Lab 2 - Linguistics Data, Stat 215A, Fall 2021

Sahil Saxena

October 07, 2021

1 Introduction

Dialects, or linguistic variations, may contain interesting insights about the social structure of a group. As seen in the survey of American dialects, collected by Bert Vaux, dialects can vary according to geography, social class, sex, and age. Consequently, they can potentially reveal the underlying social identity and shared history of groups with similar dialects. The data is a result of surveying over 47 thousands individuals in all states and districts of the USA. The results in this report mostly support traditional ideas of American dialects; using dimensionality reduction and clustering methods, we can learn more about the relationship between dialect groups and regional geography.

2 The Data

The linguistics data is collected by the Harvard Dialect Survey in 2003. The survey consisted of 122 questions that explored the variations in the phonetic and lexical differences in the English language across the USA. This report focuses on analyzing only the 67 questions that explore lexical differences (e.g. "Do you cut or mow the lawn or grass?").

Two different representations of the dataset exist: LingData and LingLocation. LingData includes the categorical responses to 67 survey questions by 47,471 respondents, while also containing location data of each respondent, through state, city and zip code. The latitude and longitude (of the center of each zip code) measurements are also added to each observation.

LingLocation is a binary encoding of the categorical data described above, binned into one degree longitude by one degree latitude squares. This partitions the map of US into 781 cells; the data for each cell is encoded as the sum of binary response vectors for individuals who belong to the cell.

With this data, we can gain insights into how specific responses to specific questions vary over different regions in the USA.

2.1 Data Cleaning

Compared to the previous lab, data quality from this survey is much higher, with fewer missing values and inconsistencies.

For LingData, 1,020 observations were removed for which either longitude or latitude information was missing. This data could have been corrected by inserting the real longitude and latitude for these cities. However there are many missing value zip codes and many discrepancies in the observations that had missing lat and long, such as fictional or overly broad city and town names (e.g. "Chicagoland"). Furthermore, 107 and 98 observations were removed from Hawaii and Alaska, respectively, to restrict the analyses to the 48 contiguous states. The number of samples from these two states is too sparse to impact the entire data set (only 1.06% of total).

Also, some observations have missing responses to some of the survey questions; with a large sample size, a stringent filter was applied to remove any observations with any missing values. Remaining are a total of

39,051 observations for further analysis.

For LingLocation, cells with extreme lat/long values, i.e. longitude < -150 or latitude > 50, (cells from Alaska and Hawaii) were removed, leaving a total of 758 cells. There are also some cases where the number of people who responded to a question in a cell did not add up to the total number of people in that cell. Since this type of missing data was more difficult to filter out, these observations were left in the sample.

Consequently, LingData is used for the main analysis and LingLocation is primarily used to validate these results in the Stability section.

2.2 Exploratory Data Analysis

Here we explore a few survey questions and their potential relationship to geography. Figure 1 shows the distribution of responses, across the USA, to 2 questions: 73 - What is your general term for rubber-soled shoes worn in gym class? and 103 - What do you call the thing from which you might drink water in a school?

Figure 1.Q73 shows the distribution of the two most common answers, which are "sneakers" and "tennis shoes" (removing other answer choices to make interpretability easier). Those who answered "sneakers" are concentrated in the Northeast region while "tennis shoes" is more commonly used throughout the rest of the country; some overlap does exist, which indicates people may have moved from one region to another or began adapting to different terminology in their native regions.

Figure 1.Q103 shows the distribution of the most common answers, which are "drinking fountain", "water fountain", and "bubbler." The term "water fountain" seems to be more commonly used in the West, with a high concentration in southern California. On the other hand, "drinking fountain" seems to be more favored in the Northeast and the Southeast. The Midwest seems to use both terms at similar frequencies. Interestingly, individuals who responded "bubbler" form small isolated groups: Wisconsin, Massachusetts, and Rhode Island; this seems to support the idea that some terminology evolves from local/regional "slang" rather than an official dictionary.

