Lab 2 - Linguistic Survey
STAT215A, Fall 2014

Russell Chen

October 7, 2014

Kernel density plots and smoothing

1 Kernel density estimate of temperature distribution

Temperature data from the redwood dataset was used to explore kernel density estimation. I experimented with 4 different kernels (biweight, rectangular, epanechnikov and gaussian), each fit with 4 different bandwidths. The plots of the estimated density are shown on the next page, overlaying a scaled histogram of the temperature data for reference. Density estimates fit with smaller bandwidths correspond to the rougher curves and vice versa. This follows from the definition of bandwidth. In general, it seems that the choice of kernel does not have a large impact on the final density estimate. The bandwidth definitely affects the final estimate more since it strongly influences the amount of smoothing. The bandwidth was changed using the `adjust` parameter in `geom_density` but the values used are not the actual bandwidth of the kernel. For the data here, the 4 `adjust` parameters used were 0.2 (roughest), 0.8, 2 and 5 (smoothest). The density estimates corresponding to the smallest bandwidth is probably too rough, as can be seen by the frequent black spikes on the plots. In contrast, the density estimates corresponding to the largest bandwidth is probably too smooth, as can be seen by the yellow curves which do not fit the histogram closely. The two intermediate bandwidths used in this case are more appropriate.

2 Loess smoothers for temperature against humidity

Using the redwood dataset again, temperature and humidity values across all nodes were plotted, restricting ourselves to a single time of day. Loess smoothers were fit, with polynomials of degree 0, 1 and 2, corresponding to locally constant, locally linear and locally quadratic fit respectively. The amount of smoothing was varied using the `span` argument. As can be seen on page 3, the polynomial degree chosen does not affect the final fit nearly as much as the amount of smoothing does. As is the case with the role of bandwidth in kernel density estimation, a smaller span corresponds to a rougher fit and vice versa. The blue curves plotted are probably too smooth while the red curves are too rough. The green curves, which takes the middle ground between the two, is probably the most appropriate.

