Lab 2 - Linguistics Data, Stat 215A, Fall 2020

10/8/20

1 Introduction

Language is fundamental to the human experience. It allows us to communicate complex thoughts and ideas to each other while maintaining our identity and sense of self. Dialectology studies dialects, the different spoken varieties of the same language, and dialectometry aims to use computational methods in dialectology. This report aims to understand how clustering methods can be applied to dialectometry. To this end, we will investigate how a few questions relate to geography, apply several clustering methods, and test the stability of one interesting clustering result. In the domain context, we hope to discover groups of people with similar dialects and understand other similarities within those groups, namely geographic.

2 The Data

The data come from a survey of preferred terms, phrases, and pronunciations conducted across the United States. The survey includes more than one hundred questions, but the data contain 71 of these questions. These questions include both pronunciations and phrase usage, two important parts of distinguishing dialects. Nearly fifty thousand people were surveyed. The data are directly related to the domain by giving a look at the dialects of people across the country.

2.1 Data Cleaning

The dataset is mostly full and does not contain many severe gaps. However, there are several rows that contain little information to no information. In these rows, the survey participant did not answer many of the questions, so the row contains mostly zeros. These errors are removed in the process of analyzing data, dimension reduction, and clustering, so I did not clean the data up front. For example, when doing PCA, rows with only zeros are removed. I will describe the data processing for each task in their respect sections.

Another issue with the data is the most common answers to several questions is 'other', 'I have no preference', or something similar. The data from these rows appear as though they contribute information, but vague answers do not give a better understanding of the dialect. However, this issue cannot be rectified post-hoc.

2.2 Exploratory Data Analysis

To determine which questions I though were interesting, I looked through the list of questions in `quest.use`. The first two that seemed interesting to me were questions 66 and 121. Question 66 is 'What do you call a miniature lobster found in lakes and rivers?' and 121 is 'What do you call gawking at someone lustfully?' I found 66 interesting because I've heard the most common answers (crawfish, crawdad, crayfish) used interchangeably in the South where I'm from. For 121, I didn't think there were multiple answers, so I hoped to find out where the distinction lies. I later added question 100 ('Do you cut or mow the lawn or grass?') after doing some clustering and seeing that the questions I chose did not entirely reflect the clusters I discovered.

Figure 1: Map of the most common answers for questions 66, 100, and 121

To plot the questions meaningfully, I only plotted the most common answers which I defined to be answers with at least ten percent of survey population. We can see from the maps above that 66 and 100 have significant geographic biases. In particular, we can see a distinct split between the south, midwest, and north in 66: southerners tend to use crawfish, northerners crayfish, and midwesterners crawdad. However, in both 66 and 100 there is mixing in several states, notably the area around West Virginia, Pennsylvania, and Ohio. The answers to 121 don't seem to have much relation to geography. These maps help us see that there is some geographic encoding in some of the answers, but some answers do not depend on geography.

3 Dimension reduction methods

To start out, I decided to try as many dimension reduction techniques as I could, even techniques whose assumptions I think fail. I wanted to get an idea of how the data are structured, and figuring out which dimension reduction techniques work was a part of my process.

The first step in many dimension reduction techniques is deciding whether or not to scale the data. I decided to try both, but I feel that scaling is not necessarily appropriate in the domain context Our data are binary and scaling the data ignores this fact. It can be argued that scaling the data can help to represent the continuum of language rather than the binary that the data present, but the survey does not allow us the information to make this leap.

It is also important to note that I do not feel that the data squarely fit the assumptions of all of the techniques I used. As a result, I operated on the basis of the data being 'good enough' for the methods I used. However, 'good enough' in this report is guided more by convenience and time than by empirical rigor.

First, I tried PCA. PCA is the most commonly used method of all, and here it was the easiest to implement. I looked into both scaled and unscaled results as shown in Figure 2.

