Lab 2 - Linguistic Survey Stat 215A, Fall 2018

3034213264

October 05, 2018

1 Kernel density plots and smoothing

For this section, we use the cleaned redwood dataset from the previous lab. In Figure 1, we experiment with kernel density estimates of the redwood temperature distribution using various kernels and bandwidths.

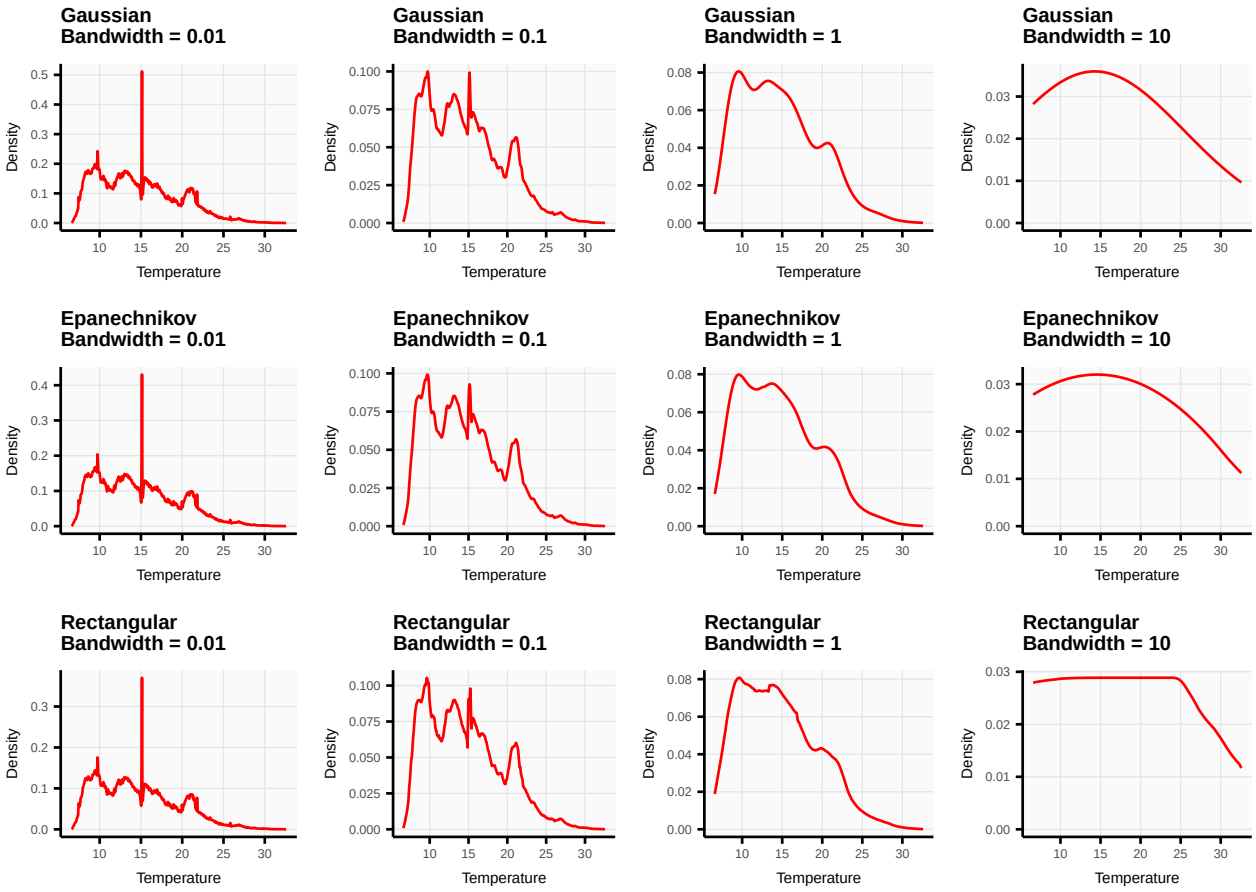

Figure 1: Distribution of temperature measurements for the entire redwood study. The kernel density subplots from left to right increase in bandwidth, and each row corresponds to a different kernel function.

We clearly observe a trade-off between bias and variance, depending on the choice of bandwidth. As we move through the subplots from left to right in Figure 1 (i.e. as the bandwidth increases), the kernel density estimate becomes much smoother, which corresponds to lower variance but higher bias. As bandwidth decreases, the density estimate fluctuates greatly and eventually overfits to spurious artifacts and noise in the data. In the case when the bandwidth is too small, bias is low while variance is high. We point out that a bandwidth of around 1 seems to be a reasonable compromise between the two extremes.

Though the bandwidth appears to heavily influence the density estimate, the choice of kernel (e.g. Gaussian, Epanechnikov, Rectangular) does not appear to impact the densities as much, particularly when the bandwidth

is small. When the bandwidth is large however, the density estimates increasingly resemble the kernel function. For example, the rectangular kernel density estimate when the bandwidth was 100 is becoming more like the rectangular kernel function. Additionally, when the bandwidth was 1, we observe kinks in the rectangular kernel density estimate while the other density estimates using differentiable kernels (e.g. the Gaussian and Epanechnikov kernels) gave smooth estimates. Overall, we conclude that the density estimates are relatively robust to the choice of kernel, but the choice of bandwidth is heavily influential in the bias-variance trade-off.

We next study the relationship between temperature and humidity at noon throughout the redwood study using loess smoothing with different bandwidths and polynomial degrees.

