Lab 2 - Linguistic Survey
Stat 215A, Fall 2017

3032998360

October 5, 2017

1 Kernel density plots and smoothing

In this section, we use the redwood data from the previous project to plot a density estimate for the distribution of temperature over the whole dataset. Kernel density estimation is a nonparametric way to estimate the probability density function. Let $(x_1, ..., x_n)$ be an independent and identically distributed sample drawn from an unknown distribution f. The estimator $\hat{f}$ is defined as

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big),$$

where K is called the kernel, which accounts for the weight given to the observations x_i at each point x_0 based on their proximity. h is the bandwidth, which greatly influences the performance of the estimator, as illustrated in Figure 1. For example, with the choice of a small bandwidth (1/10), the resulting density is bumpy with a spike at one point, making it difficult to interpret the structure. In contrast, the choice of a large bandwidth (10) results in a density that is very smooth with small variability, masking the structure of the data. Thus, we see that the bandwidth is related to the bias and variance of the estimator; bias is reduced by increasing variance, and vice versa.

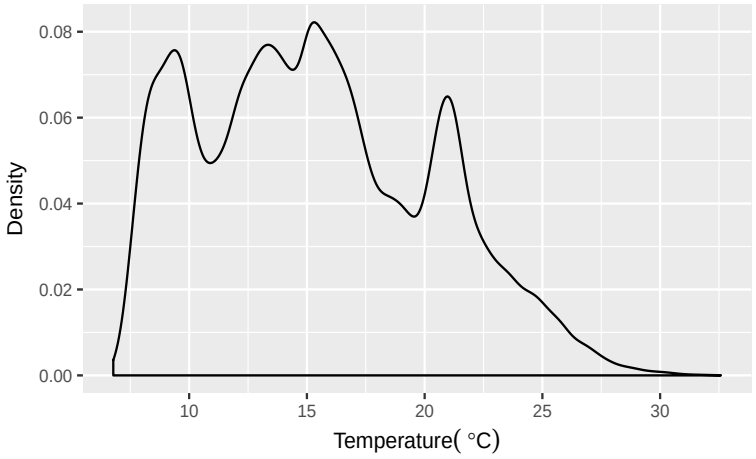

Figure 1: Kernel density plot of temperature (bandwidth = 1)

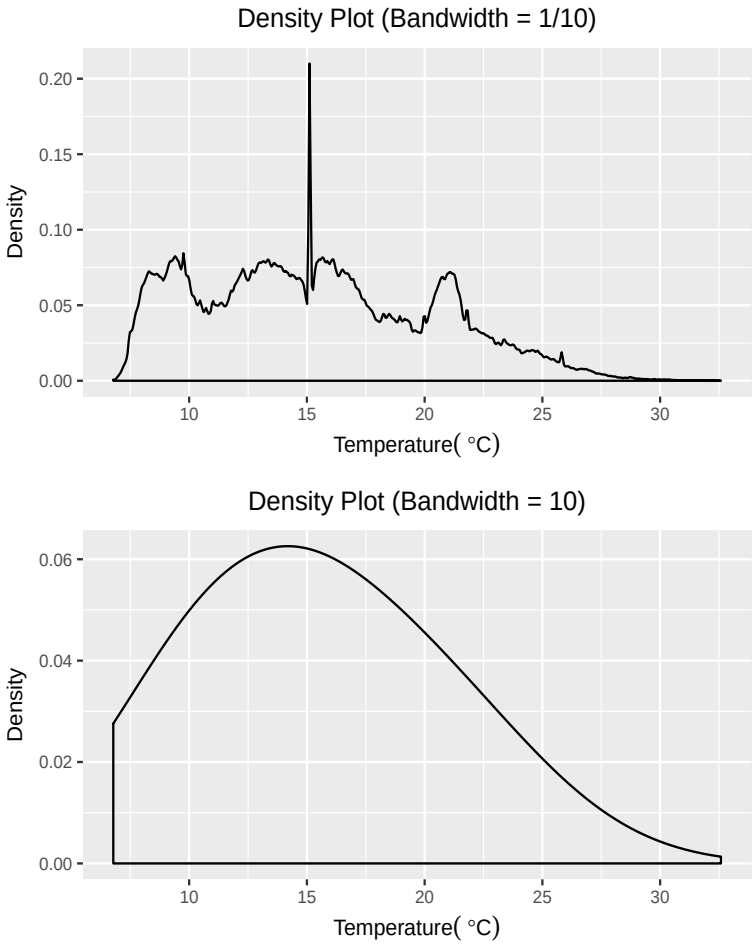

Figure 2: Kernel density plot with varying bandwidths

Next, we plot a bivariate plot of humidity and temperature at 1pm for all nodes during the entire project period (Figure 3), with locally weighted scatterplot smoothing (loess). Loess, also called a local regression, fits a smooth curve through points in a scatter plot. The algorithm takes the following steps:

1. Choose a span (aka smoothing parameter or bandwidth) s in $(0, 1]$. This represents the proportion of observations to be used in local regression. Default s used in R is 0.75.
2. Find the k nearest neighbors to x_0. If there are n observations, then $k = n * s$ points.
3. Assign weights to the nearest neighbors. The weight function gives more weight to observations whose value is close to x_0 and less weight to observations that are farther away.
4. Perform local weighted regression in the local neighborhood of x_0.

If bandwidth is small, there will be less observations near x_0 and thus the resulting curve has a large variance. On the other hand, if span is large, then the resulting curve is over-smoothed and contains a large bias. In fact, in Figure 3, we see that the second plot with bandwidth set to 0.1 is fluctuated and it may look like the curve is over-fitting the data. The third plot with span set to 0.9 is much smoother, but it fails to capture much of the low humidity observations that are clustered around temperature from 18 to 24 degrees celsius, as well as observations whose temperature are above 27. As with the kernel density plot, choice of bandwidth is associated with the amount of variance and bias of the resulting estimate.

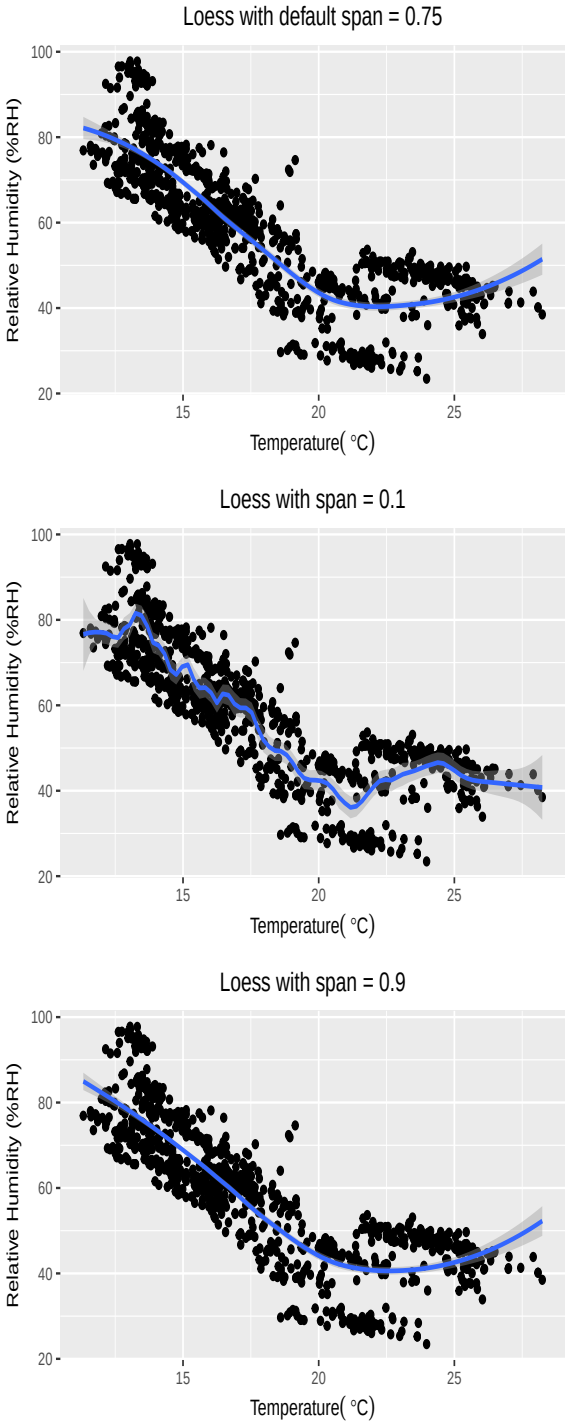

Figure 3: Humidity vs temperature at 1pm for all nodes during the entire project period

2 Introduction of Linguistic Data Analysis

It has been understood that differences in language are caused by various factors, including geographic location, social line, age, and gender. In this project, we explore linguistic variations across different regions of the United States. In Section 2, we will be discussing the quality of the survey data with dialectological

information and how we clean it, followed by an exploratory analysis that discusses the geographical distribution of different dialect questions (Section 3). Since there is a large number of questions in the survey, we will apply a dimension reduction technique and come up with new variables (components) that sufficiently capture the variability in the data (Section 4). Based on observations projected on these components, we then perform a cluster analysis to understand how responses are grouped (Section 5). In the last section, we will check the robustness of our findings.