Figure 1: One-dimensional kernel density estimation with various bandwidths

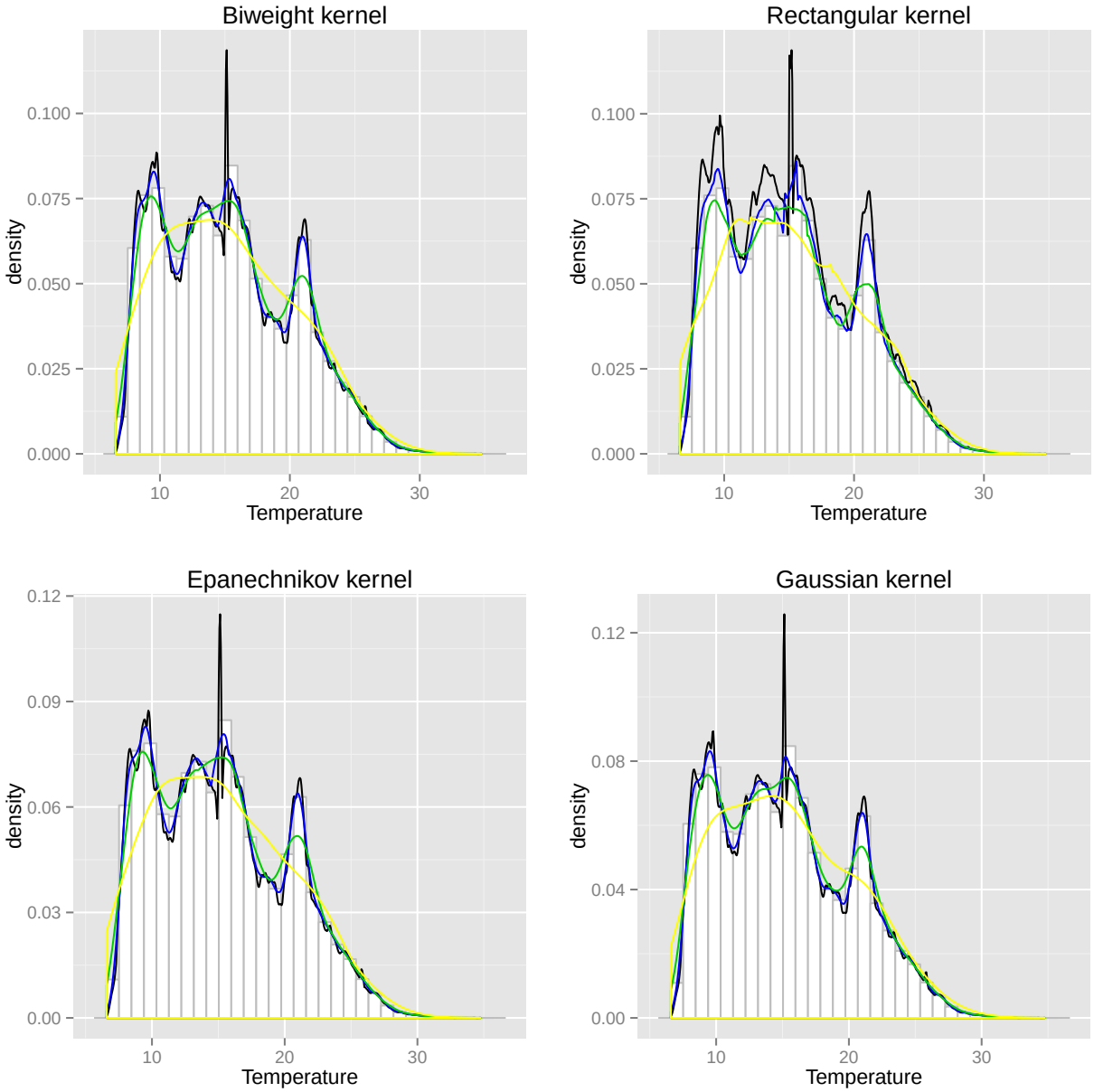
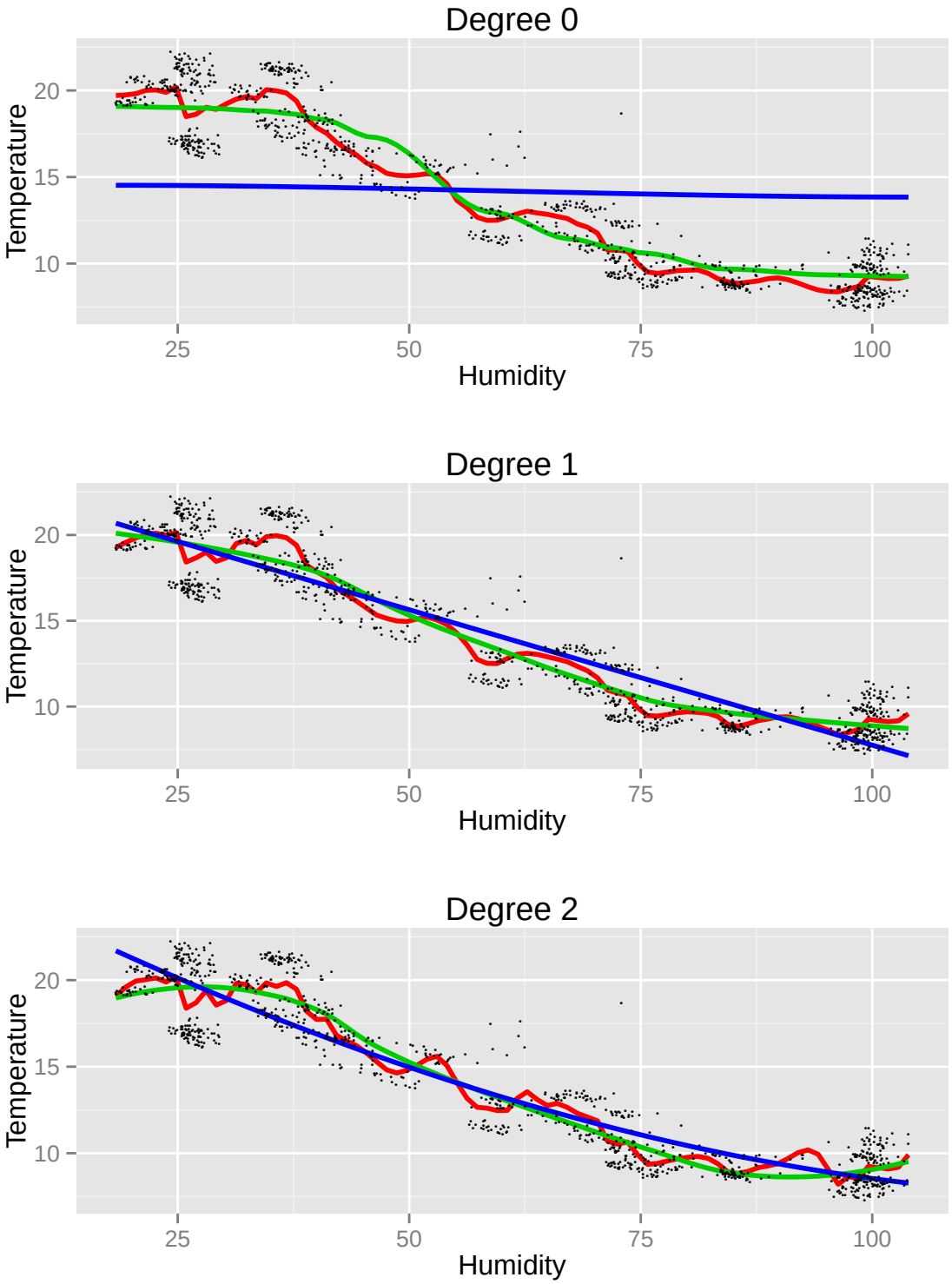