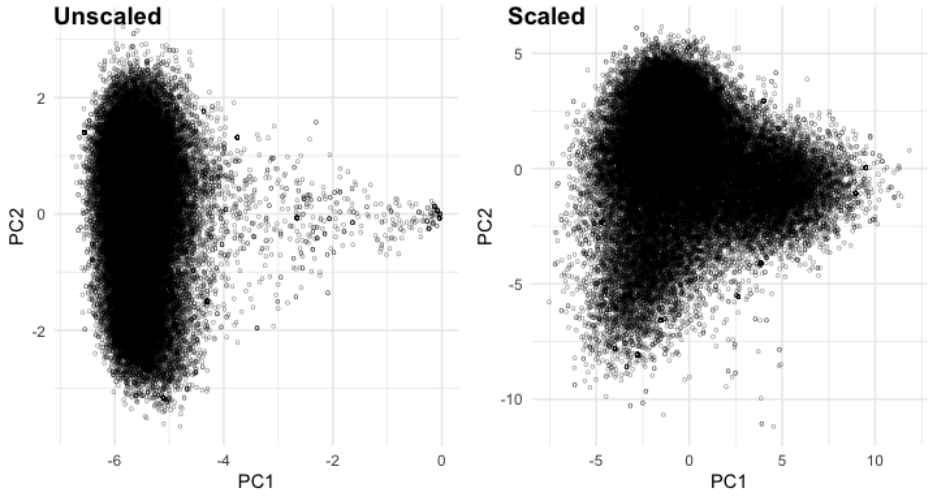

Figure 2: Unscaled and Scaled PCA

It seems that the scaled version does a better job of reducing the dimensions and showing us the shape of the data. We can see potentially three clusters already from the scaled PCA plot. The unscaled plot on the other hand does not show us much in terms of potential clusters. We can only see one central group flanked by a few dozen stragglers. For these reasons I decided to use scaled PCA to plot results in later analyses.

Next we tried t-SNE. t-SNE is more difficult to work with because it requires choosing hyperparameters, but it has shown promise when properly tuned. t-SNE is sometimes implemented with PCA, so we included both t-SNE with PCA and t-SNE without PCA.

