Machine Learning for Social Science - Lab 3

Marc Sparhuber

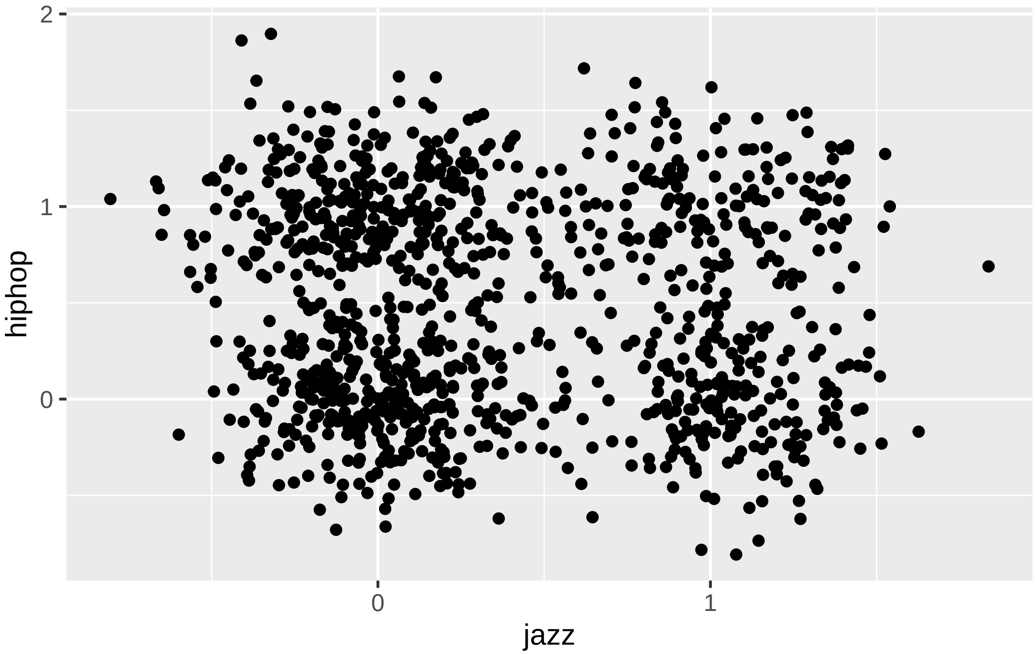
Table of contents

Part 1: Taste clustering and influence

Task 1

Begin by importing the file “taste_influence.csv”. Report the number of rows and columns of the data set, and the genres contained in it. Create a scatter-plot of two combinations of genres of your choice. Based on this, do you get any indication that the data is clustered along musical tastes?

```
[1] 1075    4
```



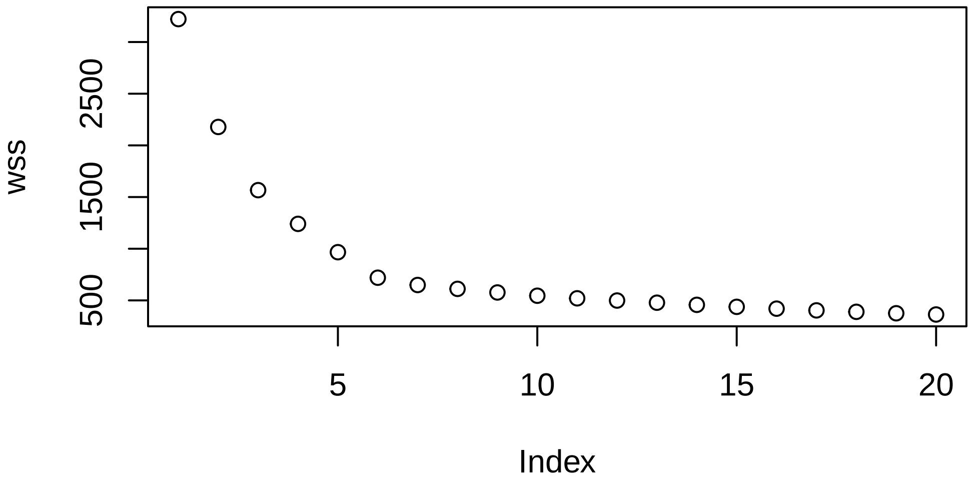
1075 observations and 4 variables. Based on this very basic visualization there seems to be a slight clustering at the poles, i.e., where the two music taste are near 1 or 0, with fewer values inbetween.

