DTU Compute
Institute for Mathematics and Computer Science

Archetypal Analysis for Ordinal Data

Archetypal Analysis of European Social Survey: ESS8 dataset

Bachelor Thesis

Archetypal Analysis for Ordinal Data
Archetypal Analysis of European Social Survey: ESS8 dataset

Bachelor Thesis
December, 2021

By
Andreas Råskov Madsen

Cover photo: Vibeke Hempler, 2012
Published by: DTU, Institute for Mathematics and Computer Science, Richard Petersens Plads , Building 324, 2800 Kgs. Lyngby Denmark
www.compute.dtu.dk
ISSN: [0000-0000] (electronic version)
ISBN: [000-00-0000-000-0] (electronic version)

ISSN: [0000-0000] (printed version)
ISBN: [000-00-0000-000-0] (printed version)

Abstract
In this thesis, I will demonstrate how Archetypal Analysis (AA) proposed by Cutler and Breiman [1], can be combined with an ordinal likelihood function described by Wei Chu and Zounin Ghahramani [2] to create and Ordinal Archetypal Analysis (OAA). I will demonstrate how OAA can be used to recreate archetypes, and find a continuous scale in synthetic ordinal data. I demonstrate how an extended version of OAA can be used to learn and adjust for response bias, by learning an individual scale for each respondent a method named Response Bias Ordinal Archetypal Analysis (RB-OAA). As a case study, I work with section H of the European Social survey round 8 data-set[3] where I compare archetypes found by OAA to Schwarts theory of basic human values, and demonstrate how a Likert scale can be converted to a continuous scale and adjusted for response bias.

Acknowledgements
Main advisor
Morten Mørup, Professor, DTU

Advisor
Mikkel N. Schmidt, Associate Professor, DTU

Advisor
Advisor: Fumiko Kano Glückstad, Lector, CBS

Contents

1 Introduction

Archetypal analysis(AA) proposed by Cutler and Breiman in 1994[1], is a method to reduce dimensionality while maintaining interpretability of a dataset. By trying to create a convex hull of data points, and then reconstruct the dataset from the convex hull. A convex hull is often decided as the result you would get by slapping a rubber band around the data such that all data point is encapsulated. AA will try to encapsulate as much as the data as possible however in multidimensional space the number of points needed to make is $\mathcal{O}(\ln(N)^{m-1})$[4] where N is the number of points and M is the number of dimensions.

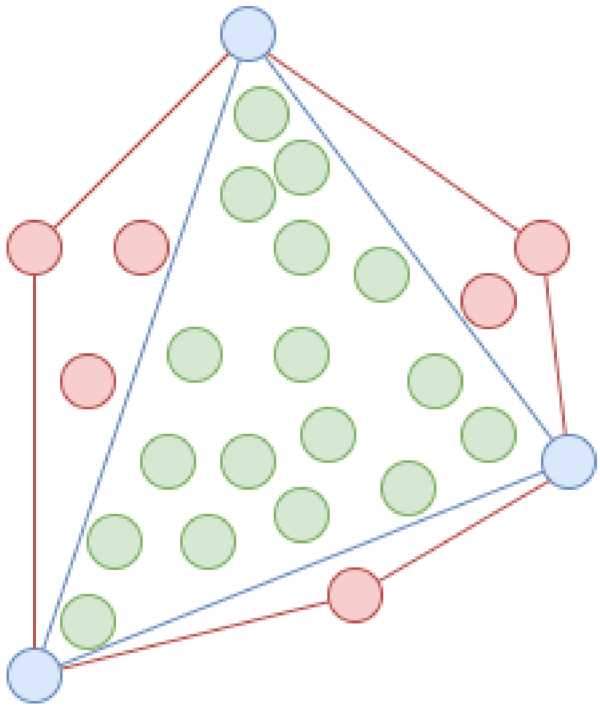


Figure 1: Illustration example of how AA works. The red lines are the real convex hull, the blue lines are the Archetype hull, and the blue points are the archetypes. The blue and green points can be reconstructed as a convex combination of the archetypes while the red ones will be projected down on the blue line. Note: the model doesn't need to select pure points as archetypes but can define an Archetype as a combination of several points

Ordinal data can be thought of as data that is ordered but the length between each ordinal can not be assumed to be equal. For example, we may define a scale where $A > B$ and $B > C$, thus it can be concluded that $A > C$ and so on. However, we can not make a statement such as $A = 3 \cdot C$,$A = B + C$ or $B = \frac{A+C}{2}$, unless we have a reliable way to convert the ordinal scale into a continuous or discrete scale.

A method for doing AA on ordinal data was proposed by Fernández, Epiganio and McMillan (2021) [5]. Using a two-step process where ordinal data was first converted to a continuous scale using an Ordered stereotype model before a traditional AA was applied. We theorise that due to the two-step process the models were not able to complement each other, and thus is not making an optimal solution.

Another method for doing ordinal archetypal analysis (OAA) was proposed to me by my supervisors Professor Morten Mørup and assistant Professor Mikkel N. Schmidt. Using an ordinal likelihood process described by Wei Chu and Zounin Ghahramani [2] we can make a hybrid model that performs AA while adjusting the scale by using the reconstruction of the AA. It was also proposed that OAA could learn an individual scale for each respondent and thus adjust for respons bias in the data set, a method we chose to call response bias ordinal archetypal analysis (RB-OAA) In this thesis I will demonstrate an implementation of both OAA and RB-OAA evaluate them on synthetic data as well as the ESS8 dataset, an discuss the advantages and disadvantages of each model. The models is implemented in Python where I use the PyTorch library to make the model run on GPU and use the Pythorch autograd function to calculate the gradient needed for training the model.

In cooperation with my external supervisor Lektor Fumiko K. Glückstad from Copenhagen Business school this project wish to investigate some of the commercial value of OAA/RB-OAA by finding archetypes in the ESS 8 data set [3], to present a suitable OAA model that can be used for analyses of the European populations, by finding archetypes. The project only works with section H in the data-set which is questions based on Schwarts theory on basic human values [6] with the focus on comparing archetypes to the personal focus vs social focus and protection vs Growth axis described by Schwartz, and to compare archetypes to external variables.

We also wished to investigate if archetypes was correlated with the external variables: country, gender, age, political leaning and subjective income, partially motivated by a study that showed that AA could find genders based on skeletons[7]

1.1 Research questions

1. Can Archetypal Analysis as described by Cutler and Breiman be implemented using PyTorch backpropagation, that can be run on GPU?
2. can ordinal data be maped to continues scale inside the model using an ordinal likelihood framework [2] with parameters found by the model with PyTorch autograd?
3. Do the archetypes found correspond with archetypes described by Schwartz theory [6]?
4. Can external variables (social class, country, etc) be linked to certain archetypes that can be used for marketing purposes?
5. Can the ordinal likelihood function be used to account for annotation bias or response style of group, of participants?

1.2 report outline

I will start by presenting the ESS-8 dataset. While introducing the data-set I will also explain Schwartz's theory of basic human values since the theory also explains the structure of the dataset. In the method section, I will explain traditional Archetypal analysis and the Ordinal likelihood function as two separate models before explaining how they can be combined into one Ordinal Archetypal analysis. I then explain the extended model that can adjust for response bias RB-OAA. To explain how the model was tested I demonstrate how a synthetic dataset can be made from the assumptions in the model implementation. And I present performance parameters, such as the correlation between archetypes the normalized mutual information of the reconstruction matrix and the difference between two scales, can be calculated. In the result section, I start by demonstrating how OAA can be used to learn an ordinal scale where the ordinal don't have an even distance between them. I then test OAA and RB-OAA on different levels of noise and response bias before testing them on the ESS8 dataset. The model with 4 archetypes is then selected to demonstrate how archetypes are modelled. I also include a visual representation of the archetypes compared with the mean of external variables in order to demonstrate the lack of correlation. In the discussion, the implementing issues of the model and the correctness of the found solution will be discussed, and qualitative analysis of the RB-OAA's relations to Schwartz's theory will be performed. I also include a short discussion of the commercial value of OAA and RB-OAA. Before presenting the conclusion.

2 Data

2.1 ESS8 Dataset

The ESS8 dataset is collected by the European Social Survey, an organisation that every 2 years collect data from countries in and around Europe for academic purposes. According to their website [3] the main objective of the European Social Survey is:

- to chart stability and change in social structure, conditions and attitudes in Europe and to interpret how Europe's social, political and moral fabric is changing
- to achieve and spread higher standards of rigour in cross-national research in the social sciences, including, for example, questionnaire design and pre-testing, sampling, data collection, reduction of bias and the reliability of questions
- to introduce soundly-based indicators of national progress, based on citizens' perceptions and judgements of key aspects of their societies
- to undertake and facilitate the training of European social researchers in comparative quantitative measurement and analysis
- to improve the visibility and outreach of data on social change among academics, policymakers and the wider public

The survey was collected by an interviewer asking the questions and writing down the answers. The ESS8 round was made in 2016 and involved following countries [8]:

- Austria
- Belgium
- Czechia
- Estonia
- Finland
- Germany
- Hungary
- Iceland
- Ireland
- Italy
- Lithuania
- Netherlands
- Norway
- Poland
- Portugal
- Russia Federation
- Slovenia
- Spain
- Sweden
- Switzerland
- United Kingdom

The ESS8 investigate the Schwartz values using a method called portrait values Questionnaire (PVQ). In PVQ participants will be given a short deception of a person and is then ask "How much like you is this person?", the response was noted on the Likert scale presented in table 1 The questions will be phased so the gender of the person matches the respondent. The Questions given to a male respondent is given in table 2

Very much like me	Like me	Some-what like me	A Little like me	Not like me	Not like me at all
1	2	3	4	5	6

Table 1: Options given to respondent. The questioner also had the responses (refusal) and (Don't know) those data-point has been removed.

