Mathematics of Data Science: A review of the Big Five

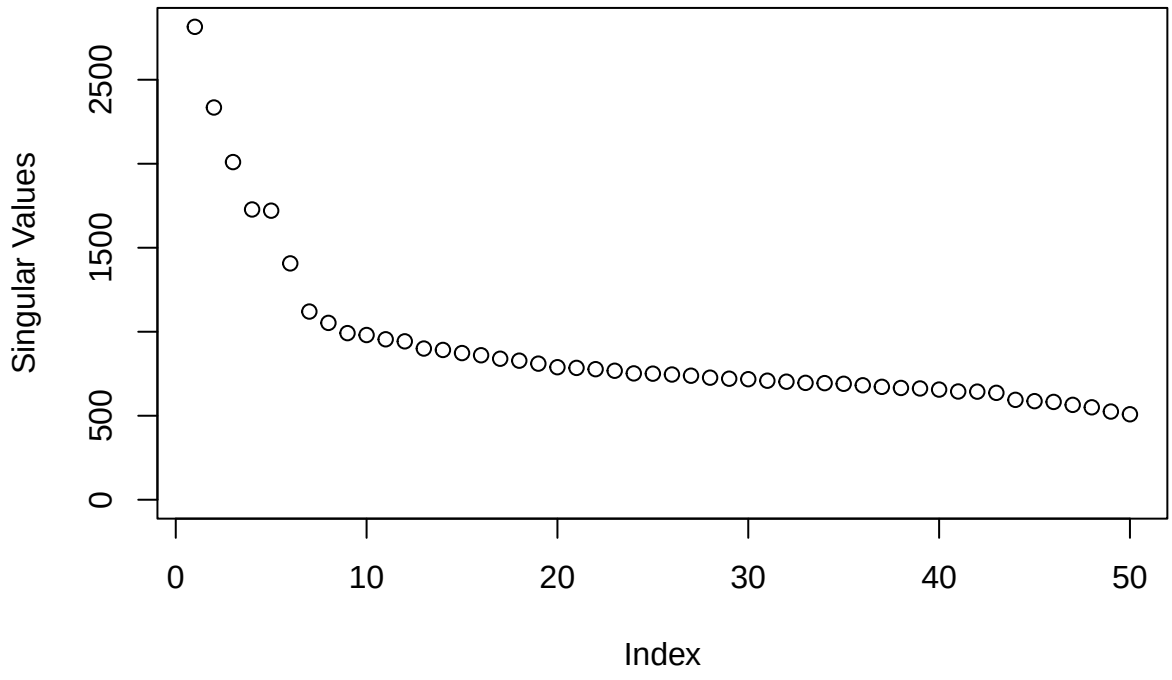
Introduction

The Big Five is a model in psychology that attempts to describe a person's personality by breaking personality down into 5 dimensions. The basic assumption is the lexical hypothesis, which states that the adjectives of a language fundamentally describe the personality.

For personality assessment, a questionnaire with 50 questions (10 questions per factor) was developed, which was filled out over 500000 times (see last chapter for a description).
We want to examine whether this data is consistent with the theory of the "Big Five".

Analysis

To get a low-rank approximation of the personalities, we choose the approach of a truncated SVD. For this, we first form the SVD and inspect the singular values:



Contrary to our hopes, the singular values do not converge (very quickly) towards 0. Therefore, we conclude that we can approximate the data only moderately well.

To find the appropriate number of dimensions to use when approximating the data, there is no obvious

method. Different techniques (see Wikipedia) yield different recommendations:

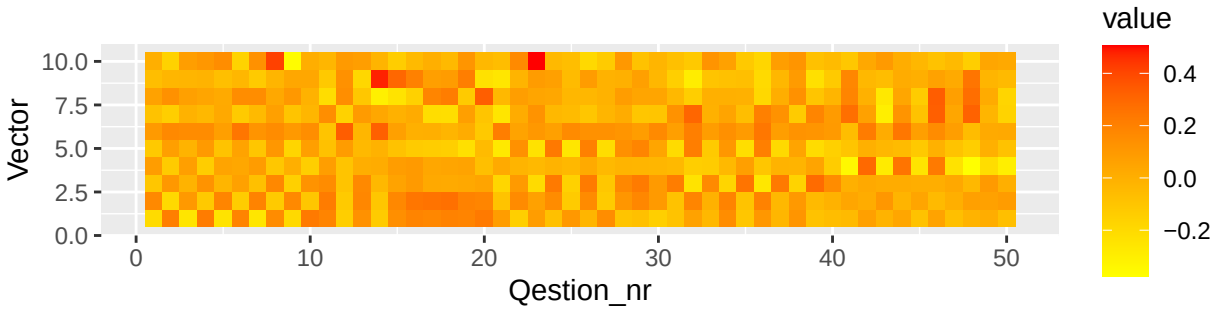
```
## # Method Agreement Procedure:
##
## The choice of 8 dimensions is supported by 3 (21.43%) methods out of 14 (Optimal coordinates, Parall
```

```
##    n_Factors              Method              Family
## 1          1 Acceleration factor               Scree
## 2          3                 CNG                 CNG
## 3          4                beta Multiple_regression
## 4          6                   t Multiple_regression
## 5          6                   p Multiple_regression
## 6          8 Optimal coordinates               Scree
## 7          8   Parallel analysis               Scree
## 8          8    Kaiser criterion               Scree
## 9         14          Scree (R2)            Scree_SE
## 10        37          Scree (SE)            Scree_SE
## 11        47             Bentler             Bentler
## 12        49            Bartlett             Barlett
## 13        49            Anderson             Barlett
## 14        49              Lawley             Barlett
```

