Analysis of categorical sequences

Adriana Clavijo Daza

2023-02-15

Optimal Matching

Optimal Matching Analysis (OMA) is a technique used in social sciences for the comparison of sequences with applications on different areas, in particular, life course and career path analysis. It is based on the Needleman-Wunsch algorithm, which is used to align protein sequences and is an application of dynamic programming, an iterative method that simplifies an optimization problem by breaking it into a recursion of smaller problems that are simpler to solve.

The algorithm

Given a set of n states, say, $S = \{s_1, \dots, s_n\}$ a sequence of size $t > 0$ can be denoted as $X = (x_1, \dots, x_t)$, where $x_i \in S$ for $i = 1, \dots, t$. Also, the set of all possible sequences with states belonging to S is denoted by $\mathbf{S}$.

Let $X, Y \in \mathbf{S}$ be two sequences of size t_1 and t_2, respectively, that we want to align. In order to align the sequences, we define a matrix F of size $(t_1 + 1) \times (t_2 + 1)$ structured as follows:

$$F(1,1) = 0, \tag{1}$$

$$F(i,1) = F(i-1,0) - d \text{ for } i = 2, \dots, t_1, \tag{2}$$

$$F(1,j) = F(1,j-1) - d \text{ for } j = 2, \dots, t_2, \tag{3}$$

$$F(i,j) = max\{F(i-1,j) - d, F(i,j-1) - d, F(i-1,j-1) + k(x_{i-1}, yj-1)\}, \tag{4}$$

Where d is the cost of inserting a gap, and $k(x_{i-1}, yj-1)$ is the cost associated to change from the state x_{i-1} to y_{j-1}, which is defined in a K of size $n \times n$, known as the cost matrix.

Equations 1-3 correspond to initialization and equation 4 shows the row-wise recursion to fill the matrix and its terms represent inserting a gap in X, inserting a gap in Y and match or mismatch, respectively.

Cost matrix

There are several ways of obtaining the cost matrix.

Transition rates The substitution cost between states s_i and s_j, $1 \le i, j \le n$, is calculated as:

$$K(s_i, s_j) = c - P(s_i|s_j) - P(s_j|s_i), \tag{5}$$

where $P(s_i|s_j)$ is the probability of transition from state s_i in time t to s_j in time $t+1$ and c is a constant, set to a value such that $0 \le K(s_i, s_j) \le 2$).

Example

Data from the 40+ Healthy Aging Study

About the data

As part of the Women 40+ Healthy Aging Study, a large study that was conducted by the Department of Clinical Psychology and Psychotherapy of the University of Zurich, a psychometric instrument was developed in order to obtain information about the history of romantic relationships of women. The study was conducted between June 2017 and February 2018 with women between 40 and 75 years who (self-)reported good, very good or excellent health condition and the absence of acute or chronic somatic disease or mental disorder. The participants who reported psychotherapy or psychopharmacological treatment in the previous 6 months were excluded as well as habitual drinkers. Other exclusion criteria were pregnancy in the last 6 months, premature menopause, surgical menopause, intake of hormonal treatment (including contraceptives), shift-work and recent long-distance flight. The participants were recruited from the general population using online advertisement and flyers.

The questionnaire asked the participants to provide information about relationship phases starting from the age of 15 years until the current age at the time of the data collection. The phases were defined by the start and end age and for each phase and information about civil status, relationship status, living situation, children and quality of the relationship was collected. Before including the data corresponding to their own history, the participants were prompted to answer some of the questions based on an example. Some of the participants were excluded when the example entries were not correctly filled. In total 250 individuals were considered in the analysis.

