Lab 2 - Linguistic Survey, Stat 215A, Fall 2017

SID:3032096549

October 5, 2017

1 Introduction

In this report, in second section, the kernel density and Loess smoother applied on Redwood data is discussed. Then the linguistic geography of the United States by analyzing the Harvard Dialect Survey conducted in 2003 is studied. In Section 3, data cleaning and preliminary data exploration is performed. In this part, question 74, Question 76 and Question 86 are chosen for investigation. In Section 4, several classical data reduction methods are applied to explore the dimensional characteristics of the dataset as well as to eliminate noise. In Section 5, several clustering methods are combined with data reduction methods on the linguistic data. The number of clustering is determined by gap statistics visualization. The geographical pattern shown in the results are discussed. Four dialect regions are found after clustering, i.e., eastern part, middle part, southern part and western plus northern part of the U.S. Then several critical questions which dominates clustering results are detected and discussed. In Section 6, the raw data are added Gaussian noises to test the robustness of the clustering methods.

2 Kernel density plots and smoothing for Redwood data

For this part of re-analyzing the Redwood data, the version after cleaning procedures is used, where missing values, mismatching samples, outliers and samples from problematic sensors are removed. Then in estimating the distribution of temperature over the whole dataset, six different kernels are used, i.e., Biweight kernel, Cosine kernel, Epanechnikov kernel, Gaussian kernel, Optcosine and Rectangular kernel. For each kernel, the bandwidth are adjusted by scale 0.2, 2, 4, 8 with respect to the standard estimated scale defined as $0.9 \times min\{\hat{\sigma}, \hat{(q)}/1.34\}$. Here σ is the estimated standard deviation and q is the interquartile. Notice that larger bandwidth may lead to more smoothness but higher bias, and vice versa. This is also known as the bias-variance tradeoff. From the kernel density plots in Figure 1, it's hard to tell which kernel outperforms the others. But it's obvious that as mentioned above, larger bandwidth leads to more smoothness in the density line.

Then a fixed time point for each day during the project is chosen to analyze the correlation between temperature and humidity for all nodes. Since time points are uniquely represented by epoch in the data and there are 288 epochs each day, the samples with epochmod288=66 is chosen as the subset. Loess smoothers with different bandwidth and degrees of the polynomials are experimented to fit the subset of data, and the results are shown in Figure 2. Again, with larger span, when fitting each point, more neighbouring samples are included, resulting in more smoothness but higher bias, and vice versa. This is similar to the bia-varaince tradeoff issue mentioned above. As for the degree of polynomials, notice that higher degrees would bring in more paramters into the model, increasing the complexity of the parametric model and might suffer from fitting the noise other than signal. To be concrete, as shown in the plots, higher degree of polynomials are more curved in order to better fit this training dataset, and might suffer from over-fitting issues.

