Formal Paper Draft

Influence Diagnostics for High-Dimensional Ordination Techniques

Paul Harmon

12 January, 2022

Introduction

Consider the problem of visualizing multivariate (or potentially high-dimensional) data. In many cases, it is useful to generate a 2 or 3-D mapping of some higher-dimensional space, projecting the high-dimensional data into an ordination that can be directly visualized. There are many methods that exist to accomplish this task, ranging from tools such as classicial multidimensional scaling (MDS) to more modern tools like t-distributed Stochastic Neighbor Embedding.

Unlike regression or classification, most tools for dimension reduction and higher-dimensional data visualization are unsupervised, meaning that they do not take advantage of some marked response variable; rather, they capitalize on the signal present in a set of features and produce a data-driven result.

Despite those differences, tools like t-SNE and MDS are susceptible to aberrant data values, much in the same way that a regression model might be susceptible to outliers and influential points. In a regression model, the estimated coefficients or predictions might be affected in meaningful fashion by the inclusion or exclusion of a single point - such points are referred to as influential (Cook, 1977). A suite of tools have been developed to identify and measure the impact of influential points on the results of regression models.

The goal of this method is to develop a diagnostic that can be used to help identify influential points and potentially quantify how much they impact the resulting mapping, irrespective of how that mapping was created. Additionally, this method can be used to assess the sensitivity of each method to influential points and the impact that sensititivy might have on how an end user might want to interpret the resulting map produced. Note that while robustness to influential points might at first seem like a positive trait for an ordination method, masking distinctly different observations in a mapping may have negative consequences as well.

This document overviews the statistical methodology we have developed to identify influential points that, when excluded from a dataset, can meaningfully impact the shape of an ordination created from data. We overview our methodology step by step.

Tools for High-Dimensional Data Visualization

t-SNE

One of the notable new methods for high-dimensional data visualization is t-distributed Stochastic Neighbor Embedding, or t-SNE (Van der Maaten, 2011). Since the time that Laurens Van der Maaten published the first t-SNE paper in 2011, it has been cited more than 21,000 times in use cases ranging from biology (Amir et al, 2013) and genetics () to economics (https://economics.yale.edu/sites/default/files/files/Faculty/Tsyvinski/tSNE_draft15.pdf) to natural language processing (https://aneesha.medium.com/using-tsne-to-plot-a-subset-of-similar-words-from-word2vec-bb8eeaea6229) and machine learning. The method's

prevalence is astounding - it is used not only in academia but in industry problems, including tools like FoodGenius (https://github.com/foodgenius/tsneplot).

The method centers on calculation of a balance between two similarity metrics. The first, called p_{ij},is a conditional probability based on the Euclidean distance between pairs of points x_i and x_j in the high-dimensional space. In the lower-dimensional representation, this similarity is defined as q_{ij}.

A good representation of the high-dimensional space should mean that p_{ij} and q_{ij} look reasonably similar to each other; since these are both simply joint probabilities, the objective function of t-SNE is simply a symmetricized Kullbeck-Leibler divergence between the two, defined below:

$$Cost = \sum_i \sum_j p_{j|i} log(\frac{p_{j|i}}{q_{j|i}})$$

t-SNE is an improvement over Stochastic Neighbor Embedding (SNE), which makes use of Gaussian distributions to model both p_{ij} and q_{ij}. The use of a light-tailed distribution for the lower dimension leads to a phenomenon known as the "crowding problem" where the 2-dimensional map cannot accommodate many neighbors in high dimensions (Van der Maaten, 2011). By utilizing a t-distribution for the q_{ij}, t-SNE better preserves the local structure of the data. Most of the minimization of objective functions is done via gradient descent (and some methods require simulated annealing because the algorithm can get caught in local optima). Critically, because the optimization is non-convex, and as a consequence of the gradient descent methods used to optimize it, t-SNE has the potential to provide different maps when run on the same data at the same perplexity.

Reduction of Initial Dimensionality with PCA

Additionally, t-SNE is subject to a pre-processing step involving principal component analysis (PCA). In most implementations, "whitening" via PCA is performed on the initial dataset (or input distance matrix) to reduce the dimensionality to a reasonable value. For instance, in Van der Maaten's initial paper, several examples reduced the initial dimensionality of the data to 30 for computational gains (Van der maaten, 2011).

Depending on the size of the initial dimensionality of the data, this step may drastically reduce the number of dimensions that t-SNE needs to deal with. However, it has some potential to impact the resulting ordination, particularly if the number of initial dimensions kept is relatively low or PCA does not do a great job of fitting the original dataset. At best, the t-SNE mapping is only as good as the initial PCA is. For the simulations shown in this paper, no whitening was performed, and the original distance matrix was passed to t-SNE directly.

Campell, Caudle and Hoover (2019) explored different methods than PCA as an intermediate dimensionality reduction step in t-SNE. They compared PCA, Independent Component Analysis (ICA), Linear Discriminant Analysis (LDA), Sammon Mapping, and Locally Linear Embedding (LLA) as preliminary steps in t-SNE and found that resulting maps tended to form different groupings with variation in separation. While we did not consider assessing the impact of different intermediate dimensionality reduction techniques in this analysis, it does present an interesting route for additional research.

Multidimensional Scaling

Classical multidimensional scaling (MDS) is closely related to PCA (an in some cases, equivalent) (Gower, 1966). MDS takes as input a distance or dissimilarity matrix representing the distance between pairs of observations. This distance does not have to simply be Euclidean (although it often is). Classical multidimensional scaling returns maps in the same metric as the original distance or dissimilarity matrix; in general, the more dissimilar two observations are in the high-dimensional space, the farther apart they should be in the resulting MDS ordination.

Consider a multivariate dataset with data points x. For distinct pairs of points x_i and x_j, multidimensional scaling attempts to minimize a metric called "stress" which can be defined generally as follows:

$$S_M(x_1, x_2, ..., x_N) = \sum_{i \neq j} (d_{i,j} - \|z_i - z_j\|)^2$$

Different versions the stress function exist for different variations of MDS, including methods like Sammon mapping and Classical MDS, which frames the problem from a similarity perspective:

$$S_c(x_1, x_2, ..., x_N) = \sum_{i,j} (s_{i,j} - (z_i - \bar{z}, z_j - \bar{z}))^2$$

Classical MDS can be expressed equivalently as Principal Component Analysis when distances are Euclidean (Elements, XXXX).