2.2 Schwartz Theory of Basic Values

While the survey collects data on a wide range of different topics the only data-set relevant for this thesis is section H that's about the social values impotent to the participant according to Schwartz Theory of Basic Values. Schwartz identifies 10 basic human values, to explain peoples beliefs and goals[6]. The values and their definition according to Schwartz[6] are the following:

- Self-direction: Independent thought and action–choosing, creating, exploring.
- Stimulation: Excitement, Novelty and challenge in life.
- Hedonism: pleasure or sensuous gratification for oneself.
- Achievement: Personal success though demonstrating competence according to social standards
- power: Social status and prestige, control or dominance over people and resources
- Security: Safety, harmony, and stability of society, of relationships and of self.
- Conformity: Restrains of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms.
- Tradition: Respect, commitment, and acceptance of the customs and ideas that one's culture or religion provides.
- Benevolence: Preserving and enhancing the welfare of those with whom one is in frequent personal contact.
- Universalism: Understanding,appreciation, tolerance, and protection for the welfare of all people and for nature.

Schwartz also identifies two axes to put the values on, one being Personal focus versus social focus and the other protecting and Anxiety-avoidance versus growing and Anxiety-free. See figure 2. Thus the values will often conflict along those axes, where selfish actions to get achievement or power may conflict with benevolence or universalism, or Hedonism, Stimulation or self-direction will conflict with conformity or security.

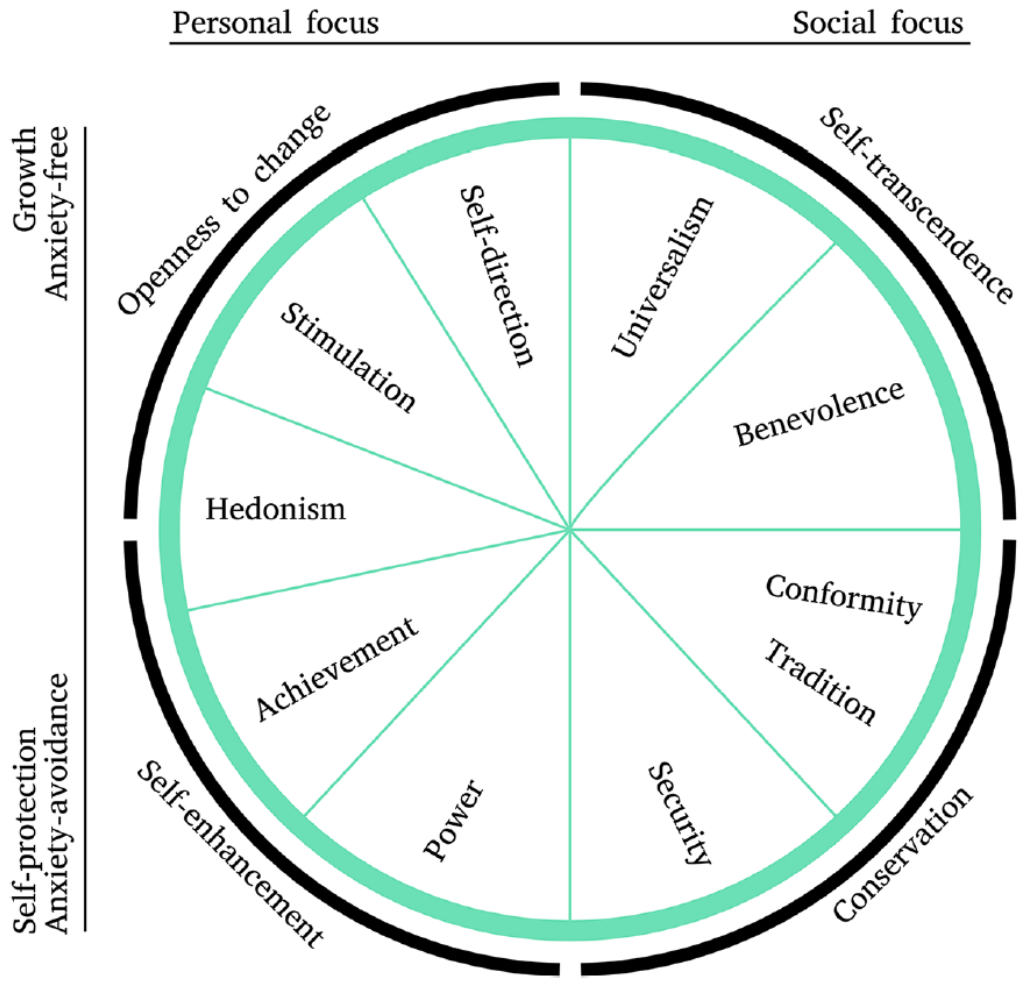


Figure 2: Schwartz theory of values visualised. Image credit [9]

2.2.1 Pre-processing

The data used in this thesis comes from a project conducted by Mikkel N. Schmidt, Daniel Sedding, Eldad Davidov, Morten Mørup, Kristoffer Jon Albers,Jan Michael Bauer, Fumiko Kano Glückstad. On making LPA on the ESS8 dataset [9]. Proposing consists of excluding 3094 responses out of the original 44387 responses due to incomplete responses. Thus the final data set has the dimensions N=41293, M=21.

When data enter the model it will be transformed to a scale going 0,1,2,3,4,5 because the model sees the ordinal values as indexes. This has the nice property that the model doesn't care about the original scale as long as it is ordinal.

Question	Encoding	Category	Qestion test
A	SD1	Self-direction	Thinking up new ideas and being creative is important to him. He likes to do things in his own original way.
B	PO1	power	It is important to him to be rich. He wants to have a lot of money and expensive things.
C	UN1	universalism	He thinks it is important that every person in the world should be treated equally. He believes everyone should have equal opportunities in life.
D	AC1	achievement	It's important to him to show his abilities. He wants people to admire what he does.
E	SC1	security	It is important to him to live in secure surroundings. He avoids anything that might endanger his safety.
F	ST1	stimulation	He likes surprises and is always looking for new things to do. He thinks it is important to do lots of different things in life.
G	CO1	conformity	He believes that people should do what they're told. He thinks people should follow rules at all times, even when no one is watching.
H	UN2	universalism	It is important to him to listen to people who are different from him. Even when he disagrees with them, he still wants to understand them.
I	TR1	tradition	It is important to him to be humble and modest. He tries not to draw attention to himself.
J	HD1	hedonism	Having a good time is important to him. He likes to "spoil" himself.
K	SD2	self-direction	It is important to him to make his own decisions about what he does. He likes to be free and not depend on others.
L	BE1	benevolence	It's very important to him to help the people around him. He wants to care for their well-being.
M	AC2	achievement	Being very successful is important to him. He hopes people will recognise his achievements.
N	SC2	security	It is important to him that the government ensures his safety against all threats. He wants the state to be strong so it can defend its citizens.
O	ST2	stimulation	He looks for adventures and likes to take risks. He wants to have an exciting life.
P	CO2	conformity	It is important to him always to behave properly. He wants to avoid doing anything people would say is wrong.
Q	PO2	power	It is important to him to get respect from others. He wants people to do what he says.
R	BE2	benevolence	It is important to him to be loyal to his friends. He wants to devote himself to people close to him.
S	UN3	universalism	He strongly believes that people should care for nature. Looking after the environment is important to him.
T	TR2	tradition	Tradition is important to him. He tries to follow the customs handed down by his religion or his family.
U	HD2	hedonism	He seeks every chance he can to have fun. It is important to him to do things that give him pleasure.

Table 2: List of questions that was ask to participants in section H1 female participants was ask H2 that consisted of same questions but female pronouns.

