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 necessary to generate a 2 or 3-D mapping of some higher-dimensional space, projecting the high-dimensional data into an ordination that can be printed on paper. 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.

High-dimensional Data Visualization Tools

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 need not simply be Euclidean. Classical multidimensional scaling returns maps in teh same metric as the original distance or dissimilarity matrix.

Sparse t-SNE

Cross Validation and Mantel-esque Test

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 jacknifing observations is done for all the observations in the data.

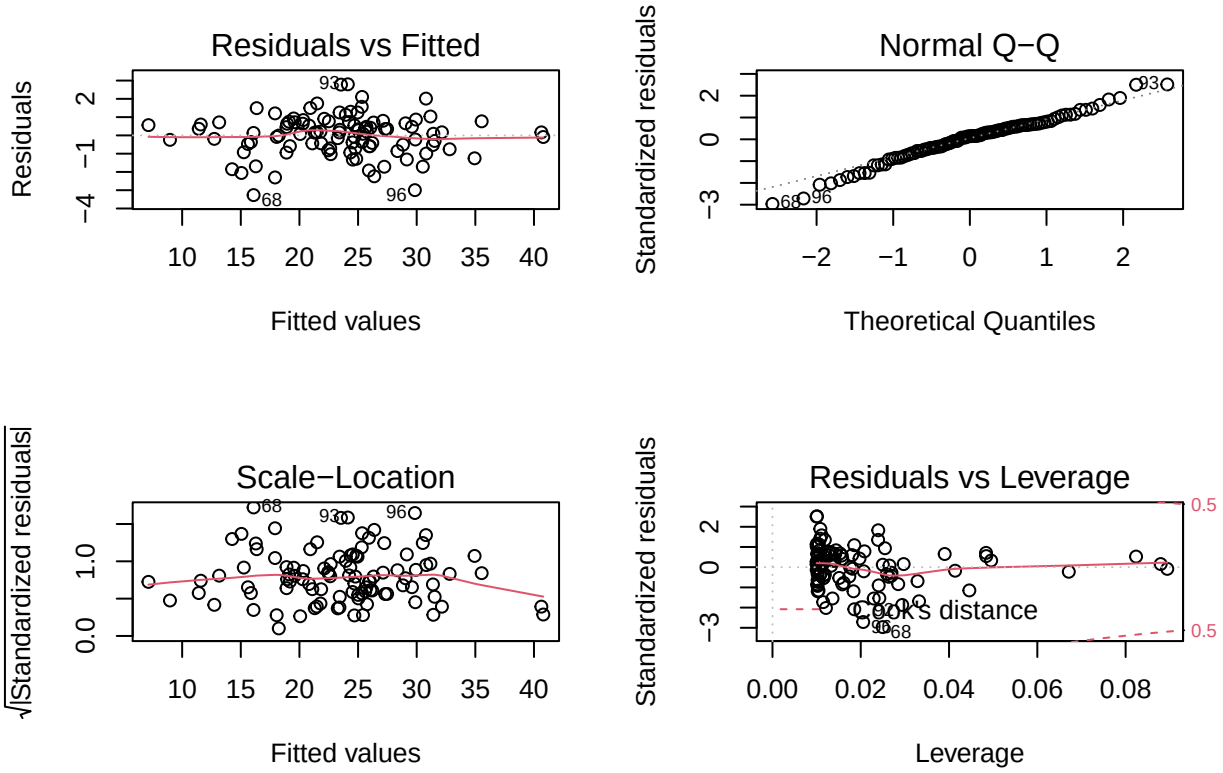
Cook's distance is defined as follows for the ith observation, with$\hat{j}_j$ the fitted response value for the model withthe ith observation omitted, s^2 defined as the mean-squared error for the model and p the number of variables for each observation:

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

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

Similarly, 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.

Difference between outliers and influential points.

talk about influence functions (Joliffe and cited papers)

Joliffe (2002) overviews tools for identifying influential observations in Pricipal Component Analysis, which is a related dimension-reducting technique to MDS (at least in a Euclidean setting). 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.

This section overviews some work into identifying multivariate outliers that we've looked into. We have not found much in the way of users applying Mantel tests or Permanova-type tests on t-SNE or Classical MDS results in the way that we are thinking about things. In fact, much of the literature around robustness and 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.

Influential Points in Maps

Method Overview

This document assesses the effect of multivariate outliers on the maps produced by two methods: classical multidimensional scaling and t-SNE. In general, we seek to identify either **multivariate outliers** (or groups of them) or alternatively single influential points that, for some reason, cause large differences in the shape of the resulting ordination.

Our Permanova-based approach is as follows:

1. For each observation in the dataset, remove it, and generate a t-SNE map in 2 dimensions from the resulting (n-1) observations. (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 $1 - \|\sqrt{stuff}$.

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). 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

Overview of Real Data Results

Penguin Data

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.

Discussion

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 the case of other methods, like MDS, influential points change the configuration of the ordination significantly.