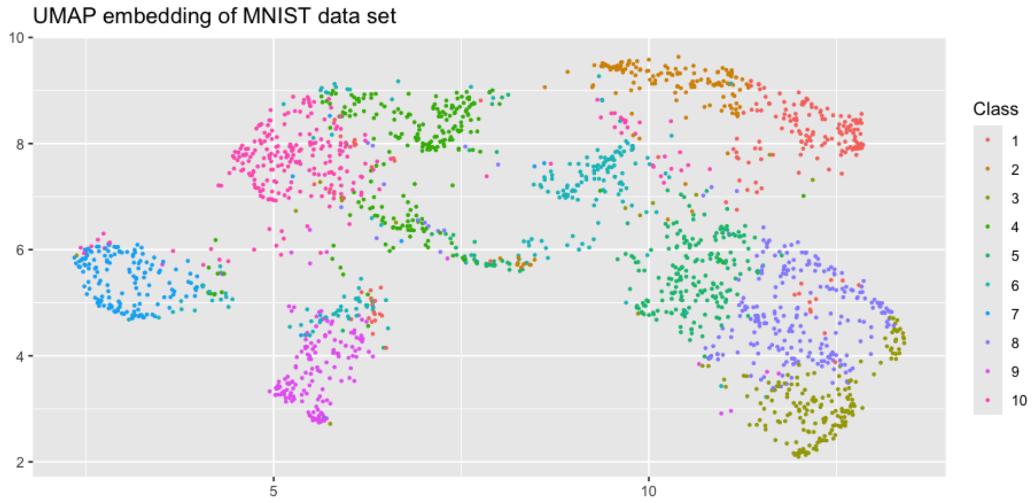

Figure 2: MNIST Embedding with k-Means Coloring

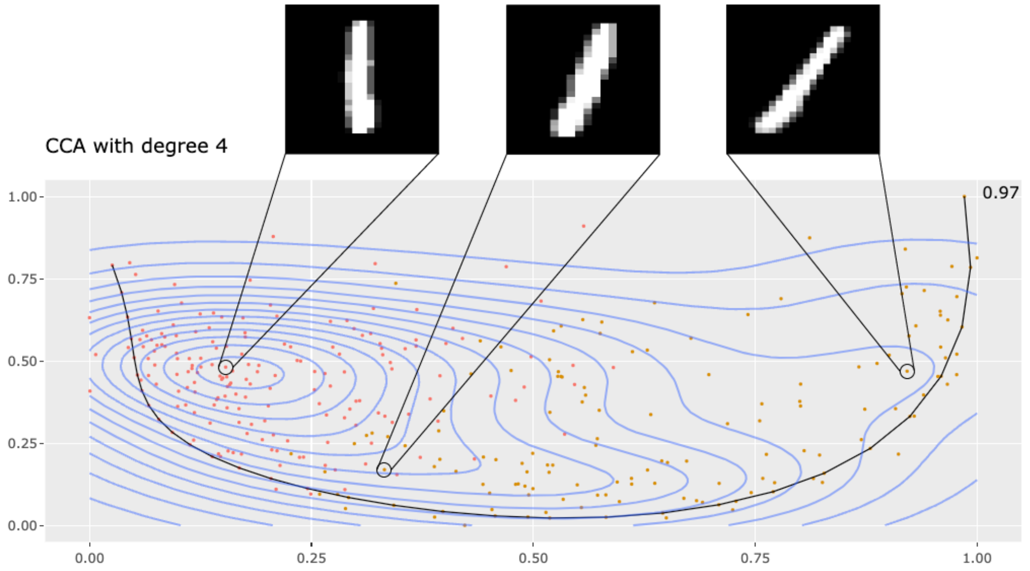

Figure 3: Projection of Path Between Classes 1 and 2

Inspection of the handwritten digits themselves reveals an interesting trend. While the majority of samples from both classes depict the digit one, the angle of the stroke differs drastically between the two classes (Figure 3). Following the path from class 1 to class 2, the angle of the stroke becomes less and less steep. Both the MST path and MST test were able to detect this phenomenon, even though the two classes technically corresponded to the same digit.

Overall, the analytical plots provide context that the UMAP embedding alone does not. There is a gradual decline in density as you move across the combined cluster, which corresponds to increasingly slanted one digits. The MST test was able to detect the difference in penmanship with the near-significant p-value of 0.06.