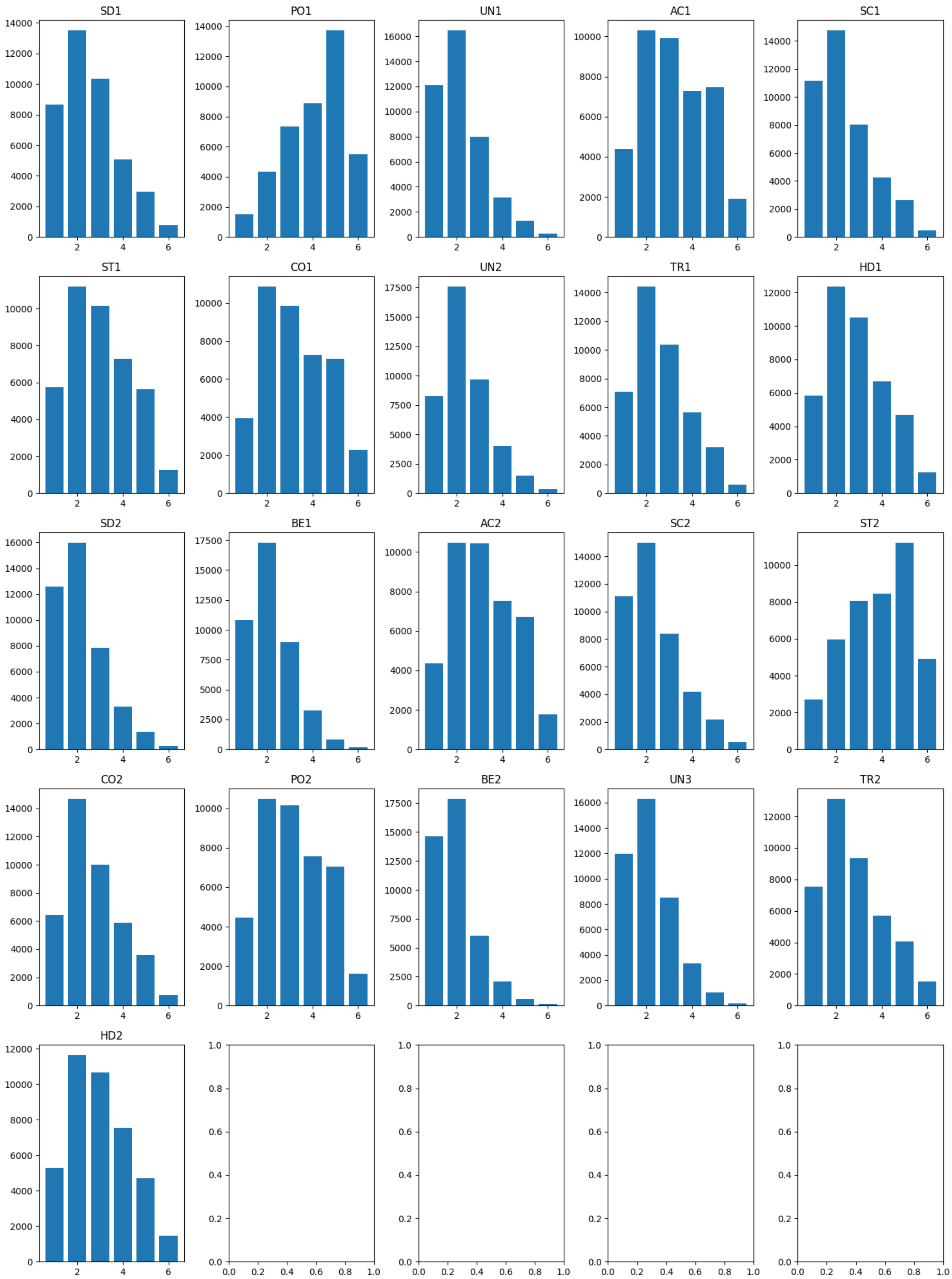


Figure 3: Distribution of answers to each question from the data-set

3 Method

The proposed model is a combination of Archetype Analysis as described by Cutler and Breiman[1] and the Likelihood for Ordinal variables method described by Chu and Ghahramani [2] I

3.1 Archetypal analysis

Let $\tilde{X}_{N\times M}$ be a data-set with N samples on a continuous scale and K be the number of Archetypes to find.

$C_{K\times N}$ is trying to find the point or combination of data-point that can be used to make the archetypes. Thus C must have following properties $\sum_{k=0}^{K} C_{k\times n} = 1$ for all n and all entries must be equal or greater than 0.

The archetypes $A_{m\times k}$ is then given as:

$$A_{K\times M} = C_{K\times N} \cdot \tilde{X}_{N\times M} \tag{1}$$

$S_{N\times M}$ is trying to reconstruct the data points from a combination of the archetypes where all the new data points are within the convex hull. Thus S must have following properties $\sum_{n=0}^{N} S_{k\times n} = 1$ for all k and for all entries must be equal or greater than 0.

$$\hat{X}_{N\times M} = S_{N\times K} \cdot A_{K\times M} \tag{2}$$

In Archetypal analysis the loss can be defined as $Loss = ||X - \hat{X}||$ where $||\cdot||$ the squared Frobenius matrix norm [10]. However, since this project is dealing with ordinal data the loss function will be defined by the Ordinal Likelihood function.

3.2 Likelihood for Ordinal variables

Given an ordinal observation x on an ordinal scale with J ordinals, a continues scale defined $\beta_0, \beta_1 ... \beta_J$ where β_j can be considered the cut between the part of the scale where each ordinal would map to, with $\beta_0 = -\infty$ and $\beta_J = \infty$. By using the function $f(x)$ where $f(x)$ is a latent function that maps from an ordinal scale to a continuous scale by deconstructing and reconstruct the data. Then according to [2] the likelihood of observing x in the interval $[\beta_{j-1}, \beta_j]$ can be described as:

$$Z_j = \frac{\beta_j - f(x)}{\sigma} \tag{3}$$

$$P(\beta_j, f(x), \sigma) = \Phi(Z_{j+1}) - \Phi(Z_j) \tag{4}$$

where $\Phi()$ is the cumulative density function of a standard normal distribution, and σ can be considered the standard deviation of the likelihood function. If j is the ordinal category that X comes from, the loss function for the β scale and $f(x)$ for a single x is then:

$$loss = -\ln(P(\beta_j, f(X), \sigma) \tag{5}$$

3.3 Ordinal archetypal analysis (OAA)

3.3.1 Initialisation

Let $X_{N\times M}$ be the original ordinal data, J is then the number of unique values in the data-set, a sort preprocessing is done such that all values in X maps to the ordinal scale 0,1,2..J-1.

γ used to calculate the new, is random initialised as J long array.

The noise parameter $\tilde{\sigma}$ is random internalised.

$\tilde{C}_{N\times K}$ is initialised by selecting K random points, such that each column in $\tilde{C}$ contains one entry of $log(N)$ and 0 for all other entries. $log(N)$ is used to ensure that C still is a use full combination of points after equation 9
$\tilde{S}_{N,K}$ is random internalised.

3.3.2 train loop

First the scale is defined by finding all β values as softmax function of γ such that $\beta_0 = 0$ and $\beta_J = 1$ while all other values is and ordered scale between 0 and 1.

$$\beta_j = \sum_{i=1}^{j} \frac{e^{\gamma_i}}{\sum_{l=1}^{J} e^{\gamma_l}} \tag{6}$$

α can be considered the new value we give each ordinal when placing it on the scale defined by sigma. While transforming from an ordinal scale to a continuous scale means that α could be anywhere between the two β cuts, in order to make the best transformation we will place α where the likelihood of observing the value is biggest, which is when the distance to the two β values is equal. α is calculated using equation:

$$\alpha_j = \frac{\beta_j + \beta_{j+1}}{2} \tag{7}$$

The transformed matrix $\tilde{X}_{n,m}$ is made using the α such that $O : X \rightarrow \tilde{X}$ where O do: $j_{n,m} = X_{n,m}$ and $\tilde{X}_{n,m} = \alpha_{j_{n,m}}$

S and C is calculated using Softmax on $\tilde{S}$ and $\tilde{C}$ such that they fulfil the properties described in section 3.1

$$S_{n,k} = \frac{e^{\tilde{S}_{n,k}}}{\sum_{i}^{K} e^{\tilde{S}_{n,i}}} \tag{8}$$

$$C_{k,n} = \frac{e^{\tilde{C}_{k,n}}}{\sum_{i}^{N} e^{\tilde{C}_{k,i}}} \tag{9}$$

The archetypal Archetypal analysis is then:

$$\hat{X}_{N\times M} = S_{N\times K} \cdot C_{K\times N} \cdot \tilde{X}_{N\times M} \tag{10}$$

σ needs to be positive non zero and can thus be calculated as:

$$\sigma = \ln(1 + e^{\tilde{\sigma}}) \tag{11}$$

Z is defined as:

$$Z_{j_{n,m}} = \frac{\beta_{j_{n,m}} - \hat{X}_{n,m}}{\sigma} \tag{12}$$

with the exceptions that $Z_0 = -\infty$ and $Z_J = \infty$ so the likelihood that and observation fall outside the scale is given to the first or last ordinal.

The loss function for the entire model is then:

$$loss = \sum_{n}^{N} \sum_{m}^{M} - \ln(\Phi(Z_{j_{n,m}+1}) - \Phi(Z_{j_{n,m}})) \tag{13}$$

where $j_{n,m} = X_{n,m}$

A gradient for γ sigma $\tilde{C}$ and $\tilde{S}$ is calculated using Pythorch autograd. Then a gradient step is made using the ADAM optimiser with a 0.01 learning rate.

3.4 Response bias ordinal archetypal analyses (RB-OAA)

The responses bias model assumes that each respondent has its own scale and noise. To avoid randomly selected responses dominating the archetype analysis the RB-OAA is initialised by running OAA. So we get the found values for $\tilde{C}$, $\tilde{S}$, γ and $\tilde{\sigma}$

γ and $\tilde{\sigma}$ is the duplicated N times such that γ is a $N \times J$ matrix and σ is a N long list.

The train loop is the computed in the same order but now β,α ,and σ in equation 6,7 ,and 11 is calculated row-vise. Z is then given as:

$$Z_{j_{n,m}} = \frac{\beta_{n,j_{n,m}} - \hat{X}_{n,m}}{\sigma_n} \tag{14}$$

and the loss function is the same as in equation 13

3.5 Synthetic Data

To test the model synthetic data was sampled to simulate ordinal data with an unknown true scale, and with known archetypes.

For a data set with M parameters, N samples and J ordinal the cuts between ordinals β is then by a J long list called $\beta parm$. $\beta parm$ can also be a single value, in this case it will be duplicated into a J long list of the same value. If the sampler is set to sample whiteout annotation bias the cuts is defended as:

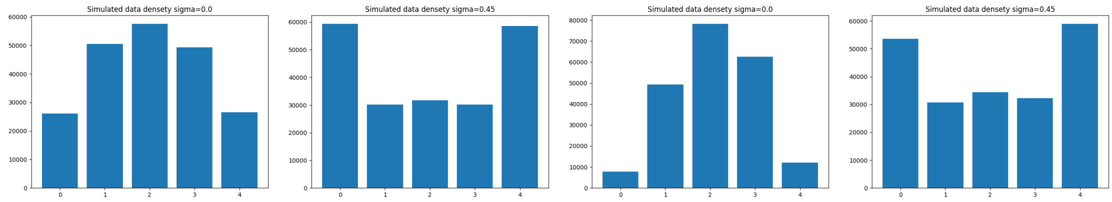
$$\beta_j = \sum_{i}^{j} \frac{\beta parm_i}{\sum_{l=0}^{J} \beta parm_l} \tag{15}$$

If annotation bias is to be sampled the way to do it is by using a Dirichlet distribution where β is constructed as a $N \times J$ matrix where each row is sampled from $Dir(\beta parm)$. If $\beta parm$ is a list the mean value for the cuts when $N \xrightarrow{\lim} \infty$ would the same as stated in equation 15. If $\beta parm = 1$ it means that the probability a sample would be heavily biased is the same as the sample being unbiased. The Higher $\beta parm$ is will then create a less bias data set.