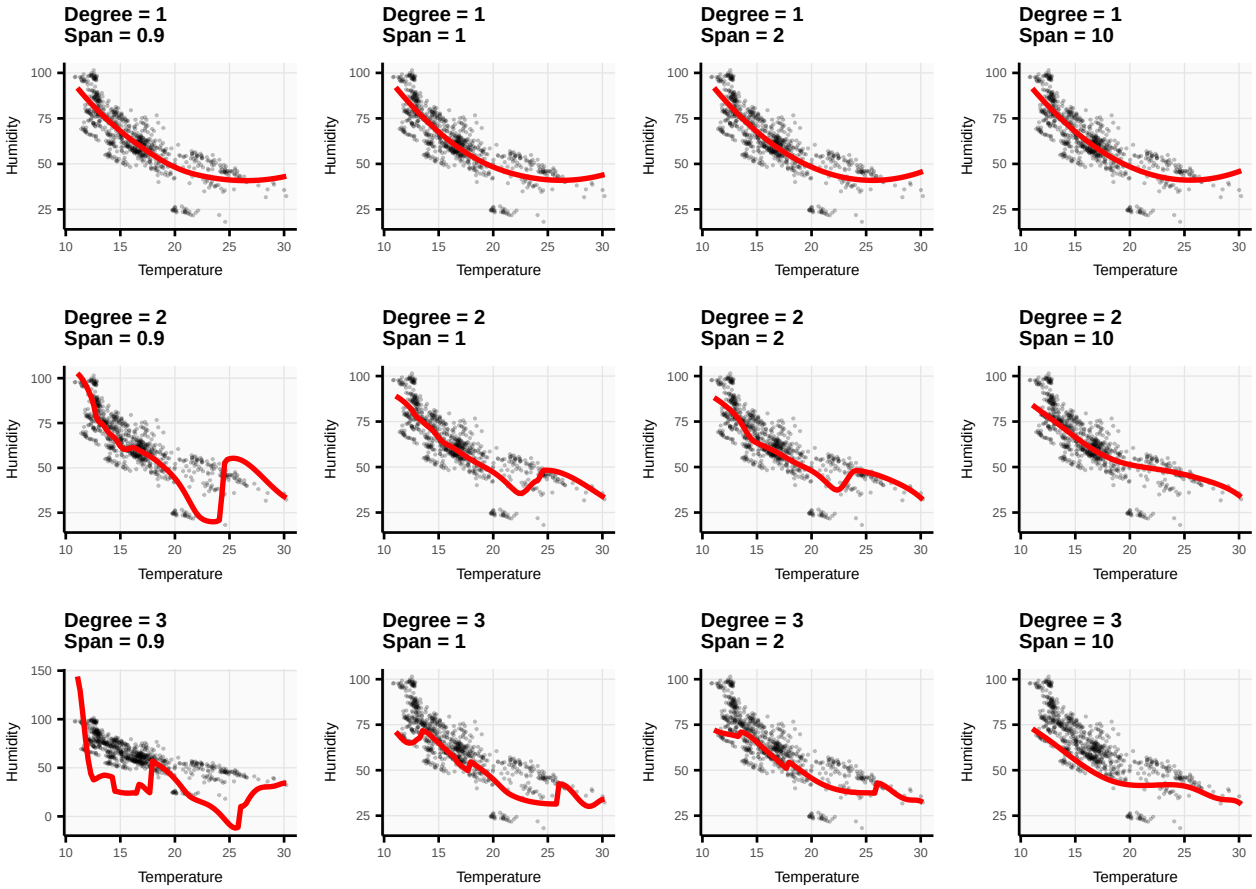

Figure 2: Relationship between temperature and humidity at noon for the entire redwood study. The amount of smoothing used in the loess increases as we move from left to right. The degree of the polynomial used in the loess increases as we move from top to bottom.

As in the previous investigation, we see a trade-off between bias and variance. As the degree increases (i.e. as we move down the columns in Figure 2), the fitted loess is more flexible, and variance increases. As the span increases (i.e. as we move from left to right in Figure 2), the fitted loess becomes smoother so that the bias increases while variance decreases. We point out that in most cases, the loess has a negative slope, suggesting a negative relationship between humidity and temperature. However, when the degree is 2 or 3 and the span is small enough (e.g. 0.9 or 1), the loess overfits to a group of outlying points with abnormally low humidity levels. This disrupts the negative pattern that we observe in all other subplots, illustrating that overfitting is a concern when selecting loess parameters. On the other hand, underfitting is also a valid concern. After fitting loess with different degrees and spans, we conclude that a degree 2 polynomial with span 10 appropriately fits the data, and it suggests that there is a negative, slightly nonlinear relationship between temperature and humidity.

2 Linguistic Data

2.1 Introduction

The study of language variation is a valuable field of linguistics, which can provide unique insights into historical, social, and geographical factors in society. For instance, shared linguistic traits between two groups of people may suggest interpersonal contact at some point in history. Language variation among social classes may signal social identity and prominence, and it is well known that geographical distances can influence how one speaks. Nevertheless, there are many challenges in studying linguistic variation, a gradual and slow process with many irregularities. In this report, we focus upon the study of dialects, or dialectology, and the measurement of dialect differences, or dialectometry. We consider a Dialect Survey conducted by Bert Vaux [Vaux and Golder, 2003], and we aim to (1) understand whether there are clusters of people with similar survey responses and (2) determine whether these clusters correspond to geographical locations. Before proceeding though, we cautiously note that language is known to depend on sex, age, occupation, social class, as well as geography, so while we study the relationship between geography and language here, further investigation is required to explore more complex relationships between language and other covariates.

2.2 The Data

In this study, we analyze data from the Dialect Survey conducted by Bert Vaux [Vaux and Golder, 2003], and in particular, we focus on the survey questions regarding lexical differences, rather than phonetic differences. The overall survey consisted of $p = 67$ questions of interest and $n = 47,471$ respondents from across the United States with their geographical location encoded by city, state, and ZIP code.

2.2.1 Data quality and cleaning

As with any survey data, there are always misspellings and typos of cities, states, and ZIP code. While we are aware of this issue, it does not seem to be a large obstacle for our downstream analysis, so we deal with this issue as it comes up. The more pressing issues are missing locations and unanswered survey questions.

We found that there were three entries with missing state identifications. Two of the three listed cities which appeared to be outside of the US, and the third listed the city to be "nowhere." We thus deleted these three observations. There were also missing latitude and longitude values for 1020 observations. To reduce this number, we merged our given survey data with the zipcode data from `library(zipcode)`. However, in this process of merging, we noticed that many ZIP codes in the survey data had only four digits rather than the usual five digits. We believe that a leading 0 in the ZIP code was dropped when the ZIP code was converted from a string to numeric, so we padded the four digit ZIP codes with a leading 0. Then we merged the survey data with the zipcode data by ZIP and by state. We included state as a precaution in case the ZIP code was inputted incorrectly. After this merge, there were 648 observations with missing latitude and longitude data. Careful investigation into the ZIP codes with missing latitudes and longitudes reveals that several of these ZIP codes appeared with high frequencies (e.g. the ZIP code 95411 was entered 86 times but did not have a corresponding location). We speculate that there may have been a change in the ZIP code recently, so removing these observations may introduce some location bias. However, as there is not much we can do, we were forced to omit the 648 observations with missing latitude and longitude data.