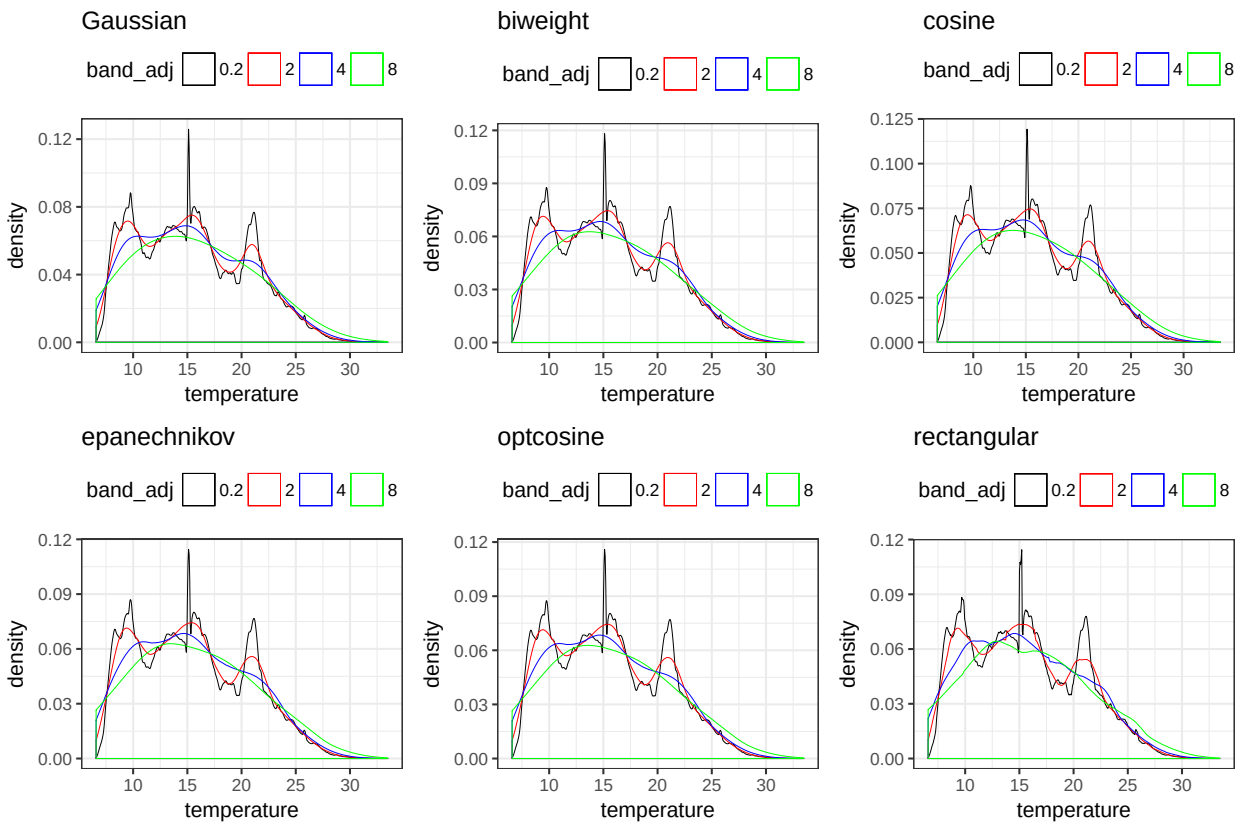

Figure 1: Kernel density for for temperature, using different kernels and different adjusted bandwidths

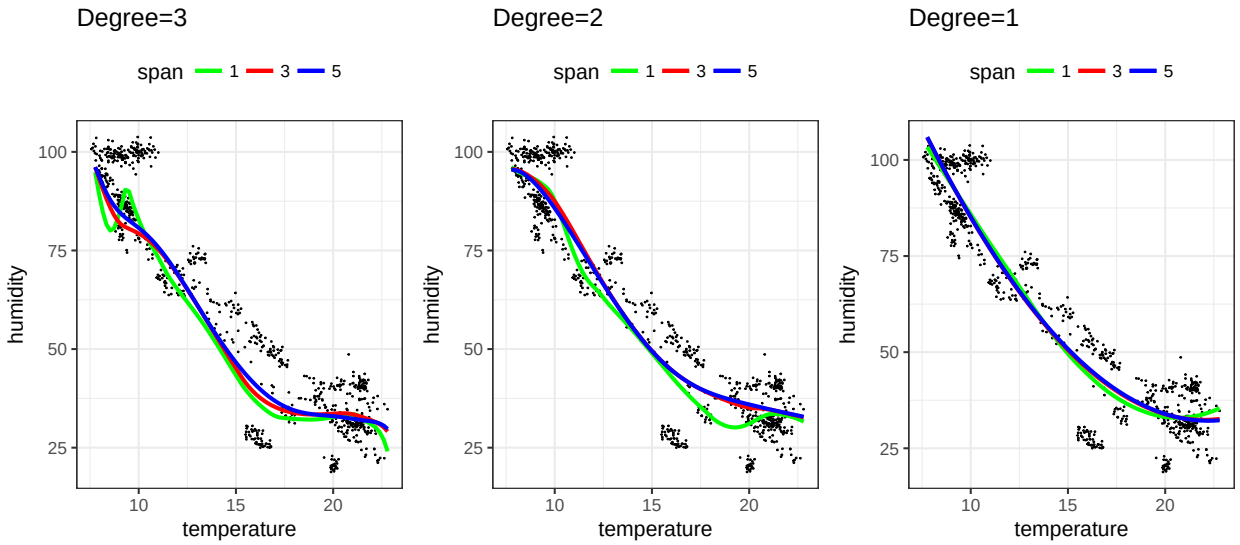

Figure 2: LOESS smoothing on the relationship between temperature and humidity fora certain time point of a day throughout the whole project period

3 Data from the linguistic suevey

In total three datasets are provide: lingData, lingLocation and questionData. 47471 people from all parts of the U.S were surveyed on 122 questions. A subset of 67 questions are available in the given datasets. In dataset lingData, for each sample, the ID, CITY, STATE, ZIP, lat (latitude), long (longitude) and their responses to questions are provided. ID is the individual index while CITY, STATE, ZIP, lat, long describe the location. Dataset lingLocation aggregates Dataset lingData by put people with the same lat, long and responses together. The questionData gives content of each question, but dose not provide information for the answer choices. For the choices for each question, one has to refer to the orginal Harvard Dialect Survey.

Since one of the targets for this project is to represent the results using map plots, extra geographical information are needed. Since each location could be identified by ZIP code, I referred to the FIPS county code to match and add county information for each sample. The reason is that in showing the results of clustering analysis, representing individuals' clusters on a map is comlicated, so the idea is to group the samples according to geographical closeness. So the choices could be city, county or state. Notice that city is too small a unit, and state could be too large, so the county is used as the basic unit to represent geographical patterns of dialects in following analysis.