Figure 2: Loess smoothers with varying amount of smoothing



Linguistic Data

1 Introduction

Data from 47471 respondents to an internet survey of 122 questions about how they use the English language was analyzed. We are only interested in 67 of the questions, which deal with lexical differences rather than phonetic differences. After cleaning and exploring the data, dimension reduction was carried out.

2 The Data

2.1 Data quality and cleaning

Of the files provided, I mainly checked the raw dataset `lingdata` for consistency while assuming `lingloc`, `all.ans` and `quest.use` were constructed correctly from this raw data or scraped correctly from the web. In its raw form, `lingdata` has 47471 observations of 73 variables. Each observation records one person's answers to 67 questions together with his City, State, ZIP code, latitude, longitude as well as an ID number.

I checked that the question columns were all encoded with numeric data and that there were no missing values in those columns. There are also no duplicate ID numbers. The first issue that came up was the missing latitude and longitude for 1020 observations (rows). While city and state data were available for some of these rows, I decided not to write a script to scrape the latitude and longitude of these places from the internet due to time constraints and concerns about accuracy. Instead, I decided to remove the offending rows, leaving 46451 rows. Next, I checked for observations which recorded no response to any of the 67 questions of interest. There were 1015 such observations and these were removed since they contain no linguistic information and there is absolutely no way to recover the missing data. 45436 observations are left at this stage. The state column is also severely corrupted, with values such as XX, 94, C) and !L. A brief look at observations with the state labelled XX shows that they are from different states so this is not a case where all observations from a certain state were mislabelled. The state column was removed.