Finally, because a non-negligible number of people answered only a few questions, we decided to omit those who left more than ten questions unanswered. We chose the threshold of ten after looking at the distribution of the number of unanswered questions per person and saw that ten was a reasonable balance between removing too many observations and keeping too many non-informative observations. After this pre-processing, we were left with $n = 45,332$ samples.

With regards to the data quality, we note that the sampling is not necessarily representative of the US population. It seems to favor responses from the northern US. For instance, 2,506 responses were from

Texas (population 28.3 million) while 2,575 responses were from Pennsylvania (population 12.8 million). Additionally, there were very few responses from the western US, aside from California. We keep this sampling bias in mind throughout the analysis.

2.2.2 Exploratory Data Analysis

In our exploration of the linguistics data, we choose to compare survey questions 65 and 66. That is, "What do you call the insect that flies around in the summer and has a rear section that glows in the dark?" and "What do you call the miniature lobster that one finds in lakes and streams for example (a crustacean of the family Astacidae)?" For both questions, there were three answers which were overwhelmingly popular (encapsulating over 90% of the surveyors in each case). We henceforth only consider the three most popular answer choices for each question to enhance clarity.

Remark: Throughout this report, we include Hawaiians and Alaskans in all statistical analyses, but we refrain from plotting their responses on the maps to avoid clutter.

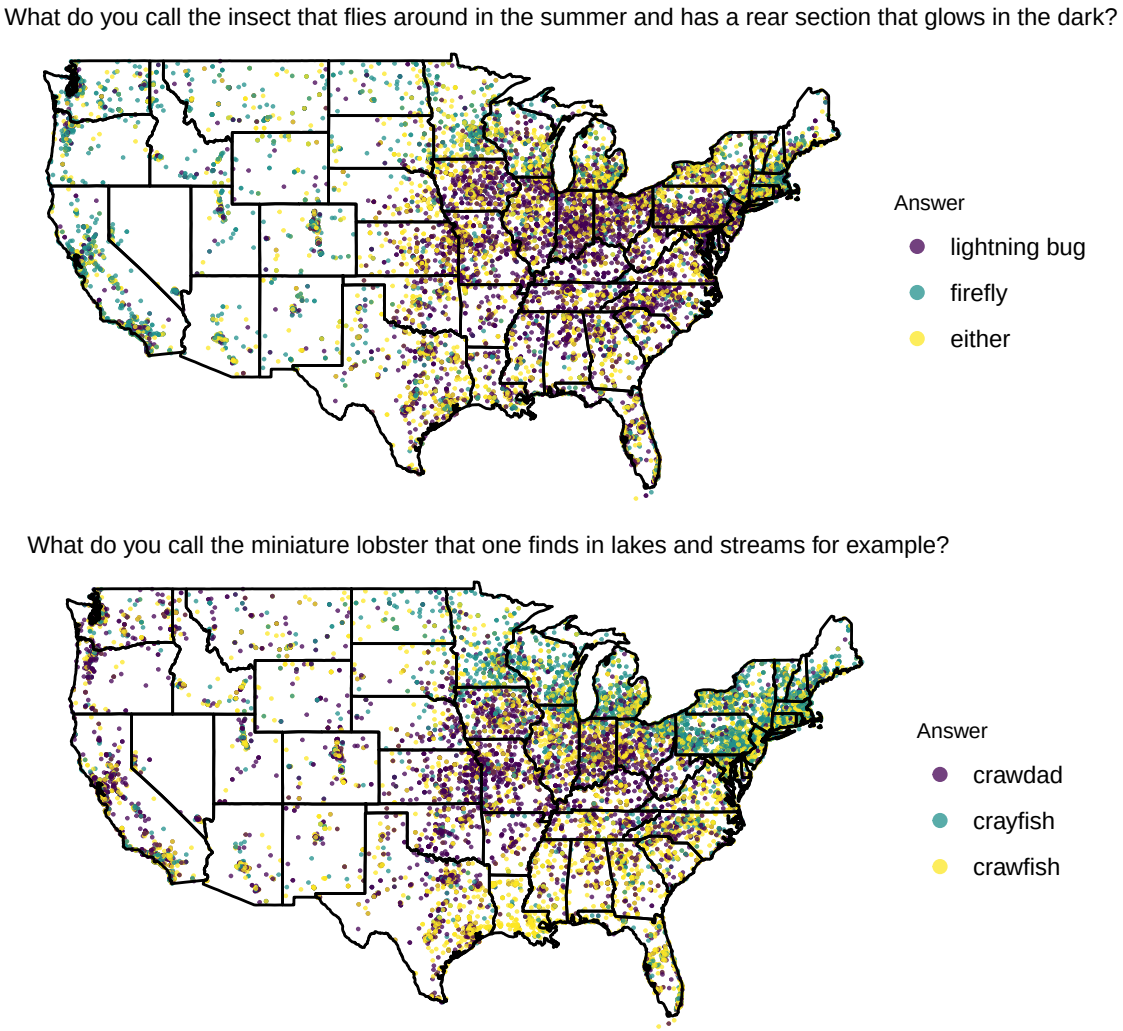

Figure 3: Responses to questions 65 and 66 according to geographical location.

From Figure 3, we observe that responses tend to cluster around certain geographical locations. In the top panel of Figure 3, lightning bug is clustered around the mid west while firefly is used more the in the west.

The bottom panel of Figure 3 shows a slightly different pattern with crawdad being used frequently in the mid west, crawfish in the south, and crayfish in the north/northeast.

Table 1: Confusion matrix for survey questions 65 and 66 (in %)

	crawdad	crayfish	crawfish
lightning bug	11.6	8.0	15.2
firefly	5.9	10.3	10.8
either	8.3	12.1	17.7

Looking at the confusion matrix in Table 1, we see that though people who answered crawdad are more likely to answer lightning bug than the other two choices, the responses appear fairly evenly distributed across the confusion matrix. Consequently, if we were to use the crustacean answer to predict the insect answer, the optimal predictor (measured via classification accuracy) would be as follows: if x = crawdad, then predict lightning bug; if x = crayfish, then predict either; and if x = crawfish, then predict either. (This can be seen by finding the largest percentage in each column of Table 1.) Under this prediction rule, the *maximum* classification accuracy we can obtain is $11.6 + 12.1 + 17.7 = 41.4\%$. Since we used all of the data to obtain this estimate, 41.1% is most likely an overestimate of the true population classification accuracy. Moreover, 41.1% is only slightly higher than random guessing (33%). If we were to use the insect answer to predict the crustacean answer, we would see a similar story. This suggests that while the two questions appear to define geographical regions, the questions are not highly predictive of one another.