An archetypes matrix $A_{K \times M}$ can be sampled by uniformly drawing $M \cdot K$ samples from α where alpha is given as

$$\alpha_j = \frac{\beta_j + \beta_{j+1}}{2} \tag{16}$$

A construction matrix $S_{N \times K}$ is made by sampling N rows from a Dirichlet distribution $Dir(Aparm)$ where $Aparm$ is a list of length K with the same value, in this thesis we only



(a) $\beta parm = 1$ (b) $\beta parm = 1$ (c) $\beta parm = 5$ (d) $\beta parm = 1$

Figure 4: Examples of destitution's of synthetic data for different β and $sigma$

explore the case where $Aparm = 1$ meaning that all archetypes are equally represented, and some sample will be a fit one archetype while others are a mix of all archetypes. If $Aparm \in [0, 1]$ a data-set where all sample is more likely to be one archetype will be generated, and if $Aparm > 1$ a data-set where all samples is a mix of different archetypes will be made. If γ of different K values is a list then low values will mean an archetype will be more represented in the data-set than one with a higher value. A data-set of continuous data can then be constructed as:

$$\hat{X}_{N\times M} = S_{N\times K} \cdot A_{K\times M}$$

$$\tilde{X}_{n,m} = \begin{cases} j & \text{if } \beta_j < \hat{X}_{n,m} < \beta_{j+1} \end{cases}$$

If a noise term σ is given then a probability function $P(\tilde{X}|\hat{X})$ is defined as:

$$Z_{n,m,j} = \frac{\beta_j - \hat{X}_{n,m}}{\sigma}$$

$$P(\tilde{X}_{n,m}|\hat{X}_{n,m})_j = \Phi(Z_{n,m,j+1}) - \Phi(Z_{n,m,j})$$

Where $\Phi(\cdot)$ the is cumulative distribution function of a standard normal distribution. $\tilde{X}$ is then found by sampling integer in the range $[0, J[$ from the distribution $P(\tilde{X}|\hat{X})$. If provided a K long list of α^{false} the data is transform to the final data-set X by using $\tilde{X}$ as an index. With the underline assumption that α^{false} is a custom ordinal scale such as 1,2,3.. 10,20,30.. a,b,c... etc.

3.6 Evaluation methods

The performance of the model was measured using Archetype correlation coefficient and Normalised mutual information. To measure the ability to find response bias the mean distance between β values was calculated. For the simulated data the found attribute's was compared with the true attribute's. For the real data, several models will be trained by random initialisation and mean value of the inter-model correlations, NMI and difference in β was calculated under the assumption that if several models with different initialisation come to the same result it will indicate the models found and underlying pattern in the data.

3.6.1 Archetype correlation coefficient

Given two archetype matrices A_1 and A_2 the Pearson correlation coefficient between each row and compared to each row in the other matrix is found, giving a $K \times K$ matrix. Since archetypes are randomly indexed it can be assumed that the archetypes with the highest correlation coefficient are paired. Thus by taking the maximum correlation coefficient in each row a K long list of the archetype's correlation coefficients is found. And by taking the mean of that list a quantitative measure two compere two archetype's matrices is found.

3.6.2 Normalised mutual information (NMI)

Normalised mutual information (NMI) is a way to evaluate the performance of archetypal analysis[11]. It is based on compering the mutual information in between reconstruction matrix S construction matrix for simulated data S^{true} described in section 3.5. Where:

$$NMI(S, S') = \frac{2I(s, s')}{I(S, S) + I(S, S')}$$

$$I(S, S') = \sum_{k}^{K} \sum_{k'}^{K} \frac{S_k \cdot S_{k'}}{N} \ln\left(\frac{S_k \cdot S_{k'}}{P(S_k)P(S_{k'}) \cdot N}\right)$$

$$P(S_k) = \frac{\sum_n^N S_{k,n}}{\sum_{n'}^N \sum_{k'}^K S_{n',k'}}$$

3.6.3 Response bias analysis

To evaluate the models ability to measure response bias distance between the values was calculated.

$$\Delta\beta = |\beta 1 - \beta 2| \tag{17}$$

To calculate the mean value over all $\Delta\beta$ the first and last value in each row is excluded that since those values is 0 and 1 by definition.

4 Results and analysis

The models were trained on different data sets, all the models were trained by running 1000 epochs each with a learning rate of 0.01. An impotent detail when comparing the two models is to remember that RB-OAA is initialised with OAA reported, thus the lower loss function of the RB-OAA should not be considered evidence of the RB-OAA superiority but rather the consequence of adding more parameters to a model already converted. This in practice also means that the RB-OAA runs for 2000 epochs because it is initialised by a model that has already completed 1000 epochs.

It should be noted that I discovered a mistake when saving σ so the reported is from the last epoch not the best one, thus σ may be inaccurate for models who's best epoch is not the last.

4.1 Experiments on synthetic data

All data-set was sampled with 10000 respondents with 21 categories, the dataset was sampled with 4 archetypes and the models was set to reconstruct the 4 archetypes. For each experiment, the models were trained with different random initialisation.

4.1.1 Learning a new scale

A mini experiment was made to demonstrate OAA's ability to find a new scale where the distance between ordinal was not the same, the data was sampled without annotation bias by a scale where $\beta^{true} = [0, 0.3, 0.5, 0.6, 0.9, 1]$

	0	1	2	3	4	5
0	0.0	0.249125	0.471052	0.589380	0.923262	1.0
1	0.0	0.253736	0.476009	0.594316	0.930327	1.0
2	0.0	0.247567	0.468943	0.585420	0.922040	1.0

	0	1	2	3	4	5
0	0.0	0.235469	0.449829	0.559299	0.886768	1.0
1	0.0	0.275500	0.501121	0.614784	0.961141	1.0
2	0.0	0.254795	0.463252	0.568525	0.891250	1.0

	0	1	2	3	4	5
0	0.0	0.294865	0.472285	0.562103	0.826504	1.0
1	0.0	0.413322	0.592468	0.681914	0.947942	1.0
2	0.0	0.332655	0.477698	0.550748	0.768028	1.0

