MSCR 509: High Dimensional Analysis

Homework 11

Anish Shah

April 20, 2020

Question 1

The file places.txt contains data from Places Rated Almanac data (Boyer and Savageau) which rates 329 metropolitan areas of the United States according to nine composite variables:

- Climate and Terrain
- Housing
- Health Care & Environment
- Crime
- Transportation
- Education
- The Arts
- Recreation
- Economics

Each composite variable is constructed from available data, for example, housing costs are the sum of 3 components: utility bills, property taxes, and mortgage payments. Utility bills are a function of gas and electric prices, heating degree days (when the temperature is less than 65 0 F), air conditioning degree days (when the temperature is more than 800), and how common it is for houses to be heated electrically rather than by gas.

Notes:

- The data for many of the variables are strongly skewed to the right.
- The natural log transformation was used to normalize the data.
- The places_data.sas file has code to read dataset and log transform 9 variables.

Researchers are interested in learning ranking of places to live in the United States. One way of doing this is to principle component analysis.

Part 1

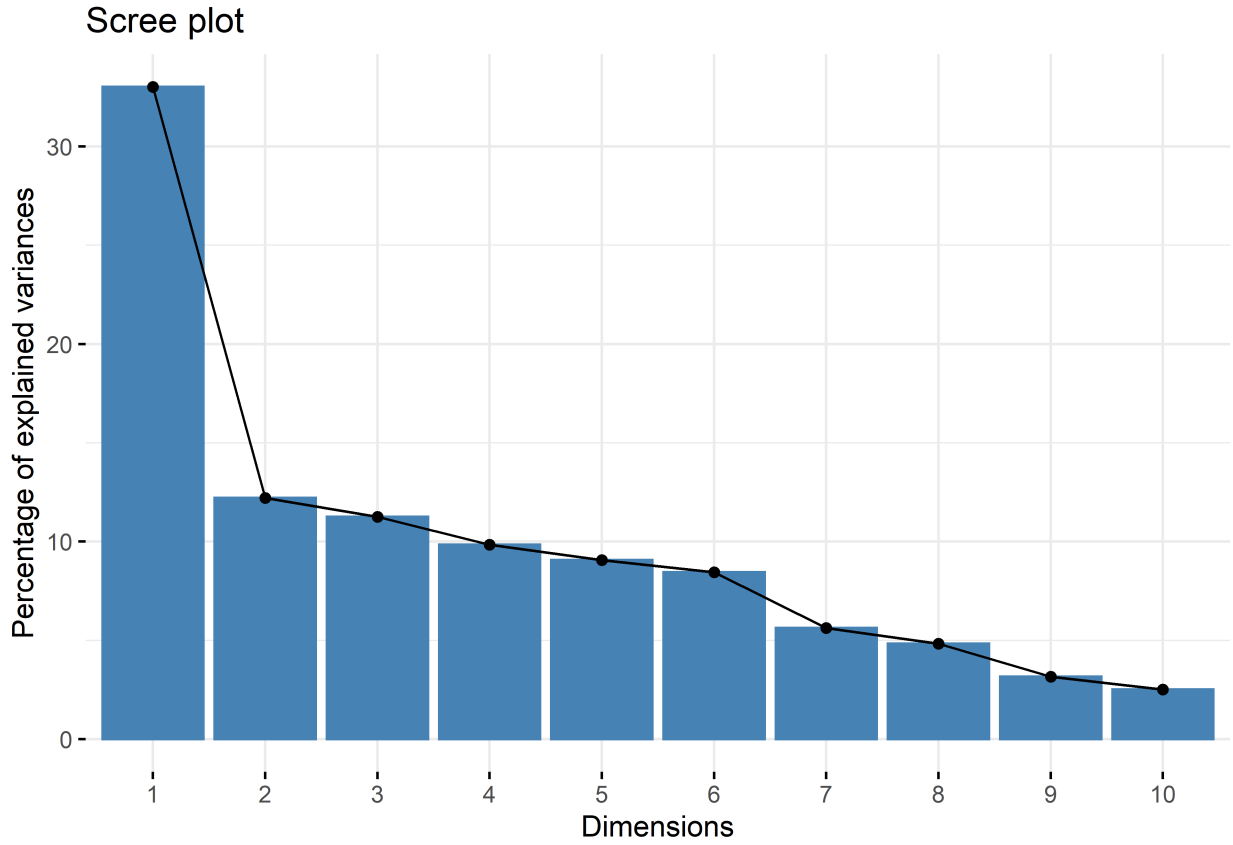
Report the principal components (i.e., the loadings associated with each PC) of these variables based on correlation matrix (i.e. without COV option).

Table 1: Loadings of PC

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
climate	0.1585	0.1359	-0.7152	0.3318	-0.3873	0.0349	0.2160	-0.1451	0.3391	-0.0291
housing	0.3849	0.1569	-0.0853	-0.1397	-0.1766	0.5623	-0.0827	-0.2752	-0.6074	0.0396
health	0.4092	-0.3679	-0.0362	-0.0492	-0.1118	-0.0370	-0.5347	0.1354	0.1516	-0.5936
crime	0.2590	0.4572	-0.0018	0.3289	0.0141	-0.6377	-0.1394	0.1043	-0.4193	-0.0526
trans	0.3737	-0.1658	0.1377	0.0522	0.4224	-0.1837	0.3232	-0.6788	0.1246	-0.1346
educate	0.2734	-0.4677	0.1582	-0.1150	-0.4595	-0.2377	0.5272	0.2571	-0.2109	0.1094
arts	0.4731	-0.1080	-0.0140	0.0268	0.1454	-0.0070	-0.3211	0.1241	0.2551	0.7475
recreate	0.3537	0.2942	-0.0063	-0.0207	0.4130	0.2891	0.3944	0.5560	0.1329	-0.2267
econ	0.1636	0.4768	0.5584	-0.1164	-0.4727	0.0439	-0.0013	-0.1433	0.4153	-0.0462
id	0.0448	0.2017	-0.3518	-0.8539	0.0370	-0.3143	0.0038	-0.0451	0.0451	0.0085

Part 2

How many principal components would you select to summarize this data? Explain why?



Based on the scree plot of the variance, we can see that by the **PC4** we have dipped below the eigenvalue of 1, thus limiting the amount of novel information we are receiving. The first three PC would explain approximately 56% of the total variance, thus may be sufficient in our analysis.

Part 3

What are the interpretations of the first three PCs?

When we look at the contributing loadings of PC1-3, we see that PC1 and PC2 use a linear combination of

all the variables for the most part. However, PC3 only relies on several variables: *climate, trans, educate, econ, and id*. We can however summarize and say that hte first 3 PC explain almost 6% of the variance, and thus may be a sufficient to start modeling to predict rankings of places to live.

Part 4

Explain why PCA is a good technique for this problem.

PCA is a good technique for this problem because there are 10 attributes that are contributing to rankings, however each of them are likely to have some level of correlation. A regression model for rankings will likely be biased and underpowered at estimating the best rankings because of overfitting (with too many variables). With just PC1-3 we can make a fairly well fitting model.