If one were to add another question into the mix, then the predictive value can increase. However, after trying many different combinations, it was difficult to achieve greater than 50% test classification accuracy using a simple multinomial regression to predict the insect response from the crustacean response and one other survey response.

2.3 Dimension reduction methods

To gain a better understanding of dominant patterns in the high-dimensional linguistics data, we experiment with several dimension reduction techniques, the first being PCA. However, it does not make much sense to perform PCA on the categorical linguistics data because the survey responses are unordered. We thus transform the categorical survey responses to binary indicator variables and apply the dimension reduction techniques to this binary dataset.

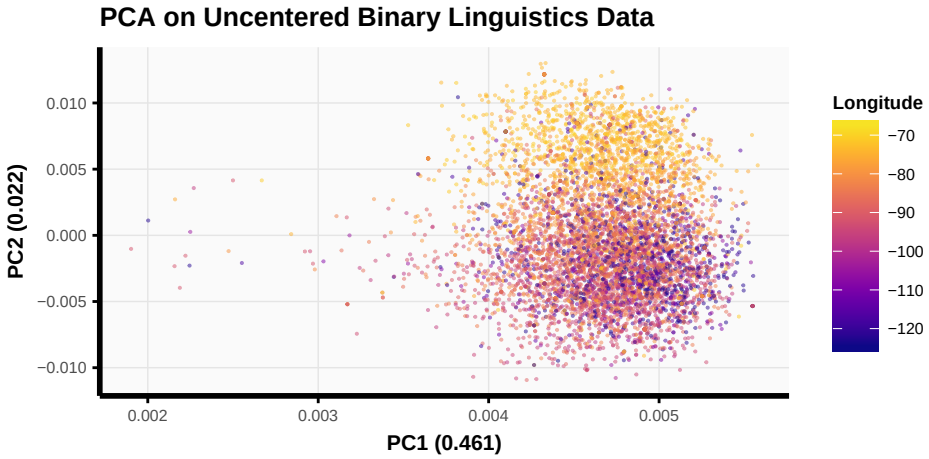

Figure 4: Here, we show the first two PCs, colored by longitude, after applying PCA to the uncentered binary linguistics data.

In Figure 4, we apply PCA to the uncentered binary linguistics data. From Figure 4, we see that the first PC explains a lot of the variation in the data (46.1% as displayed in the x-axis title) while the second PC explains only 2% of the variation in the data. The scree plot is not displayed, but it is clear from the first 2 PCs that there is an elbow in the scree plot after the first PC. This is an interesting observation, but further investigation reveals that the first PC is not associated with geographical location in any way. Rather, the variation explained by PC1 corresponds to the variation within questions, which had a single dominant response. That is, the top 2 features with the largest PC1 loadings (in magnitude) corresponded to Q63, answer choice 1 and Q67, answer choice 1 - both of which received over 90% of the votes for their respective question. Since the variation within questions is not of interest to us, we next try PCA on the centered binary linguistics data, as shown in Figure 5.

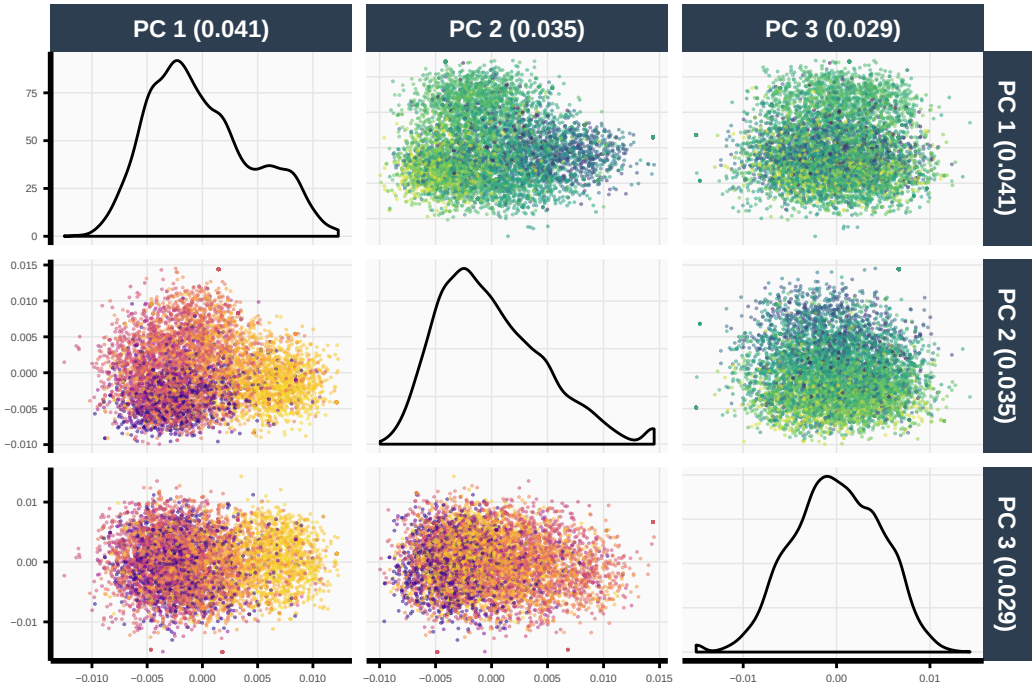

Figure 5: PCA on the centered binary linguistics data. The lower triangle of PC plots are colored by longitude while the upper triangle of PC plots are colored by latitude

Pair plots for the first 3 PCs are shown in Figure 5. The lower triangle of PC plots are colored by longitude with yellow corresponding to the northern US, orange being the middle US, and purple the southern US. The upper triangle of PC plots are colored by latitude with yellow being the eastern US, green the central US, and blue the western US. Our first observation is that the proportions of variance explained (shown in the parentheses in the PC titles) are very small, indicating that if one were to simply take the first three PCs and ignore the other the PCs, we would be losing a lot of variation in the data. However, even though these PCs capture a small portion of the variation in the data, they appear to be associated with geography. This is in stark contrast to PCA on the uncentered data. Specifically, in the first PC in Figure 5, the northerners in yellow are fairly differentiated from the rest of the data. By looking at the top three largest PC1 loadings (in magnitude), we are led to believe that the sneakers (Q73), soda (Q105), and sunshower (Q80) questions help to differentiate the north/northeast from the rest of the US. Then in the PC2 vs PC3 plot, there seems to be a slight gradient from yellow to green to blue, suggesting some association with the eastern vs. western dialects. The important features for PC2 were the catty-corner (Q76), water fountain (Q103), and the plural you (Q50) questions. We also try PCA on the scaled dataset, but we believe scaling is unnecessary, particularly because the features were all measured using the same scale.