Additional Mapping Techniques

While the focus of this paper is on t-SNE and MDS, there are manifold additional methods for generating lower-dimensional representations of high-dimensional data. We include some additional methods that were either compared directly with t-SNE in Van der Maaten and Hinton's original paper (2008) or because they represent the cutting edge of high-dimensional data visualiztion tools. They range in terms of complexity, with some adding relatively minor changes to the MDS structure to those that share similarities with other branches of mathematics, such as topology or topoligical data analysis. Sammon Mapping is a non-linear version of classical scaling which scales the stress metric by the pairwise distances, attempting to better maintain local structure by equalizing the impact of large and small distances (Sammon, 1969). The objective function for Sammon mapping is as follows, where y_i and y_j are the ith and jth points in the lower-dimensional representation and d_{ij} represents the Euclidean distance between points x_i and x_j in the high-dimensional space:

$$\phi(Y) = \frac{1}{\sum_{ij} d_{ij}} \sum_{ij} \frac{(d_{ij}^2 - ||y_i - y_j||^2)}{d_{ij}}$$

Isomap similarly extends MDS to be more flexible, and is a non-linear reduction technique (Tenenbaum et al., 2000). It was initially used as a comparison for t-SNE by van der Maaten (2011) and is utilized as a comparison mapping technique here as well.

Stochastic Neighbor Embedding, or SNE, was a precursor to t-SNE and relies on gaussian distributions rather than the t-distrubution specified previously. It is subject to a problem known as "crowding" which led to the development of t-SNE. It is worth noting that SNE need not be based on symmetricized KL-divergences, as t-SNE is.

Finally, UMAP is a modern competitor for t-SNE that is based topological data science techniques; chiefly, a model of a Riemannian manifold used to approximate the data (McInnes and Healy, 2018).

Defining "Influence" in High Dimensional Maps

Influence in Regression - Cook's Distance

Belsley, Kuh and Welsch define influential points as a data point "which, either individually or together with several other observations, has demonstrably larger impact on the calculated values of various estimates (coefficients, standard errors, t-values, etc.) than is the case for most of the other observations."(1980 page 11). In regression modeling, as in most statistical methods, care must be taken to properly distinguish between an outlier and a so-called influential point. An outlier is typically denoted as a point that is either

much larger or smaller than similar observations on a specific scale. However, an outlier may not necessarily be influential; a truly influential point must be outlying enough relative to similar observations that it impacts the fit of the regression line. The identification of influential points that can impact the results of a model is commonplace and has been the subject of substantial research for regression-based methods.

Cook (1977) proposed a measure of influence called "Cook's Distance" that has since become one of the most common tools for diagnosing influential points and outliers. The idea is predicated on the concept that by leaving an out an influential observation, the results of the model should change. This process of jackknifing observations is done for all the observations in the data.

Cook's distance is defined as follows for the ith observed data point, with$\hat{j}_j$ the fitted response value for the model withthe ith observation omitted, s^2 defined as the mean-squared error for the model and p the number of variables for each observation:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y_j} - \hat{y}_{j(-i)})^2}{ps^2}$$

Although Cook's distance is essentially an F statistic from an $F_{p,n-p}$ distribution, and the method does not use p-values to determine how "influential" a point is. Rather, the F-statistics themselves are typically compared to cutoff values, or observations with relatively large Cook's distance values are considered for investigation of influence.

Identifying Outliers in Maps

Similar to a regression setting, the resulting maps produced by ordination methods like MDS, t-SNE and the like can be impacted by the inclusion or exclusion of points. In short, an "influential" point in a lower dimensional mapping can be defined as one that, when excluded, drives meaningful differences in the ordination method that lead to a different shaped map as compared to when that point is included. An outlier, on the other hand, might look different in the high-dimensional space (or on a subset of features), but would not meaningfully change the shape of the map when left out.

Joliffe (2002) overviews tools for identifying influential observations in Pricipal Component Analysis, which is a related dimension-reducing technique to MDS. Influence functions (Critchley, 1985; Hampel, 1974; Huber, 1981) can be defined for regression methods as well as for loadings in PCA. However, the literature on more recently-developed methods lacks a similar metric, most notably one that can be used or compared across different mapping methods.

Much of the literature around robustness of mapping results and multivariate outliers centers on dealing with the problem by modifying the dimension reduction technique rather than identifying it. Blouvshtein and Cohen-Or (2018) analyzed the effect of outliers on the maps produced by classical multidimensional scaling methods. In general, these methods are not particularly robust to the effects of multivariate outlier. They propose a method to detect outliers prior to the generation of the MDS map so that they can be removed from the data. Their method involves treating the distances input to the algorithm as edges of a complete graph that connects each data point.

These edges $d1, d2, d3$ can then be used to form triangles - in general, outliers tend to *break* the triangle inequality and inliers tend to satisfy it. They recommend a rule to examine the histogram H of b, the number of edges of broken triangles, and generate a threshold ϕ that is intended to identify inliers vs. outliers.

Two other papers, including Forrero and Giannakis (2012) as well as Kong et al. (2019) frame the problem in a slightly different way. They consider a penalized *stress* metric for optimization, utilizing regularization to better identify outlier points. In both cases, they are able to identify outliers using a majorization-minimization procedure that iteratively produce more outlier-robust maps of the lower-dimensional embedding of interest.

Alternatively, archetypal analysis (Cutler & Breiman, 1994) can be used to identify outlying points in a multivariate space. The method approximates a convex hull that lies around the boundary of a dataset based on "archetype" points; these extreme points have been used to identify potential multivariate outliers

(Catherine's PhD dissertation). While this method may be effective at identifying and categorizing extreme points on the boundary of the multivariate space, it does not necessarily answer the question of how the inclusion/exclusion of those points might change the shape of a lower-dimensional representation. This is the problem we seek to answer with our method.