Next, I looked at `all.ans` to see whether the percentages for each question summed to 100, within a small tolerance of 0.1 percentage point. All of them did. I looked to see whether there were options for which it is recorded that 0% of the respondents picked. I found 6 such questions but decided against removing them at this stage since the percentages may have been so small as to have been rounded down. After constructing the binary matrix with $n = 45436$ and $p = 468$, I checked for the empty options again by looking at the columns, each of which correspond to a single option. None of the columns sum to zero, indicating that there are in fact no empty options. The last bit of data processing I did was to remove observations with a low rate of response to the questions of interest. While these observations still do contain some data, retaining them may give spurious clustering results later since encoding all non-responses as zeroes gives the illusion of agreement on a question where we actually have missing data and thus, no infomation. 552 respondents who answered less than 60 out of the 67 questions of interest were identified and these rows removed to leave 44884 rows. Finally, a plot of latitude against longitude shows no obvious location outliers so that concludes the data cleaning.

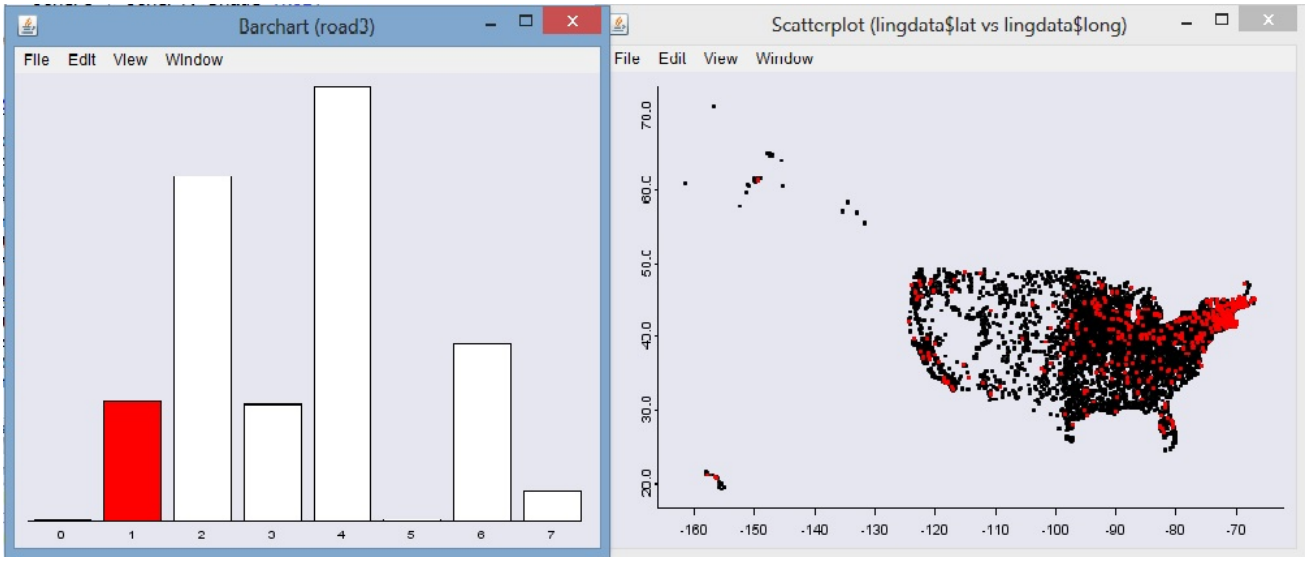


Figure 1: New Englanders refer to traffic circles as rotaries

2.2 Exploratory Data Analysis

Most of the exploratory data analysis was done with linked brushing using the `iplots` package. These cannot be reproduced here so results will be illustrated with screenshots instead. First, I will examine a couple of questions separately and investigate their relationship with location. Consider the following question:

> "What do you call a traffic situation in which several roads meet in a circle and you have to get off at a certain point?"

The options are:

- rotary
- roundabout
- circle
- traffic circle
- traffic circus
- I have no word for this
- other

A linked plot shows that people who chose the first option, rotary, overwhemingly reside in the New England area. On the left side of Figure 1 above, a bar chart of the data from this question is shown. The bar corresponding to the first option is selected (zeroes correspond to a non-response), thereby showing up filled red. The people who chose option 1 then show up in red dots on the scatterplot of latitude vs. longitude to the right. This geographic pattern is quite clear.