Experimenting with linked brushing (in other/brushing_experiment.html) shows that response to question 73 does not help predict the response to question 103. However, adding another question might give more insight. Specifically, if someone answered "drinking fountain" for question 73 and "tennis shoes" for question 103, then for question 110 (What do you call the night before Halloween?), it is unlikely to call the night before Halloween "mischief night." This suggests that the responses to some groups of questions may be correlated with each other.

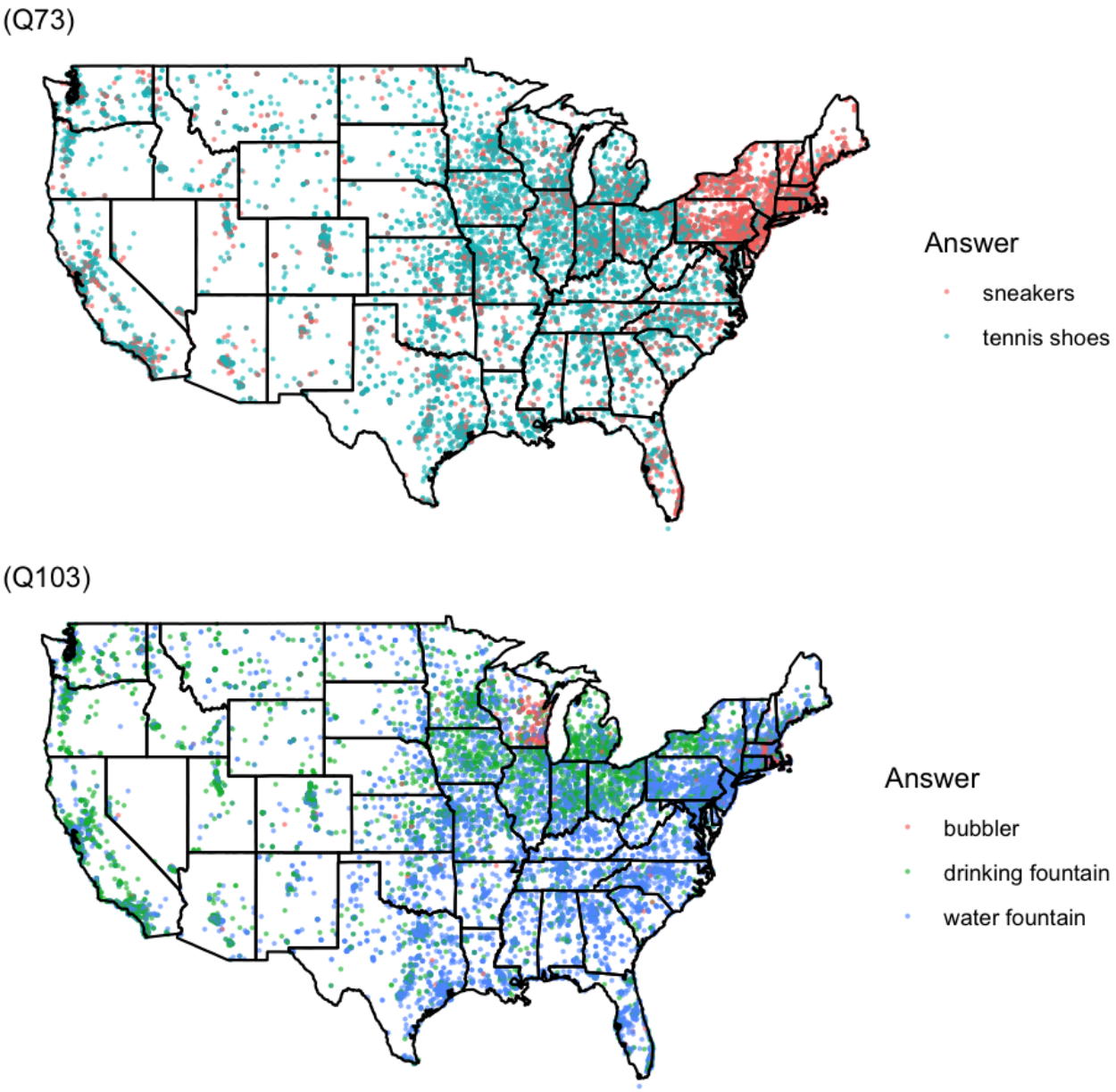

Figure 1: Distribution of responses in the contiguous states. (Q73) What is your general term for rubber-soled shoes worn in gym class? (Q103) What do you call the thing from which you might drink water in a school?