3 The Data

The dialectological data comes from the Harvard Dialect Survey project in 2002 by Bert Vaux. Participants across various regions of the United States responded to questions on lexical and phonetic differences in dialects. There are 47,471 respondents and 121 multiple choice questions. In this project, we focus on 67 of the questions that are related to lexical differences. Other variables in the data are: city, state, zip, latitude and longitude of the respondent.

The second dataset consists of essentially the same information, except that it provides us, for each square unit area, the frequency of a specific answer choice being chosen. For example, imagine that there are two respondents living in the same square unit area, and they respond to two questions. The first question has 3 answer choices and the second has 4 answer choices. Say we have a vector $(1, 1, 0, 0, 0, 0, 2)$ in our data. From this, we can tell that there is 1 respondent who answered choice A of question 1, 1 respondent who answered choice B of question 1, and 2 respondents who answered choice D of question 2.

3.1 Data quality and cleaning

The linguistic data contains some missing values. Missingness pattern for randomly sampled 1000 respondents is shown in Figure 4 (We used a subsample as creating a plot for as doing this with a full sample is computationally too intensive). We notice that some missingness appears to have a noticeable pattern. For example, from the horizontal red lines that span from Q050 to Q121, we see that there are respondents who only answer the residence-related questions and skip the majority of the dialect questions. It seems reasonable to remove those who do not answer any of the dialect questions.

Another point we notice is that some questions are more likely to be skipped than others. For example, Question 92, which asks "What do you call it when a driver changes over one or more lanes way too quickly?", has about 3800 missing out of 47,500 respondents. This is roughly three times as much as the missingness of other questions. It may be likely that respondents had harder time understanding the question.

While there is no missing values for the city variable, there are three missing for the state variable, which we tried to impute based on information in other variables. When city and zipcode agreed, we were able to fill in the state name. However, for the two cases in which the city name and zipcode did not agree, we kept their state as missing.

City and state variables have multiple misspelled entries. An attempt was made to fix erroneous city and state names as much as possible, but not all were fixed. There are also city names that are outside of the states (e.g. Mumbai, St. Petersburg, Calgary, Trinidad). It is likely that these surveys were taken by international residents or immigrants who were born outside of the U.S. Due to misspelling and out-of-the-states cities, there are nearly 1000 observations with missing latitude and longitude information.

In addition to missing responses, there are also 5 missing questions: Q108, 112, 113, 114, and 116. One possibile reason for the removal may be due to low variability in the responses, but we do not know the exact reason.

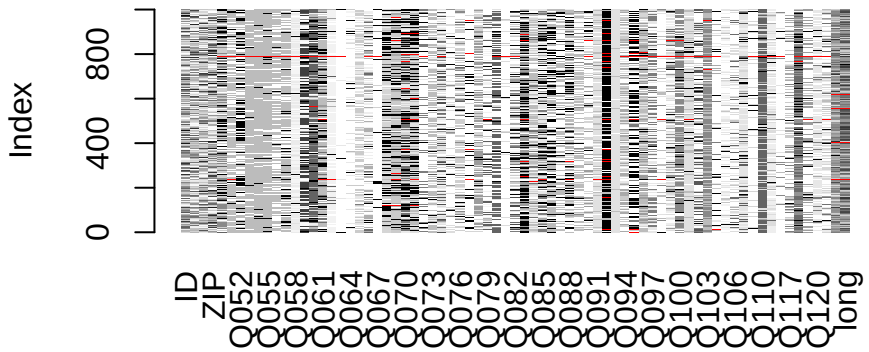

Figure 4: Pattern of missing values. Red indicates missing. Index is the ID of respondents.

3.2 Exploratory Data Analysis

We explore the distribution of answers to each of the 67 dialect questions. The map of the 67 questions can be seen https://giphy.com/gifs/map-l1J9wwozq5yEkIKic. From these plots, we learn that answers to some questions have similar distributions.

For example, responses to questions 105 and 106 appear to have a similar geographical pattern. For both questions, dialects distinctly differ in the Northeast, the Midwest, and the South. Let us examine what the questions are. Question 105 asks "What is your generic term for a sweetened carbonated beverage?" with 10 different answer choices, most frequent ones being "soda" (approximately 53%), "pop" (25%), "coke" (12%), and "soft drink" (6%). Question 106 asks "What do you call the act of covering a house or area in front of a house with toilet paper?", with 8 answer choices, most frequent ones being "tp'ing" (approximately 58%), "toilet papering" (21%), "I have no word for this" (8%), and "rolling" (6%).