Figure 5: Examples of new scales learned with noise 0,0.10 and 0.25

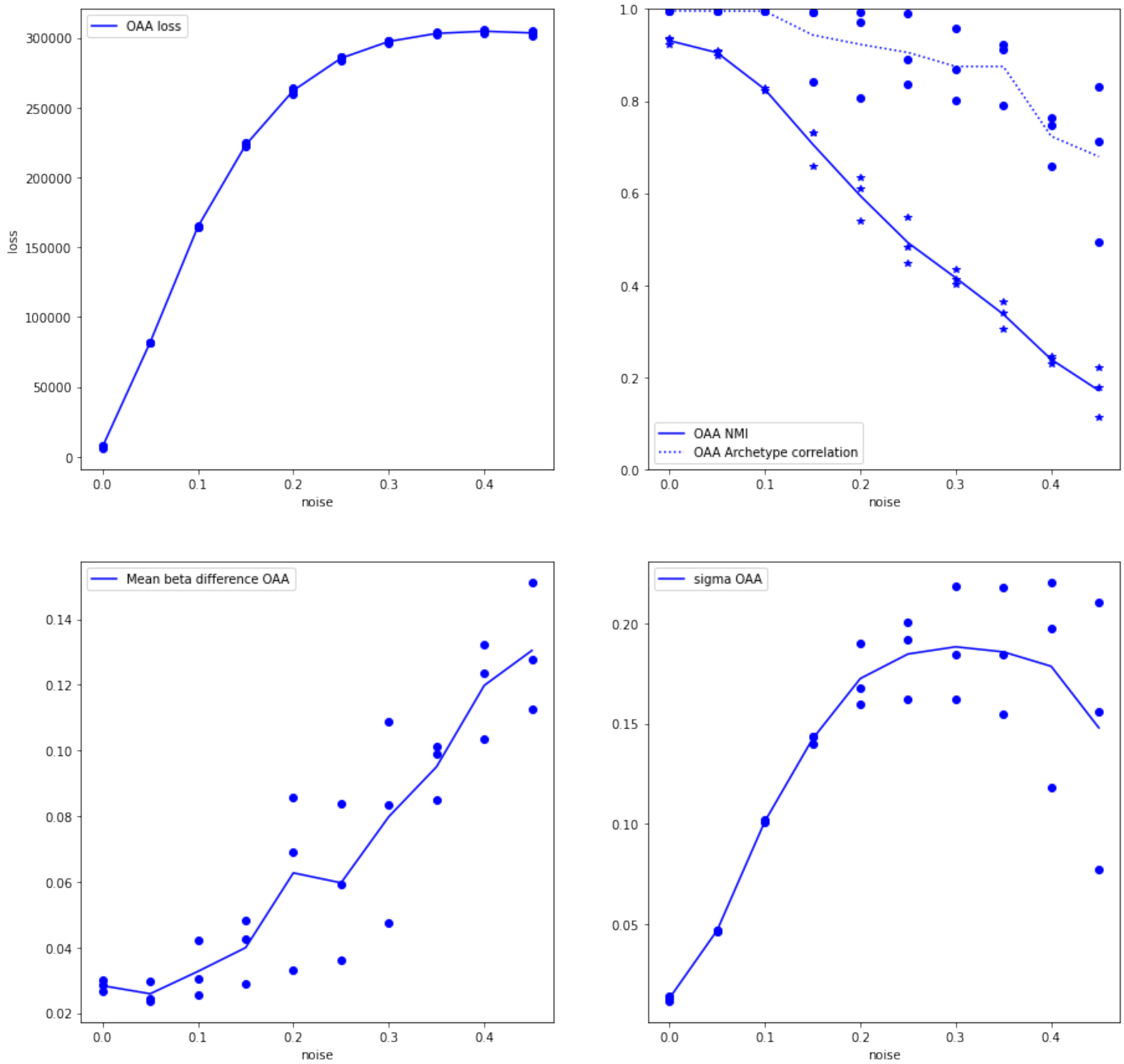
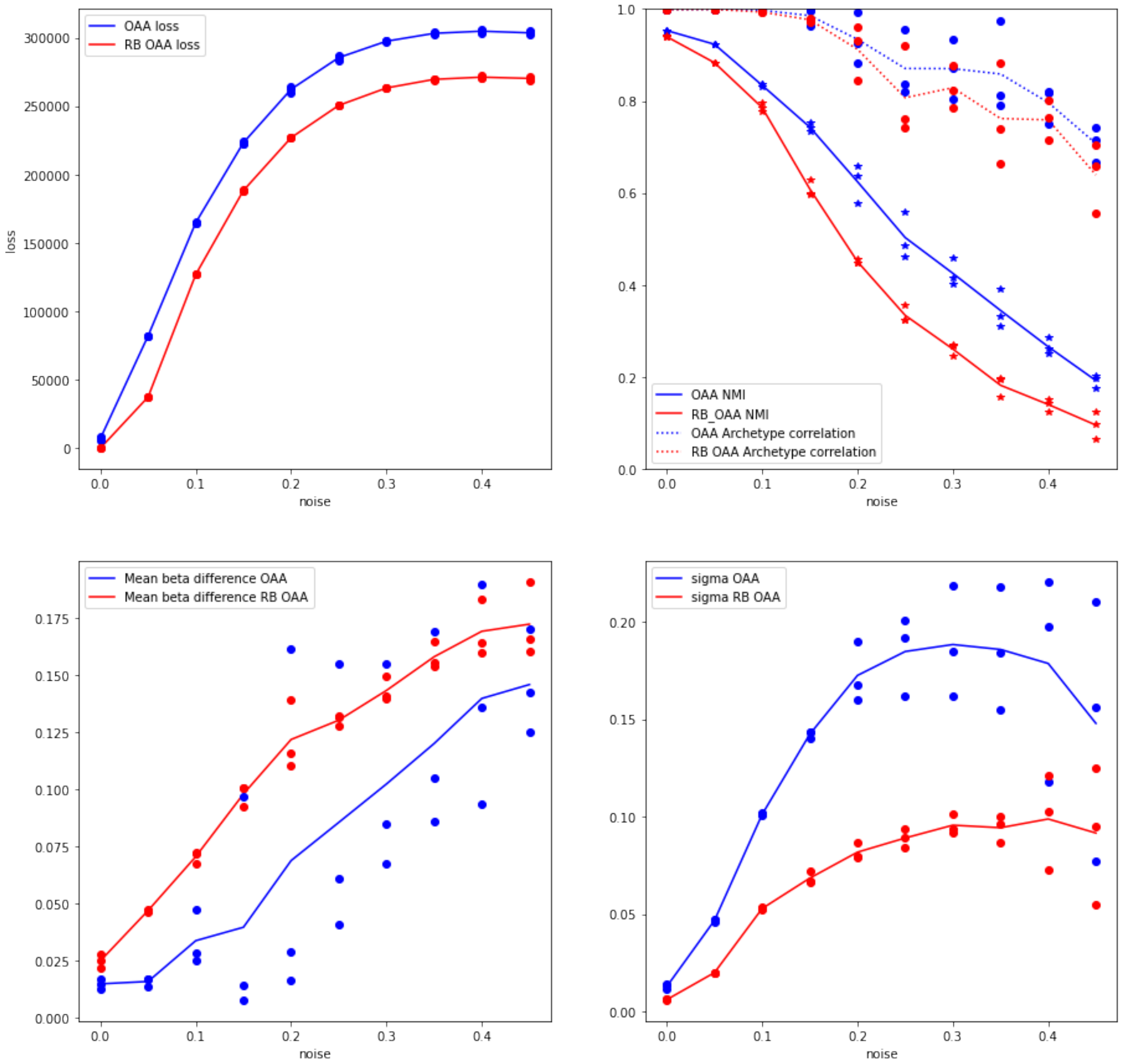


Figure 6: OAA performed on an ordinal scale

4.2 model comparison

The two models was tested on different noise levels, with a 2 selected beta parameters and one sample data-set sampled without annotation bias



[H]

Figure 7: Model performance on data sample without response bias.

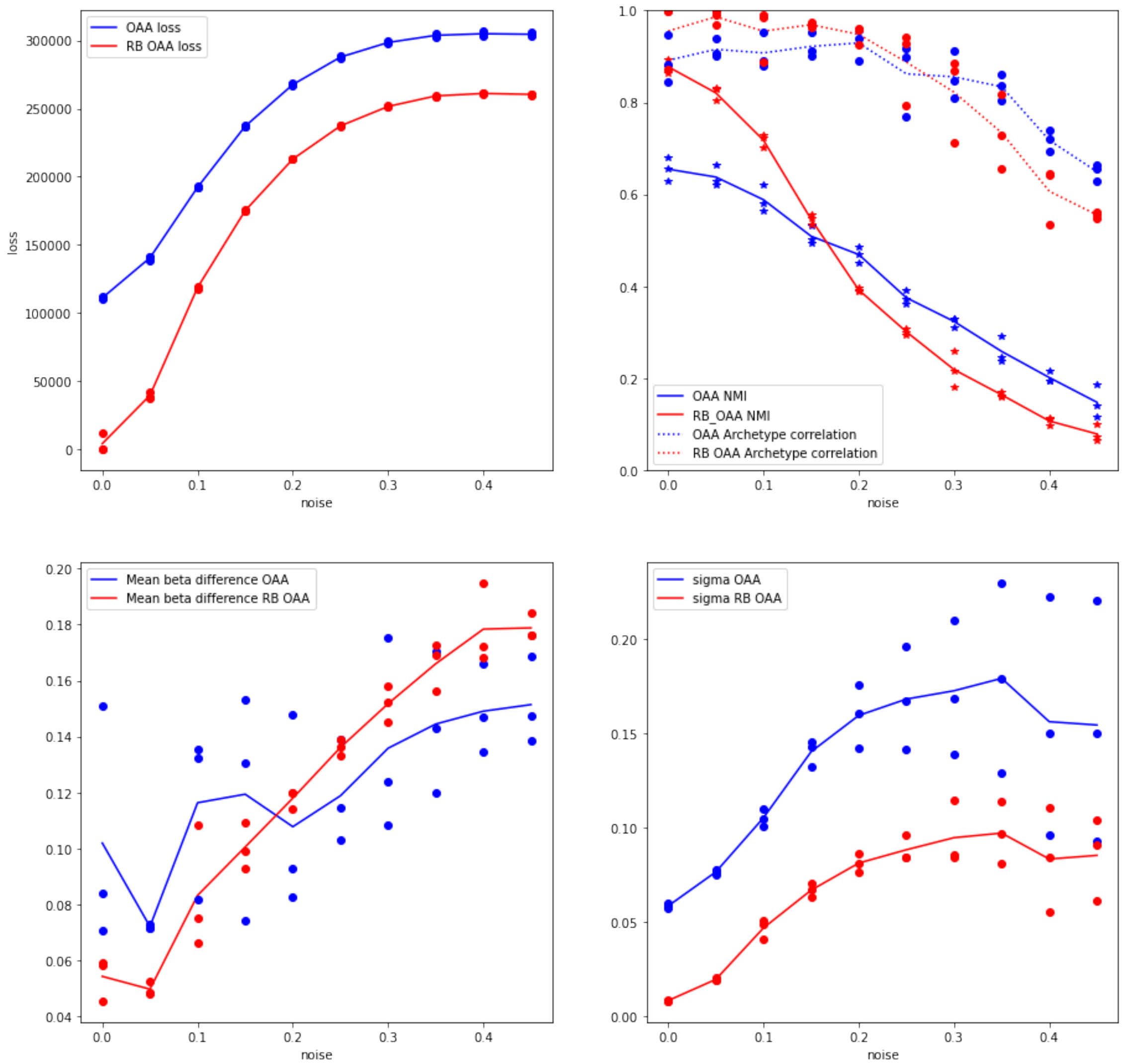


Figure 8: Model performance on date sample with $\beta parm = 5$.