3 Dimension reduction methods

Through the dimension reduction method of PCA, Figure 2 shows the projection of the binary representation of LingData onto the first two principal components. Here, three loose clusters according to longitude can be seen, which may be relating the linguistics results to regional geography. Furthermore, Figure 3 shows that the first ten principal components explain only a modest proportion (about 20%) of the total variability, which indicates that the underlying cause for linguistic variations is not simple; in other words, it is not accurate to portray the data's complexity in just a few features.

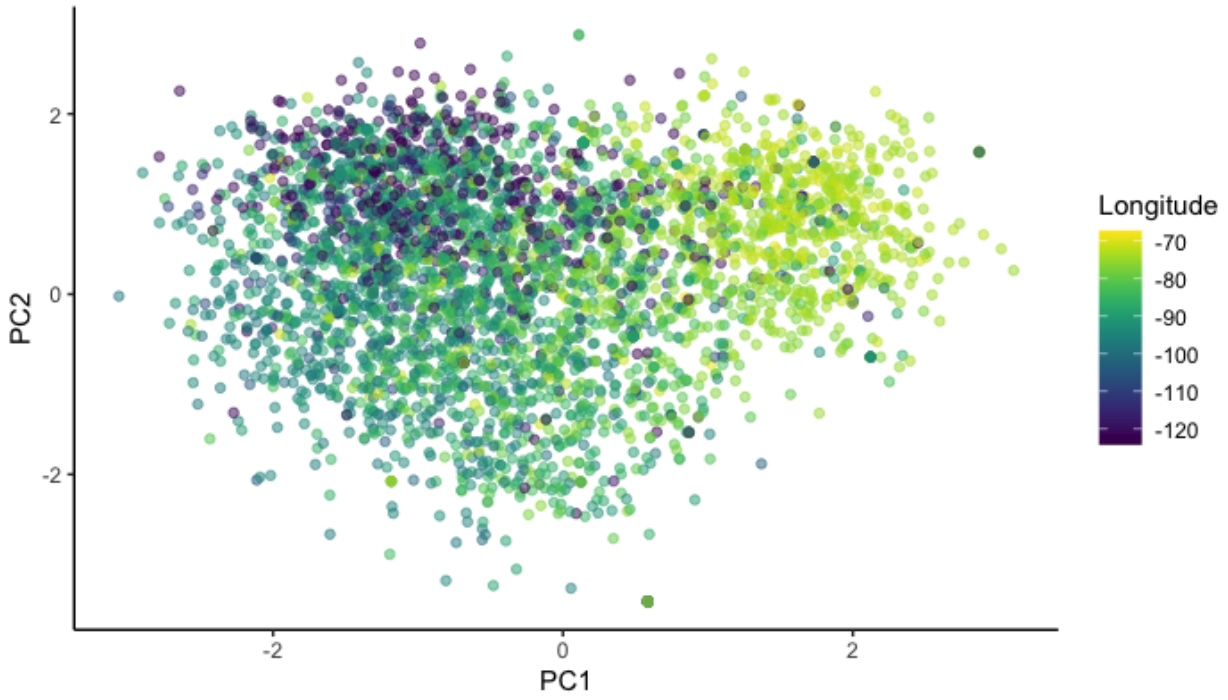

Figure 2: Projection of LingData onto the first two principal components.