Task 2

Now you shall do some clustering. To prepare the data, do the following: (i) store/copy the data to a new R object, and subset it so that it only contains the three “taste columns”—these are the columns you will cluster based upon, (ii) standardize this data table (hint: you can e.g., use scale() for this purpose), (iii) transform it into a matrix (hint: e.g., by using as.matrix()).

Task 3

Having formatted the data according to #2, you shall now use the kmeans algorithm to cluster your data. Recall that a requisite for running kmeans is that the parameter k has been specified. In practice—and as is the case here—we often do not know the appropriate number of clusters a priori. Therefore, you shall implement a loop that, at every iteration, runs kmeans with a different number of clusters, and extracts the total within cluster sum of squares (hint 1: which can be extracted using $tot.withinss | hint 2: set the argument nstart=100 to ensure robustness of the local optima you find). Consider no. clusters ranging from 1 to 20, with an interval of 1. Plot k against tot.withinss. Which number of clusters do you find appropriate? Motivate.



Using the scree plot we can see that after k = 6 only very small gains in the accounted for variance are made so that the trade-off with added complexity becomes less worthwhile. I think it might even be reasonable to go for a smaller k than this.

Task 4

For the specification (of k) that you decided on in #3, extract the centroids and interpret each cluster in terms of what distinguishes it from the rest. Do the clusters seem meaningfully distinct?

1. Jazz dislikers, pop lovers, hiphop appreciators.

2. Jazz lovers, pop & hiphop dislikers.
3. This cluster is very similar to 1., indicating that perhaps I chose too large a k.
4. Jazz lovers, pop dislikers, hiphop appreciators.
5. Doesn't like any of these genres.
6. Jazz & pop disliker, hiphop lovers. All in all, I think k = 5 would have been a better fit, looking at cluster 3.

Task 5

To get a feeling for the role that the choice of k plays, estimate another kmeans model but this time with k = 2. Inspecting the centroids, how does your clustering change; how does it alter your understanding of the population?

> If this had been my choice I would've overlooked different parts of the population entirely because now I can only see: 1. People who dislike Jazz, kind of like Pop and are neutral toward hiphop, 2. and people who really like Jazz, dislike Pop and are also neutral toward hiphop. So this kind of clustering kind of fails to take into account people at the extremes of the hiphop distribution as their own groups which kind of shows how big of an influence the "manual" choice of k can really have on your results.