While PCA has enabled us to gain some intuition into the data, PCA is limited in the sense that it is a linear projection method. This motivates us to try another dimension reduction method - a nonlinear method - to

see if we can explore more complex relationships. In this light, we try a popular nonlinear dimension reduction method called t-Distributed Stochastic Neighbor Embedding (t-SNE) [Maaten and Hinton, 2008], but to apply t-SNE, we must reduce the size of the data for computational feasibility. This is the computational price we must pay for nonlinearity. (Note that PCA is computationally advantageous for large datasets because one only needs to compute the first few eigenvectors of interest, which can be done very efficiently. This illustrates a trade-off between computation and model complexity.)

We choose to reduce the size of the data by aggregating the responses by county instead of the 1 degree latitude-longitude squares. We believe that there is a greater sense of community and cohesiveness within a county, rather than a 1 degree latitude-longitude square that ignores state borders and community lines. To aggregate by county, we first mapped each person's latitude-longitude to a particular county, then summed the binary data for each county, and finally divided each county's data by the number of observations in that county. After binning observations in this way, we ended up with $n = 2372$ counties. We use this binary county data for the remainder of the report to speed up computation.

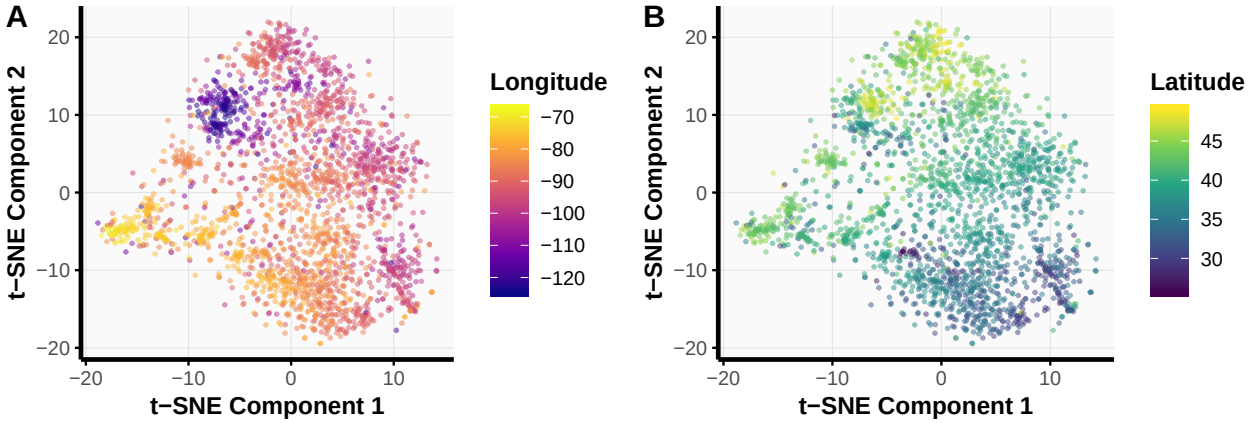

Figure 6: t-SNE on the centered county-level binary linguistics data. Each point represents a county, colored by longitude in (A) and by latitude in (B).

Similar to PCA, t-SNE appears to do a respectable job at separating the northern US from the rest of the US (i.e. clustering the yellows in Figure 6A). There is also some differentiation between the eastern US and western US as we see the yellows separate from the dark blues in Figure 6B. It is difficult though to visually compare the t-SNE results with PCA, so we prepare maps in Figure 7, where each county is colored by the weight of the PCs and t-SNE components. From these maps in Figure 7, it looks like PC1 and t-SNE component 2 are capturing the same geographical patterns and similarly for PC2 and t-SNE component 1. Because PCA and t-SNE are similar, we prefer PCA over t-SNE since PCA is by far more interpretable.

2.4 Clustering

While we have seen PCA capturing some geographical patterns from the survey data, we next try to make these patterns more concrete by clustering the county-level survey data. We begin by trying k-means. Since k-means is known to perform poorly in high dimensions, we choose to apply k-means to the first 79 PCs, which explained 75% of the variation in the data. To select k, we looked at the within sum of squares and the silhouette information, as shown in Figure 8. The large peak in the silhouette plot at $k = 2$ is highly suggestive, and since there is not clear elbow in the within sum of squares plot that contradicts choosing $k = 2$, we decide to go with $k = 2$. The k-means clusters are depicted in Figure 9, and it shows that the two clusters correspond to the northern and southern US with a few exceptions. We will examine the stability of these k-means clusters in the next section.