By looking at the responses to the two questions together, we learn that popular responses are "soda and toilet rolling" in the northeast, "pop and tp'ing" in the midwest, and "coke and rolling" in the south. Florida, even though it is in the south, agrees with the northeast on both questions. Thus, a response to one question seems to predict the other quite well. There are a few exceptions, though. For example, California agrees with the northeast for question 105 on "soda", and agrees with the midwest for question 106 on "tp'ing". Texas, although it agrees with majority of states in the south for question 105, it does not for question 106. In fact, Texas, near Houston area, seems to be the only region where "wrapping" is the popular response for question 106.

We explore one more question, which is question 66, "What do you call the miniature lobster that one finds in lakes and streams for example (a crustacean of the family Astacidae)?". Popular choice in the southeast appears to be crawfish, while in the northeast and the upper midwest crayfish is popular. Crawdad is the third popular choice, and is mostly concentrated in the lower midwest and upland south. Distributional pattern of this question is different from those of the previous two questions. One main difference is for this question, there are two competing answer choices within people in the south; upland south answers crawfish and deep south answers crawdad.

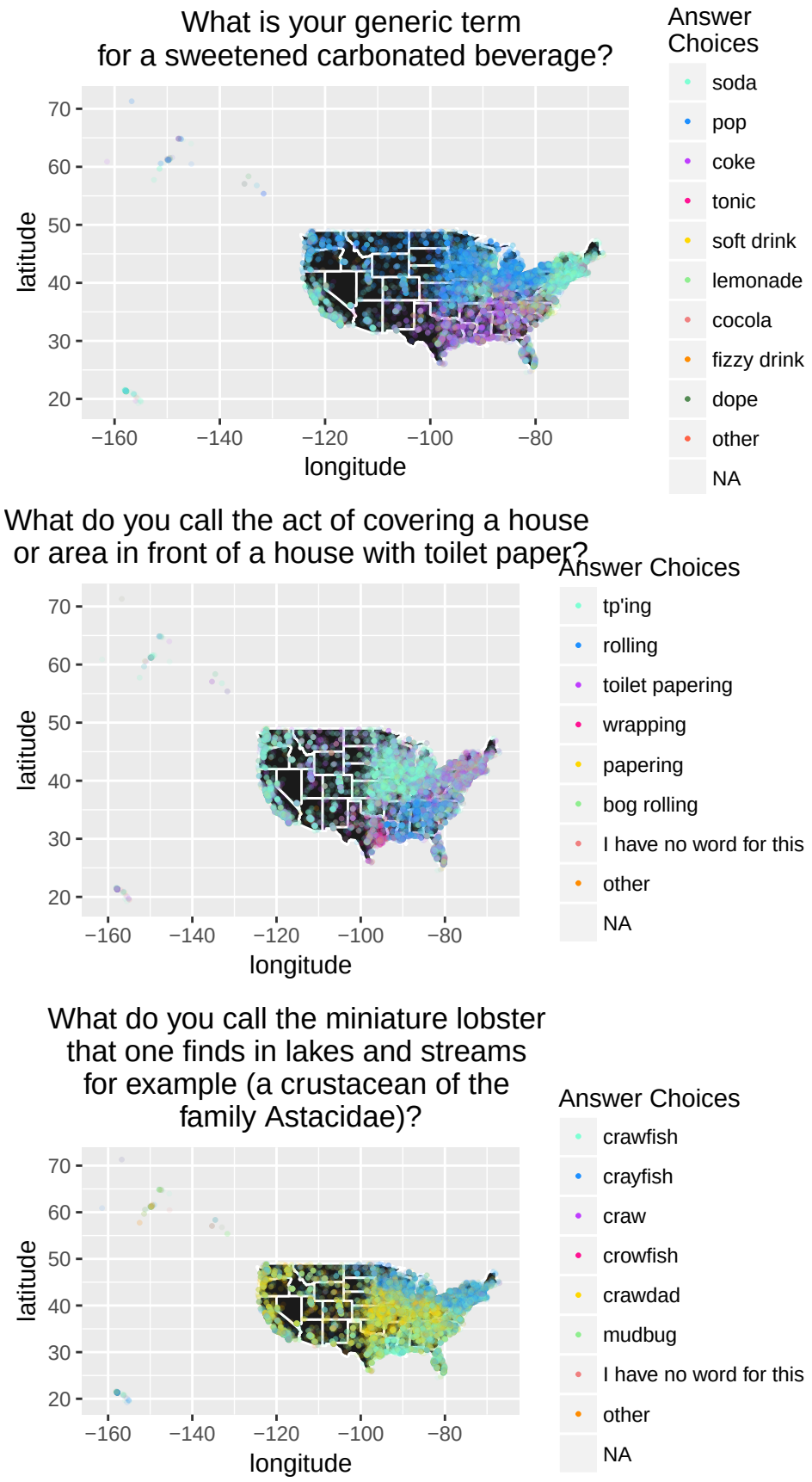

Figure 5: Geographic distributional pattern of different dialects

4 Dimension reduction methods

4.1 Reduction of Dimension via PCA

In previous section, we have explored spatial pattern of responses to three questions, individually. However, there are 67 questions and 468 answer choices in total, thus it is difficult to compare patterns for all 67 questions. In this section, we are goign to reduce 468 answer choices by using principal component analysis (PCA). By doing so, we are able to obtain principal components, i.e. linear combinations of the answer choices, each of which contributes an different amount to the component.