Permutational Multivariate Analysis of Variance

Although the previous methods give rise to identifying extreme points, they do not clearly inform the influence of a given observation on the resulting shape of the ordination. Since the goal of this research is to identify a method for identifying and quantifying influential observations in ordination problems, our goal is to extend the ideas that underlie Cook's Distance from a regression framework to a multivariate response framework. This necessitates the use of distance matrices rather than raw observations, and requires the use of multivariate Analysis of Variance methods to generate something similar to a F-statistic.

Permutational Multivariate Analysis of Variance (PERMANOVA) is a tool for allows for partitioning variation across multivariate data (Anderson, 2017) in other words, it allows for generalization of ANOVA-type tools on multivariate datasets. Much like some of the multivariate extensions of ANOVA, PERMANOVA allows for testing of differences in means while taking into account the correlation structures between variables.

A critical difference between some of the standard tools for performing MANOVA and PERMANOVA is that the latter method does makes no assumption of multivariate normality. It only requires an assumption of exchangability between Instead, PERMANOVA relies on the use of permutation tests to generate a pseudo-F test statistic. In a similar vein as ANOVA, PERMANOVA requires the calculation of among-group sum of squares (SSA) and residual within-group sum of squares (SSW). The total sum of squares is calculated as follows:

$$SS_T = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{d_{ij}^2}{N}$$

Partitioning can be obtained via subtraction of the residual sum of squares SS_W from the total sum of squares (as in ANOVA), which is defined as the sum of each of the groups' squared distances between the observed data and the group centroid, as defined below for the l^{th} group and g is total of number of groups present in the data:

$$SS_w = \sum_{l=1}^{g} W_l$$

The W_l is thus defined as follows, with n_l defined as the the group sample size (such that $\sum_{l=1}^{g} l_g = N$) and where $\epsilon^{[l]}$ is a binary indicator between the ith and jth samples that equals 1 when samples are in the lth group and 0 otherwise:

$$W_l = \sum_{i=1}^{N-1} \sum_{j=(i+1)}^{N} \epsilon_{ij}^{l} \frac{d_{ij}^2}{n_l}$$

The resulting partitioning of $SS_T - SS_W$ allows for calculation of a pseudo F-test statistic (which is a multivariate analogue for a standard F ratio). This tests the null hypothesis that there are no differences in the positions of the group centroids in the distance matrices is as follows, where N is the number of observations and g is the number of columns in the design matrix:

$$F = \frac{SS_A}{SS_R} * (\frac{N-g}{g-1})$$

P-values are then obtained in a typical permutation-based method, meaning that they are directly calculated based on the number of observed permutations that are more extreme than a given result. For the purposes of this paper, all PERMANOVA models were performed using the R package 'vegan' (Oskanen, et al).

Method Overview

We instead propose a method for identifying influential observations that is agnostic to the mapping method. In addition, our method follows a philosophical framework similar in flavor to Cook's distance for mapping methods; however, it relies on a different set of tools than its regression analogue.

Similar to Cook's Distance, we propose calculating hold-out maps, with each of the 1...n observations held out singly. Rather than basing the test on a standard F distribution, as is Cook's distance, our method utilizes permutation-based MANOVA (Permanova) tests to calculate pseudo-F test statistics. Our method borrows from some of the machinery of the mantel test (Mantel, 1967), which is a method for comparing correlation between two matrices.

In general, we seek to identify either **multivariate outliers** (or groups of them) which, for some reason, cause large differences in the shape of the resulting ordination. The methods rely on a few different ideas. First, PERMANOVA (Anderson, 2001) is a non-parametric, multivariate version of Analysis of Variance (ANOVA) that relies on a pseduo-F test to compare within-group and between group similarities based on a specified distance (dissimilarity) measure. Unlike ANOVA, PERMANOVA makes use of permutation tests to draw inferences without assuming distributions of test statistics.

Second, distance matrices contain information on point-to-point distances (or dissimilarities) and, when vectorized, the correlation of the distance matrices measures similarity of two configurations Mantel's test (Mantel, 1967) makes use of this approach to assess correlation between distance matrices; in this case, we simply compare the similarities in the different resulting map distances.

Third, similarities (or correlations) can be converted to distances; one such metric for doing this is $\sqrt{2(1-r_{ij})}$ (James et al, 2013) where r_{ij} is the correlation between x_i and x_j.

Finally, correlations can be computed using pairwise complete observations to leverage all available information to estimate correlations in the presence of some missing data values.

Then, we utilize PERMANOVA to assess the statistical significance of the differences in our resulting distance matrices. Note that while Anderson (2017) points out PERMANOVA is less robust than its competitors to choice of dissimilarity, this is not of importance for our method as we utilize only Euclidean distance measures in this step. Our PERMANOVA-based approach is as follows:

1. For each observation in the dataset, remove it, and generate a map in 2 dimensions from the resulting (n-1) observations for a selected ordination method. (For accounting purposes, it may make sense to keep the nx1 structure but replace the holdout observation with NA).
2. Generate a distance matrix of the nx1 by 2 matrix. Do this for each of the held-out observations.
3. Generate a pairwise-complete correlation matrix from the combination of these vectorized distances that measures similarity of distances (where available).
4. Convert that correlation matrix (of all the hold-out vectorized distance matrices) into a distance matrix using a
$$\sqrt{2(1-r_{ij})}$$
(James et al, 2013).
5. Apply non-parametric PERMANOVA tests for each holdout observation - this means that on a 100-observation dataset, you apply 100 individual tests where each test uses an independent variable that is equal to 1 for the index of the holdout observation and 0 elsewhere. The hypotheses are specified as such:

- H_0: No differences in map with observation j removed vs. those without j removed
- H_a: Some difference in map with observation j removed

Generation of P-Values

In a typical PERMANOVA model, the number of permutations that could be generated from a covariate could be quite large, given that in most cases the majority of the values of the covariate may be nonzero. However, because of the structure of the method presented here, the predictor variable is always a vector with a single indicator 1 in the holdout index and 0's in all of the remaining indices. Thus, there are only n possible permutations for the predictor variable in the PERMANOVA model.