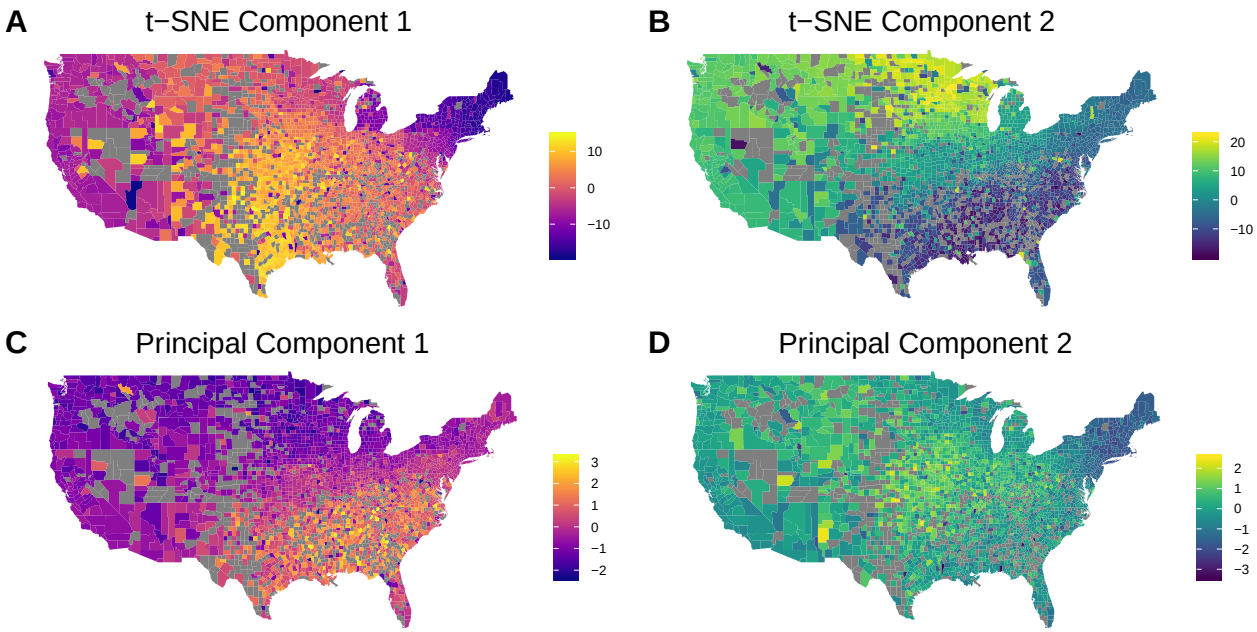

Figure 7: Here, we color each county according to (A) the weights of the first t-SNE component, (B) the second t-SNE component, (C) the first PC, and (D) the second PC.

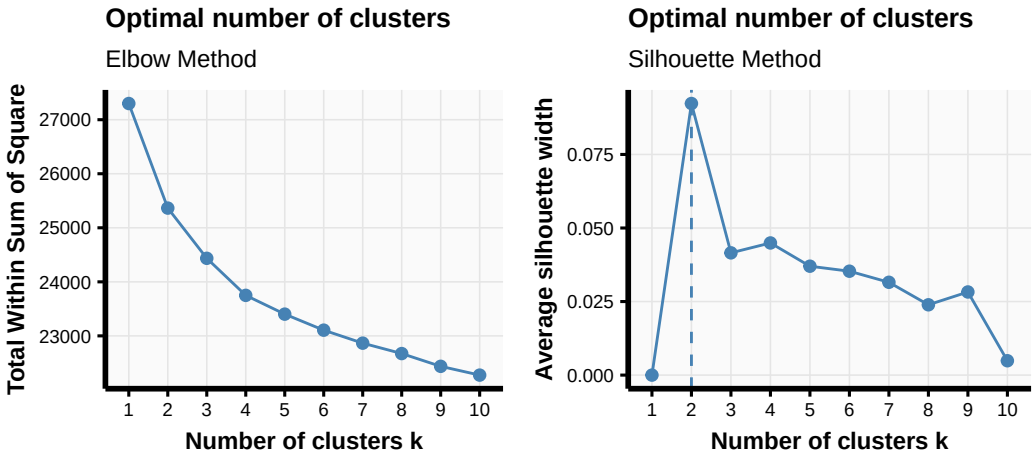

Figure 8: Metrics for choosing the number of clusters k in k-means. From the within sum of squares and the silhouette information, it appears that k = 2 is an appropriate choice of k for the binary county-level linguistics data.

We next tried hierarchical clustering with various distance metrics and linkages, but for brevity, we only include the results using Euclidean distance with Ward's linkage. We note that from our exploration of hierarchical clustering which is not shown here, the clusters obtained from hierarchical clustering are *heavily* dependent on the choice of linkage and distance metric, which is not a desirable property for stability.

Similar to k-means, we choose to cut the tree by examining the elbow and silhouette plots for hierarchical clustering with Euclidean distance and Ward's linkage. Because of the similarities to the previous case with k-means, we omit the diagnostic plots and simply report that the silhouette plot showed an obvious peak at $k = 2$, and the elbow method agreed, so we choose to cut the tree to obtain $k = 2$ clusters. The results from hierarchical clustering with $k = 2$ clusters are shown on the US map in Figure 10 (We omit the hierarchical

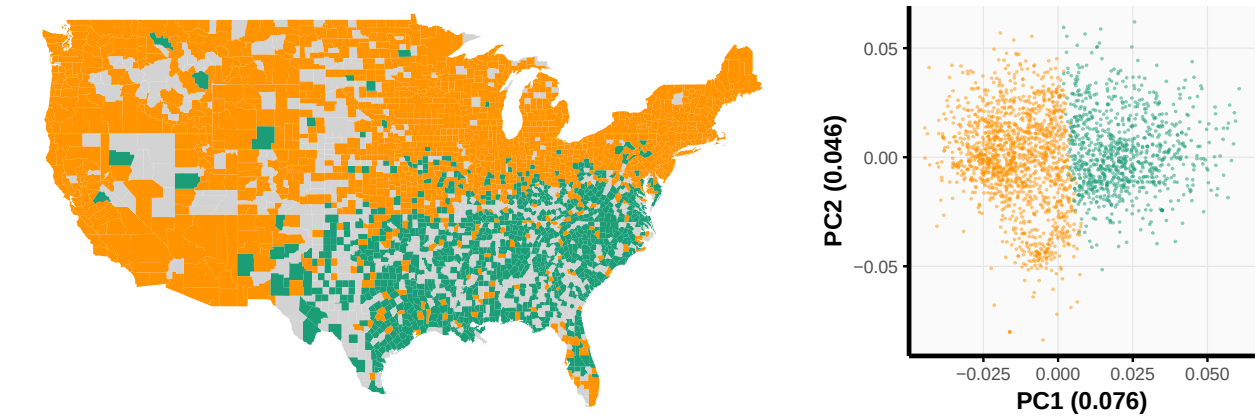

Figure 9: Clusters given by k-means using k = 2. The left panel shows the clusters on the US map while the right shows the clusters on the PC plot. Note that the grey counties do not have any survey responses.

