Formal Paper Draft

Influence Diagnostics for High-Dimensional Ordination Techniques

Paul Harmon

8/21/2021

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 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:

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

t-SNE is an improvement over Stochastic Neightbor Embedding (SNE), which makes use of Gaussian distributions to model both p and q make preserving local structure somewhat difficult. 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 datato 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.

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.

Multidimensional scaling attempts to minimize a metric called "stress" which can be defined generally as:

$$S_M(z_1, z_2, ..., z_N) = \sum_{i \neq i'} (d_{i,i'} - \|z_i - z_{i'}\|)^2$$

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:

$$S_c(z_1, z_2, ..., z_N) = \sum_{i,i'} (s_{i,i'} - (z_i - \bar{z}, z_{i'} - \bar{z}))^2$$

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

Additional Mapping Techniques

Placeholder section in case we want to add another mapping technique in.

- SNE
- Sammon Mapping
- Isomap

Defining "Influence" in High Dimensional Maps

Influence in Regression - Cook's Distance

In regression models, 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 need not be an influential point. An influential point simply needs to 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.

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 observation, 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, cutoff values are typically assessed, or observations with relatively large Cook's distance values are considered for investigation of influence.

```
#image of an influential point
x = rnorm(100, 10, 3)
y = 3 + 2*x + rnorm(100, 0, 1)

lm1 = lm(y~x)
par(mfrow = c(2,2))
plot(lm1)
```

Outliers and Influential Points 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.

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).

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

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.

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

We will discuss correcting p-values for multiple testing later, although currently our plan is to assess using Benjamini-Hochberg False Discovery Rate or Bonferroni corrections (or something similar, assuming that the results are based on p-values as in the Bonferroni Outlier test (Cook & Weisberg, 1982)). We would expect that if the maps are suitably distorted, or *influenced* by a single point, the resulting p-value for the jth adonis test would be small.

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 (contaminated) dataset, runs the t-SNE and MDS mapping procedure, and tests the consistency of the configurations for each of the n holdout points. Several versions of the simulations were assessed, including 1) 1 group with 3 "influental" points, 2) simulation with 3 groups with 3 "influential" points with different variances, and 3) simulation with 3 groups with 3 influential points with same variances.

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

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 scaled version of the data 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 (Cite Harmon, Kosar).

For this analysis, we instead look at 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.

```
CC2015 <- read.csv("https://raw.githubusercontent.com/paulharmongj/Carnegie_SEM/master/data/CC2015data.

CC2015_filter <- CC2015 %>% filter(BASIC2015 %in% c(15,16,17)) %>% select(PDNFRSTAFF, S.ER.D, NONS.ER.D
perplexity = 30

## NEED TO STANDARDIZE THE DATA FIRST (MAKE UNITLESS)
CCscale <- CC2015_filter %>% lapply(scale) %>% as_tibble()

out <- MultiPermanova(CCscale, maptype = "MDS")
```

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.

Key Questions to Answer

Is there a way to look at the individual contribution of each feature to the influential points? (simulate this?)