4.1.2 Class 4

Class 4 is split between two different clusters in the UMAP embedding. We use the drawing tool to select the two clusters as our groups. The path projection settings are calibrated to 100 dimensions and a CCA degree of three. We also select the Group Coloring setting so the points are colored according to group, rather than class. Analysis of the plot and estimated density does not provide evidence of separation. The MST edges, however, are more revealing after close inspection. There are very few inter-group edges, even in overlapping regions (Figure 4). Unexpectedly, however, the MST test counts a much larger number of edge crossings. 14 are counted when only 12.02 are expected with a standard error of 3.378. This discrepancy is due to the counting procedure used by the MST test. To account for the presence of clusters other than the two groups of interest, the MST test aims to count the number of paths between the two groups, rather than edges. Now because the two class 4 clusters are joined via the class 1 cluster, there are many more inter-group paths than inter-group edges. Hence, the MST test suggests both class 4 clusters, along with the class 1 cluster, may belong to one large cluster.

According to the true class labels, these clusters correspond to distinct digits (Figure S1). The top class 4 cluster corresponds to the digit eight, while the bottom class 4 cluster corresponds to the digit five. The connecting class 1 cluster corresponds to the digit three. All three of these digits are written in a very similar way and can be easily confused for one another, leading to blurred boundaries between their respective clusters.

4.1.3 Class 9

Class 9 is well-separated, but its cluster also contains some points from other classes, mainly class 6. To determine if these points should belong to the same class, we use the drawing tool to select the class 9 points and the remaining points in the cluster as our groups. The path projection settings are calibrated to 100 dimensions and a CCA degree of five. Together, the points form a unimodal cluster, as shown by the approximate density calculated with a bandwidth of 1.3 (Figure 5). Visually, there is also a consistent density of MST edges throughout the cluster, even where the two groups meet. The MST test agrees. There are 16 crossings counted, just below the expected value 16.37 under the null hypothesis. All evidence points towards the merging of these two groups.

According to the true class labels, this entire cluster corresponds with the digit six (Figure S1). The k-means clustering incorrectly scattered the points into multiple classes.

4.2 Mass Cytometry Data Set

We now explore a mass cytometry data set [19] covering 35 samples originating eight distinct human tissues enriched for T and natural killer cells. The data is processed and labeled inline with the procedure used in [2]. To replicate a real use case, we explore a k-means clustering instead of the true class labels (Figure 6). The reader may follow along using the `run_example(example="Wong", cluster="kmeans")` function.