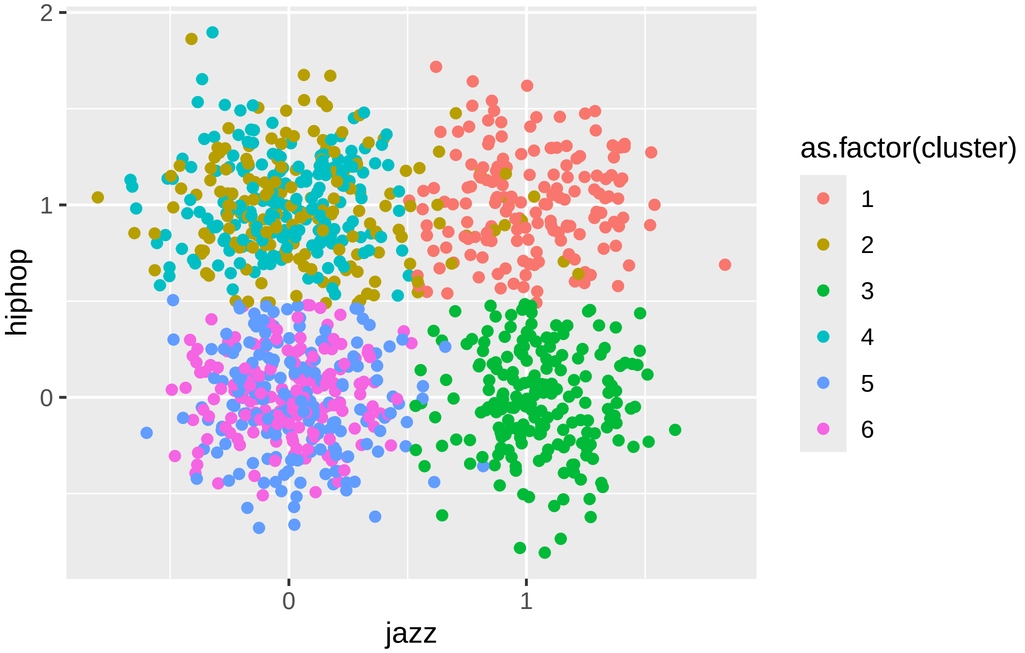
Task 6

Clustering provides a tool for discovering underlying structures in our data. Once these structures have been discovered, they can be studied in separate analyses. That is what you shall do now. We want to examine whether different "taste types" have differential degree of influence on others. To do so, (i) create a new column in your original data set storing the the retrieved cluster assignments (hint: you find the cluster assignments using $cluster). Then (ii) estimate a linear regression with the influence score (infuence) as the outcome variable, and the clusters (formatted as a factor) as predictors. Interpret the results: are there any difference in influence between the clusters?

> It seems that clusters 3 and 5 are significantly negatively associated with influence (those who like pop, some hiphop but no jazz, and people who like none of the 3 genres), whereas cluster 4, the jazz lovers who also like some hiphop, are significantly positively associated with it. Also in cluster 2 (jazz lovers, other things dislikers), a less but still significant negative effect is apparent. I wouldn't really dare to draw any conclusions from this because the findings seem not to be generalizable along the lines of how many things you like etc. so maybe people who like jazz and hiphop are just cooler than others?

Task 7

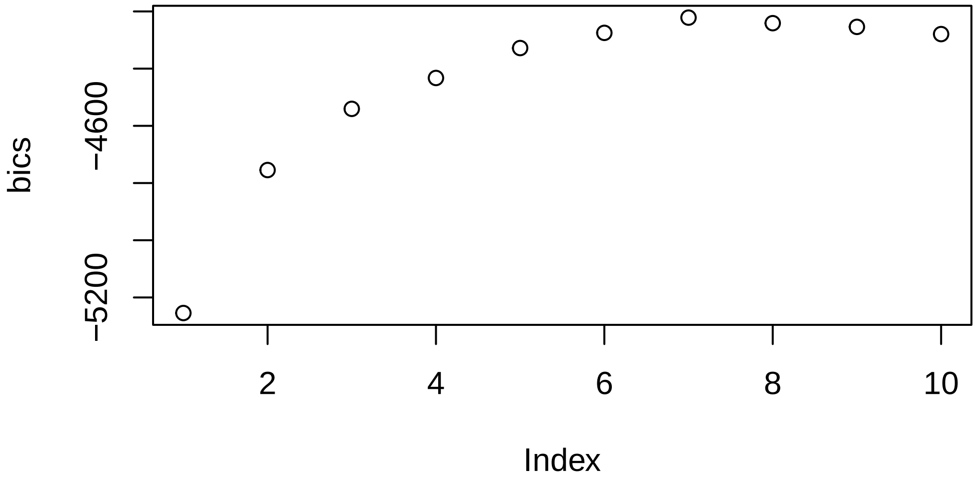
Now that you have merged the cluster assignments to the original data, produce the same plots as you did in #1, but now colored by the cluster assignments. Does it look like kmeans have picked up on the patterns you observed in #1? Further—what you think of the separation between the clusters? Is there clear spacing betweeen the clusters, or are the borders almost touching each other (note that there will be certain overlap due to plotting the data in 2D)?