The p-value for each test is then directly calculated as the number of observations that could be more extreme than a given observation, meaning that a lower bound on the p-value is $1/n$; roughly, the set of obtainable p-values would then be directly calculated as $\frac{k}{n}$ where k, an integer value between 1 and n, refers to the number of cases that are more extreme than the observed result.

In some cases, it may make sense to consider a multiple testing correction for the p-values generated, considering that there are n tests being performed (one for each holdout). In large data settings, it may be feasible to use such methods to control for the rate of spurious detections; however, in small data settings this may be impractical. Indeed, in a very small data setting, assuming a typical $\alpha = 0.05$, if $n < 20$, it would be impossible to obtain a p-value of significance at all!

We recommend using Benjamini-Hochberg False Discovery Rate (or something similar, assuming that the results are based on p-values as in the Bonferroni Outlier test (Cook & Weisberg, 1982)). A more conservative approach, such as a Bonferroni correction, would require a significance level to be divided by the number of simultaneous tests performed, N, meaning that a Bonferroni-based approach would yield a significance threshold of $\alpha^* = \frac{\alpha}{N}$ which is at least as small as the smallest p-value that is possible to obtain.

Instead, it may make more sense to use the p-values based on their rank regardless of their statistical significance. The p-values can instead be construed as providing a sorting of the observations based on their relative extreme-ness. It stands to reason that, much in the way that Cook's Distance uses F-statistics to identify abberant observations, the pseudo-F statistic or sorted p-value can be compared based solely on relative magnitude. In a setting where only a single influential point is present, we would expect to see that influential point have the smallest p-value (or, equivalently, the largest pseudo-F statistic) of those obtained.

Simulation Studies

We conducted a series of simulations to assess the efficacy of the method at identifying influential points in various different situations. In all cases, we simulated data from multivariate normal distributions. Additionally, we did not use any of the data pre-processing available in any t-SNE maps; no PCA was used to pick an initial set of dimensions to reduce.

In the simulation, each iteration simulates a new dataset, runs the t-SNE and/or MDS mapping procedure, and tests the consistency of the configurations for each of the n holdout points. Observations can be considered to be influentieal in several ways. First, a point may be considered influential in the mean structure, indicating that it may bounce between several groups formed by the rest of the data across one or many of the variables simulated. We consider this case in numerous simulations, and test the impact of observations that differ in mean structure on a single feature ranging to all the features simulated.

Additionally, the influential point may be influential in the correlation structure. In this case, a point may not be visibly different from the rest of the data; however, such an observation might indicate highly correlated measurements on metrics that are not correlated in the rest of the data, or vice-versa. Finally, it is possible that a point could be outlying both in the mean structure and the correlation structure. We additionally consider this case in simulations below.

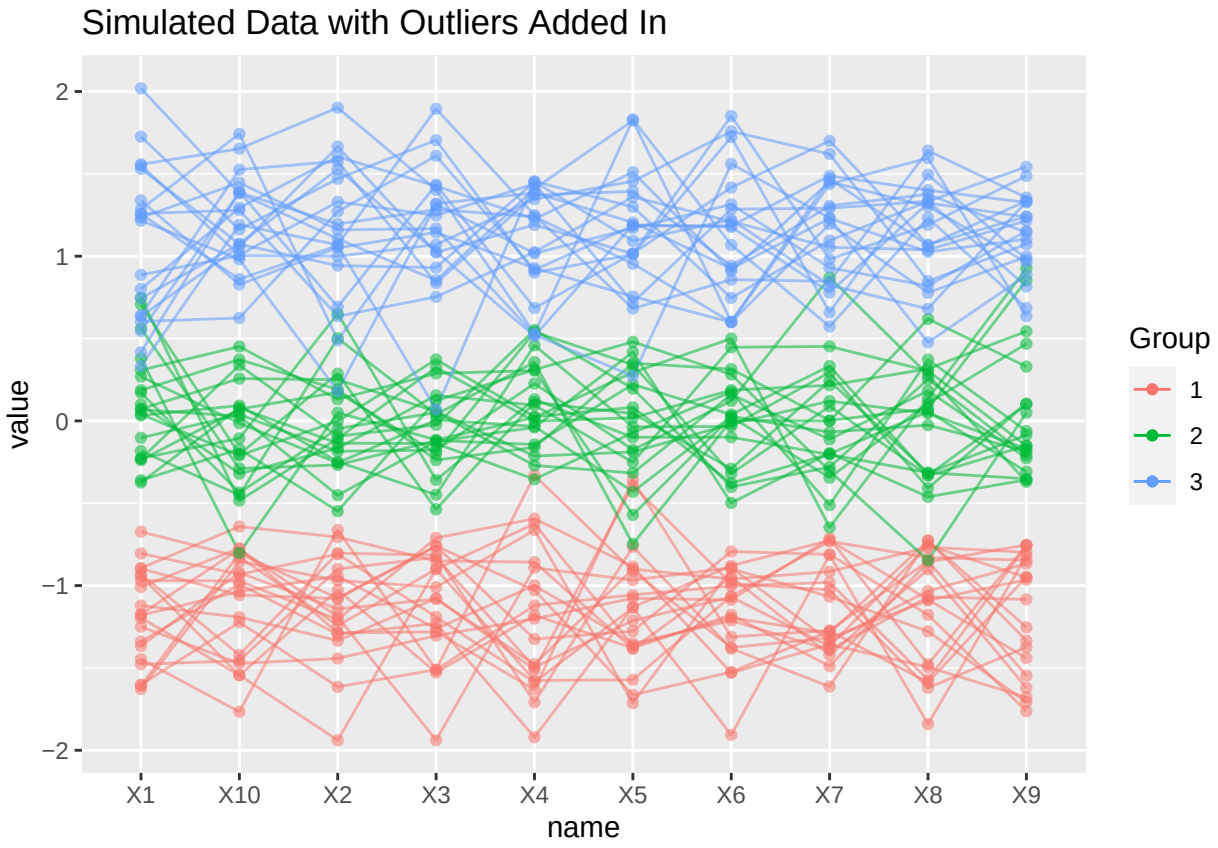
The following scenarios were considered:

- 1. No influential points

- 2. Single influential point bouncing between 3 groups
- 2. Large or Small influential point (3 groups)
- 4. Highly correlated data with 3 group means
- 5. Single Influential point on varying number of covariates