In order to create a sequence for each participant the information about civil status, relationship status, living situation and the matenity is taken into account. A yearly sequence is created and the states considered are the following:

- 1 = Single + no children
- 2 = Single + children
- 3 = Changing relationships + no children
- 4 = Changing rel. + children
- 5 = Relationship + living apart + no children
- 6 = Relationship + living together + no children
- 7 = Relationship + living apart + children
- 8 = Relationship + living together + children
- 9 = Married + no children
- 10 = Married + children

Additionally, personality scores for the women included in the study are available. Personality refers to the enduring characteristics and behavior that comprise a person's unique adjustment to life, including major traits, interests, drives, values, self-concept, abilities, and emotional patterns. These scores are obtained via psychometric instruments and evaluate the main personality traits:

- Agreeablenes
- Conscientiousness
- Extraversion
- Neuroticism
- Openness

Optimal matching analysis is performed with the aim to obtain clusters of sequences that are similar and characterize the most common relationship history profiles.

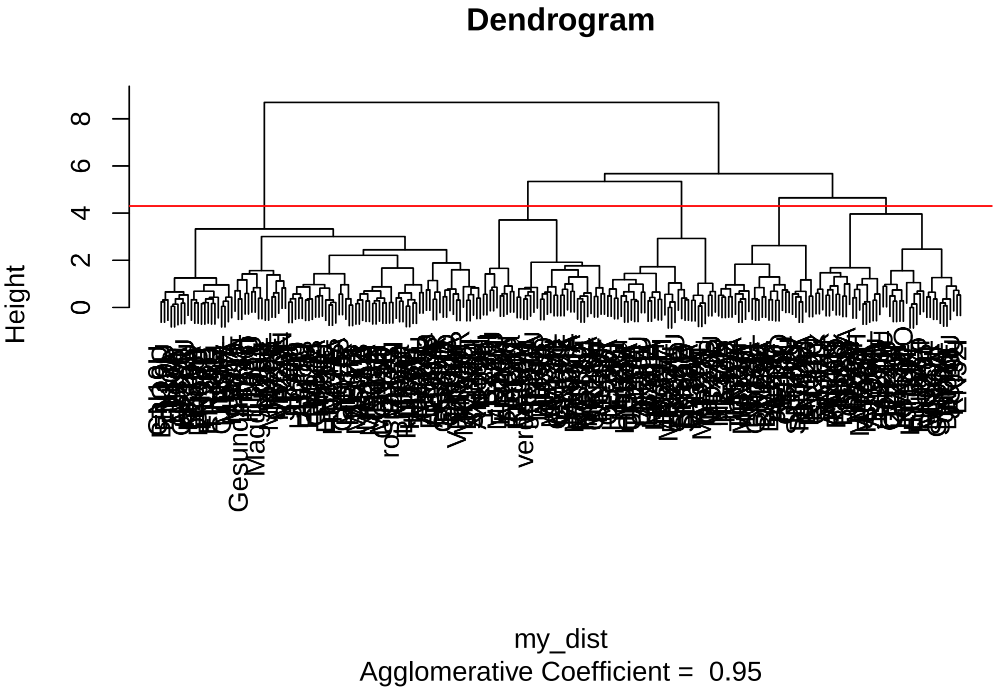
Application

Using the `R` package `TraMineR` the cost matrix is calculated with transition rates between states.

Status	1	2	3	4	5	6	7	8	9	10
1	0.0000	2.0000	1.9845	2.0000	1.9246	1.9462	2.0000	1.9950	1.9846	1.9836
2	2.0000	0.0000	2.0000	2.0000	2.0000	2.0000	1.9552	1.9224	2.0000	1.9985
3	1.9845	2.0000	0.0000	2.0000	1.9409	1.9198	2.0000	1.9982	1.9736	1.9765
4	2.0000	2.0000	2.0000	0.0000	1.9878	2.0000	1.9550	1.9512	2.0000	1.9965
5	1.9246	2.0000	1.9409	1.9878	0.0000	1.9522	1.9772	1.9903	1.9808	1.9745
6	1.9462	2.0000	1.9198	2.0000	1.9522	0.0000	2.0000	1.9971	1.9766	1.9583
7	2.0000	1.9552	2.0000	1.9550	1.9772	2.0000	0.0000	1.9708	2.0000	1.9965
8	1.9950	1.9224	1.9982	1.9512	1.9903	1.9971	1.9708	0.0000	2.0000	1.9836
9	1.9846	2.0000	1.9736	2.0000	1.9808	1.9766	2.0000	2.0000	0.0000	1.9862
10	1.9836	1.9985	1.9765	1.9965	1.9745	1.9583	1.9965	1.9836	1.9862	0.0000

From this cost matrix it is possible to calculate pairwise distances between sequences using the algorithm previously described. A correction of the distances is done to account for the differences in size of the sequences. This is done dividing the obtained distance by the length of the longest sequence.