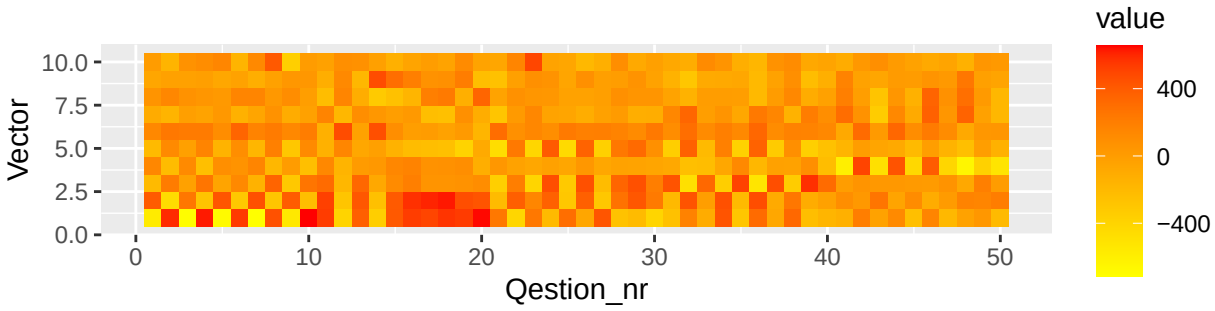
Surprisingly, in this dataset (created with the assumption of 5 factors) it is recommended to use a 6 dimensional approximation. Indeed, there is also a more recent 6-factor model (c.f. HEXACO).

Analyzing right singular vectors

Now we want to inspect whether our low dimensional approximation also recovers the Big five. Since in our data set each observation corresponds to a row of $X = U\Sigma V^T \in \mathbb{R}^{n\times p}$ (not column), the principal factors are given by the columns of V. Thus, we plot the columns of V which are associated with the 10 largest singular values.



To adjust for the fact that the factors have a different influence, we multiply each vector by its singular value:



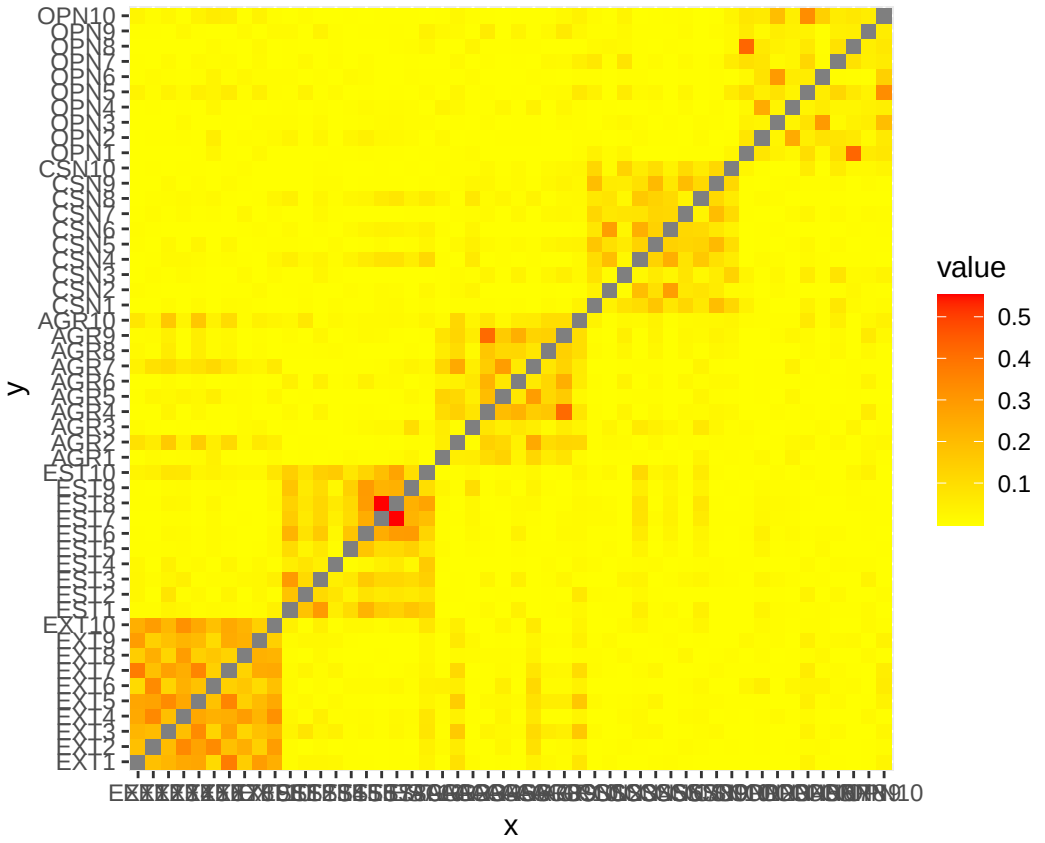
We recall that the first 10 questionnaire questions were designed for the first category, the next 10 for the second, and so on. Moreover, we notice that the questions are not asked in a uniform way. For example, EXT1: "I am the life of the party." and EXT2: "I don't talk a lot." are similar questions, but the answers are (expected to be) exactly opposite. Hence assuming the BigFive model, we expect to recover the categories in the vectors.
For the first vector we can observe a salient signal in the first category, for the third vector in the fourth category, and for the fourth vector in the fifth category. For categories two and three, on the other hand, we cannot clearly identify a vector.