No Outliers

In the most basic simulation, we consider 60 observations on 10 features, each taken from multivariate normal distributions in three distinct groups. The data are simulated to be uncorrelated, and the resulting data is standardized so that overall, the resulting three-group cloud has mean 0 and standard deviation 1. (Note that because of the way the data are standardized, each group's standard deviation is necessarily less than 1). An example of the scenario we consider in the simulation study is shown below:



We identify similar performance of both mapping methods when there are no true influential points present in the data. As expected, there are some spurious detections - both methods falsely detect influential points at about a 3% rate, less than expected given that we are using an alpha value of 0.05. Given the relatively low percentage of spurious detections, it may not be necessary to consider a multiple testing correction; in particular, a conservative approach like Bonferroni is likely not appropriate as none of the p-values that can be obtained in a dataset with only $n = 60$ observations would be small enough to be deemed significant.

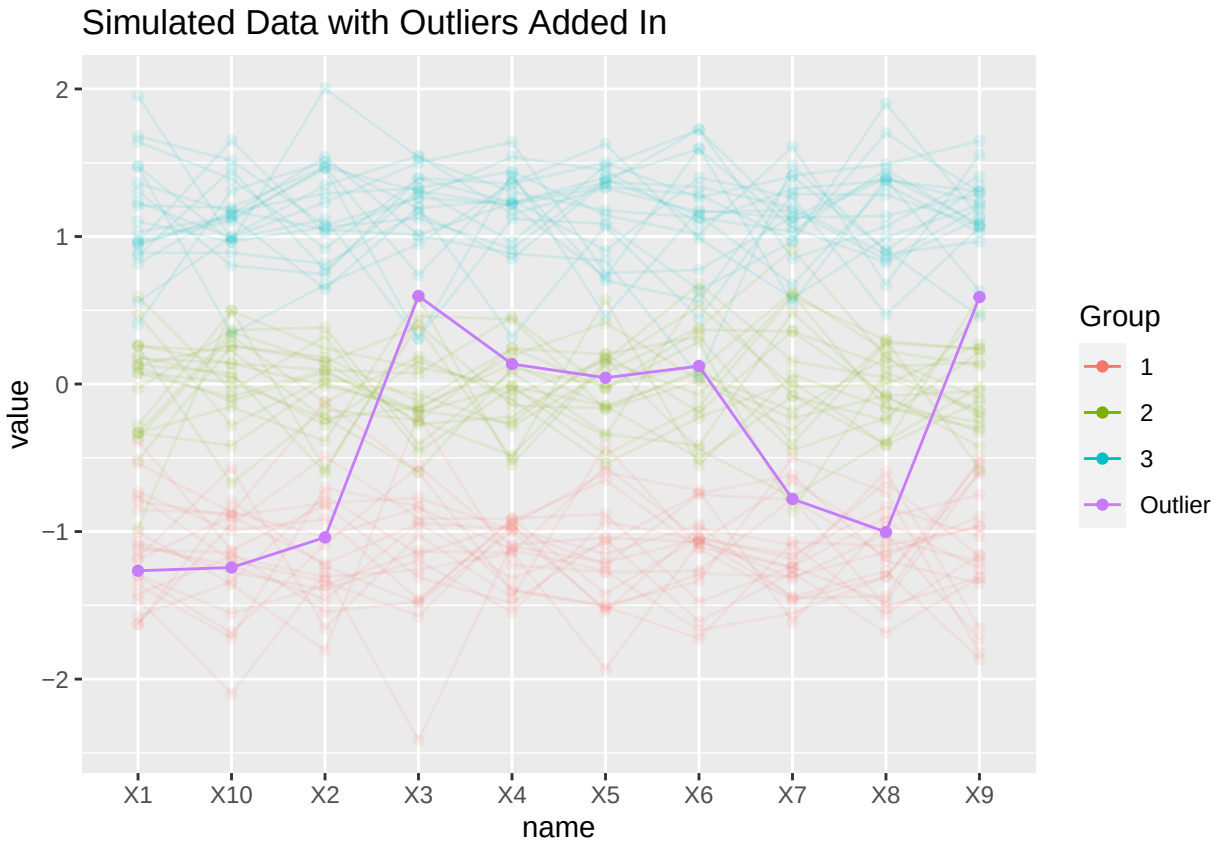
	Correctly Identified Influential	Correctly Non-Identified Non-Influential
Category	Correctly Identified Influential	Correctly Non-Identified Non-Influential
TSNE	0.033	0.967

	Correctly Identified Influential	Correctly Non-Identified Non-Influential
MDS	0.033	0.967
Isomap		
Sammon	0.033	0.967

Single Influential Points

We retain the same three group structure but add several types of influential points in the following simulation. In the first, we consider an influential point with mean 0 and standard deviation 1, indicating that the point can vary across the groups and clearly break the group structure across all 10 of the potential covariates. An example observation is shown below.

```
## Warning: The ‘x‘ argument of ‘as_tibble.matrix()‘ must have unique column names if ‘.name_repair‘ is
## Using compatibility ‘.name_repair‘.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_warnings()‘ to see where this warning was generated.
```