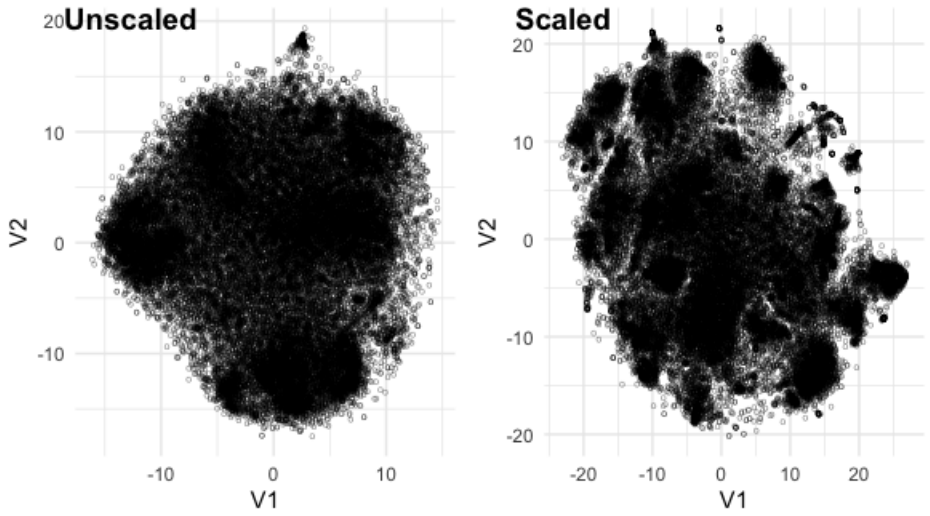

Figure 3: t-SNE w/ PCA

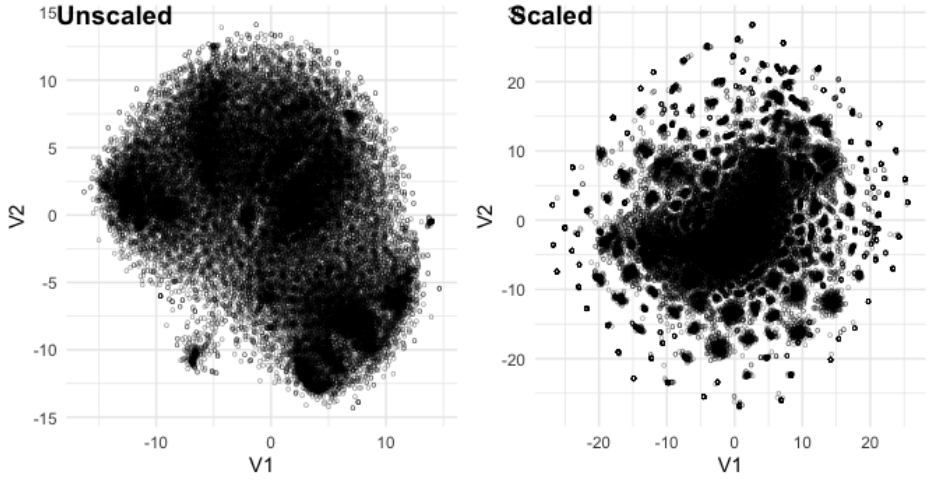

Figure 4: t-SNE w/o PCA

From Figure 3 we can see that t-SNE is attempting to find sub-groups within the larger overall group. In the scaled plot we can see a few groups separating away, and in the unscaled plot we can see some areas with higher density. If we compare these result with Figure 4, we can see a stark difference. Scaled t-SNE without PCA seems to be finding groups and pulling them out of the large central group. Because t-SNE does not preserve local structure, there may be hundreds of points in what seems like only a few dozen spots. PCA may be doing some regularizing here, preventing t-SNE from looking too deeply into the data.

4 Clustering

I tested clustering using a variety of methods, but the only method that was both scalable and stable was k Means. Because of this, I decided to try k Means in conjunction with different dimension reduction techniques. To this end, I tested k Means on the data without any dimension reduction, the data with PCA, and the data with t-SNE (without PCA).

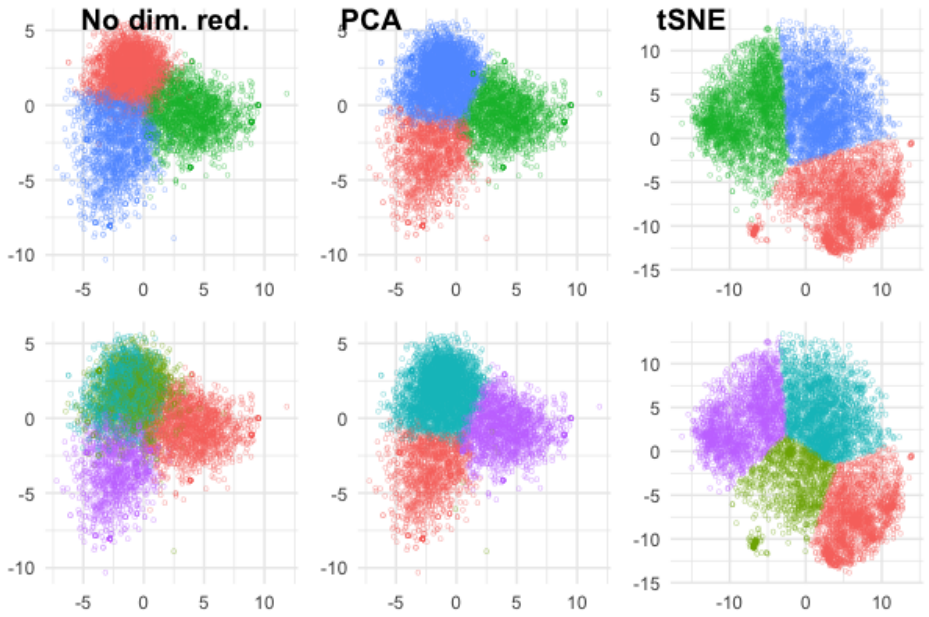

Figure 5: Clustering Results

I initially chose four clusters based on my knowledge of US dialects. I believed that there would be a cluster for the south, the northeast, the midwest, and the west. After testing all the methods, it seemed that three clusters would be more appropriate. Figure 5 shows us that the raw data and PCA cannot produce 4 clusters reliably, however t-SNE can.