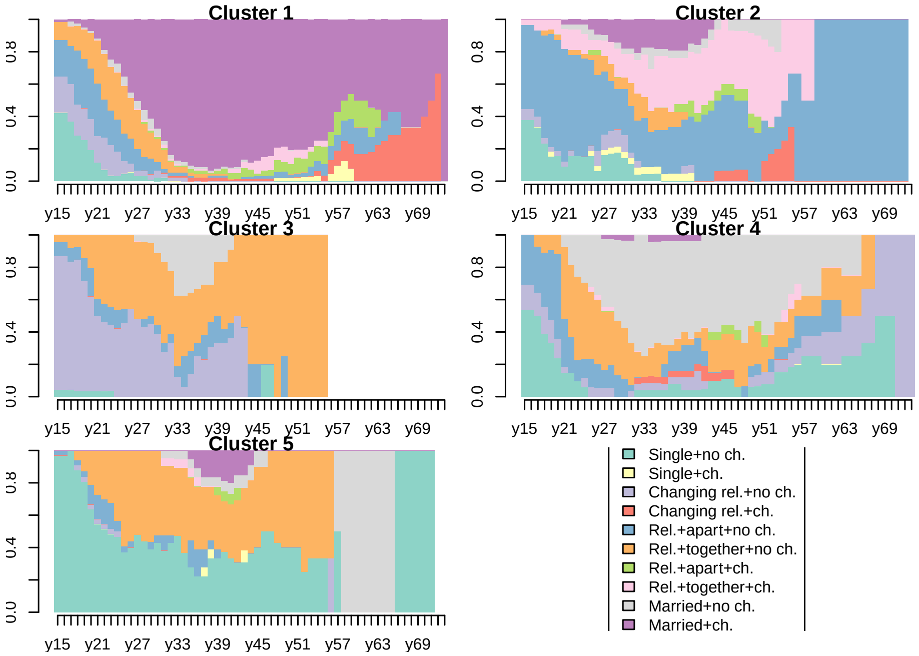
We then use this sequences to apply a hierarchical agglomerative clustering method called AGNES. The following figure shows the dendrogram. In this case, we decided to cut at 5 clusters in order to preserve enough individuals in each cluster.