3.1 Data cleaning and processing

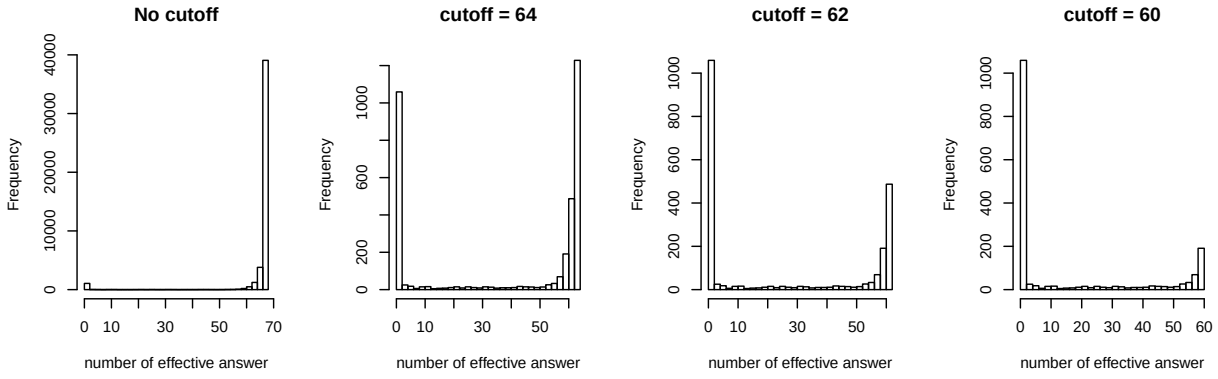

Figure 3: Histogram of effective number of answers under different cutoffs

The data cleaning process involves the following procedures.

1. Matching each individual to the county using the zip code. The individuals with zip code not matching to any county are discarded, since eventually the clusters are visualized with counties as basic unit, so each sample must have a county.

2. Samples with zip code matching to counties outside mainland U.S are excluded, i.e Alaska, Hawaii, since the focus would be the dialect pattern in mainland U.S. After this cleaning procedure, approximately 46500 samples are left in the data, matched to 1482 different counties.

3. Some samples have tremendous missing values, i.e their answers to many of the 67 questions are missing. Since many clustering algorithms and dimension reduction techniques do not accept missing values, individuals who have valid answers for less than 60 of the 67 questions are excluded. The justification for the cutoff is as follow. If we take a look at the distribution of number of effective answers under different cutoffs (Figure 3), it's obvious that if an individual has less than 60 responses, he/she is likely to have zero effective reponse.

4. To plot the U.S county map in ggplot, a standard county map data file is needed. So the linguistic data is merged with the county map data, and unmatched samples are discarded. After this step, 44707 samples are left in the data.

5. Since the question response belongs to categorical data and cannot be compared directly between questions, they must be transformed into binary coding. For example, if one question has 4 responses one person chose the second option, his response is expressed as (0,1,0,0). Finally, each person has a response vector of length of 468 (the sum of the number of options for all questions), with each element being the binary indicator if the corresponding option is chosen. Here we get a response matrix with a size of 44707 by 468. By simply taking an average of the response vectors among people in the same county, the data matrix for county is 1482 by 468.

Also, instead of doing individual-wise clustering, county-wise clustering might also be informative. So the individual data matrix is also aggregated into county level. The distribution of effective samples for each county in mainland U.S is plotted in Figure 4. Notice that there are many counties with no or few samples in the middle part.

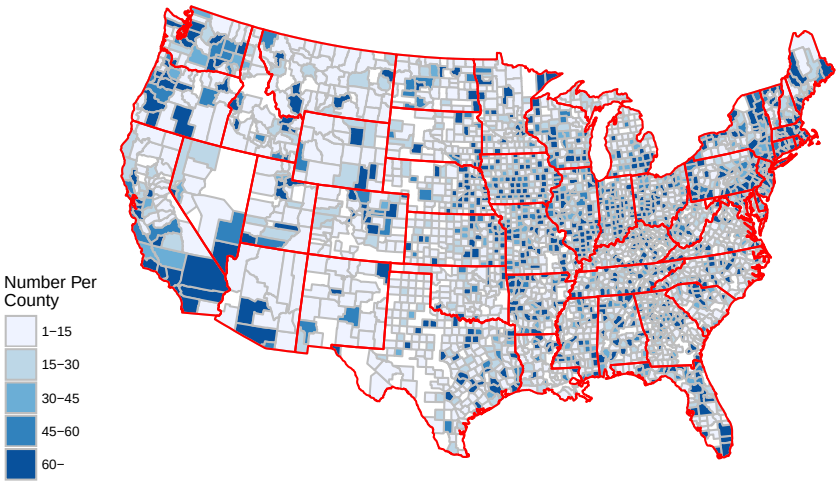

Figure 4: The distribution of sample size on counties

4 Exploratory Data Analysis

In this part, three survey questions are picked:

- Q074: What do you call the little gray creature (that looks like an insect but is actually a crustacean) that rolls up into a ball when you touch it?
- Q076: 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)?
- Q089: Can you call coleslaw "slaw"?

They reason for the choice of these three questions is that they focus on the way people call several common things. Question 74 has 14 choices, Question 76 has 9 and Question 89 has 5, which reflects the diversity of dialects. The geographical implications are obvious: Question 76 and Qeustion 89 share a similar pattern that they separate individual from the others, while Question 74 isolates the northern part and eastern part

(Figure 5). Its easy to interpret the isolated eastern and northern part with responses like no idea as the cold environment offer them few opportunities of seeing such bugs. But other questions may involve complicated cultural and historical factors. See the interactive map BY CLIKCING HERE.
Then all three questions are examined to see if there is obvious association. The correlation of the pairwise questions is calculated by performing a kernel density estimation on the two dimension distribution, as shown in Figure 5. Most individuals incline to response roly poly to Question 74 and kitty-conner toQuestion 76 simultaneously or take roly poly and catty corner as a pari. Such connection can help predict the response for each other due to some strong correlations in specific answers to different questions.