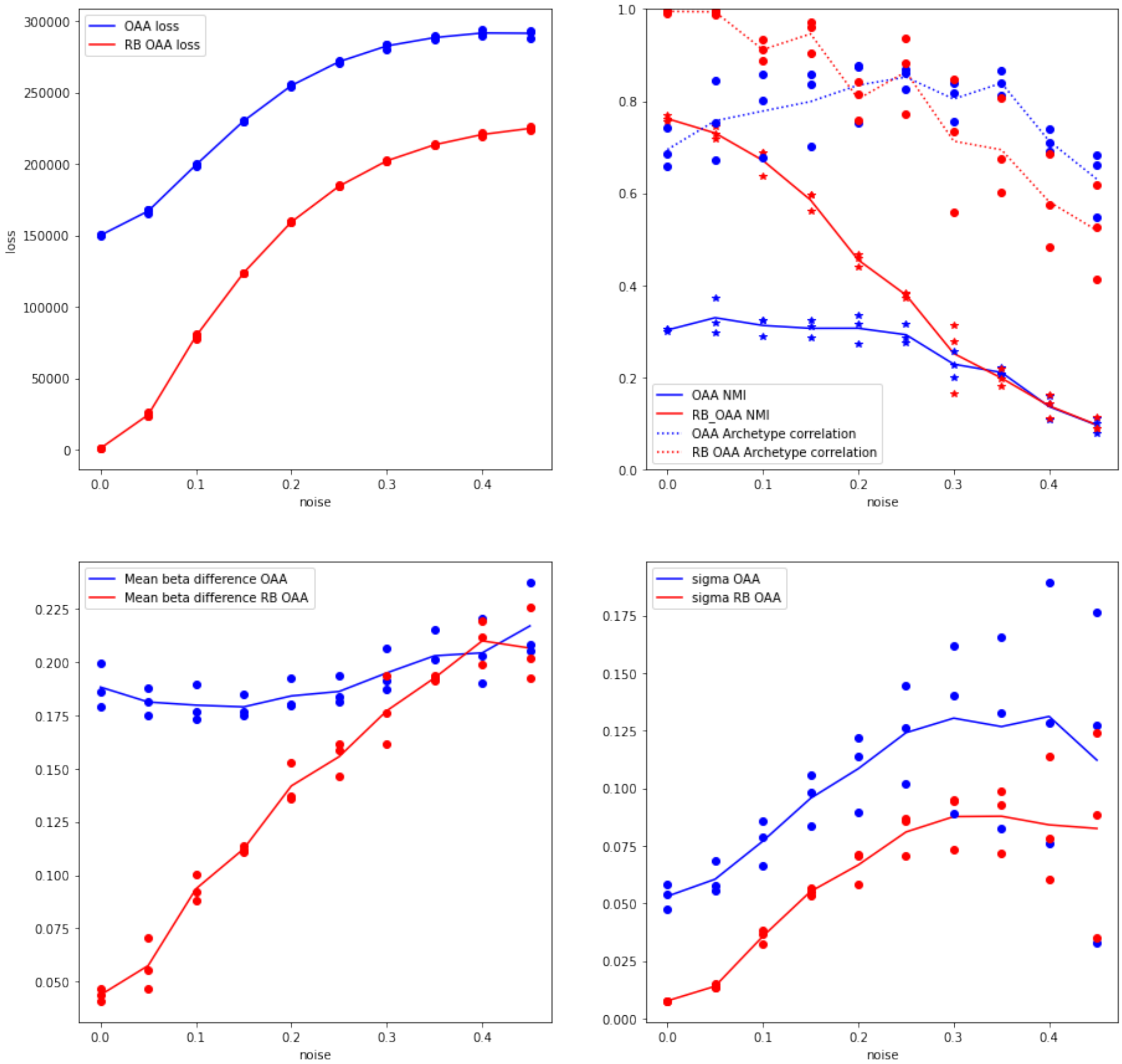


Figure 9: Model performance on date sample with $\beta parm = 1$.

4.3 Experiment on real data

on the real ESS-8 data-set experiments were made for different numbers of archetypes in the range 2 to 20, for each experiment 3 models were trained.

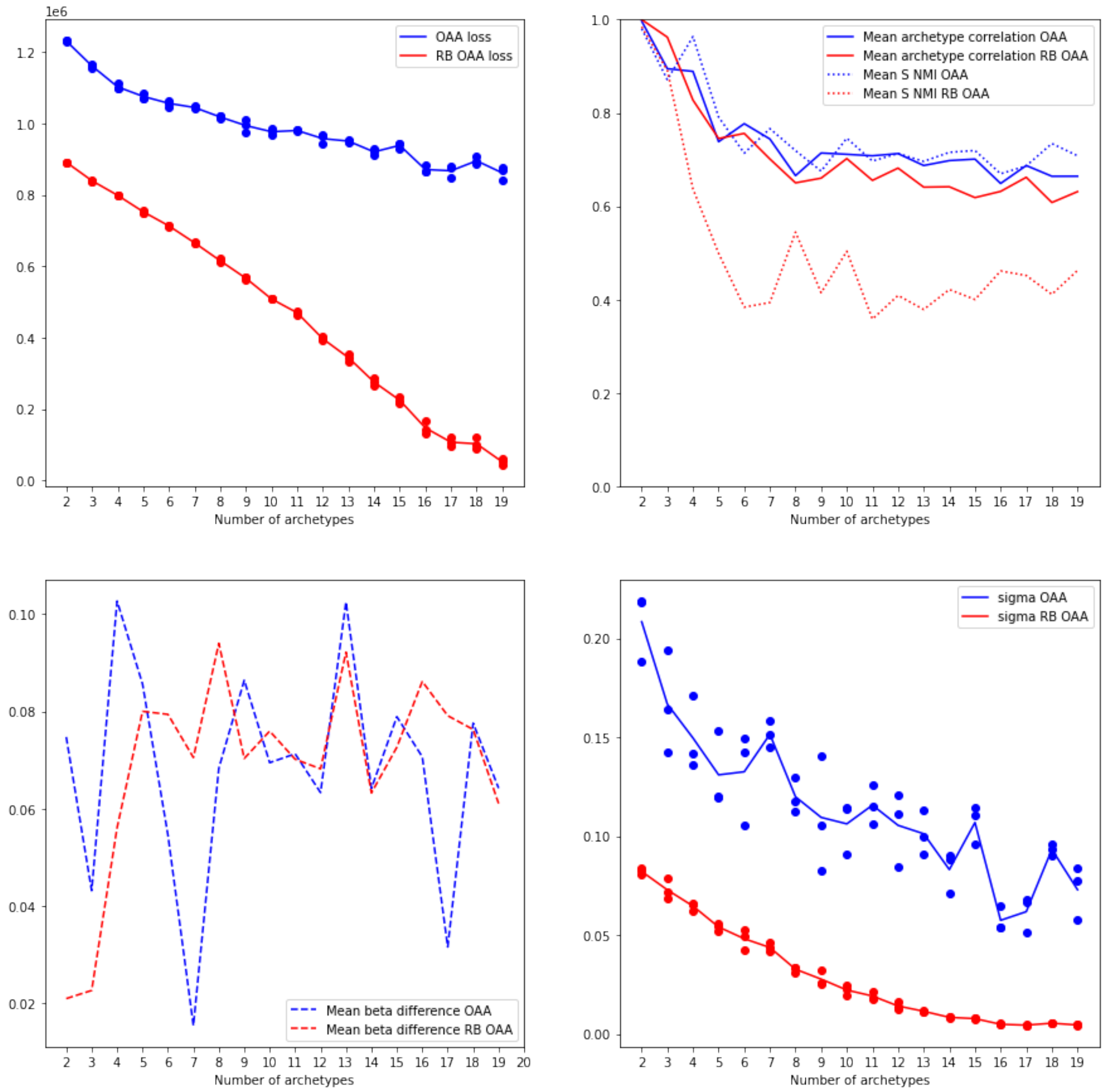


Figure 10: Model performance on the ESS8 dataset

Figure 10 should be used to select the best numbers of archetypes interesting to look at for further analysis, and ideally say something about the numbers of archetypes in the data set. First by looking at the OAA model based on the elbow criterion[5] the point where the loss function change direction should be considered. Thus 4 would be a point of interest, OAA also makes has good inter-model arability in point 4.

4.3.1 4 archetypes

Since 4 archetypes was a point of interest results from this experiment is reported.

model	OAA1	OAA2	OAA3	-	RB-OAA1	RB-OAA2	RB-0AA3
Loss	1.09 e6	1.11 e6	1.09 e6	-	797836	800118	797638
Σ	0.14	0.17	0.13	-	0.066	0.062	0.065
best epoch	1000	1000	100	-	1000	999	959

Table 3: Training summery from models with 4 archetypes

model	OAA1	OAA2	OAA3	-	RB-OAA1	RB-OAA2	RB-0AA3
compered to	OAA2	OAA3	OAA1	-	RB-OAA2	RB-OAA3	RB-0AA1
AC	0.83	0.85	0.98	-	0.77	0.72	0.98
NMI	0.49	0.47	0.94	-	0.66	0.62	0.94
$\Delta\beta$	0.1	0.15	0.04	-	0.071	0.08	0.017

Table 4: Inter model correlation

4.3.2 Correlation with external variables

When looking at the data in 12 it appeared that there is none to little correlation, between external variables and archetypes, or at least it can be said that external variables correlated more with each other than with archetypes. Thus further investigation into the relationship between archetypes and meta variables was not prioritised.