PCA was conducted on a dataset with responses encoded to binary. This dataset is 46,431 by 468. Rows represent the number of respondents (as mentioned in the data cleaning section, respondents who did not answer any of the questions were removed from the data). Columns represent the number of answer choices for 67 questions. Say the first question has 3 answer choices and the second questions has 4 choices. A respondent chooses first answer choice for both questions. Then the first row of the dataset would be $(1, 0, 0, 1, 0, 0, 0)$.

Usually, the first step of the principal component analysis is to scale the dataset. Let's consider a case in which variables are measured in different units and for this reason some variables have much higher covariance values than others. Then PCA will project the first component in the direction of those variables, even if doing so does not provide us with a meaningful interpretation of the component. Unstandardized data is also problematic when conducting cluster analysis later in the section. Large covariance can pose unnecessarily large weights to certain variables. In this dataset, however, we do not see a need to standardize data as all the variables are binary. We will only center the data.

To determine the appropriate number of components, we first look at the scree plot (Figure 6). The plot shows the amount of variance explained by each component. The idea is to select the number of components at which there is a clear break of explained variance. While it is difficult to identify the exact number of components from the plot, it appears that it is at around the 8th component that the variance begins to level out. The first 8 components explain, respectively, 4.1%, 3.5%, 2.9%, 2.1%, 2.0%, 1.8%, 1.8%, 1.7%, and 1.6% of the variance. The 8 components together explain 20% of the variance.

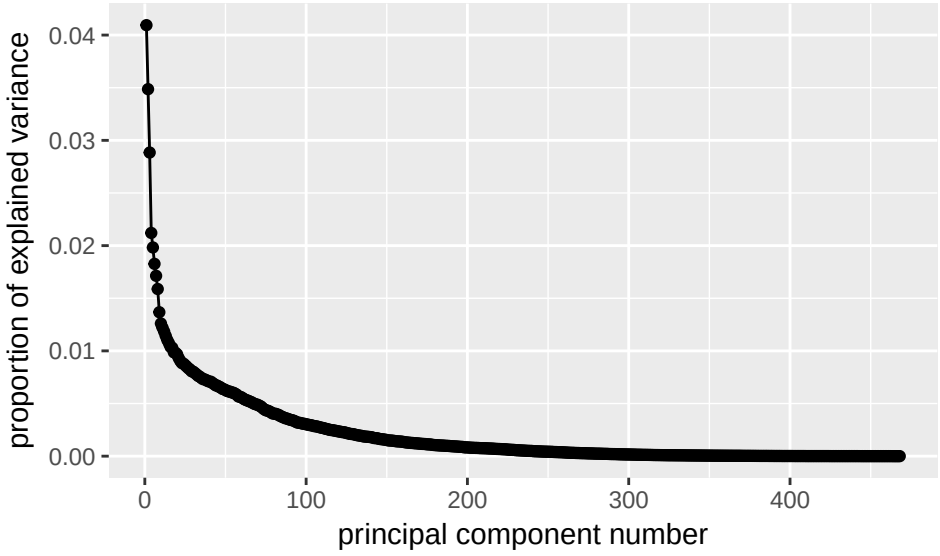

Figure 6: Cattell scree plot

Another rule of thumb for determining the number of components is to select components that cumulatively

explain more than 80% of the variance. However, we see from the eigenvalues that more than 100 components are needed to cumulatively explain 80% of the variance. However, 100 components is too many to have a meaningful interpretation and thus we prefer a smaller number of components. We will for now proceed with the first 8 components suggested by the scree plot.