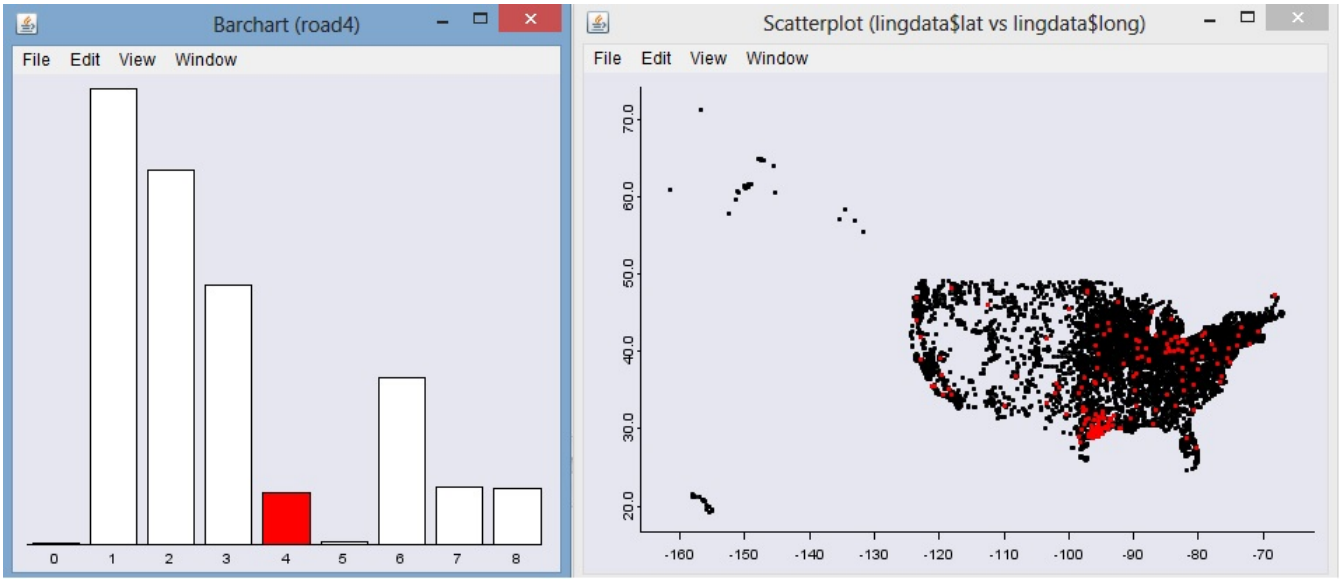


Figure 2: Use a feeder road to get onto the highway in Houston

Consider another question:

"Which of these terms do you prefer for the small road parallel to the highway?"

The options are:

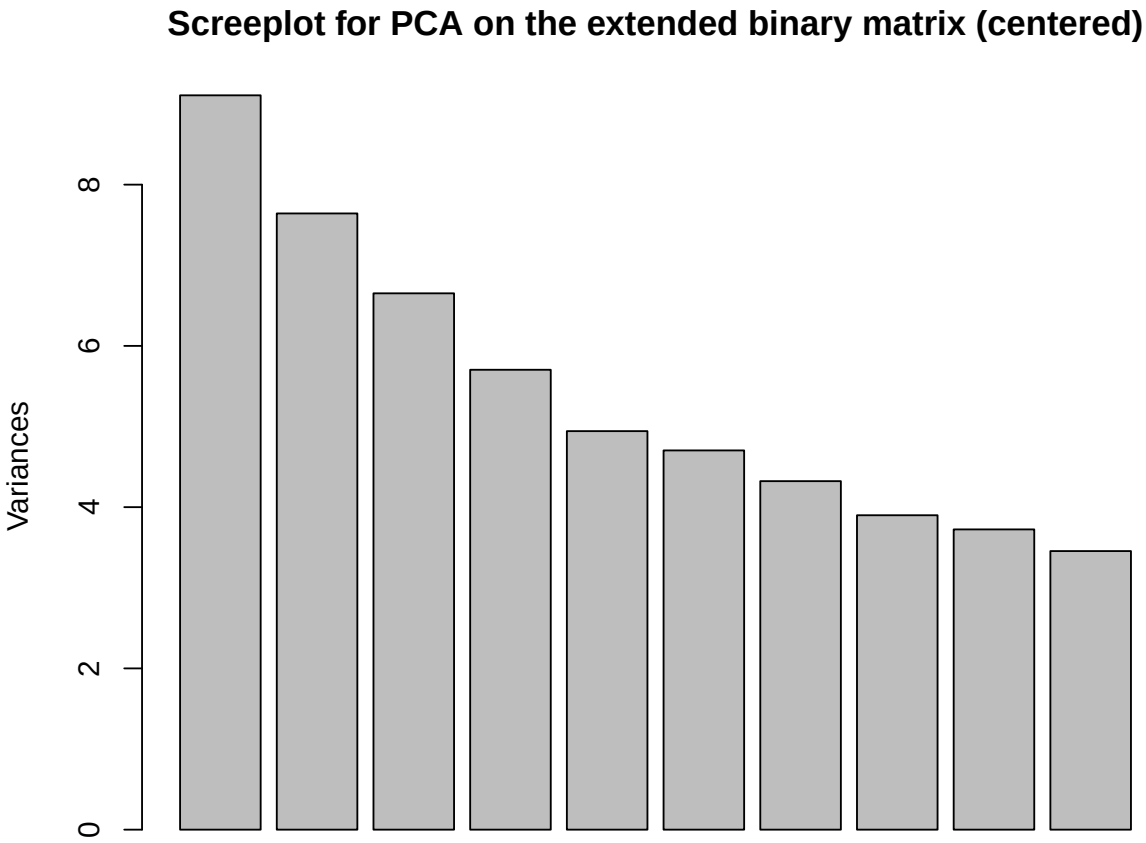
- frontage road
- service road
- access road
- feeder road
- gateway
- we have them but I have no word for them
- I've never heard of this concept
- other