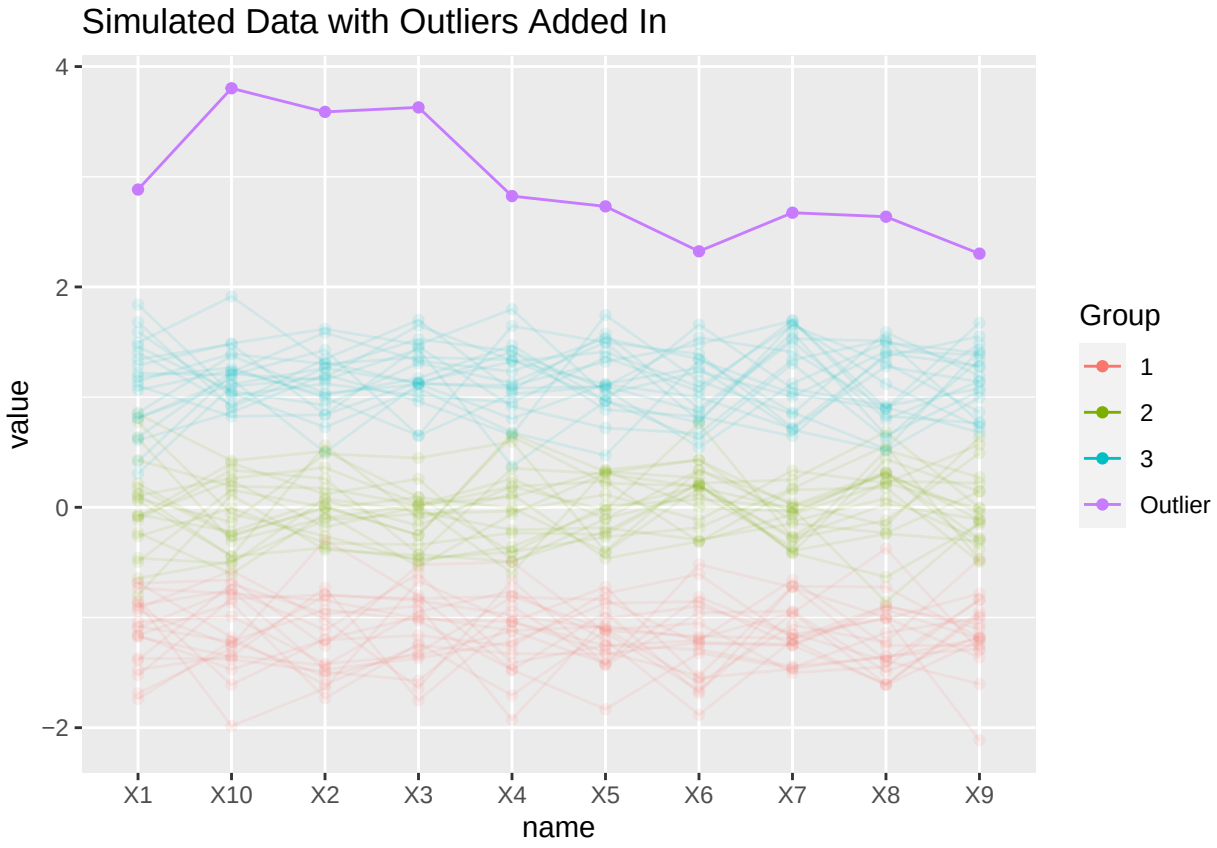


The results of the simulation with 100 simulations per ordination type is shown in the table below. The diagnostic method identifies the influential point about 90% of the time when assessing MDS maps, as compared to only 2% of the time when assessing t-SNE maps. The ineffectiveness of the diagnostic for t-SNE is not overly surprising given that in many of the maps, the influential point ended up near the middle of the map rather than on the periphery, meaning that leaving it out had a neglible effect on the hold-out t-SNE maps.

In both cases, the rate of spurious detections was very small- i.e., the number of non-influential points that were detected with small p-values (<0.05) was 3.3% for t-SNE and 1.8% for MDS mapping techniques, respectively. It's worth noting that in both cases, we would expect about a 5% spurious detection rate, but it may be that the lack of distinct permutations in the PERMANOVA tests actually leads to a slightly decreased spurious detection rate because very small p-values are difficult to obtain in a relatively small dataset.

	Correctly Identified Influential	Correctly Non-Identified Non-Influential
TSNE	0.02	0.967
MDS	0.94	0.982
Isomap		
Sammon	0.17	0.97

The results for either very large or very small observations (relative to the three groups) are quite similar - but it is interesting to note that the influence diagnostic tool seems to treat these observations differently than it does the influential point that bounces across group means. An example of a "large" observation is given below.



In both settings, the power to detect the influential point decreases relative to the observation that breaks group structure. Detection rates for the influential point fall to the 28-31% range for MDS maps, and 1-6% range for t-SNE. This is again unsurprising in the t-SNE setting because the t-SNE maps tend to place the very large or very small point in the center of the map. Interestingly, the relative magnitude of the outlying observation tends to be less important than the structural differences, at least when it comes to identifying the influence in MDS maps.

	Correctly Identified Influential	Correctly Non-Identified Non-Influential
Category	Correctly Identified Influential	Correctly Non-Identified Non-Influential
TSNE	0.01	0.967
MDS	0.27	0.971
Isomap		
Sammon	0.14	0.969

Power to Detect Differences

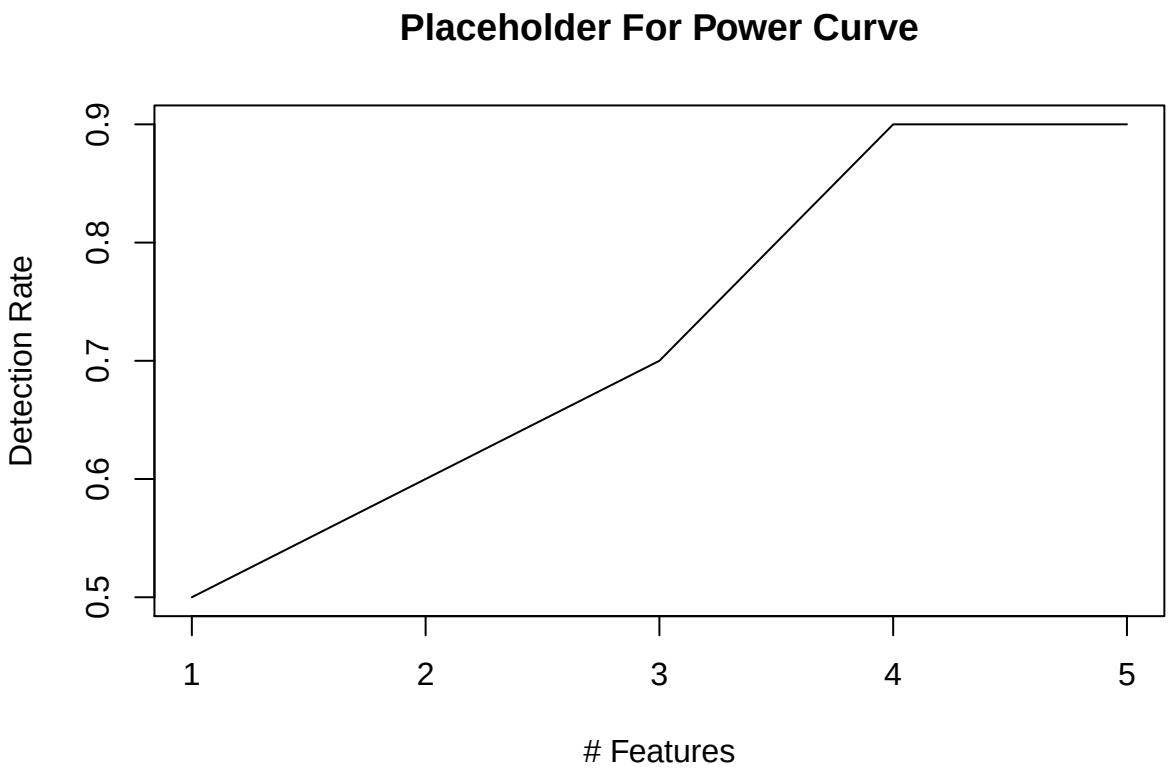
In order to determine a notion of how severe an outlier/influential point might be, we investigated several methods for quantifying the extremeness of an observation in the multivariate or high-dimensional space. We calculate the Mahalanobis distance (Mahalanobis, 1936) for the influential point as compared to the original distribution of multivariate normal points (without regard to the different group labels).