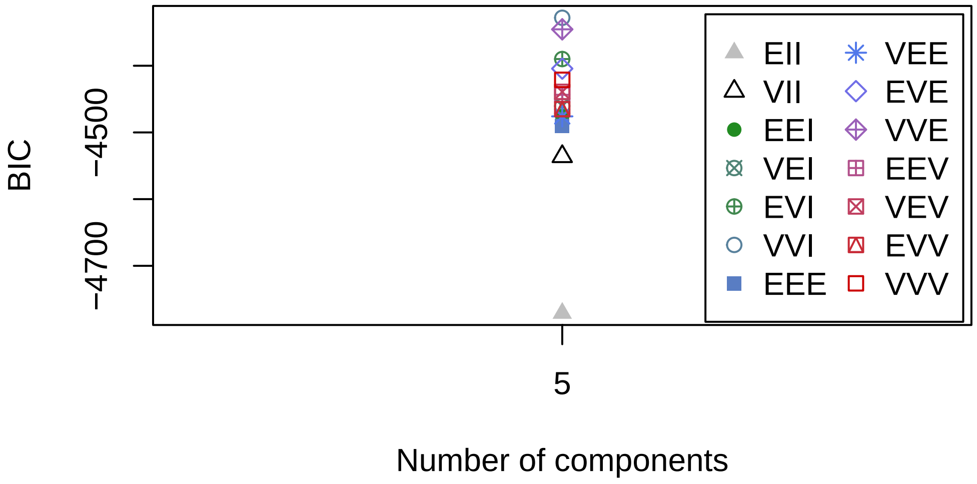
> I think it picked up the split visible in #1, though more than what I hoped for due to choosing k = 6. Some of the clusters seems fairly distinct, whereas there is substantial overlap between 3 and 5 and 1 and 6 which makes sense as for example 1 and 6 are to a large extent differentiated from one another by their differing opinions on pop music.