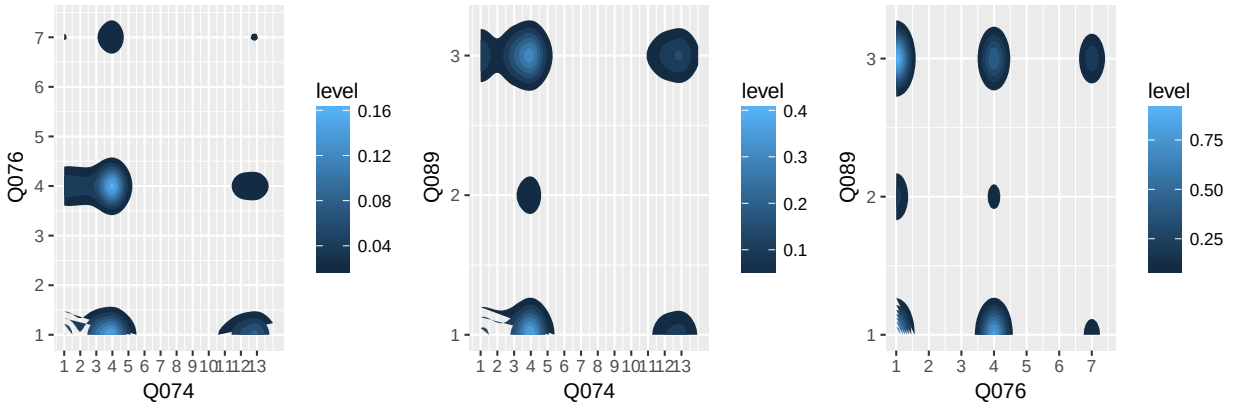

Figure 5: Pairwise 2D Kernel densitiy estimation for the correlation among the three question

5 Dimension Reduction

After the data cleaning process, we have a data matrix of dimension 44707×468. Since the categorical variables are all transformed into binary coding, the data is of extremely high dimension. Since the ultimated goal is discover the dialect pattern via clustering techniques, and while most clustering algorithms are distance-based (which means the similarity of two samples are based on some distance measure), extra noisy dimensions could easily bring in noise with magnitude comparable to signal. Notice that it is unlikely that all 468 features here reveal the dialect patterns. So before doing clustering, dimension reduction is of demand. The overall goal is to reduce the ambient dimension from 468 to a manageable number and thus reduce noise during this process.
Various classical dimension reduction techniques are experimented, including principal component analysis (PCA), independent compoennt analysis (ICA), and random projection. Notice that all of these methods falls into unsupervised learning category, that is, no prior labelled data would be used in learning the underlying lower dimension structures, so it's hard to evaluate their performance. However, since the ultimate goal for doing dimension reduction is to reveal dialect geographical pattern via clustering, the quality of clustering results could be used as an indicator for the performance of dimension reduction methods.
One way to do this is via subjective inspection at the resulting plots of clusters. For instance let c^1 and c^2 denote two possible clusterings of n data points. That is, c_{1i} is an integer denoting the cluster to which ith individual has been allocated in the first clustering, and inspect the closeness of samples belonging to the same cluster. But this method could be time consuming and tedious, since there are about 40,000 samples. Another way is to use the notion of Rand index to compute the similarity of two clusterings. The Rand index is defined as

$$R(c^1, c^2) = \frac{a+b}{C_n^2} = \frac{|(i,j)|c_i^1 = c_j^1, c_i^2 = c_j^2| + |c_i^1 \neq c_j^1, c_i^2 \neq c_j^2|}{C_n^2} \tag{1}$$

where where a is the number of pairs of individuals who are clustered together in both c^1 and c^2 and b is the number of pairs of individuals who are clustered separately in both c^1 and c^2. Note that the denominator is

simply the total number of pairs of individuals.
Each of the above mentioned dimension-reduction techniques will be evaluated using Rand index, after doing the basic K-means clustering. Also, the number of clusters existed in the original whole dataset is determined using Rand Index, which will be mentioned in the next section. For now, let's just assume as given that the number of clusters in the dataset is 4. This number is arrived at by gap statistic analysis and visual inspection of the result of k-means on full data.
Plotting clusters on map One challenge in plotting the clustering result onto the map is that the resulting clusters would be consisted of individuals. Visualizing every individual separately in a US map is problematic and clunky at best. Instead, the individual-level clusters are aggregated into county-level clusters. The counties would be used to show the clustering results, by following the majority voting rule i.e a county belongs to the cluster which contains the most number of samples from that county. So county level maps will be generated after embeded into the contour of U.S. For illustration, the clustering result of using the whole dataset is plottedg in Figure 7a.

5.1 Principal Component Analysis

The first technique to investigate is PCA. To decide on the number of PCs, fistly for all principal components, their variance and cumulative variance are ploted in Figure 6.