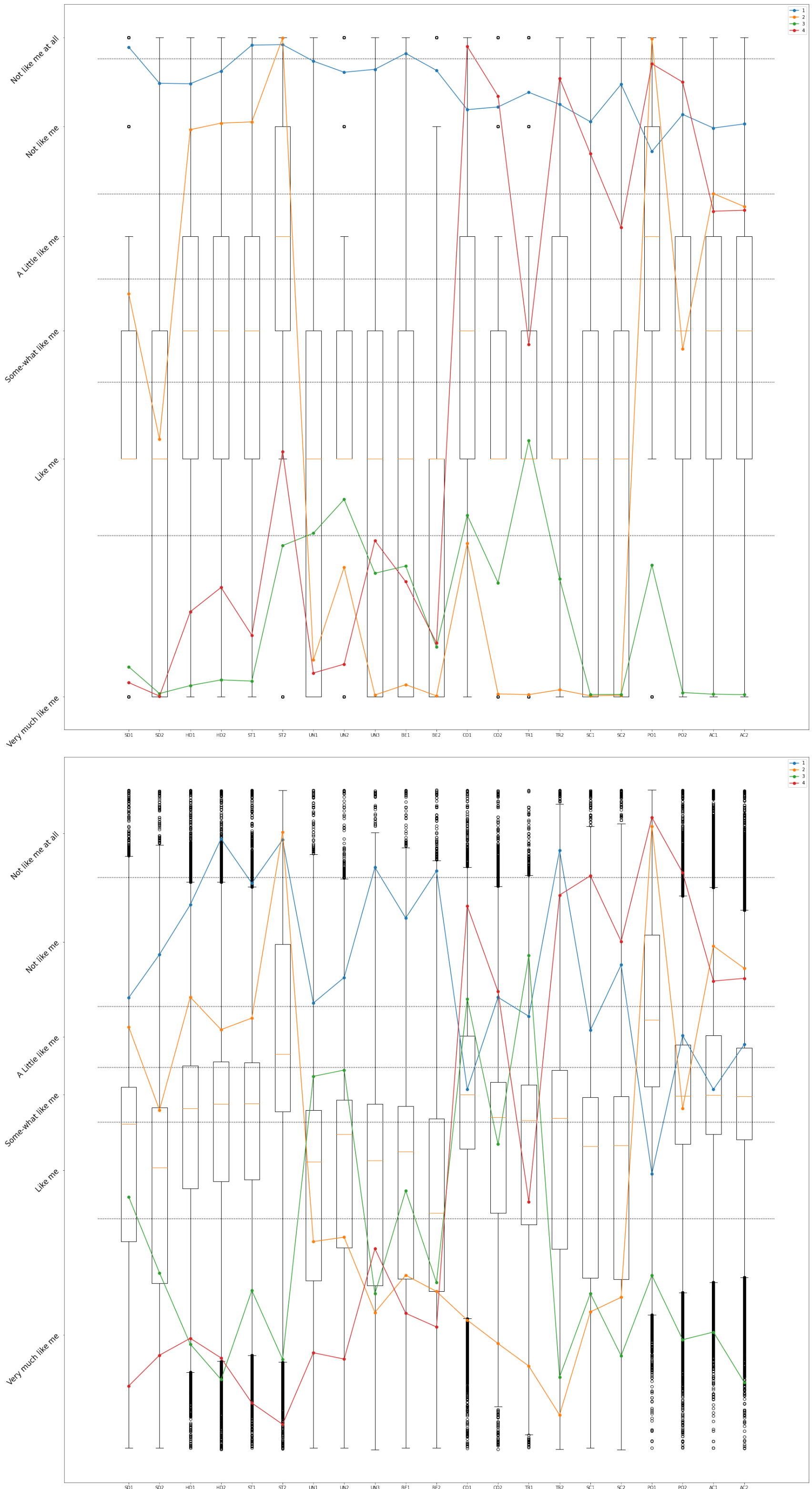
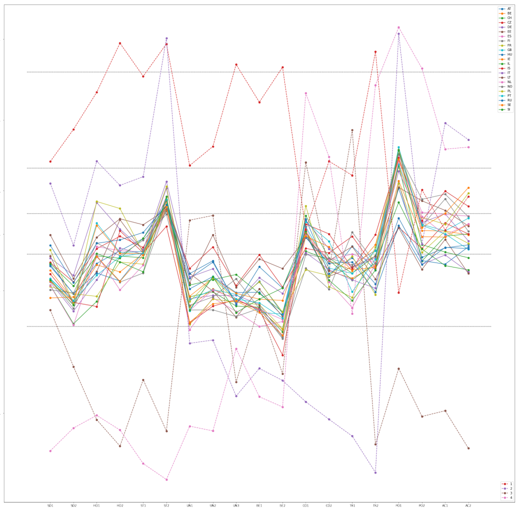
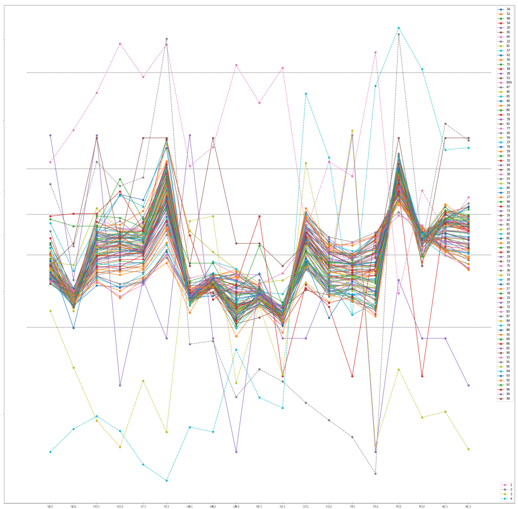


Figure 11: Archetypes and data plotted on the new scale($\tilde{X}$) OAA to the right and RB-OOA to the left

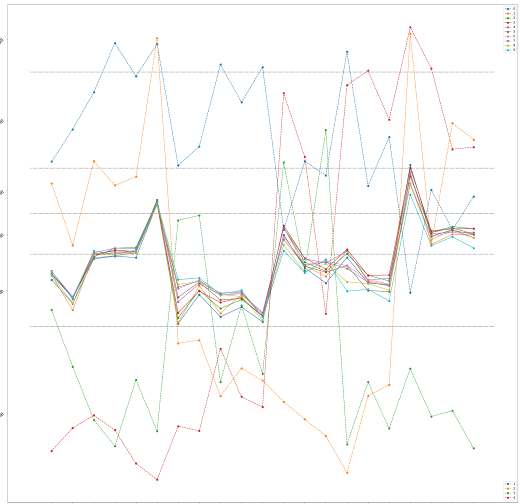


(a) Archetype versus country's

(b) Archetype versus gender 1 being male and 2 female

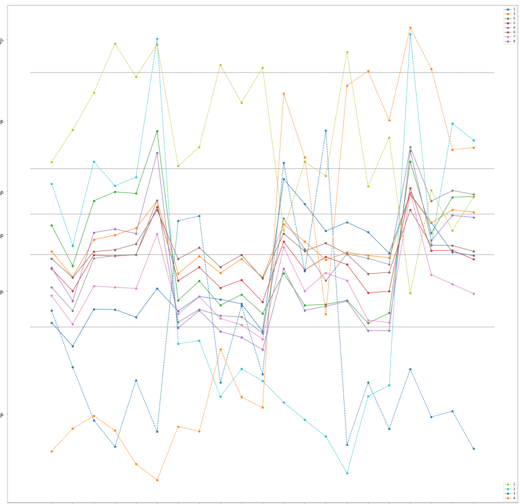


(c) Archetype versus age



(d) Archetype versus political learning 0 left to 10 right

(e) Archetype versus subjective income in the range 0 to 10



(f) Fumiko thought it would be interesting to compare RB-OAA to the results made by LPA[9] SB8 since those results was include in my data set.

Figure 12: Archetypes dotted lines plotted versus mean of $\tilde{X}$ for different external variables. The model used is RB-OOA shown in figure 11

5 Discussion

5.1 Implementation and computational limitations

Choosing the right hyperparameters for training a machine learning model is often based on intuition and experience unless active learning is used. The choice of the Adam optimiser [12] was based on what seem like best practice in the machine learning community because of its ability to automatically tune momentum for each parameters. Looking at figure 13: it should be noted that better results can be generated using more epochs for cases where the curve doesn't flatten out. In the cases where the loss function flicker or even increases a lower learning rate should have been used. While the choice of running 1000 epochs mostly was to produce the quantity of data reported in realistic computing time, its also due to a known bug in the code that allows data from previous epochs to accumulate cursing the model to run out of memory around 1300 epochs in for the ESS8 data-set.

5.2 Model comparison

Since I am proposing two models, it should be discussed which advantages and disadvantages of both models. Based on the simulated experiments it seems like an assumption on the nature of the noise in the data-set should be made. If the data-set has a great amount of random noise the RB-OAA seems to overfit on the noise and find bias where it doesn't exist. This also seems to happen even in cases where a small annotation bias may exist. Thus the RB-OAA should only be used on data where it can be assumed that i sizeable portion of the noise can be explained as annotation bias. While the OAA can be assumed to work on all ordinal data sets.

While the inter-model agreement on RB-OAA is lower than for the OAA there is still indicators that RB-OAA is the best choice for the ESS-8 data-set, the fact that OAA seems to construct archetypes that is always positive or negative indicates that it learns a response bias instead of archetypes, this also means that the models mean inter-model agreement will be artificial high by always finding those two archetypes. This could be compensated for by comparing OAA to RB-OAA with two less archetypes.
It can also be argued that inter-model agreement is a bad way to evaluate AA since AA is prone to find local minimums and looking at table 4 It seems that RB-OAA is even more prone to over-fitting giving a low inter-model agreement even though two models find the same minimum this problem may have been smaller had more than 3 models been trained per experiment.

5.3 Learning a new scale

While the results shows the model learns a new scale the model still seems to be "off" in the first and last cut in figure 5. For the OAA model, a scale seemed fitted to the destitution of the data-set in figure 11, underlining the fact that it was formed by an ordinal likelihood function. When taking the scale found by taking the mean of all the RB-OAA we see a different scale where the portion of the scale reserved for the two edge values "Very much like me" and "Not like me at all" is much bigger than in OAA see figure 11. The way this can be explained is that the model tries to define the scale such that 0 is the person most agreeing and 1 is the person most disagreeing. When the model adjusts for response bias it will find that a person who answered "Not like me at all" in only one question is identifying more with the statement than one who answered "Not like me at all" in all questions. The first person can thus be considered more extreme than the second and in order to fit both of them into the same category the scale needs to be expanded.