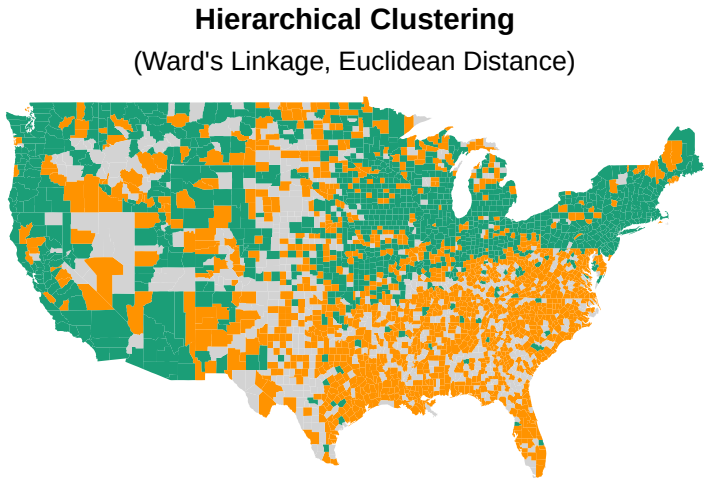

Figure 10: Clusters given by hierarchical clustering with Ward's linkage and the Euclidean distance for k = 2, as shown on the US map. Note that the grey counties do not have any survey responses.

tree for brevity). We see that like k-means, hierarchical clustering also forms clusters that correspond to the northern and southern US. However, hierarchical clustering appears to be less stable and forms more heterogeneous clusters than k-means.

Now, one drawback of k-means and hierarchical clustering is that we can no longer extract information about the features, which contribute to the patterns found by the clustering algorithm. Additionally, we can think of k-means and hierarchical clustering as "hard" clustering methods - an observation must be in one cluster or another; it cannot be in both. This is problematic in our linguistics analysis since both domain knowledge and the PC plots suggest a continuum of dialects, rather than completely disjoint groups.

We can overcome these two challenges using non-negative matrix factorization (NMF). As a matrix factorization method, NMF allows us to extract both sample patterns and feature patterns simultaneously. It can also be viewed as a "soft" clustering algorithm - that is, the weights in the NMF factors act as "probabilities" of the observation belonging to a certain cluster. Furthermore, NMF appears promising because the binary county-level linguistics data is entry-wise positive to begin with, so the non-negativity constraint will enforce sparsity and thus increase interpretability. For these reasons, we next try clustering the binary county-level linguistics data using NMF.

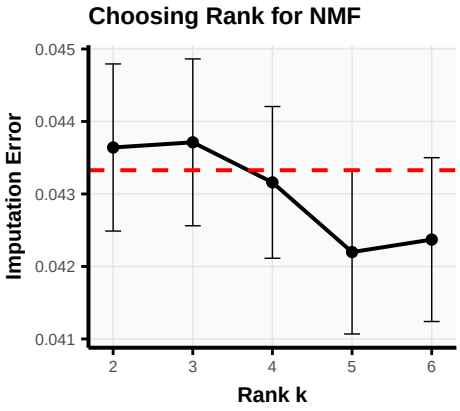

Figure 11: We plot the mean imputation error +/- 1 standard error for various ranks of NMF (averaged over 20 trials). The dashed red line represents the threshold for the 1 standard error rule. Thus, the rank which satisfies the 1 standard error rule is 4.

To choose the rank k for NMF, we adopt a missing data imputation approach. The idea is to randomly leave out scattered missing elements of the data. Then, for each potential rank k, we apply NMF to the data with the missing values, and NMF outputs the two rank k matrix factors W and H. We impute the missing values via $X_{miss} = [WH]_{miss}$, and then we compute the imputation error (i.e. the difference between the imputed values and the observed data values). We repeat this process 20 times and choose k according to the 1 standard error rule. As shown in Figure 11, $k = 4$ is the optimal rank for NMF according to the 1 standard error rule.

Using $k = 4$, we next factorize the binary county-level linguistics data $X \in \mathbb{R}^{n \times p}$ into two rank 4 factors $W \in \mathbb{R}^{n \times 4}$ and $H \in \mathbb{R}^{4 \times p}$ via NMF. Since W corresponds to the observation-level factors, we plot the weights of each column of W on the US map to find clusters among the counties (see Figure 12). On the other hand, since H corresponds to the feature-level factors, features with the largest magnitudes in each row of H correspond to the most important survey questions associated with the observed patterns in W. In Table 2, we list the questions which were found to be most important for each of the four clusters found my NMF.

Table 2: Top Questions and Responses Corresponding to NMF Clusters

Cluster 1	Cluster 2	Cluster 3	Cluster 4
y'all (Q50)	bag (Q109)	acceptable (Q56)	daddy long leg(s) (Q67)
water fountain (Q103)	sneakers (Q73)	heel (Q111)	no (Q53)
in line (Q93)	soda (Q105)	crawdad (Q66)	in line (Q93)

After examining the clusters in Figure 12 and the corresponding features in Table 2, the four clusters from NMF are intuitive and reasonable. NMF appears to divide the US into four geographical regions - the south, northeast, mid west, and northwest, and the important features (i.e. question/responses) corresponding to the clusters also seems to match with intuition. For example, southerners are known for saying "y'all". Overall, this NMF analysis seems superior over k-means and hierarchical clustering. NMF allows us to capture a gradient between regions, rather than assigning hard clusters to the US. One could argue that this soft clustering technique enables NMF to select more clusters than what we have seen previously in k-means and hierarchical clustering. Lastly, the fact that we can extract meaningful patterns among counties *and* the corresponding important questions/responses is a nice property of NMF.