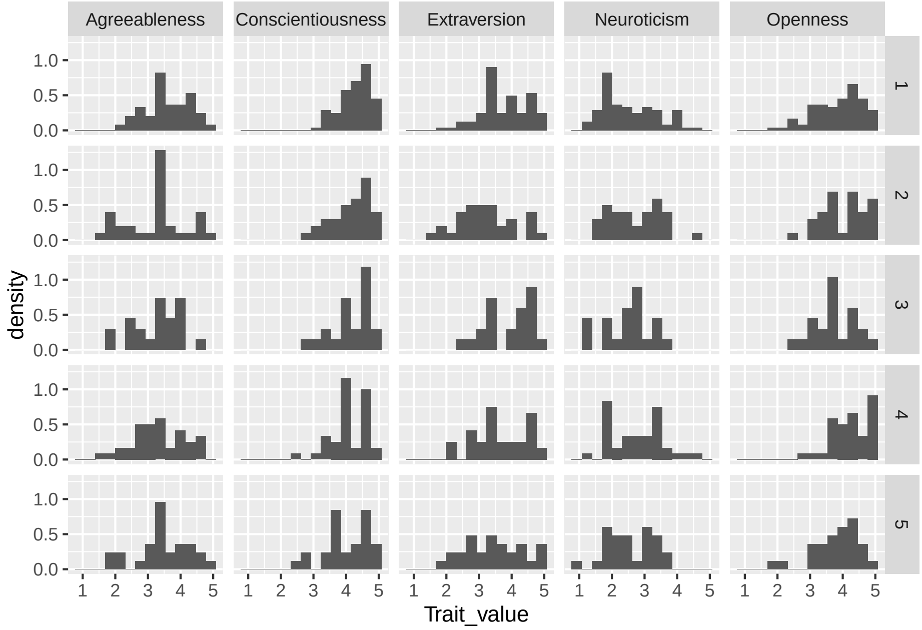
Cluster	n
1	96
2	37
3	29
4	44
5	33

We can visualize the clusters of sequences to and try to identify common features to describe them. It is important to consider that this description is subjective but can be useful to characterize the groups.

- Cluster 1: Married with children then divorced/widowed
- Cluster 2: Sequences with more changes (unstable)
- Cluster 3: Younger, mostly not married without children
- Cluster 4: Older, without children
- Cluster 5: Married late then divorced/widowed, without children



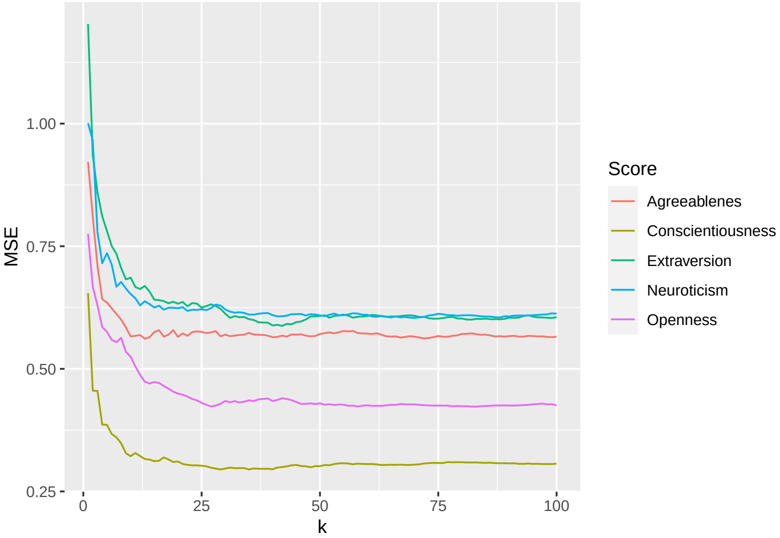
Now, we are interested in predicting the personality scores based on the relationships history of the women. The following figure shows the distribution of the score for each trait by cluster.



No difference is obvious at a first glance. However, we can also use the distance matrix to obtain predictions of the personality traits using k-nearest neighbors and compare them with a measure such as the mean squared error (MSE).

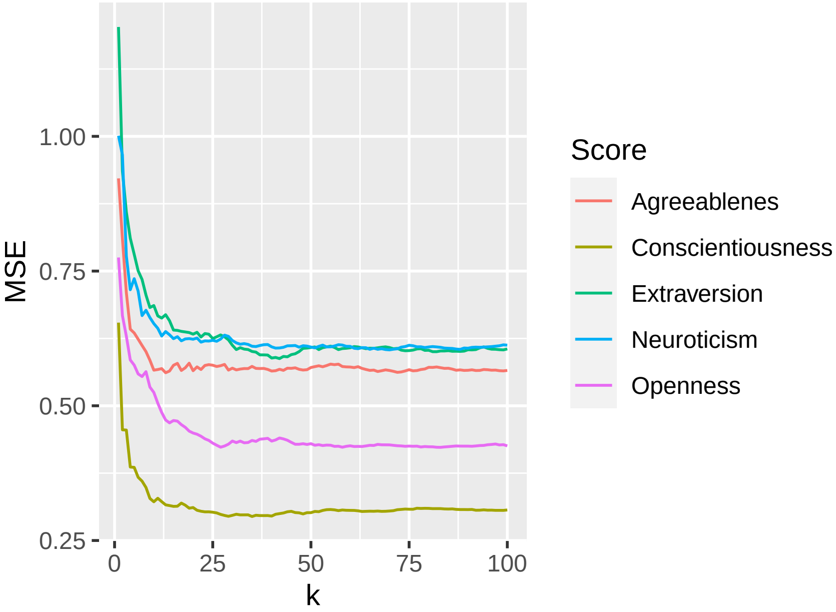
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad (6)$$

The following figure shows that the MSE for different values of k decreases at the beginning and stays mostly flat after $k = 25$, meaning that the prediction is not very good and probably not much better than just taking the average value of the scores.