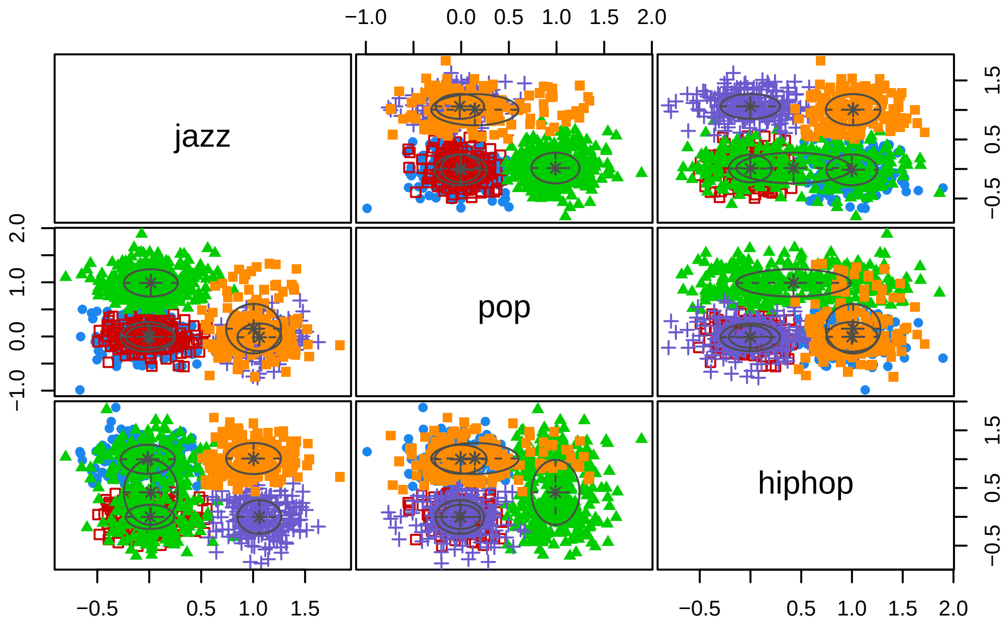
Task 8

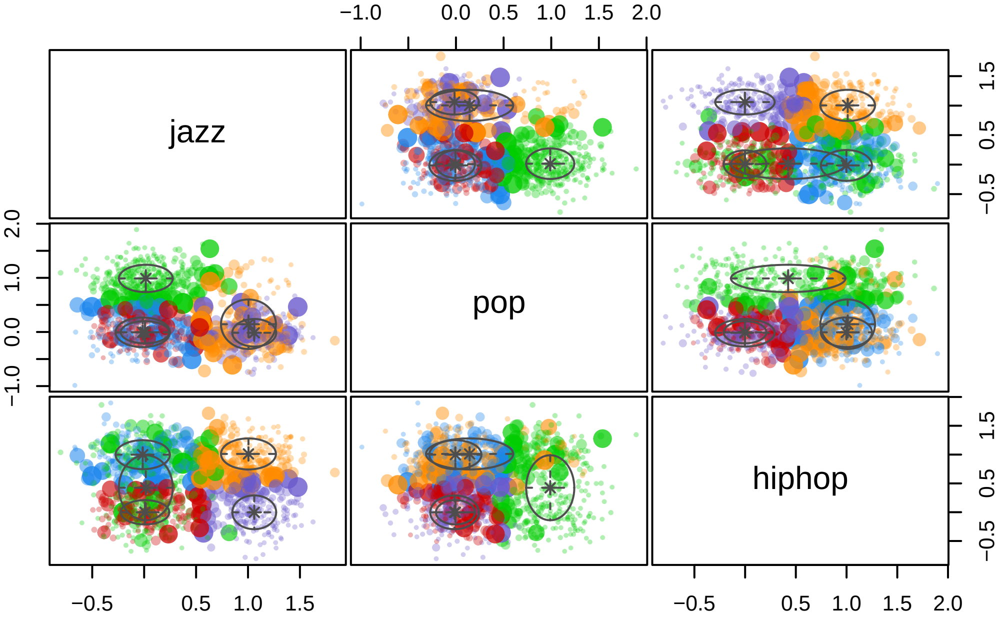
Repeat step 3–7 (but skip #5) using now instead a Gaussian mixture model. For this, you may use mclust's function Mclust() (specifying the number of components with the argument G). For #3: Note that, because this is a probabilistic model, we retrieve a likelihood score (or, more specifically BIC which is based upon the likehliood score but also penalizes for complexity) to measure its performance instead of total within cluster sum of squares (hint: you can extract the BIC by `$bic` on the model object). For #4, you can use $parameters$mean to extract the means/centroids of each cluster. For #6, you shall extract the hard cluster assignments (which you can do using $classification).1