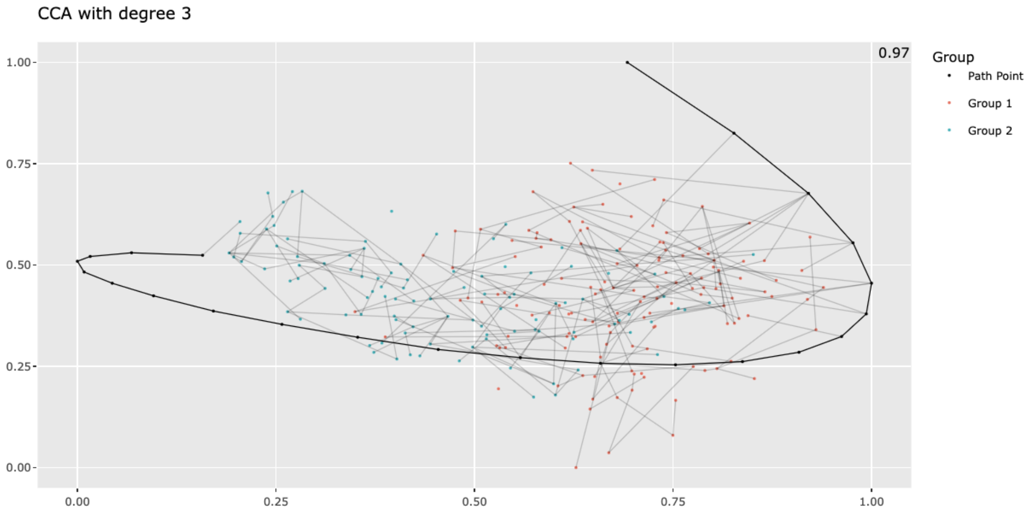

Figure 4: Projection of Path Between Class 4 Clusters

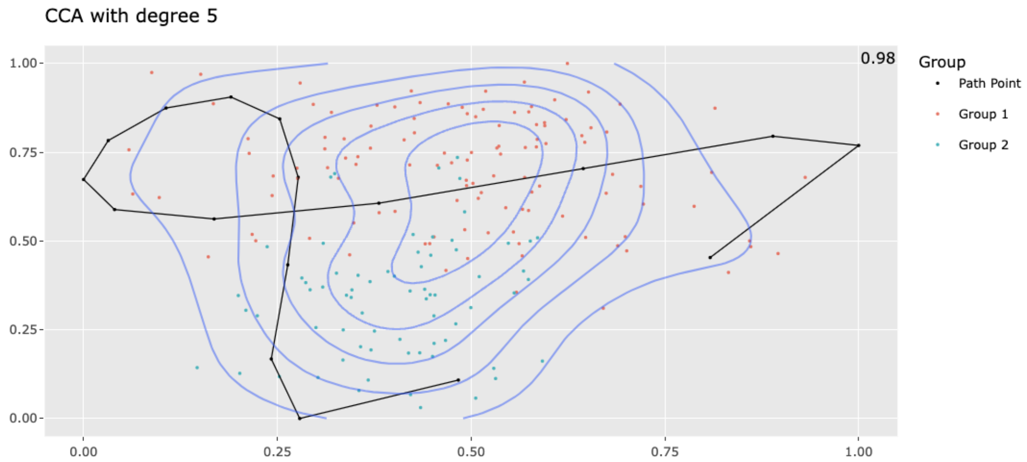

Figure 5: Projection of Path Between Class 9 and Remainder of Cluster

Most of the k-means clustering seems to agree with the UMAP embedding. However, classes 4 and 8 are both split between two distinct clusters. Class 3 is also separated into three smaller sub-clusters.

4.2.1 Class 4

Class 4 is split between two separate clusters in the UMAP embedding. To diagnose, we select the two custom clusters using the drawing tool. The path projection settings are calibrated to 20 degrees and a CCA degree of two. We also select the Group Coloring setting so the points are colored according to group, rather than class. The plot along with the estimated density does not provide any evidence of separation (Figure 7). The two selected groups also have 18 crossings, larger than the null expectation of 15.62. All evidence indicates the two groups were sampled from the same population, in agreement with the k-means clustering.

To better understand why these two clusters are separated despite minimal evidence of separation, we reference the heatmap and meta data. According to the heatmap, the two groups differ most in CD8 T cell counts. This is confirmed by the cell labels provided by [2], which were passed to the tool as meta data. So while separation wasn't observed by the MST, the discrepancy in CD8 T cell counts accounts for the splitting of the class 4 points.

Turns out, the k-means cluster got it right. Together, these two groups make up the cells sampled from skin tissue (Figure S2). Within the skin tissue cells, however, exist two subgroups differentiated by