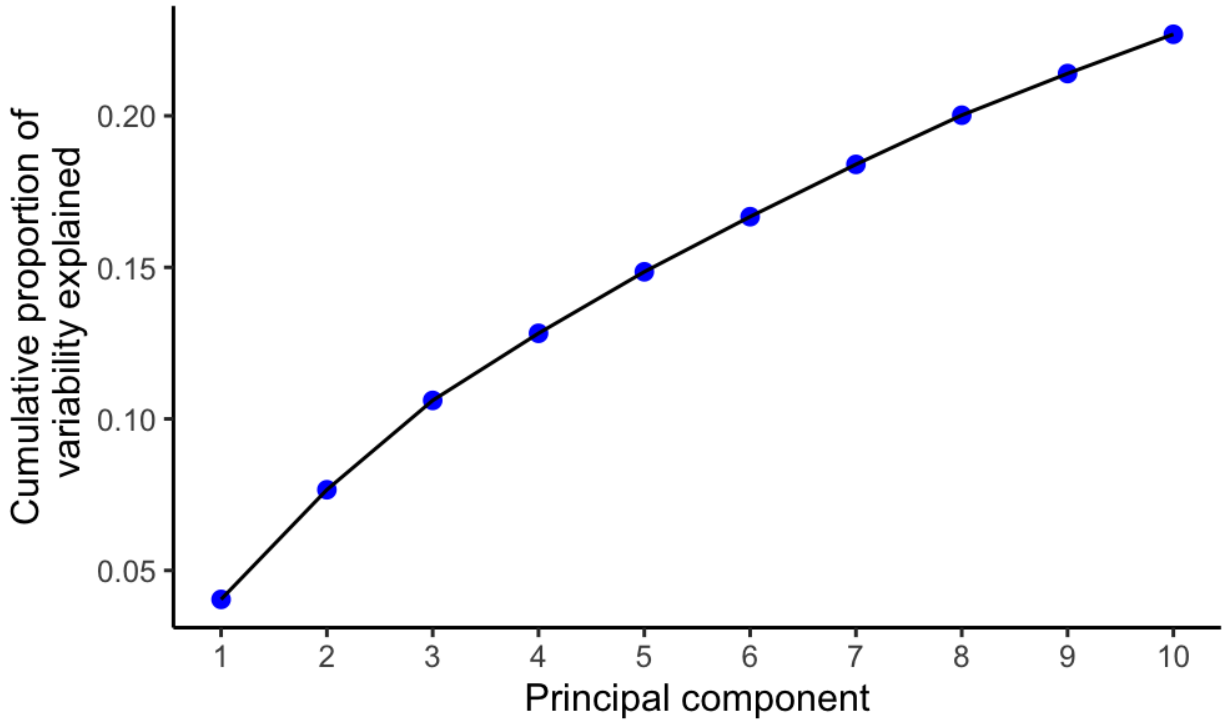

Figure 3: Amount of variability explained by the first 10 principal components.

The data was not centered as covariance matrix calculation in PCA already "centers" the data. The PCA

was tried with and without sacling. With scaling, the first 10 components cumulatively accounted for about 12% of the variance. However, without scaling, the first 10 components cumulatively accounted for more than 20% of the variance. Clearly, PCA done without scaling has less representative dimensions; this could be due to the fact that the observations are already on the same "scale" as they are binary. With naturally unstandardized data (if the data was not binary encoded), then PCA will project the first component in the direction of the questions with higher covariance, which does not make sense. Given that answer choices are just mapped to numbers, this does not provide a meaningful interpretation.

It is not a good idea to do PCA on the original dataset because the values are inherently insignificant. Answer choices are mapped to a particular number, so for different questions, many individuals will have the "same" answer number; this does not really make sense. By encoding in binary, every answer choice is treated as its own dimension. From there, the dimensionality reduction takes place.

4 Clustering

- This is where you discuss and show plots about the results of whatever clustering methods your tried - k-means, hierarchical clustering, NMF, etc.

5 Stability of findings to perturbation

- What happens to your clusters when you perturb the data set?
- What happens when you re-run the algorithm with different starting points?

6 Conclusion

- Discuss the three realms of data science by answering the questions in the instructions pdf.
- Come up with a reality check that would help you to verify your clustering. You do not necessarily have to perform this reality check, but you can if doable.
- What are the main takeaways from your exploration/clustering/stability analysis?

7 Academic Integrity Statement

- Please include your statement here. Do NOT include your name, to avoid putting it in the lab2_blind.pdf document.