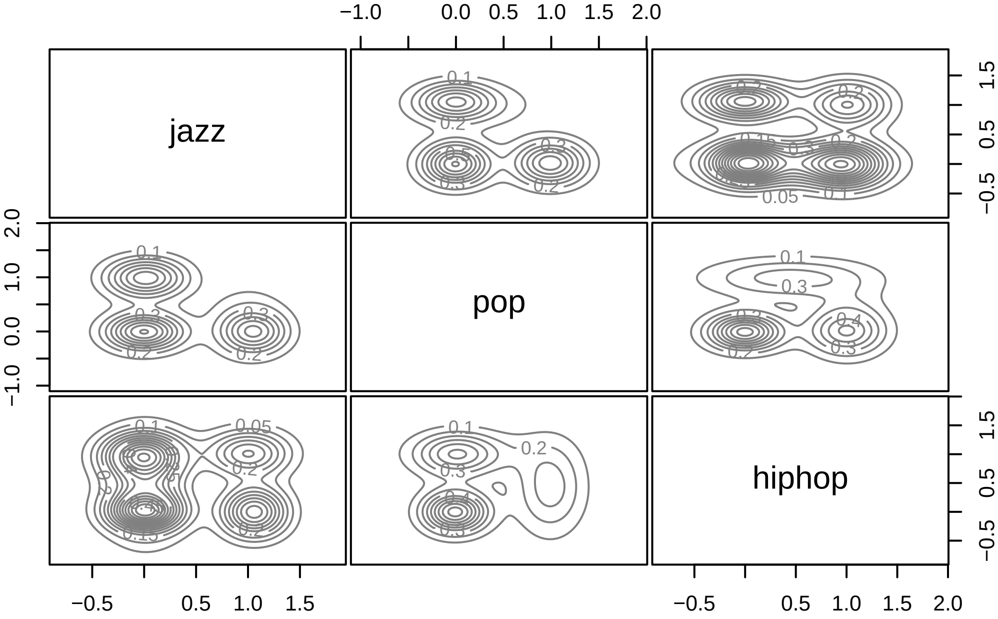


As BIC plots can be read a bit like scree plots I choose an index of 5, as after that only marginal gains in BIC are made at the expense of an increase in complexity.


BIC
−4500
−4700
5
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
VEE
EVE
VVE
EEV
VEV
EVV
VVV


jazz
pop
hiphop
−1.0
0.0
0.5
1.0
1.5
2.0
−0.5
0.5
1.5
−1.0
0.0
1.0
2.0
−0.5
0.5
1.0
1.5
−0.5
0.5
1.5
−0.5
0.5
1.0
1.5
2.0

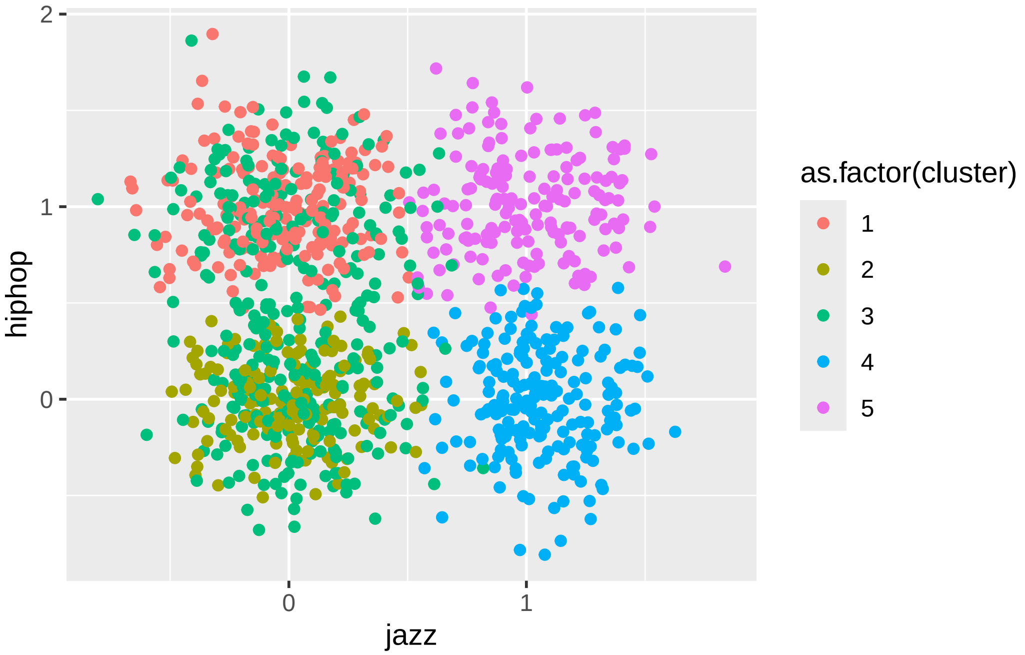




1. Hiphop lovers and neutral toward other things
2. True neutral
3. Pop lovers, hiphop appreciators, neutral toward Jazz
4. Jazz lovers, slight dislike toward pop and neutral toward Jazz

5. Jazz and Hiphop lovers, slightly positive toward pop. Some of these clusters also existed in the kmeans version though the True Neutral category is new and perhaps replaced the dislike everything category/overlap category.