Given a vector $\vec{x}$ with p elements,the Mahalanobis distance is calculated as follows, given mean mu and covariance matrix Σ:

$$D^2 = (\vec{x} - \mu)^t \Sigma^{-1} (\vec{x} - \mu)$$

The mahalanobis distance can be calculated across each of the simulated datasets and compared to get a sense for how the efficacy of the PERMANOVA-based method changes when influential points are only mildly outlying compared to very extreme points. *Ideally, we would see detection rates that are substantively larger when assessing influential points that have a larger mahalanobis distance than those that have a smaller Mahalanobis distance. This allows for us to assess a power curve to determine how the method is able to detect differences of different magnitudes.*

```
#show plot with number of features on X axis and detection rate on y axis

plot(c(1,2,3,4,5), c(.5,.6,.7,.9,.9), xlab = "# Features", ylab = "Detection Rate", type = "l", main =
```

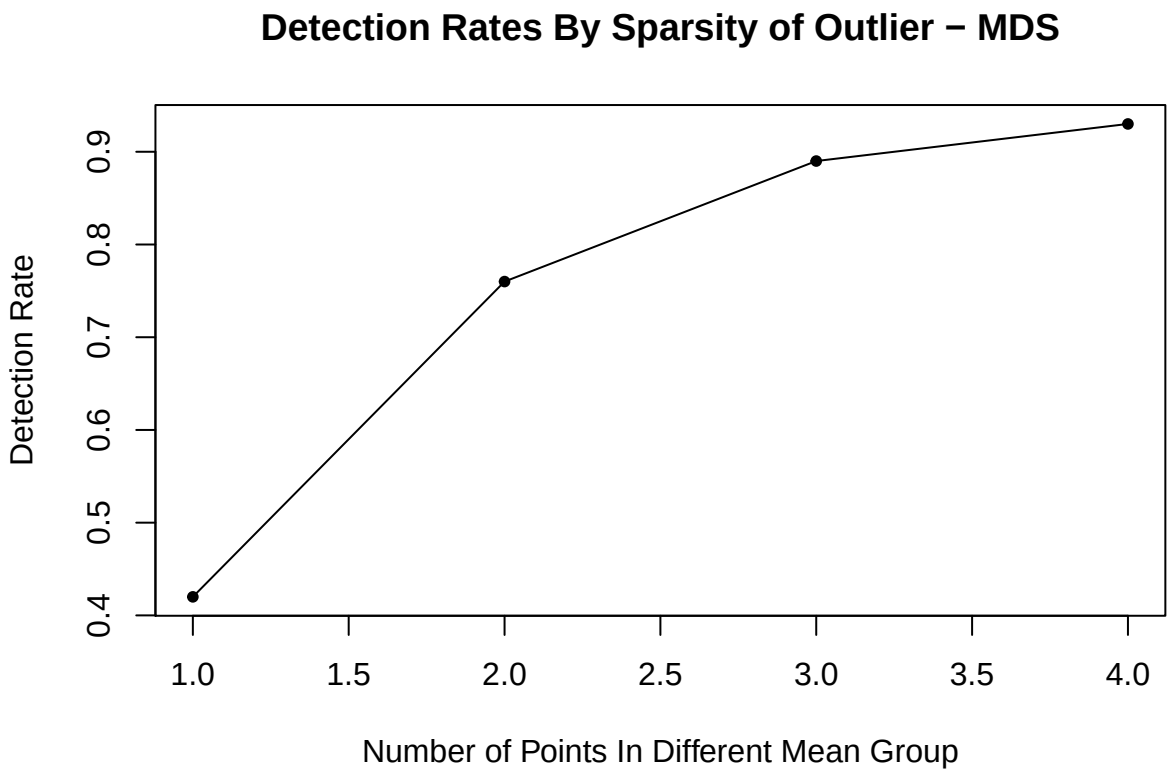


Sparsity of Influential Features

The number of features on which an observation is anomolous is potentially an additional critical element, and can impact how the method detects influence. Per the previous simulation set up, we simulated observations from the central group with means that differed on a single feature. We then increased the number of features that differed from the central group mean up to $p - 1 = 9$ features to see how the method would detect differences based on influential points that are only influential on a small number of covariates.

```
#show plot with number of features on X axis and detection rate on y axis

df3list <- readRDS("C:/Users/paulh/Documents/Doctoral Work/InfluenceInMaps/2021-11-21_SparsitySimulatio

TF <- matrix(TRUE, nrow = length(df3list[[1]]), ncol = length(df3list))
templist <- list()

for (i in 1:(length(df3list)-1)){

  temp <- rep(TRUE, 100)
  ## Identification Rate at given K
  for(j in 1:length(df3list[[1]])){
    temp[j] <- tail(df3list[[i]][[j]]$PValues,1)<0.05

  }
  TF[,i] <- temp
  templist[[i]] <- sum(temp)/100
```

```
}


plot(x = c(1,2,3,4), y = unlist(templist), type = "l", xlab = "Number of Points In Different Mean Group
points(x = c(1,2,3,4), y = unlist(templist), pch = 20)
title("Detection Rates By Sparsity of Outlier - MDS")
```



In truly high dimensional data, it is possible that an observation may only differ substantively on a small number (say 1 or 2) of features in a dataset with potentially thousands of dimensions. The computational cost of PERMANOVA on such a large dataset would be expensive in such a case.