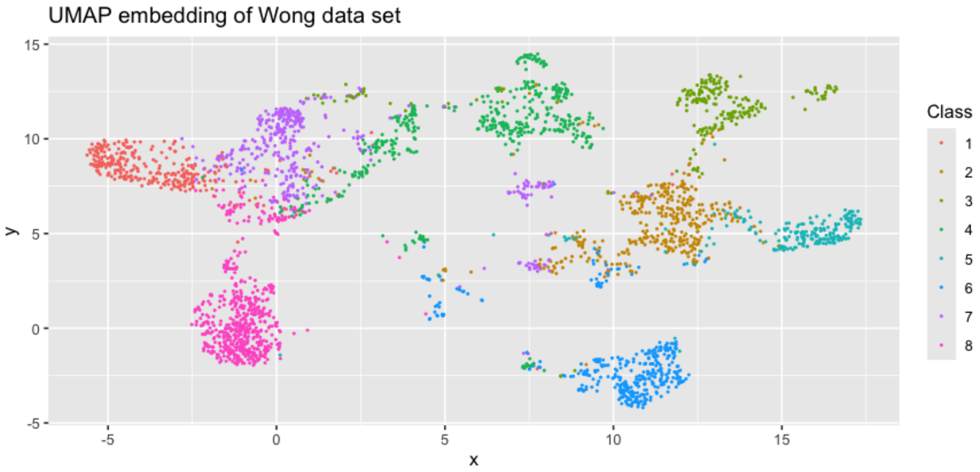

Figure 6: Wong Embedding with k-Means Clustering

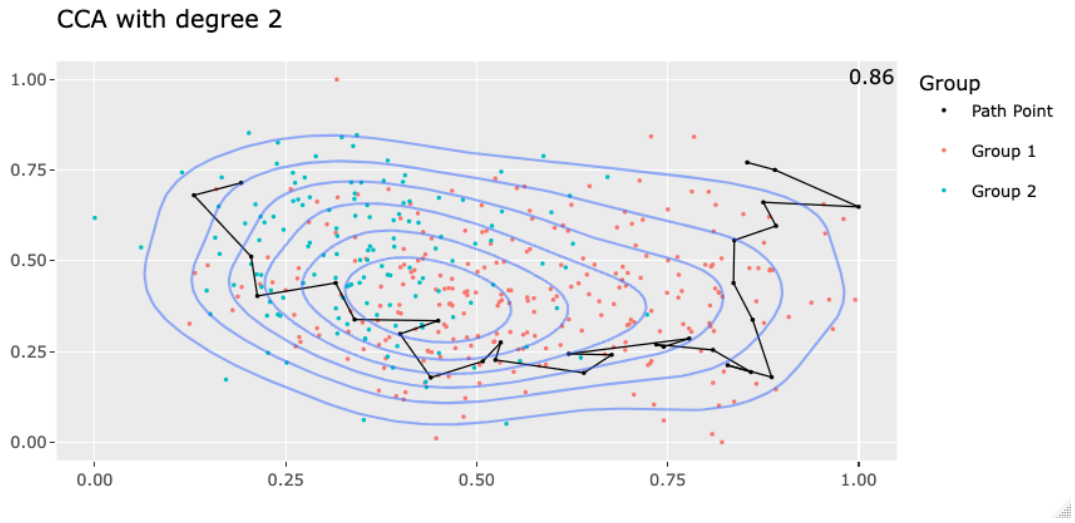

Figure 7: Projection of Path Between Class 4 Clusters

CDB T cell count.

4.3 Class 8

The majority of class 8 points lie in a self-contained cluster. However, the rest lie in a separate nearby cluster. We select the two class 8 clusters as our two groups and study the projection of the path between them. The path projection settings are calibrated to 20 dimensions and a CCA degree of three. Similar to the UMAP embedding, the path projection plot depicts one dense cluster containing a majority of the points (Figure 8). The remainder of the points fall to one side in a low-density region. There is certainly not enough evidence to conclude the two groups belong to separate clusters from this plot alone.

Running the MST test returns interesting results. 20 crossings are counted when only 16.11 are expected with a standard error 4.44. The number of crossings are well above the number expected under the null hypothesis. This is due to the disparity in group densities. In order to preserve the size of the test, the null distribution must be constructed using a unimodal distribution whose density is similar to the that of the lesser-dense cluster (Section 2.4.2). This is why the expected number of crossings is so low. This phenomenon, however, has statistical meaning. The low relative density of group 1 provides minimal evidence from which we may draw conclusions. This lack of confidence is reflected in the low null expectation, and thus, the inflated p-value.

Overall, there is not enough evidence to declare the two groups are correctly separated. This it not quite correct according to the true labels. The two groups were not sampled from the same type of tissue, but the situation is slightly more complicated (Figure S2).

4.4 Class 3

The class 3 points are separated into three clusters – two larger, elongated clusters to the left, and one smaller cluster to the right. Analysis of the path projection plots and MST tests does not capture any evidence of separation. However, the heatmap and meta data provide much-needed context. The two larger clusters to the left are completely disjoint in their CD161 gene counts, but very similar in all other variables. Meanwhile, the third smaller cluster is distinct from the other two larger clusters in its TCRgD counts. The meta data also reveals the left two clusters consist of CD8 T cells, while the right cluster consists of Tgd cells. So while the disparity in one isolated variable between each pair of clusters was captured by UMAP, but not the MST, the heatmap and meta were able to recoup the missing information.