Component 1 and 2 are plotted against each other to see if there are certain groups of individuals that lie heavily on one axis or the other (Figure 7). While the individuals are not largely separated from each other, when they are labeled according to their region, we do see some groups. For example, component scores for people from the midwest tends to be lower than others, the south has higher scores for component 2, and northeast has higher scores for component 1. Scores for people in the west do not appear to be separated from other regions. Similar comparison is made for component 1 and 3, component 2 and 3, and so on. People from the west is separated the most when component 2 and 4 are plotted against each other (Figure 7).

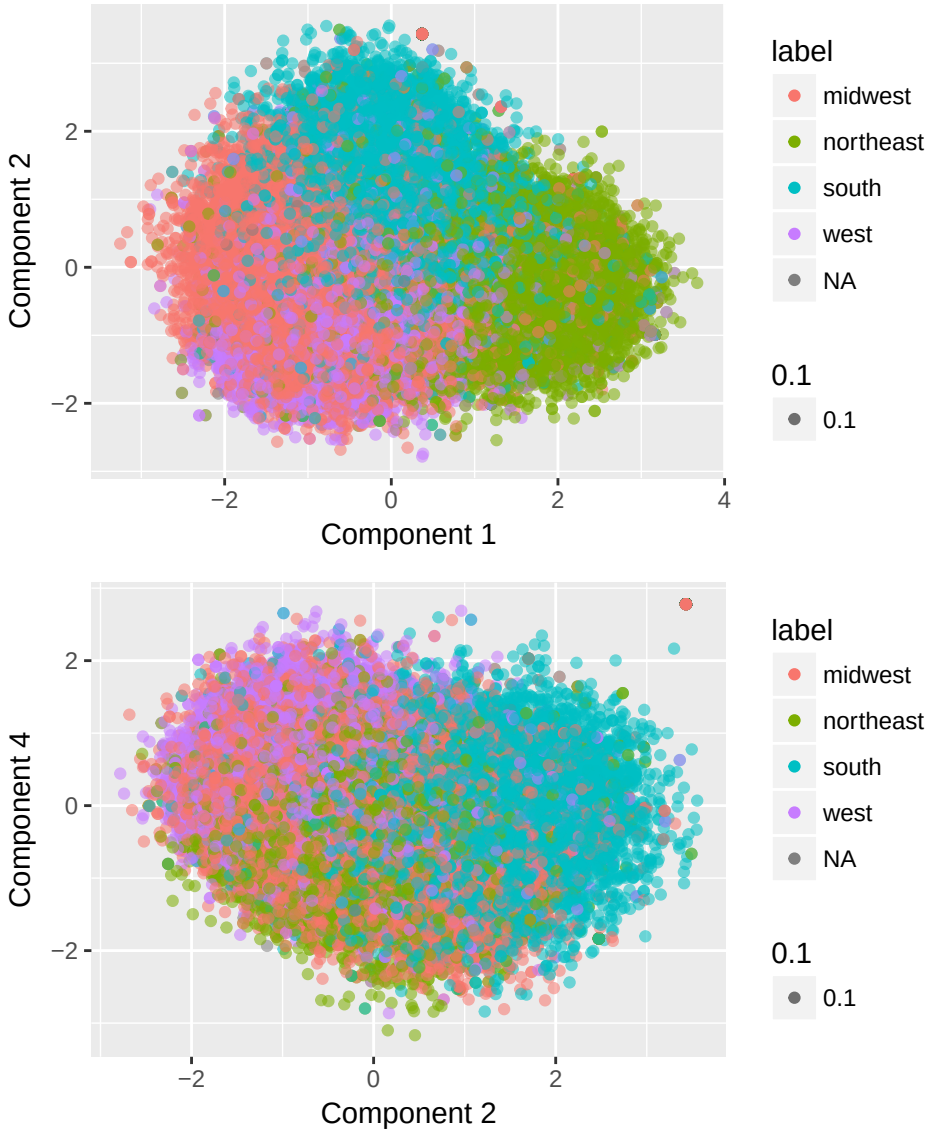

Figure 7: Bivariate plots of the selected two components, labeled by area regions

Now, we look in closer to learn which responses to the survey question are largely contributing to the impor-

tant components. Specificaly, we look at the loadings, which represent the correlation between the response and the component (Table 1). For the first component, we see that response 165 and 170 have high loadings (the values are 0.27 and 0.24, respectively). Response 165 refers to individuals that answered "sneakers" to question 73: "What is your *general* term for the rubber-soled shoes worn in gym class, for athletic activities, etc.?" Response 170 refers to individuals that answered "tennis shoes" for the same question. Since we know from Figure 7 that component 1 differentiates individuals from the northeast and the midwest, we can think of question 73 as a key question that can help differentiate people from these two regions.