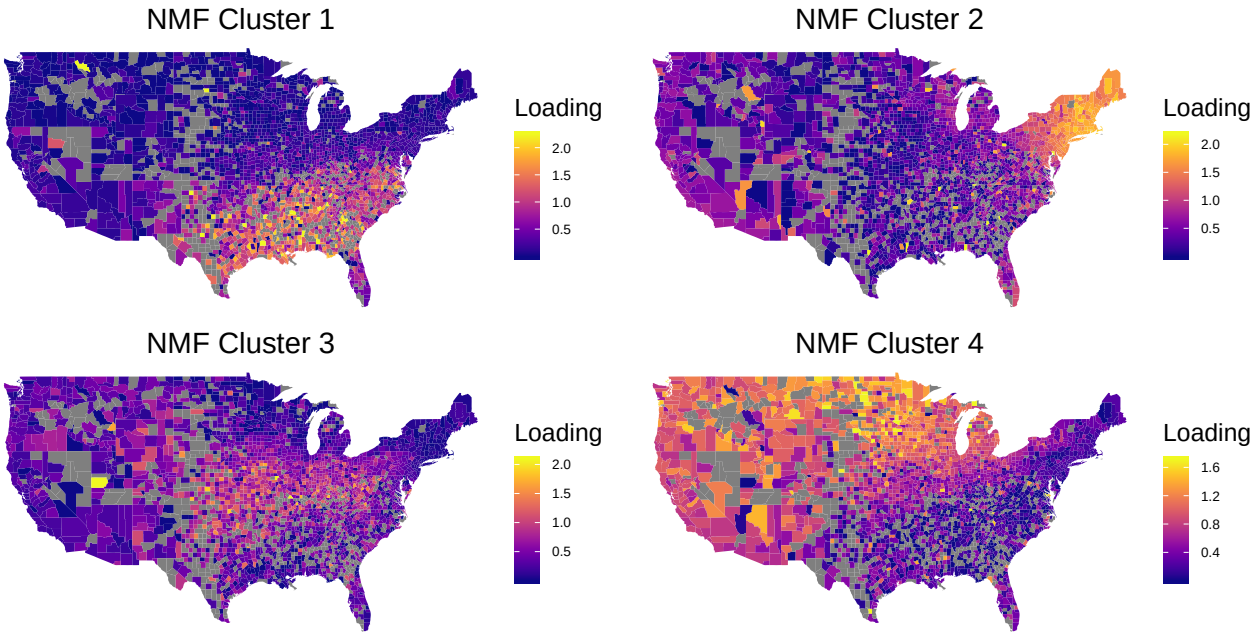

Figure 12: In (A)-(D), we color each county by the loadings given in W. For each subplot, the counties with the largest magnitudes form a soft cluster.

2.5 Stability of findings to perturbation

One of the key findings above is that the clustering of dialects seems to be a continuum rather than a discrete clustering problem. We further investigate this finding by understanding whether or not the clusters from k-means (with $k = 2$) are stable. We consider instability from two sources: 1) being a local solution of a non-convex problem and 2) data perturbations.

We first checked stability of the k-means solution for different initializations. After running k-means 25 times, each time with a different initialization, we found that every run of k-means converged to one of three different solutions. The clusters given by these three solutions overlapped for all but 4 counties, and all three solutions gave a within sum of squares value equal to 25364.41. This indicates that k-means is stable with respect to different initializations.

A perhaps more important source of instability comes from perturbations of the data. We next consider what happens to k-means when we randomly subsample 70% of the data and apply k-means to this subsample. Note that we randomly subsample the county-level data to save on computation, but this can also be done using the original person-level survey data.

To make this stability comparison, we first run a baseline k-means on the full county-level dataset. Then for each of the $B = 200$ subsamples, we ran k-means on the subsampled data and reported the cluster labels on the subsampled data. We next counted the number of counties which were clustered differently in the subsample k-means than in the baseline k-means.

We plot the distribution of these differences in Figure 13, and we see that there are numerous instances where the clusters differed by more than 100 counties. From this, we looked more closely into a case where the clusters differed by 200 counties and plotted this in Figure 14. We notice that k-means had difficulties clustering the counties near the northeast. Overall, from this preliminary stability analysis, we see that k-means can be variable and yield different clusters when the data is perturbed, but it is not too severe. We hypothesize that if we had taken $k = 3$, then the stability of k-means would be much worse.

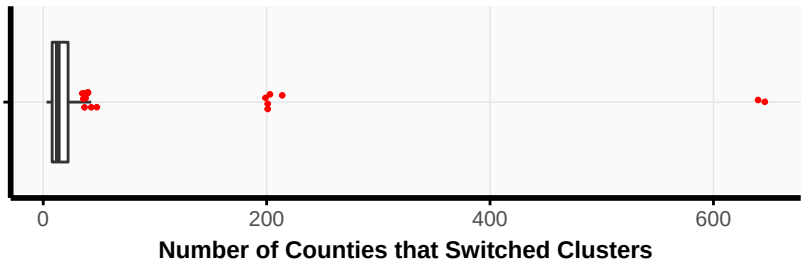

Figure 13: We plot the distribution of the the number of counties which were clustered differently in the subsampled k-means compared to the full k-means. The outliers are plotted in red.

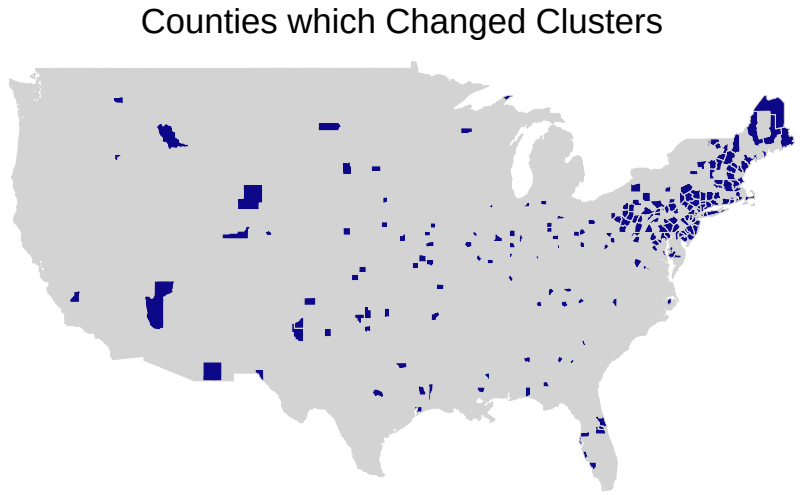

Figure 14: In this map, we colored the 200 counties, which were clustered differently by the full k-means and an instance of the subsampled k-means algorithm.

2.6 Conclusion

From our analysis of the county-level linguistics data and the success of soft clustering via NMF, we conclude that linguistic variation should be viewed as a continuum over space, rather than disparate regional identities. NMF was also useful from an interpretation standpoint and gave meaningful feature patterns which intuitively meshed with the county clusters. Lastly, stability of NMF can likely be improved with something like staNMF, but in this report, we looked into the stability of k-means with $k = 2$ and found that it was mostly stable but can drastically change in certain situations.

Bibliography

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Bert Vaux and Scott Golder. The harvard dialect survey. *Cambridge, MA: Harvard University Linguistics Department*, 2003.