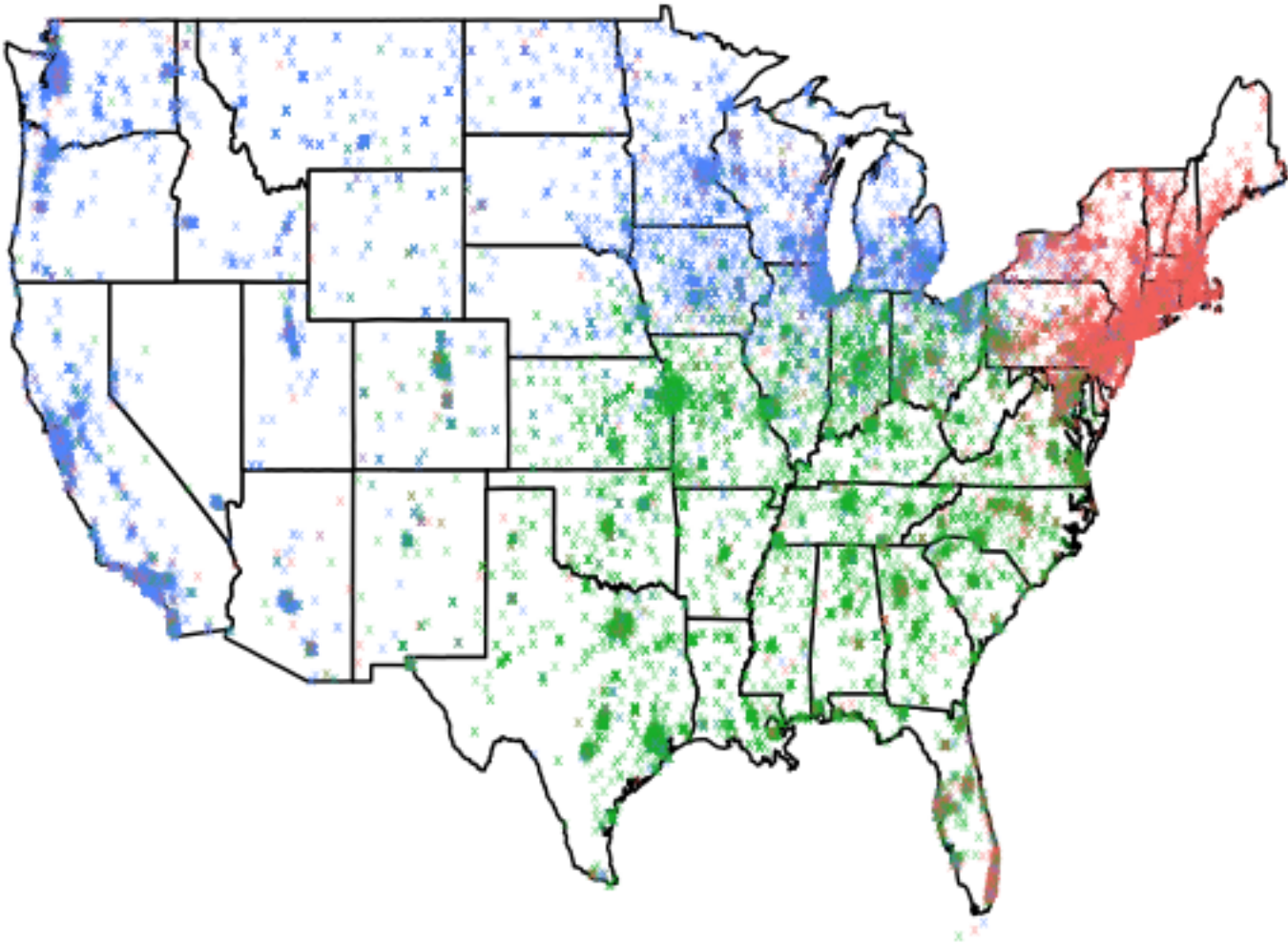

Figure 6: Clustering Map for tSNE

To investigate how people group together, I plotted the clusters over a map of the US. We can see from Figure 6 that there is noticeable separation between the south, northeast, and midwest/west. There are some exceptions (e.g. southern Florida), but the separation holds true for the vast majority of cases. The gradient between these region lies where one might expect: Ohio, Indiana, Illinois, Missouri, etc.