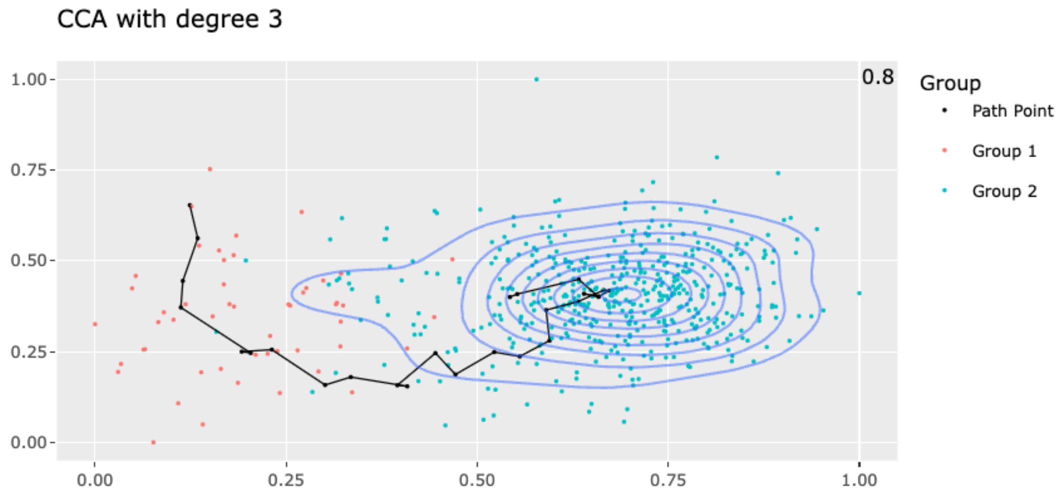

Figure 8: Projection of Path Between Class 8 Clusters

5 Discussion

We have introduced our R package, *DRtool*, and exemplified its use cases. The MST serves as an effective medium for understanding high-dimensional relationships and structures. The various analytical tools provided by the package allow the user to extract a maximal amount of information from the MST by providing multiple prospectives. Such a multi-faceted view is necessary to understand contemporary dimension reduction methods that are trying to fit hundreds, or even thousands, of dimensions-worth of information into only two dimensions. Advances in multiple fields have lead to a surge in complex data, necessitating tools such as ours that help analysts assess and affirm their dimension reduction results.

Further works should explore alternate methods for projecting paths into two dimensions. The goal of the projection is to “unwind” the path, which is a non-linear transformation, but non-linear methods could pose two problems. One, most non-linear methods do not have a natural out-of-sample extension that can be used to project points of interest other than the path points. And two, non-linear methods can be prone to overfitting, especially when the path only contains a handful of points. On the other hand, linear methods define a linear transformation on the entire data space, so the projection naturally extends to points not on the path. Their rigidity also prevents overfitting. The downside is linear methods are known to fail in high dimension due to the near-orthogonality of high-dimensional data. They also shrink space, which may obscure fine structural details that only non-linear methods are capable of capturing.

Further works should also explore alternate methods of estimating cluster volume when calculating cluster density during the MST testing process. The product of singular values works well for clusters that are generally ellipsoidal or rectangular, but can fail for irregularly shaped clusters. A better estimate of the density could increase the power of the MST test.

6 Code Availability

All data and code a freely available at `https://wwww.github.com/JustinMLin/DRtool`.

References

[1] Martin Wattenberg amd Fernanda Viégas and Ian Johnson. How to use t-SNE effectively. `https://distill.pub/2016/misread-tsne/`.

[2] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel Kwok, Lai Ng, Florent Ginhoux, and Evan Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, 37:38–44, 2019.

[3] Bhaswar Bhattacharya. A general asymptotic framework for distribution-free graph-based two-sample tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(3):575–602, 2019.

[4] Hao Chen, Xu Chen, and Yi Su. A weighted edge-count two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 113(523):1146–1155, 2018.

[5] Hao Chen and Jerome Friedman. A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 112(517):397–409, 2017.

[6] Andy Coenen and Adam Pearce. Understanding umap. `https://pair-code.github.io/understanding-umap/`.

[7] Li Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6), 2012.

[8] Persi Diaconis and David Freedman. Asymptotics of graphical projection pursuit. *The Annals of Statistics*, 12(3):783–815, 1984.

[9] Jerome Friedman and Lawrence Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Annals of Statistics*, 7(4):697–717, 1979.

[10] JC Gower and GJS Ross. Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 18(1):54–64, 1969.

[11] Bracken King and Bruce Tidor. MIST: Maximum information spanning trees for dimension reduction of biological data sets. *Bioinformatics*, 25(9):1156–1172, 2009.

[12] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(3), 2018.

[13] Daniel Probst and Jean-Louis Reymond. Visualizing very large high-dimensional data sets as minimum spanning trees. *Journal of Cheminformatics*, 12(12), 2020.

[14] DF Robinson and LR Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.

[15] GPM Rozál and JA Hartigan. The MAP test for multimodality. *Journal of Classification*, 11:5–36, 1994.