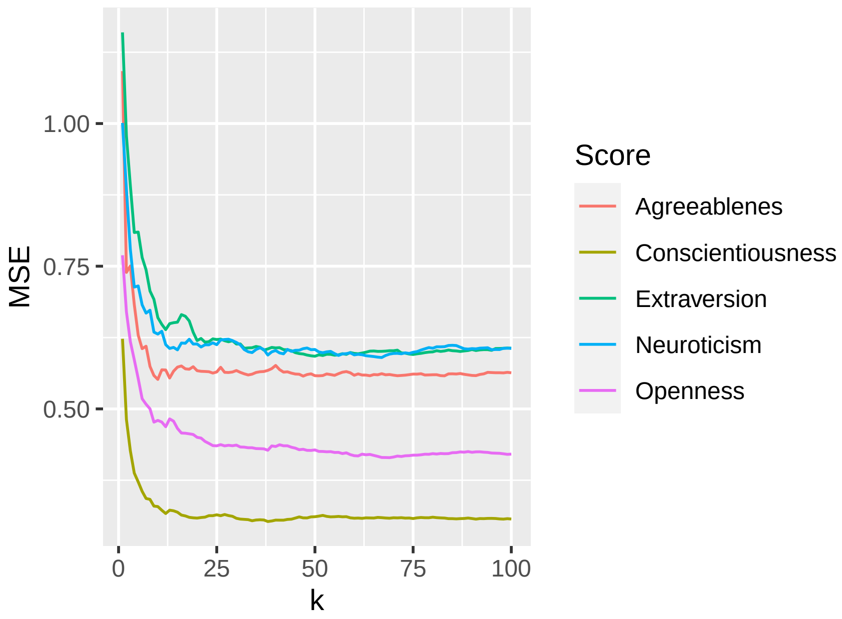


Other experiments

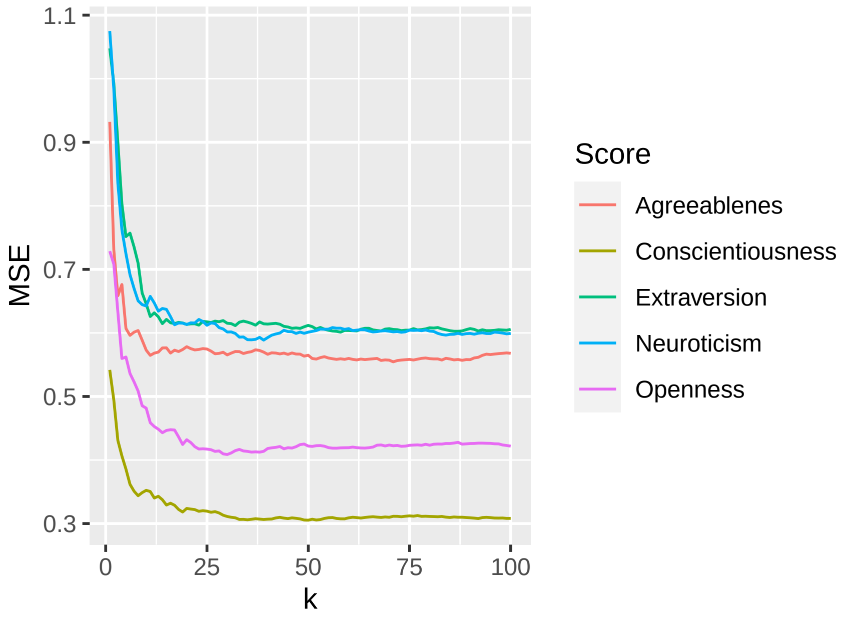
Setting the constant c as the maximum of $2 - P(s_i|s_j) - P(s_j|s_i)$.