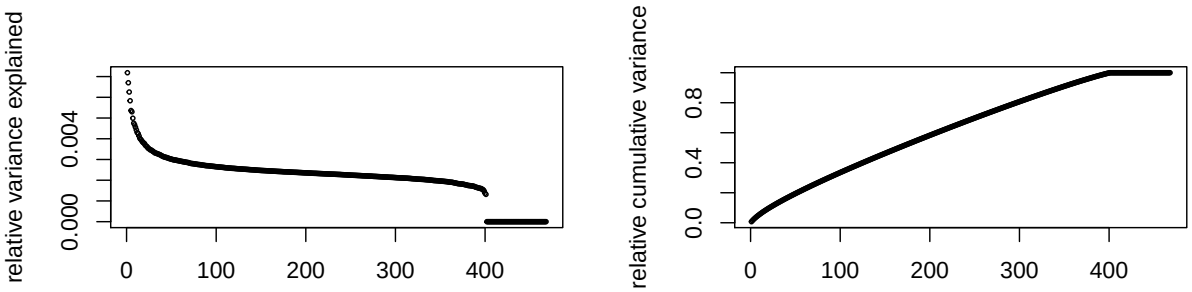

Figure 6: Variance (left) and Cumulative Variance of Principal Components (right)

Notice that the principal components do not have a sharp dropoff in variance, so there is no lower dimensional affine subspace which captures most variance of the whole data set. As a result, dimension reduction by PCA is not so helpful in this case, since a large number of principal components is needed to capture the variation of the data.
By extracting top k principal components $k \in \{10, 50, 100, 200\}$, the Rand index between the clustering result using K-means on whole dataset and the result for each of the 4 cases above are computed. Notice that from Table 1 we see that it is favorable to use the top 10 PCs, since although they explain less than 20% of total variance, the similarity between this clustering result matches the clustering reuslt using whole dataset well (Rand index=0.96). This also suggests that the whole dataset possibly retains a fair amount of noise, and using principal components can reduce the noise, which can be also justified by result in Table 2, where we see that smaller number of top PCs actually lead to a better Rand index even though they capture only a small proportion of the total variance For illustration, the clustering result of using top 10 principle components is plotted in Figure 7b.

Table 1: Rand index for clustering reuslt using k top principal components

k	10	50	100	200
Rand index	0.96	0.99	0.83	0.81

Table 2: Rand index for clustering reuslt using top principal components that explains $k\%$ of totla variance

k	50	80	100
num. of PCs	44	109	157
Rand index	0.94	0.97	0.94

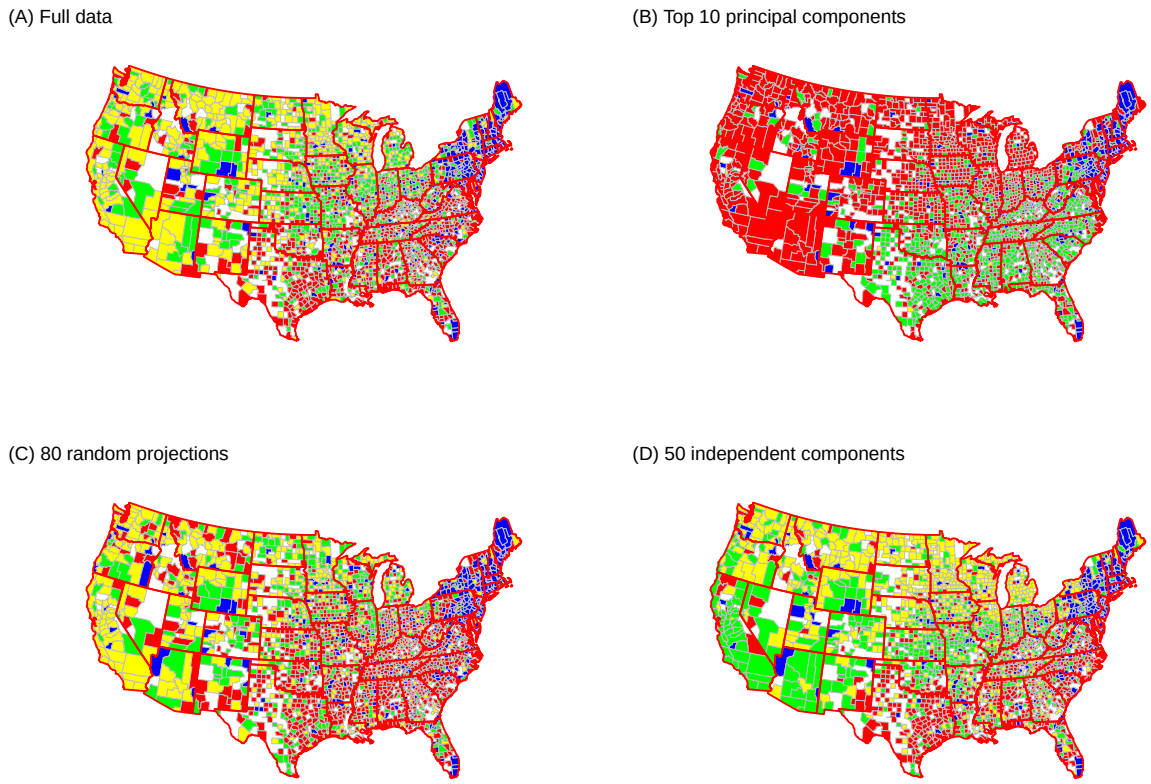

Figure 7: Comparison of 4-means clustering results on full data and various dimension reduction ap- proaches.