[16] Werner Stuetzle. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of Classification*, 20:25–47, 2003.

[17] Joshua Tenenbaum, Vin de Silva, and John Langfor. A global geometric framework for nonlinear dimensional reduction. *Science*, 290(2319), 2000.

[18] Elena Tuzhilina, Leonardo Tozzi, and Trevor Hastie. Canonical correlation analysis in high dimensions with structured regularization. *Statistical Modelling*, 23(3):203–227, 2023.

[19] Michael Wong, David Ong, Frances Lim, Karen Teng, Naomi McGovern, Sriram Narayanan, Wen Ho, Daniela Cerny, Henry Tan, Rosslyn Anicete, Bien Tan, Tony Lim, Chung Chan, Peng Cheow, Ser Lee, Angela Takano, Eng-Huat Tan, John Tam, Ern Tan, Jerry Chan, and Evan Newell. A high-dimensional atlas of human T cell diversity reveals tissue-specific trafficking and cytokine signatures. *ScienceDirect*, 45(2):442–456, 2016.

Supplementary Information

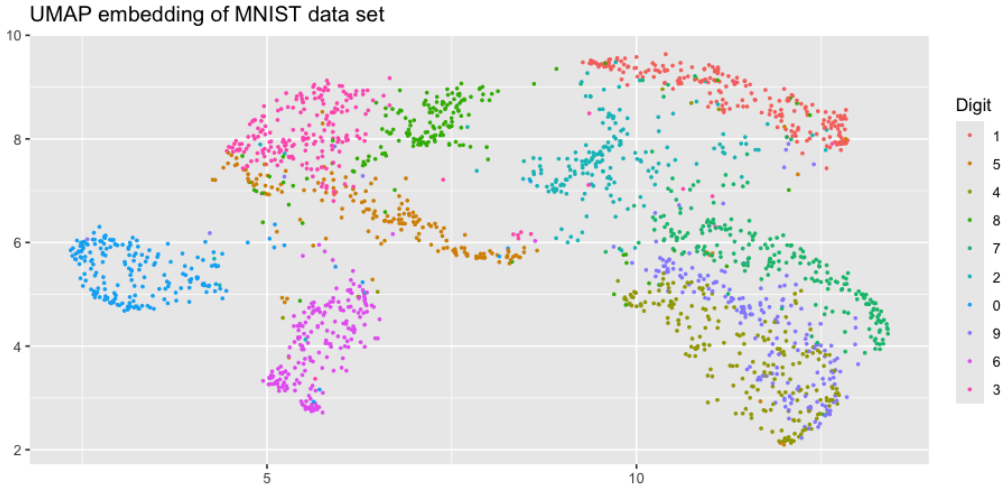

Figure S1: MNIST Embedding with True Labels

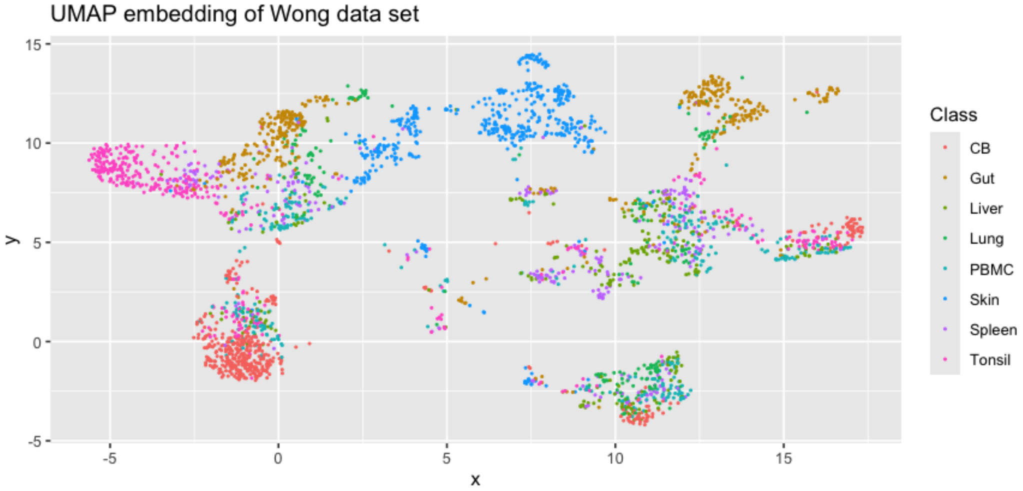

Figure S2: Wong Embedding with True Labels