When the model was originally proposed the α values was learned and the β values was defined by α while me and Morten later agreed that the scale should be defined by β a

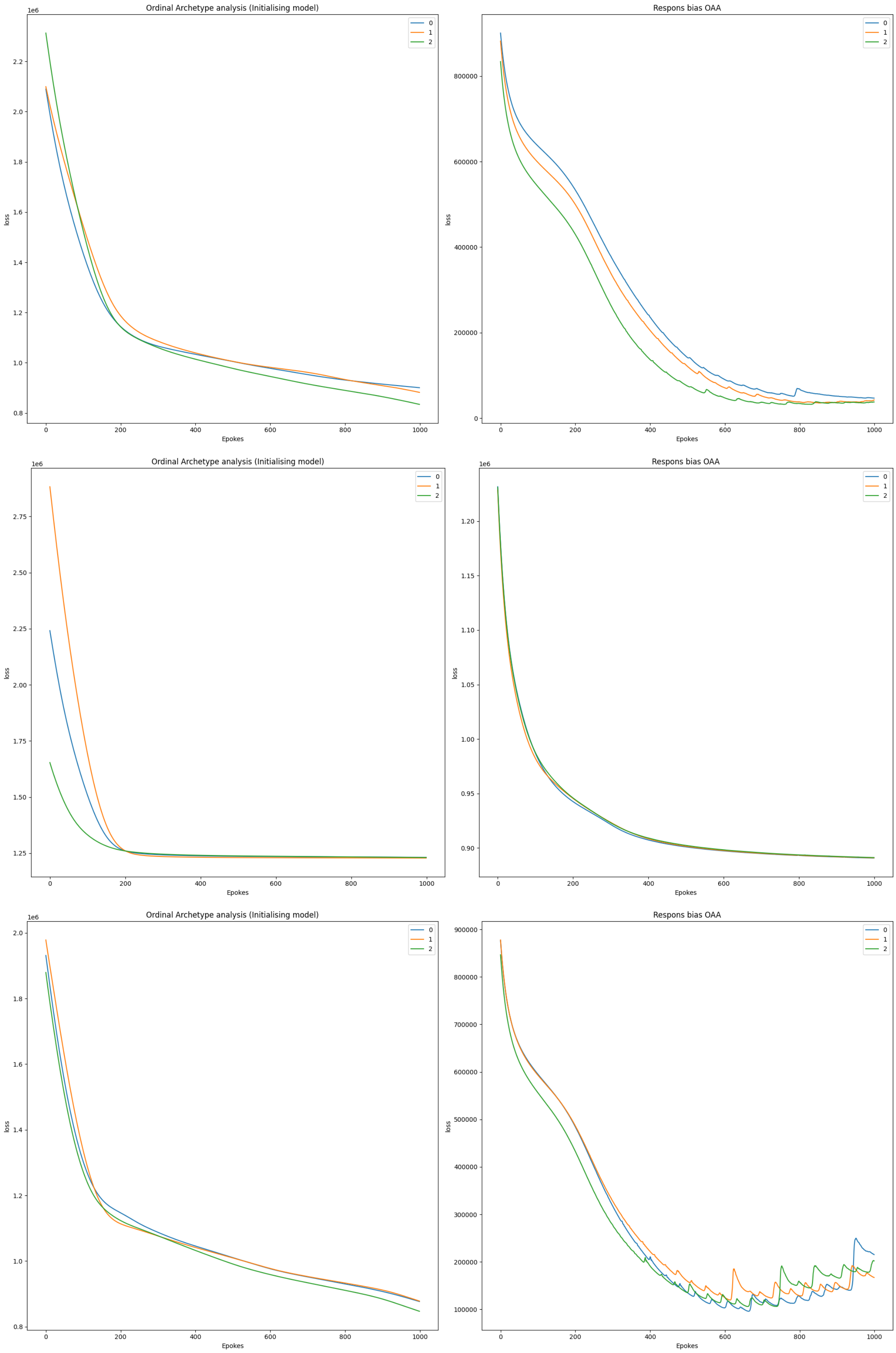


Figure 13: Loss function under training, top is synthetic data without noise and $\beta parm = 1$ middle is real data for 2 archetypes and bottom is real data for 17 archetypes

model where α as a parameters is limited, but not defined by β. This may improve the model in the critical areas.

While the model seems to models the noise in the data-set it should also be notice that they seems to underestimate the noise indication over-fitting. We also see that estimation of noise seems to droop after 0,25. However if looking at figure 4 we see that when sampling for high values of σ the data concentrate in the edges, arguably making the data appear less noise-full, this can be explained by the way the data is sampled where high noise means more of the distribution being given to the edge cases. Unfortunately due to the coding error that saved the wrong σ value all of this should be taking with some uncertainty, it should be noted that since OAA ran for less epoch-es than RB-OAA the sigma value should be more reliable since the last epoch mostly was the best epochs, see figure 13 for examples.

Further researcher to clearly demonstrate the correctness of the model is needed with the obvious first step to compare my OAA model with the one proposed by Fernández, Epiganio and McMillan [5]

5.4 Model finding related to Schwarts Theory

I would like to make a qualitative comment on Schwarts theory and the European Social Surveys implementation of the theory. For the OAA model archetypes, 1 and 3 should be disregarded as always positive and always negative. First, we should note that all questions with the same label seem correlated (archetypes answer the same or almost the same on HD1 and HD2 and so on) However I noted some exceptions, some archetypes answer very different on TR1 vs TR2 and PO1 vs PO2. If we look at table 2 we see that TR1 is about being humble and modest while TR2 ask more about traditions, which could indicate that humble and traditional is not the same value. The same can be found in the power category where PO1 is related to wealth while PO2 is about people doing what the person tells them again raising the question: is power and wealth two distinct values.?

It can also be seen that the archetypes align well with the self vs social and growth vs protective axis seen in figure 2. For the OAA we clearly see that archetype 2 answers very much like me on all social questions while being more moderate or disagreeing on self-directed questions and archetype 4 is very growth focus while disagreeing with protective questions. The same can be said for the RB-OAA where most archetypes make a cut along the axis but where the model allows more for more complex archetypes.

5.5 OAA's commercial value

The OAA and RB-OAA showed no reassembling between the mean of data from external variable and the archetype's, suggesting that archetypes are not linked to gender, nationality, age or social-economical status. However different approach such as a using the S matrix as a classifier may prove a connection between external variables and archetypes. After a meeting with Furmiko, we concluded that OAA didn't have much marketing value when applied to basic human values, but could have commercial value when applied to other forms of consumer surveys. Examples could be the original AA paper[1] where AA was used as a tool to product development of face-mask. In the same way, RB-OAA may be helpful in designing products or marketing campaigns based on surveys. Using AA to categorise consumers has already been tried in a paper making archetypes on consumers ethical values when shopping [13] a paper that may have benefited from RB-OAA.

6 Conclusion

Archetypal analysis can be combined with an ordinal likelihood function to create an Ordinal Archetypal Analysis that can recreate archetypes from a data set. Besides that OAA shows promising results in mapping ordinal data to a new continuous scale.

RB-OAA most likely has the ability to find and adjust for response bias in data-set making it a useful tool to analyse surveys. Archetypes found showed a high resemblance to Schwarts theory of basic human values, suggesting the correctness of both the model and the theory. However, the model may suggest that some values should be broken down into even smaller categories.

OAA didn't find the correlation to external variables that we had hoped for, but can not say if other approaches may give better results. I still think OAA and RB-OAA have a commercial value in marketing research when applied on different data sets.

Bibliography

[1] A. Cutler and L. Breiman, "Archetypal Analysis," *Technometrics*, vol. 36, no. 4, pp. 338–347, Nov. 1994, Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10 ISSN: 0040-1706. DOI: 10.1080/00401706.1994.10485840. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/00401706.1994.10485840 (visited on 10/12/2021).

[2] W. Chu and Z. Ghahramani, "Gaussian Processes for Ordinal Regression," en, p. 23,

[3] *About ESS | European Social Survey (ESS)*. [Online]. Available: https://www.europeansocialsurvey.org/about/ (visited on 10/08/2021).

[4] R. A. Dwyer, "On the Convex Hull of Random Points in a Polytope," *Journal of Applied Probability*, vol. 25, no. 4, pp. 688–699, 1988, ISSN: 0021-9002. DOI: 10.2307/3214289. [Online]. Available: https://www.jstor.org/stable/3214289 (visited on 11/30/2021).

[5] D. Fernández, I. Epifanio, and L. F. McMillan, "Archetypal analysis for ordinal data," en, *Information Sciences*, vol. 579, pp. 281–292, Nov. 2021, ISSN: 00200255. DOI: 10.1016/j.ins.2021.07.095. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S002002552100791X (visited on 10/04/2021).

[6] S. H. Schwartz, "An Overview of the Schwartz Theory of Basic Values," en, *Online Readings in Psychology and Culture*, vol. 2, no. 1, Dec. 2012, ISSN: 2307-0919. DOI: 10.9707/2307-0919.1116. [Online]. Available: https://scholarworks.gvsu.edu/orpc/vol2/iss1/11 (visited on 10/04/2021).

[7] F. Leisch and M. Eugster, "From Spider-Man to Hero - Archetypal Analysis in R," *Journal of Statistical Software*, vol. 30, Apr. 2009. DOI: 10.18637/jss.v030.i08.

[8] *Data and Documentation by Round | European Social Survey (ESS)*. [Online]. Available: https://www.europeansocialsurvey.org/data/download.html?r=8 (visited on 10/08/2021).

[9] M. N. Schmidt, D. Seddig, E. Davidov, M. Mørup, K. J. Albers, J. M. Bauer, and F. K. Glückstad, "Latent profile analysis of human values: What is the optimal number of clusters?" en, *Methodology*, vol. 17, no. 2, pp. 127–148, Jun. 2021, ISSN: 1614-2241. DOI: 10.5964/meth.5479. [Online]. Available: https://meth.psychopen.eu/index.php/meth/article/view/5479 (visited on 10/04/2021).

[10] M. Mørup and L. K. Hansen, "Archetypal analysis for machine learning and data mining," en, *Neurocomputing*, vol. 80, pp. 54–63, Mar. 2012, ISSN: 09252312. DOI: 10.1016/j.neucom.2011.06.033. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925231211006060 (visited on 10/04/2021).

[11] J. L. Hinrich, S. E. Bardenfleth, R. E. Roge, N. W. Churchill, K. H. Madsen, and M. Morup, "Archetypal Analysis for Modeling Multisubject fMRI Data," en, *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 7, pp. 1160–1171, Oct.

2016, ISSN: 1932-4553, 1941-0484. DOI: 10.1109/JSTSP.2016.2595103. [Online]. Available: http://ieeexplore.ieee.org/document/7523967/ (visited on 11/15/2021).
[12] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *arXiv:1412.6980 [cs]*, Jan. 2017, arXiv: 1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980 (visited on 12/02/2021).
[13] S. Li, P. Wang, J. Louviere, and R. Carson, “Archetypal analysis: A new way to segment markets based on extreme individuals,” in *In Proc. Australian and New Zealand Marketing Academy Conference*, 2003.