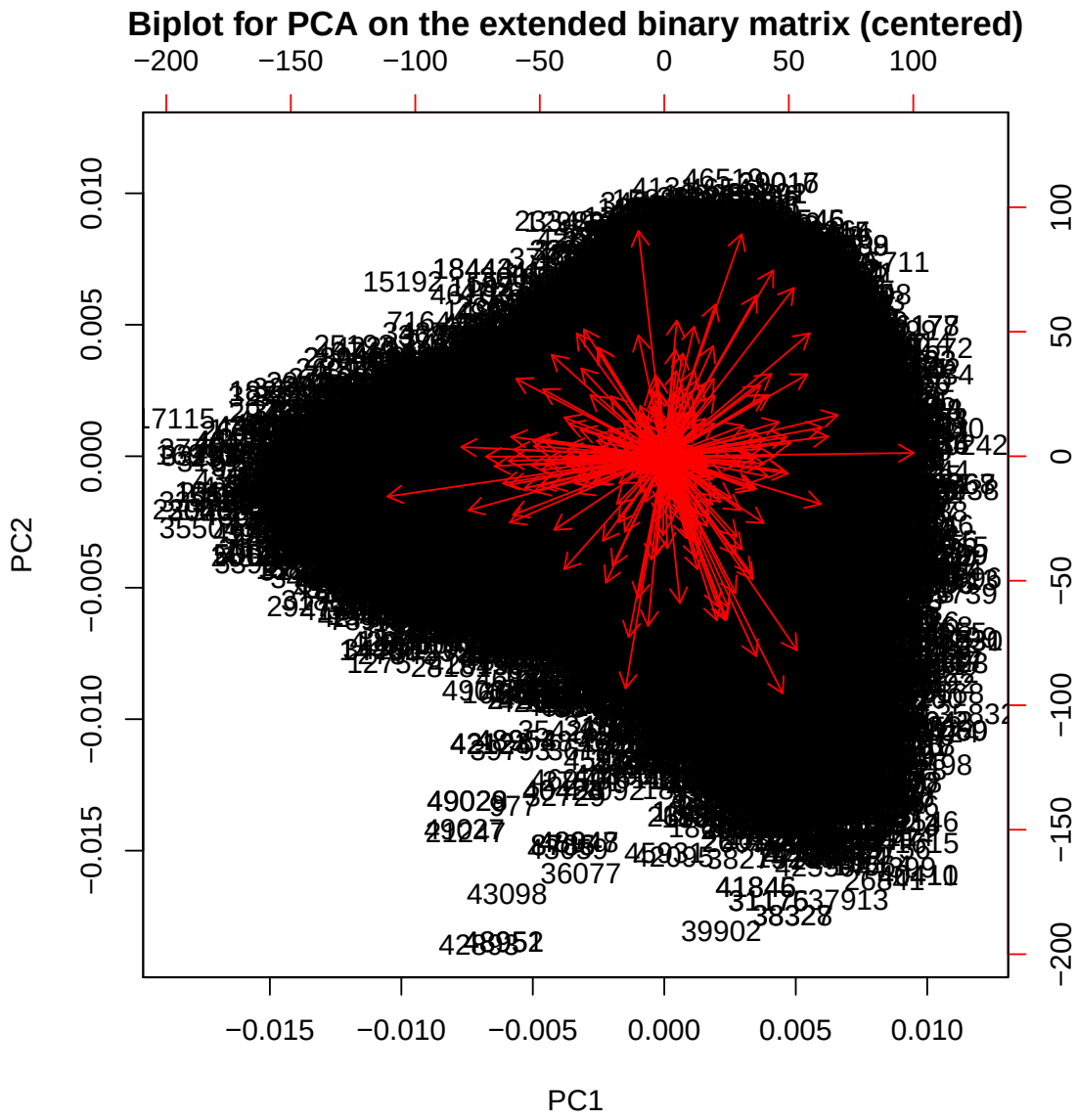
A similar linked plot shows that respondents who selected option 4 'feeder road' overwhemingly resided in the southern part of Texas. A cursory internet search shows that it is indeed characteristic of people who live in Houston [1].

3 Dimension reduction methods

Working with 468 variables is not easy; it would be great if we could reduce the dimension somewhat in order to carry out further analysis. For most dimension reduction methods, a distance matrix describing the dissimilarities between the observations is needed. Unfortunately, using the `dist()` function in R to calculate the manhattan distance (equal to the hamming distance in this case) between rows of the extended binary matrix was not practical due to the memory issues in creating a 44884×44884 distance matrix. As a result, I was unable to try nonlinear dimension reduction methods like ISOMAP [2]. This is unfortunate since it has had previous success in revealing manifold structures embedded in high dimensions, as one might suspect is happening here where a geographic relationship between respondents is embedded in high-dimensional linguistic data.

Principal components analysis (PCA) was carried out after centering the data and the screeplot is shown below. Unfortunately, it is not obvious how many principal components should be kept since the eigenvalues decrease very smoothly. The biplot for the PCA is shown on the next page




Biplot for PCA on the extended binary matrix (centered)
−200
−150
−100
−50
0
50
100
0.010
0.005
0.000
−0.005
−0.010
−0.015
PC2
100
50
0
−50
−100
−150
−200
−0.015
−0.010
−0.005
0.000
0.005
0.010
PC1
36077
43098
39902

4 Stability of findings to pertubation

I ran PCA on two bootstrap samples of the extended binary matrix. The screeplot, biplots and summaries of both these PCA runs are similar to the original PCA described in the previous pages. This indicates that the PCs are relatively stable.

References

[1] https://answers.yahoo.com/question/index?qid=20070822220340AA2d5QU

[2] http://isomap.stanford.edu/