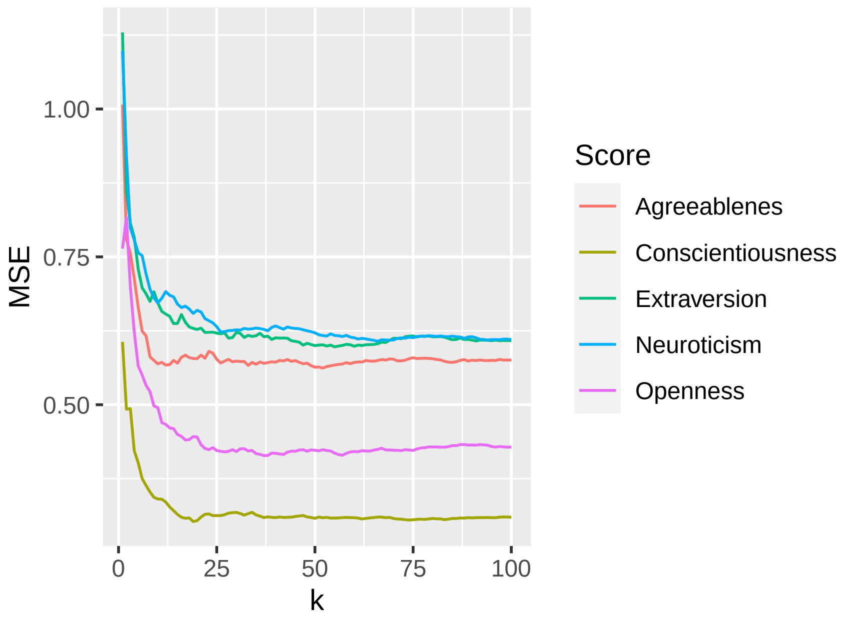
Using `norm = "gmean"` normalization in `TraMineR::seqdist`.



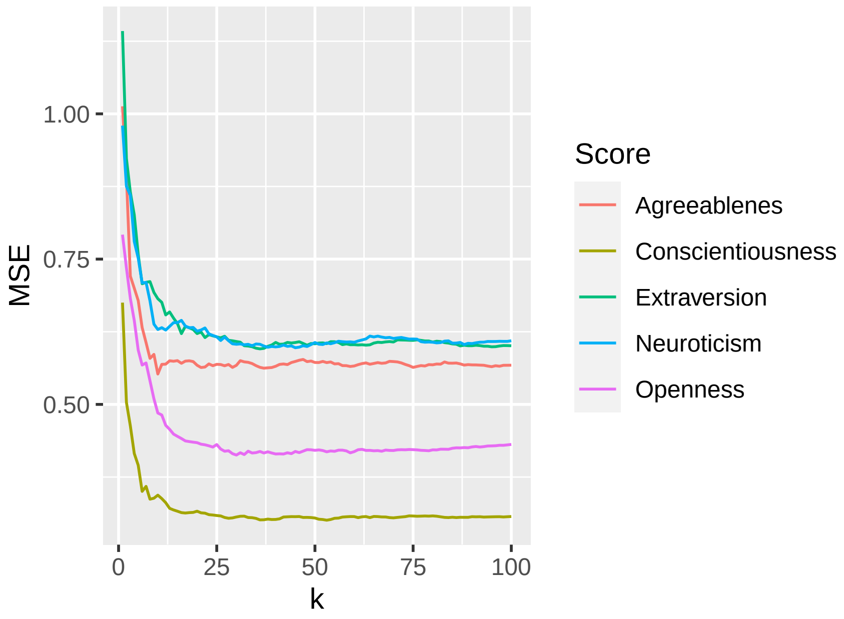
Using `method = "FUTURE"` normalization in `TraMineR::seqcost`.



Using `method = "INDELS"` normalization in `TraMineR::seqcost`.



Using `method = "INDELSLOG"` normalization in `TraMineR::seqcost`.


MSE
k
Score
Agreeablenes
Conscientiousness
Extraversion
Neuroticism
Openness