Perplexity

We also assessed the sensitivity of t-SNE's ability to identify influential points at various different levels of perplexity.

Sample Size and Dimensionality

Two factors have the potential to greatly impact the computational complexity of generating maps: the dimensionality of the dataset and the sample size considered. For each of the mapping techniques examined, we assessed the

The PERMANOVA step in the method can use either a pre-specified number of permutations, or it can be based on a permutation matrix that includes each of the permuted indices as rows. In small sample

size situations, we utilize an exact p-value that is calculated using all of the possible permutations of the indicator variable as the explanatory variable consists only of a single 1 and the rest 0's. Thus, there are only n distinct permutations to explore. In larger sample size scenarios, it may not be computationally feasible to consider every possible permutation of the indicator; instead, a raw number of permutations can be calculated and an estimated p-value can be calculated.

Overview of Real Data Results

MNIST

The MNIST dataset...

Goal is to apply the diagnostic to a sample of the MNIST dataset and identify any observations that appear to be highly influential.

Then, re-examine the images that are surfaced and see how consistent they are across both methods.

Penguin Data

The Palmer Penguin dataset (Horst et al., 2020) contains information pertaining to three different species of penguins. The dataset is intended to be an alternative to the ever-popular iris dataset (cite) as each of the penguins form relatively distinct clusters when measured on a series of biometric markers.

Overview of Penguin data.

Additionally, we simulated an additional penguin that stands out significantly from the additional penguins. The penguin is 5 standard deviations above or below the average for each of the measurements.

Carnegie Classification

The Carnegie Classification (CC) is a method for classifying institutions of higher education based on several metrics of institutional productivity. The 2015 version of CC was based on data ranging from doctorates awarded in STEM, Non-STEM, Humanities, Social Sciences, and Other (professional) fields, as well as research expenditures in STEM and Non-STEM fields and two proxies for institutional size (tenurable faculty headcount and research faculty headcount).

The Carnegie Classification is typically based on a ranked set of institutional features due to the presence of outlier institutions on several metrics. In particular, some institutions spend a great deal of expenditures on research compared to others, so the actual CC is based on two indices generated using Principal Component Analysis on the ranked data (Harmon, et al 2018; Kosar and Scott, 2017). This has a similar effect on the results as to a log transform or standardization process (Kosar and Scott, 2017).

Initially, we investigate scaled versions of the unranked data in part to assess whether or not some of the relative extreme points appear as influential relative to other universities in the group. Interestingly, even with some scaling performed, we see that for MDS, the institutions that are surfaced as "influential" are those that seem to fit the bill of being different from the average profile. The institution with the largest F statistic is Harvard University, followed by:

INSERT A PCP of the institutions identified as significant

Without Scaling or Ranking

Without scaling or ranking the data at all, the results look even more interesting.

Discussion

The proposed method uses a similar framework as to the tools available in regression in order to identify potentially abberant data points that can impact the shape of maps. Critically, the choice of ordination technique can impact the efficacy and sensitivity of maps to outliers/influential points. Some methods, like MDS, are highly responsive to influential points that do not look like the rest of the data. Maps resulting from these methods will typically make such points obvious.

Alternatively, t-SNE is not particularly sensitive to single points that are highly dissimilar to others in the dataset. It is often not obvious that such observations are present when viewing a t-SNE map. In many cases, the influential data points are pushed towards the center of the resulting ordination, making them difficult to identify.

Depending on the use case and problem to be solved, this sensitivity (or lack therof) may be benificial or potentially problematic. In some instances, it may be necessary to create maps that are relatively stable even in the presence of influential points; the results in this paper suggest that t-SNE might be more effective for creating maps in those situations. However, care should be taken when interpreting results from such maps as there may be substantively different observations that get grouped together in the resulting ordination.

Much like in regression, we recommend taking an iterative approach to dealing with influential points. It is common practice in regression diagnostics to identify the point with the highest Cook's distance, remove it, and re-fit the model to see the new dispersion of observations' impacts on the model. A similar approach might be useful when creating high-dimensional data maps using these techniques.

We recommend identifying the observation with the largest pseudo-F statistic, removing it, and re-assessing the PERMANOVA results with the reduced set. By cleaning out some of the most influential points, the resulting maps should be more reproducible, of higher fidelity, and provide potentially more meaningful results.

Future Work

There are several considerations for future work that might be worth considering. Based on the differences in t-SNE maps produced by Campell, Caudle and Hoover (2019), an exploration of the impact of different intermediate dimensionality reduction step in t-SNE could be interesting. It is certainly possible that using PCA as an intermediate method for reducing the data dimensionality prior to t-SNE, or choosing a different dimension reduction technique, could either enhance or hamper the ability of this PERMANOVA-based method to identify influential points.

For several of these methods, we initially examined metrics that assess the individual contribution of each feature to the cost function for specific mapping techniques. Interestingly, for t-SNE, the individual costs contributions for the points that were simulated as influential were relatively small, and observations that had high individual cost were often not obviously outlying in the maps created. Some additional research into the dynamics of individual cost contributions - and how tightly these tie to the resulting changes in ordinations when points are included/excluded from maps - could provide an interesting avenue for future work.

Additionally, examining additional mapping techniques may be worthwhile as well. While this paper assesses a linear method for doing dimension reduction (Classical MDS) and a non-linear method (t-SNE), there exists a bevy of other methods that might be worth examining, including UMAP, Sammon Mapping, Isomap, and others.