I attempted to use other clustering methods, but many of them were not scalable. Some gave acceptable results on sub-smps while others did not. In particular, DBSCAN worked extremely poorly. From my preliminary analysis, it seems that density-based clustering methods worked poorly while centroid-based clustering was more promising. Density-based clustering may be useful with better fine-tuning of tSNE.

5 Stability of findings to perturbation

When I ran k Means with four clusters on the raw data, I initially got four starkly separated clusters. However, after running it multiple times, these result did not stand up. I decided to use this to investigate robustness in my results.

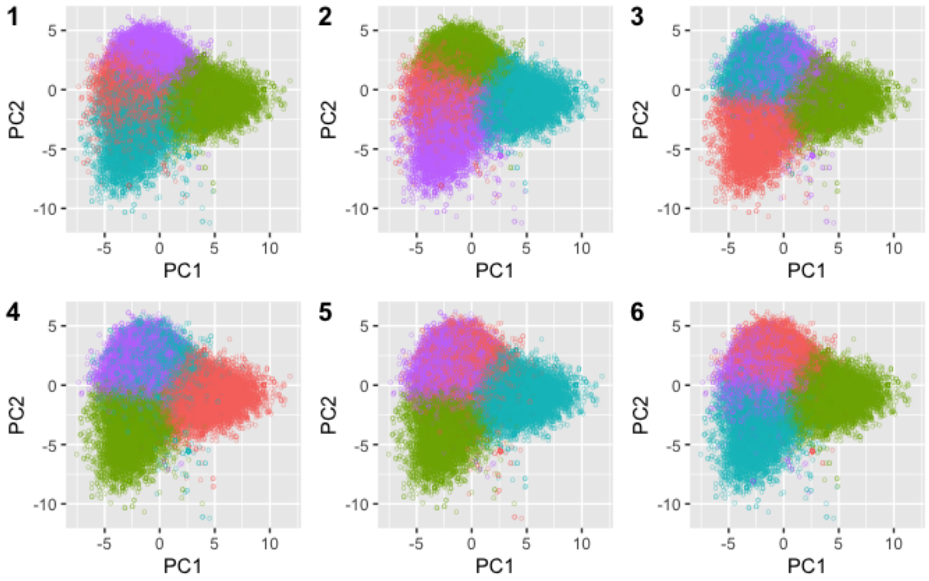

Figure 7: Cluster stability check

Figure 7 shows the six trials of k Means with four centers with the raw data. We can see that the shape, size, and boundaries of the clusters are not stable over our trials. Some trials have clusters that are fairly distinct while others have clusters that are nearly indistinguishable. We can conclude that results with four clusters are not stable with the raw data.

6 Conclusion

I do not think that this data or these results are useful for future analysis or decision-making. The results are fairly predictable; most Americans could likely reach similar conclusions without any data. Additionally, dialectologists have likely known about these results for decades. It is useful to be able to point to an 'objective' source of information, especially in the social sciences, but this is likely outweighed by the lack of detail in the results.

A good reality check could be to simply ask Americans what they think are some broad categories of accents in the US. This check is problematic, but I believe it would be accurate for a small number of groups like we see in our clusters. Otherwise we could talk to a linguist about our results. In particular, they may be able to check the geographical boundaries between clusters.

Given more time, I would have investigated t-SNE more. It seemed promising in being able to give more clusters than PCA, but running the algorithm took a significantly longer time. If I had more time and more computational resources, I would tune the hyperparameters and try to see if I could find even more clusters.

7 Academic Integrity Statement

This work is entirely my own. Academic honesty is crucial to the process of learning and growing, and it is important that all students take their role in the process seriously.