It should be said that an intuitive view is not reason enough to reject the whole theory. However, if our criticism is justified, we could reject the question list.

Remark: It is possible to successfully recover all the categories by (orthogonally) rotating the factors. However, in our case it feels like cheating, since we would not expect the answers for category i influence the j-th Factor (for $i \neq j$).

Clustering of questions

We expect answers from questions in a category to be highly correlated with each other. Consequently, we plot the correlation matrix:



Again, we can more or less recognize the categories. However, we see a surprising amount of correlation between the categories. Thus, if we estimate each factor coefficient using only questions from the corresponding category (i.e. no rotation), this contradicts a basic assumption of the BigFive (orthogonality of factors).

Survey Description

```
This data was collected (2016-2018) through an interactive on-line personality test.
The personality test was constructed with the "Big-Five Factor Markers" from the IPIP.
https://ipip.ori.org/newBigFive5broadKey.htm
Participants were informed that their responses would be recorded and used for research
at the beginning of the test, and asked to confirm their consent at the end of the test.

The following items were presented on one page and each was rated on a five point scale
using radio buttons. The order on page was was EXT1, AGR1, CSN1, EST1, OPN1, EXT2, etc.
The scale was labeled 1=Disagree, 3=Neutral, 5=Agree

THE QUESTION:

EXT1    I am the life of the party.
EXT2    I don't talk a lot.
EXT3    I feel comfortable around people.
EXT4    I keep in the background.
EXT5    I start conversations.
EXT6    I have little to say.
EXT7    I talk to a lot of different people at parties.
```

EXT8	I don't like to draw attention to myself.
EXT9	I don't mind being the center of attention.
EXT10	I am quiet around strangers.
EST1	I get stressed out easily.
EST2	I am relaxed most of the time.
EST3	I worry about things.
EST4	I seldom feel blue.
EST5	I am easily disturbed.
EST6	I get upset easily.
EST7	I change my mood a lot.
EST8	I have frequent mood swings.
EST9	I get irritated easily.
EST10	I often feel blue.
AGR1	I feel little concern for others.
AGR2	I am interested in people.
AGR3	I insult people.
AGR4	I sympathize with others' feelings.
AGR5	I am not interested in other people's problems.
AGR6	I have a soft heart.
AGR7	I am not really interested in others.
AGR8	I take time out for others.
AGR9	I feel others' emotions.
AGR10	I make people feel at ease.
CSN1	I am always prepared.
CSN2	I leave my belongings around.
CSN3	I pay attention to details.
CSN4	I make a mess of things.
CSN5	I get chores done right away.
CSN6	I often forget to put things back in their proper place.
CSN7	I like order.
CSN8	I shirk my duties.
CSN9	I follow a schedule.
CSN10	I am exacting in my work.
OPN1	I have a rich vocabulary.
OPN2	I have difficulty understanding abstract ideas.
OPN3	I have a vivid imagination.
OPN4	I am not interested in abstract ideas.
OPN5	I have excellent ideas.
OPN6	I do not have a good imagination.
OPN7	I am quick to understand things.
OPN8	I use difficult words.
OPN9	I spend time reflecting on things.
OPN10	I am full of ideas.