Once again, there is a positive association between being a lover of both Hiphop and Jazz and influence. Unlike the previous results, the neutral group is also negatively associated with influence, though this does line up if we consider that this might be due to the people who like none of the genres now being part of this cluster. Again, it's difficult to generalize from these results.



These results look identical to the previous plot due to its staying in two dimensions. Looking at the results of "plot(finalmcclust)" however, the clusters can be more clearly observed, which is neat.

Task 9

Something which Mclust() also provides is a score for each observation how uncertain we are about its assignment. As mentioned during the lecture, "border-observations" can sometimes be substantively meaningful to study. You shall do so here. Extract the vector $uncertainity from the Gaussian mixture model fit, and store it in the original data. Then yet again fit a linear regression (together with the taste variables), but this time additionally with the uncertainty variable.

> Our results from the previous regression do not change much but what can be observed is that uncertainty is signficantly negatively associated with influence. Looking at R-squared we can see that now a lot more of the variance is explained, as we went from ~0.33 to ~0.39 with the model that added uncertainty.

Part 2: Regional variation

Task 1

Begin by importing the file "neighborhood.csv". Report the number of rows and columns of the data set, and make a brief note on the types of columns contained in it.

```
[1] 200  25
```

> The dataset is split up into variables describing the neighbourhood each individual lives in as well as variables with concern the taste of the individual in music, film but also more general demographic variables such as education and income. Aside from a neighbourhood comparison, there are also some variables at city level. The variables are also variably scaled.

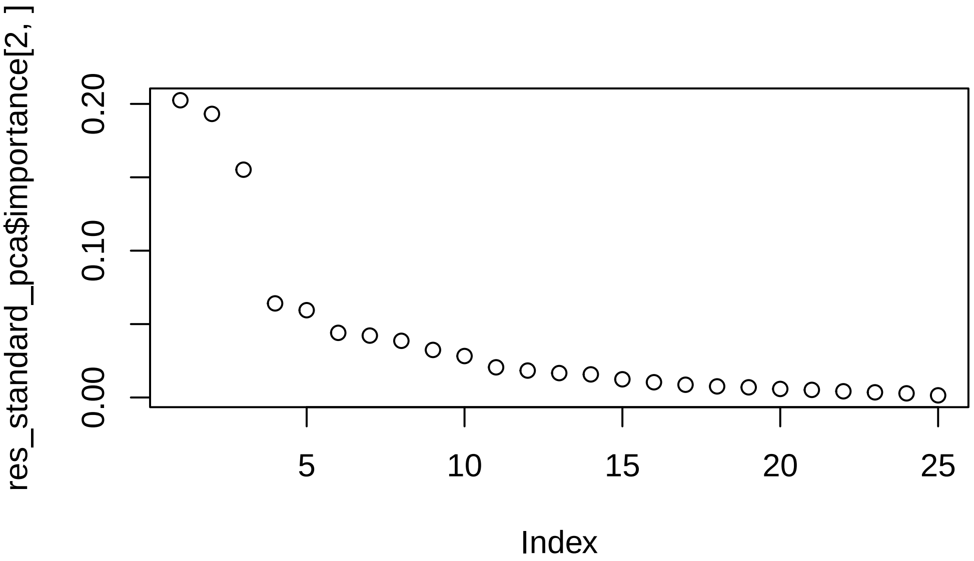
Task 2

Based on the types of variables we find, we have some suspicion that there may exist considerable correlation between different variables in this data set. To explore whether we can capture key aspects of our data using fewer dimensions, we will use PCA and its extensions. Begin by estimating a principal components model without doing any standardization (hint: to estimate a PCA, use prcomp()). Why is this problematic (hint: examine the principal loadings)

> Looking at the loadings of PC1 in comparison to PC2 & 3 we can see that all variables load quite heavily on PC1 but barely on the others. This is due to forgoeing standardization before running the PCA which leads to variables with larger scales (and therefore larger variance), such as nhbood_nr_lights to load to a much greater extent than for example average neighbourhood income.