5.2 Random Projection

Random projection is an efficient dimension reduction technique in unsupervised learning problems, under the theoretical guarantee of Johnson-Lindenstrauss lemma. In random projection, the rows (samples) of the data matrix is projected onto a lower-dimension subspace with basis constraucted as affine combinition of random variables. The general idea is that samples' pariwise "distance" are preserved after projecting into the lower-dimensional subspace. The lemma guarantees that with high probability $(1 - \delta)$, all norms and inner products after projection are preserved within an accuracy of ϵ from the original values, if the dimension of the projected subspace is at least $O(log(n/\delta)/\epsilon^2)$, where n is the original sample size. In this

case where l_2 norm is used, suppose $F : R^d \to R^m$, where d is the original dimension, m is the after projection dimension,

$$P((1-\delta) \leq \frac{\|F(u_i - F(u_j))\|_2^2}{\|u_i - u_j\|_2^2} \leq (1+\delta)) > 1-\epsilon,\ m = O(log(n/\delta)/\epsilon^2) \tag{2}$$

Notice that clustering algorithms such as K-means depends only on pairwise distances, so this justifies the use of random projection as a valid dimension reduction technique for clustering.
To investigate the performance of random projection, m = 40, 80, 120 are experimented. The Rand indices between these clustering results and that of the orginal data is given is Table 3. For illustration, the clustering result of using 80 as random projection dimension is plotted in Figure 7c. Notice that the Rand

Table 3: Rand index for clustering reuslt using k random projection

k	40	80	120
Rand index	0.70	0.75	0.78

indices for random projections are in general less than that of PCA. This is an obvious consequence since random projection aims at preserving every feature with eqaul importance and PCA aims at capturing more important features. So it's likely that both signal and noise are retained after random projection.

5.3 Independent Component Analysis

The last method to experiemnt on is Independent Component Analysis, which aims at extracting few components which contain most of the signal which follows non-gaussian distributions, while the rest are mostly noisy and consequently more gaussian. ICA works by iteratively choosing directions with maximal non-gaussianity and then projects the data along these directions. Here, I experimented on using 30, 50, 100 top independent components. The Rand indices between these clustering results and that of the orginal data is given is Table 4. For illustration, the clustering result of using 50 independent components is plotted in Figure 7d. It's obvious that the Rand indices are low for clustering results after doing ICA. This might be

Table 4: Rand index for clustering reuslt using k independent components

k	30	50	100
Rand index	0.55	0.57	0.66

justified by the fact that the signals are originally Gaussian and thus ICA couldn't tell them apart from noise, due to its working mechanism.

6 Clustering combined with Data Reduction

6.1 K-means clustering

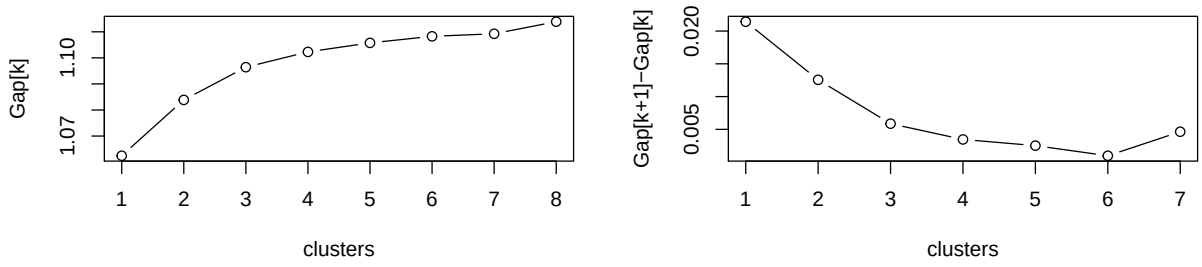

Figure 8: Gap statistics plot for the number of clusters

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Choice of the number of clusters is the first concern before application of any kind of clustering methods. In this case, this problem is solved from both the qualitative and quantitative perspectives. The Gap statistics is highly popular for selecting the number of centers for K-means.
The Gap statistics for number of cluster centers from 1 to 8 are computed and shown in Figure 8. So it's obvious that when the number of clusters goes to 4, the Gap statistics difference first falls below 0.005. Also, there is little increase after the number of clusters reaches 4.
Finally, visualization also plays an important role in determining the cluster number. In visual inspection, the aim is to find the number so that there are apparent geography groups while adding one cluster will lead to chaotic clustering caused by overfitting. It can be shown that 4 clusters make sense, as shown in Figure 7a.
With selecting the cluster number as four, affter applying k-means clustering on original data aggregated over county, the result is shown in Figure 7a. In order to see the effectiveness of the data reduction methods, k-means clustering is applied on part of those methods, as shown in Figure 7b, 7c, 7d. From the clustering results in Figure 7a, we can see New England and Florida belong to the same cluster. Many of the Northern dialects can trace their roots to this dialect which was spread westward by the New England settlers as they migrated west. It carries a high prestige due to Bostons early economic and cultural importance and the presence of Harvard University. In South Florida, there are those who consider that this region should be reclassified as part of the Northern dialect region. So many people from the North, particularly New York, have moved to south Florida that the majority of people tend to sound more Northern than Southern. The south part, is a continuous blue continuum, including South Midland, Virginia Piedmont, Southern Appalachina, and Gulf Southern. In general south, as the northern dialects were originally dominated by Boston, the southern dialects were heavily influenced by Charleston, Richmond, and Savannah. Compared with the Eastern United States, the Western regions were settled too recently for very distinctive dialects to have time to develop or to be studied in detail.