For component 2, response 198 and 201 have relatively high loadings (0.25 and 0.24, respectively). The responses are "kitty-corner" and "catty-corner" to the question, "What term do you use to refer to something that is across both streets from you at an intersection (or diagonally across from you in general)?" We have seen in Figure 7 that component 2 is somewhat successful in differentiating individuals in the south from other regions. Thus, this driving related question may be a good question to ask to learn whether the respondent is from the south. Other responses that have high loadings are "water fountain" to the question "What do you call the thing from which you might drink water in a school?" (0.21 loading), and "y'all" to the question "What word(s) do you use to address a group of two or more people?" (0.19)

However, it is difficult to understand an underlying construct of each component. In this case, we see that questions that load heavily on one component do not appear to share a common construct. Some questions are related to product name, while others are related to descriptions of a certain situation or action. Content of the questions also vary within each component, and same questions load heavily on more than one component, making it difficult to identify a shared theme within a component.

The goal of the PCA is to come up with linear combinations of original variables that maximally explain variance, and not necessarily to help ease conceptual interpretability. If variables are moderately correlated with one another, variables that correlate to each other usually end up loading on the same component. In this case, however, data we used to conduct PCA contains sparse, binary variables. Thus variables are not correlated much to begin with. This may be part of the reason why finding a shared feature between responses in the same component is difficult.

Component	Loading	Question	Response
1	0.27	Q75: What is your general term for the rubber-soled shoes worn in gym class, for athletic activities, etc.?	165: sneaker
	0.24	same as above	170: tennis shoes
	0.20	Q105: What is your generic term for a sweetened carbonated beverage?	384: soda
	0.18	Q80: What do you call it when rain falls while the sun is shining?	225: sunshower
	0.18	same as above	232: I have no term for this
2	0.25	Q198: What term do you use to refer to something that is across both streets from you at an intersection (or diagonally across from you in general)?	198: kitty-corner
	0.24	same as above	201: catty-corner
	0.21	Q103: What do you call the thing from which you might drink water in a school?	376: water fountain
3	0.27	Q59: What do you call the game wherein the participants see who can throw a knife closest to the other person (or alternately, get a jackknife to stick into the ground or a piece of wood)?	63: I do not know
	0.21	Q120: What do you say when you want to lay claim to the front seat of a car?	457: shotgun
	0.18	Q50: What word(s) do you use to address a group of two or more people?	4: you guys

Table 1: Questions that load heavily on the first three components

To see how many clusters exist in the dataset, we use K-means clutering technique based on scores for the first 8 components retrieved by the PCA. Bivariate plot of 3 cluster solution seems to correspond to the south, midwest, and northeast represented in Figure 7. In a 4 cluster solution, we see that the 4th cluster (green) overlaps the other two components, similarly to how western respondents overlaps south and midwest respondents in Figure 7.

To help choose an appropriate number of clusters, we refer to Calinski-Harabasz criterion (C(g)). The criterion is computed as the ratio of overall between-cluster variance and overall within-cluster variance. An obvious cluster would have a high C(g) value, because each cluster is contained and well separated from each other. However, C(g) should only be used as a guidance for choosing the number of clusters, because solutions with a high number of clusters can take a high C(g), even if the solution does not have a meaningful interpretation. In this case, 3 cluster solution, which appear to make sense conceptually, has the highest C(g) among 8 clusters we examined (Table 2).