Task 3

Now, standardize your data, and then fit a PCA on this standardized data set. Plot the proportion variance explained. Interpret and decide on an appropriate number of principal components.



It looks like past 4 PCs only little increases in explained variance are observed, so I choose 4.

Task 4

Interpret the retrieved principal components based on their loadings. Do they provide easy and substantively expected interpretations?

1. Demographic variables (personal and neighbourhood)
2. Personal music taste
3. City measures
4. Neighbourhood physical measures These are by no means clean interpretations but these are the trends I can visually extract from the data. That demographics seem to be most important makes sense though I do have to say I'm a bit surprised about the music taste as PC2.

Task 5

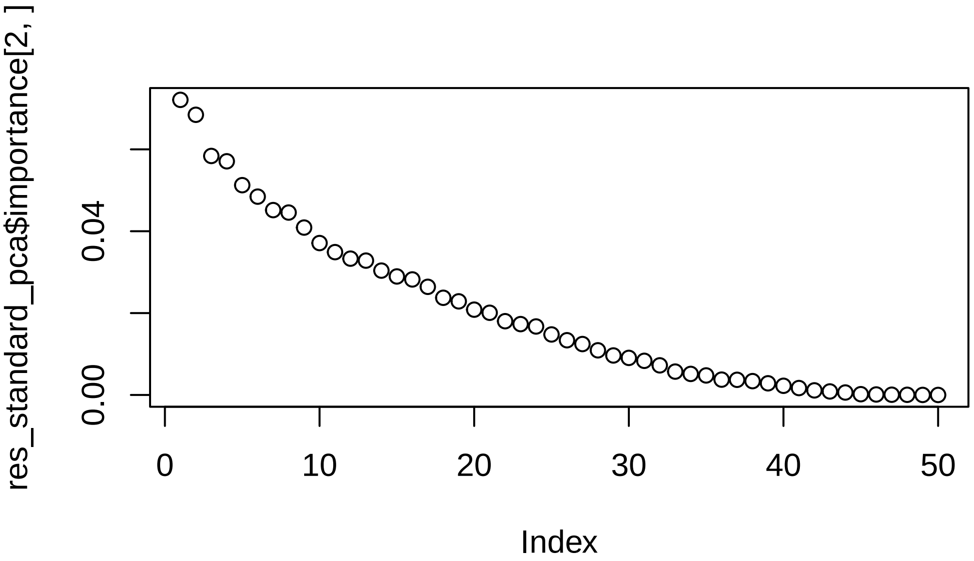
Because of the conclusions in #4, we will now consider the sparse PCA. Use the same number of principal components that you did for the standard PCA in #2. In comparison to the standard PCA, we have an additional parameter in the sparse PCA. Us the IS index to determine an appropriate . Inspect the principal loadings for the resulting configuration.

Interpret each dimension. Which do you think was easier to interpret; the sparse PCA or the standard PCA? Are there any downsides to sparse PCA?

> Looking at the IS data frame we can see that IS is highest at lambda = 80. PC1 lines up with the first dimension previously identified: neighbourhood variables (non-physical). PC2 also covers music taste. PC3 also concerns the city variables. I think sparse PCA takes away a lot of the visual clutter you get when trying to interpret the raw output of PCA. I think that this is also a potential drawback, as it might oversimplify the results, wheras normal PCA might just present them in a biplot, making them more interpretable.

Task 6

As a last exercise for today, you shall simulate your own data. Generate a dataset of 50 observations and 50 independent variables using the function provided below:



Once you have generated the data, process your data as you did above for the neighborhood data set (standardize, making into a matrix). Then, estimate a standard PCA. What do you find: could the PCA help us effectively reduce the dimensionality of our data or not? Why?

> It doesn't really help. Especially when looking at the scree plot it is hard to choose a number of dimensions according to the usual criterion. This is because PCA is built on the idea of dimensionality reduction of large data sets, not data sets

where p == n. The lack of data also makes it difficult to distinguish noise from true variance.