6.2 K-medoids clustering

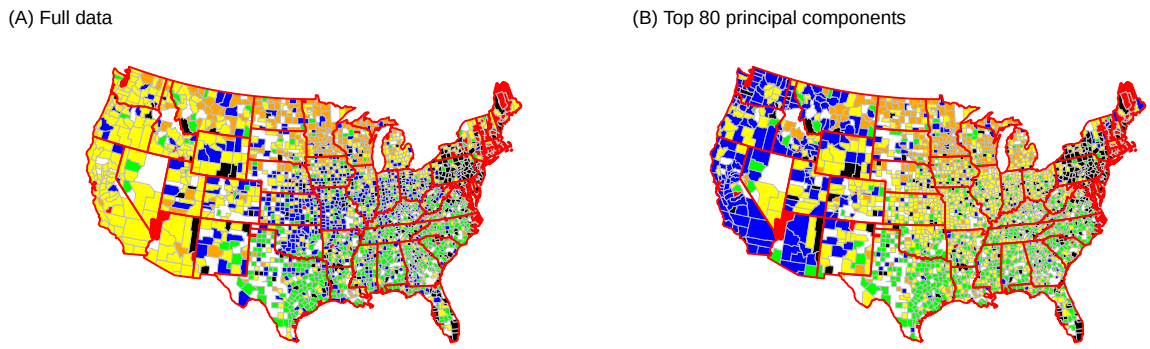

Figure 9: The clustering map for k-medoids when k = 6 on county level

The k-medoids clustering method is similar to k-means, except that the former takes k data points as centers while the latter makes use of means from k disjoint sets. The k-medoids algorithm first picks up k data points randomly, and assign other points to their nearest centers, thus forming k clusters. Within each cluster, select the data point which minimizes the sum of the distances between it and other points. Repeat the previous steps until convergence. The K-medoids clustering using here is based on the data of county level because for individual level there will be too many samples for the algorithm to work. Instead of using Gap statistics to determine number of clusters, here I rely on visualization to determine the number of clusters because both the former methods tend to be conservative, and the performance looks good when the cluster number is larger. Eventually the cluster number is decided to be six.
The performance of k-medoids on dialect clustering is tested on dataset with and without dimension reduction. Apart from checking the clustering from raw data, PCA, ICA and Random Projection are used to reduce the data dimension and eliminate noise within the data. For illustration purpose, the K-medoids county-level clustering result for raw data and top 80 pricipal components are plotted, shown in Figure 8.
Comparing and , we see that the clustering results based on 6 clusters with top 80 PCs are quite satisfactory. The violet region in (b) aggregate Rockey Mountain, Pacific Northwest, Pacific Southwest in (c). The brown region in (b) corresponds to Upper Midwestern, Chicago Urban in (c). The blue region in (b) corresponds to New England, Inland Northern in (c). The red region in (b) corresponds to North Midland in (c). The yellow region in (b) corresponds to the main part of South Midland, Southern Appalachian, Coastal Southern in (c) while the green region in (b) corresponds to the main part of Southwestern, GulSouthern in (c). The result of k-medoids can be seen as a refinement of k-means. The k-medoids has separated Upper Midwestern and Southwestern out from Western regions and Southern regions respectively. Moreover, the Mexican dialect of Spanish had an significant influence on this area because there had already been as many as ten generations of Spanish speakers live there by the time Southwestern became part of the United States