2	3	4	5	6	7	8
160	167	140	130	128	124	103

Table 2: Calinski-Harabasz criterion for 8 clusters

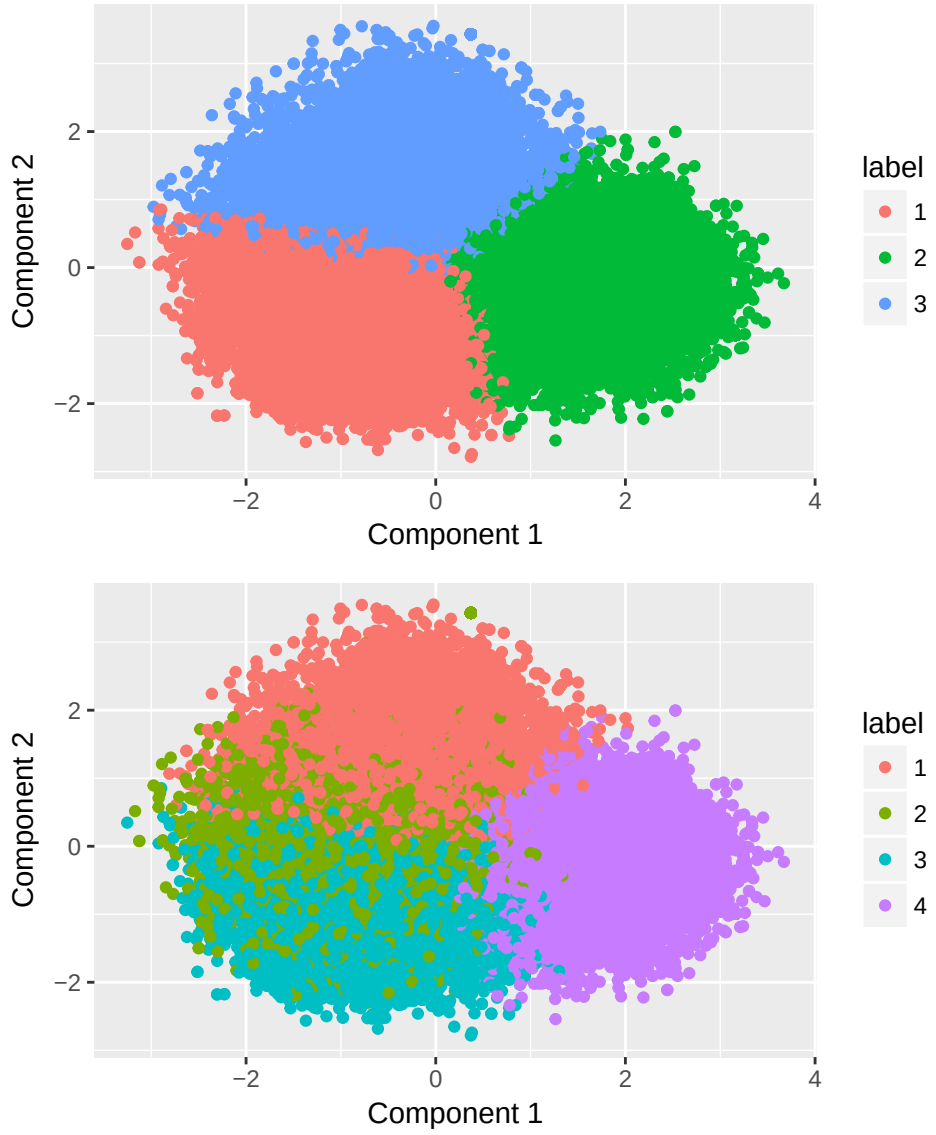

Figure 8: Bivariate plots of the selected two components, labeled by K-Means clusters

5 Stability of findings to perturbation

To see the result of the PCA and K-means clustering are stable, we perform the same analyses using a random subsample of 1000 respondents. Even though 1000 is about 2% of the sample, we find that the result of the analyses are stable. The scree plot indicates that it is at around component 10 that the variance begins to level out. Amount of variance explained by the first 8 components is respectively 4.4%, 3.6%, 3%, 2.3%, 2.1%, 1.9%, 1.8%, and 1.7%, which are very close to the result obtained previously. Questions that load heavily on each component are also unchanged, with only some differences in the amount of loading,

but the interpretation remains the same.

Stability of K-means clusters is examined in two ways: we use the subsampled data and alter the starting point of the algorithm. The result shows a very similar plot to the one obtained previously in Figure7. Further stability analysis may be conducted using a different subsample of data and different starting values.

Interpretation of data essentially does not change even if we perturb data and alter initial values of K means clustering algorithm. This suggests that the findings are robust and that we did not overfit data.

6 Conclusion

In this project, we explored how different dialects are distributed geographically in the United States. The survey dataset contained a large number of respondents and a large number of possible answer choices, requiring a need to reduce dimension to ease analysis and interpretation. Principal component analysis was conducted to come up with a few meaningful components that explain variability in data. From this, we learned the functionality of the components (e.g. first component may be used to differentiate between respondents from the northeast and midwest, based on their dialects). We also learned specific questions that contribute heavily to the components. However, it was difficult to see a common dialect feature that is shared by questions in the same component. Based on the dataset with reduced dimensions, K means cluster was conducted. Although clusters are not well separated from each other, they do appear to be related to geography. Finally, robustness of the finding was analyzed by repeating the analyses based on perturbed dataset and setting different initial values for the K means algorithm. Results were consistent with the initial analyses, suggesting that our findings are stable.

References

[1] John Nerbonne and William Kretzschmar. Introducing computational techniques in dialectometry. *Computers and the Humanities*, 37(3):245255, 2003.

[2] John Nerbonne and William Kretzschmar. Progress in dialectometry: toward explanation. *Literary and linguistic computing*, 21(4):387397, 2006.