6.3 Finding Critical Questions dominating clustering Results

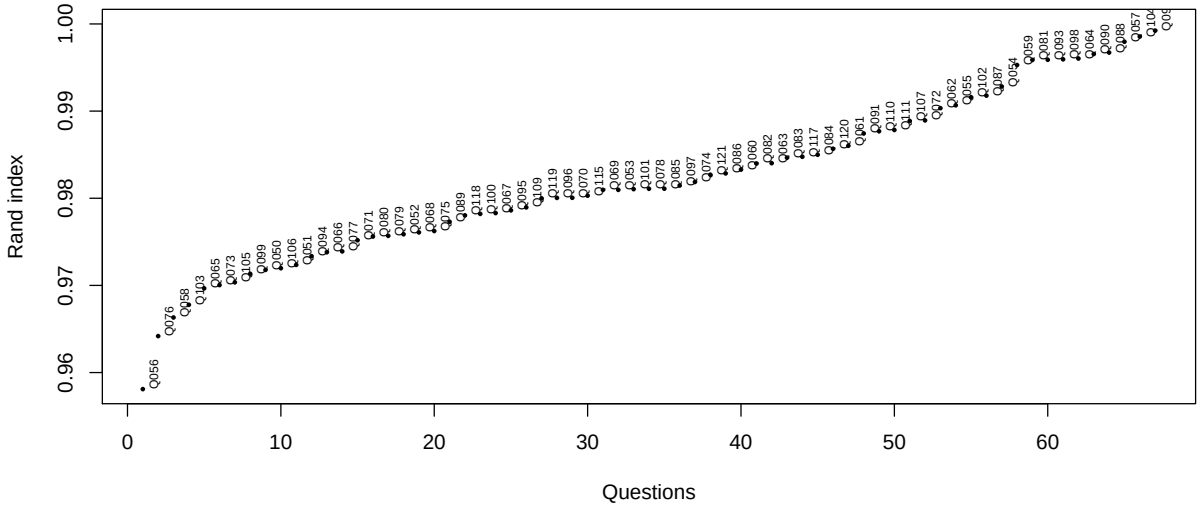

Figure 10: Rand Index for question removed clusterings

Here the target is to figure out which of those questions are most critical for clustering. To be consistent, k-means clustering is always performed on original data aggregated over county. So essentially the k-means results of the whole dataset is set as standard clustering. To see how important a question is, the idea is to remove the corresponding column for that question from the original dataset, and then apply k-means clustering on the new dataset. After that, using the method mentioned previously, the similarity measure between this clustering results and the standard clustering can be computed. This process is performed on each of the 67 questions for 10 times and the average similarity measure (Rand index) is plotted for each extracted question, shown in Figure 9.
It's obvious that the Rand index experience the most change for taking away question 56, 76 and 58, which means these questions are important for determining clustering results.

7 Stability of findings

To figure out the stability of above findings, K-means method is selected to test the robustness of the clustering results, based on county-level raw data. Several ways can be applied to perturb the data, i.e., adding noises, removing some columns (excluding some questions) or removing some surveyed people. But due to above mentioned conclusion that different questions possess distinct weights on determining separate groups or determining the continuum, its not appropriate to sample questions uniformly for testing. Besides, question design is the very thing researchers can control, so there is no need to check the influence of the missing questions. Also, excluding some individuals who took part in the survey is equivalent to adding noises to the data. So here I only consider adding some gaussian noises on the data columns and the data rows respectively, and check how robust K-means is, after different levels of gaussian noises are added to the data. The levels of noise are 0.05,0.1,0.2,0.3, 0.4, each level with 10 replicates. Here, the noise of level 0.1 means it comes from the gaussian distribution with mean of zero and standard deviation of the sample mean multiplied by 0.1.

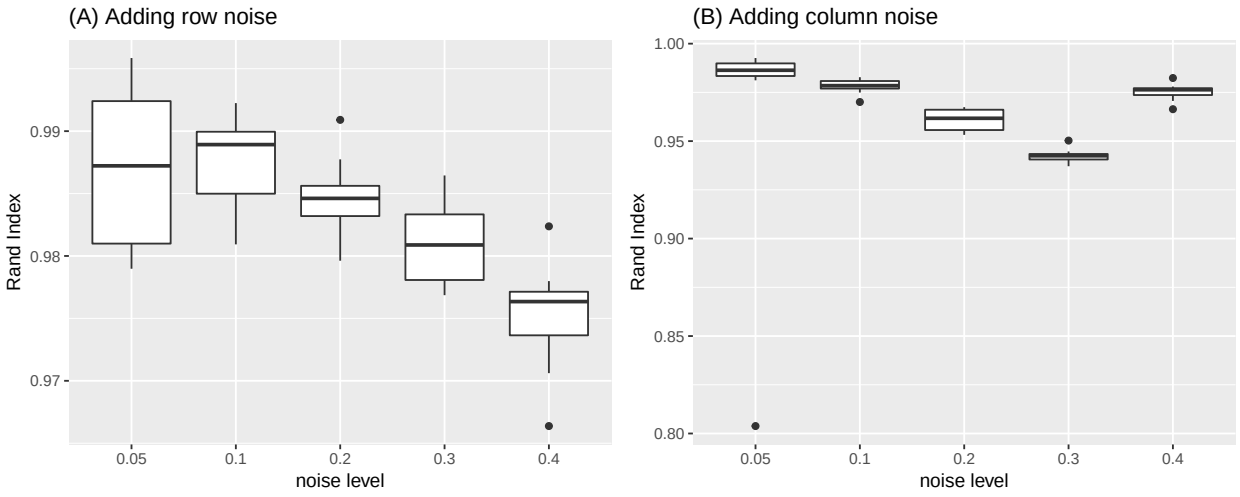

Figure 11: Rand Index between the clustering of noised data and the raw data

As the result in Figure 10 shows, the k-means clustering performs quite robustly with different levels of noises. When the noise becomes stronger, the corresponding clustering decreases from the standard one as expected. But notice that the Rand Index is always higher than 0.96 after adding row noise and higher than 0.94 after adding column noise. Slight difference is observed when adding noises to data rows and data columns respectively, that the Rand index after adding same level of noise to column would be lower than that of rows. Such phenomenon might indicate that changing questions may lay more influence on the results than changing surveyed subjects.

8 Conclusion

Harvard Dialect Survey provides us with a great deal of useful information to investigate the dialect geography of the United States. After looking into several questions, we find there are obvious geographical differences among responses. Since single question cannot reflect the dialect distribution comprehensively, we move on using some clustering methods based on data reduction to separate distinct language regions. Plenty of dimension reduction methods have been tried to eliminate the data noise, such as PCA, ICA, Random Projection, k-means, and sparse PCA. Then the dimension-reduced data are input into various clustering methods, including k-means, NMF, k-medoids and hierarchy clustering. The analysis leads to a convincing finding that the American dialect geography can be divided into multiple parts (analyzed in Section 5), which agrees with geographical partitions, and the partitions agrres with literature study, which proposes the following dialect map of American English.

Figure 12: Dialect Map of American English