Quantifying Psychological Portrait

Michał Pogoda, Jan Sieradzki

May 2020

Contents

1 Data introduction and analysis

1.1 Dataset decription

We are analyzing a dataset that was created based on a human psychology model called the Big Five model. This model states, that psychological portrait can be described by 5 variables, namely

- Surgency or Extraversion
- Agreeableness
- Conscientiousness
- Emotional Stability
- Intellect or Imagination

The questionnaire consists of 50 statements, scaled from 1 to 5 based on how much the respondent agrees with the statement. It's explicitly defined, that 1 means "strongly disagree", 3 means "neutral" and 5 means "strongly agree". There are 50 questions in total. Questions are divided into 5 groups, each composed of 10 questions. Questions from one group are tailored to measure 1 particular trait at the time.

In the raw dataset, there are additional metadata like time users spent answering each question and parameters of a device on which questionnaire was completed (IP origin, resolution, etc.). Where research into these metadata might be insightful (eg. how country affects trait distribution, are people with larger self-esteem more likely to have high resolutions?), we explicitly focused on questionnaire data.

1.2 Global statistics

In the table below, there is information about measures "mean" and "std" for each category of questions. We can see that mean of answers is roughly 3 and the standard deviation is about 1.3, so answers are on average "neutral".

trait	mean	std
Surgency or Extraversion	3.118	1.358
Emotional Stability	3.023	1.335
Agreeableness	3.155	1.394
Conscientiousness	3.123	1.314
Intellect or Imagination	3.266	1.396

Table 1: Basic statistic for each traits in data set

1.3 Correlations

In this data set, the most interesting factor is relations between questions and traits that would end up describing psychological portrait person. Heat maps were used to visualize all correlations between questions. On the main heat map from the entire set, we can notice five main areas of correlations (especially big correlations are in Extraversion(EXT) column). These areas cover each five questions category, which confirms the theory about Big Five Factor Markers. We can notice (much smaller) correlations between specific groups. There is a significant connection between columns Extraversion(EXT) and Agreeableness(AGR), which seems logical because both depend on similar character traits (for example self-confidence). On the heat maps associated with specific traits, we can notice, that the most correlated questions are in "surgency or extraversion" and in "intellect or imagination" are correlated the least.

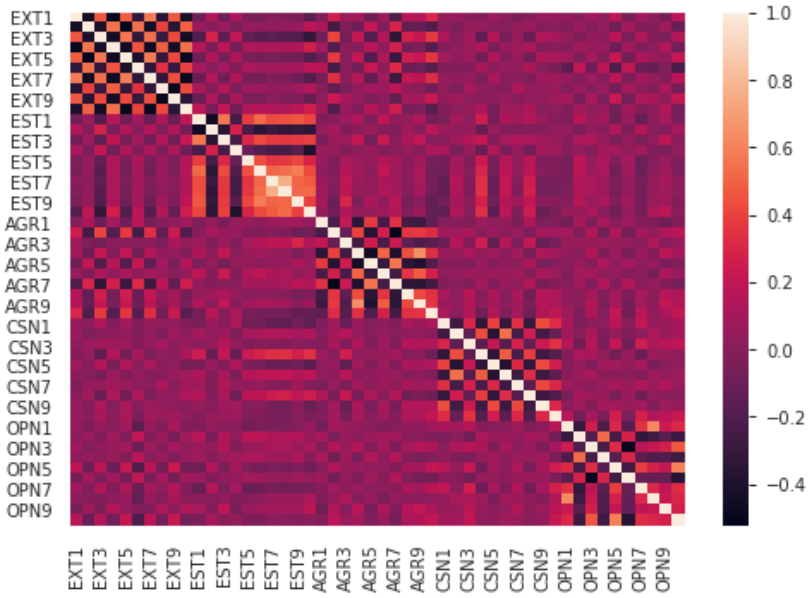

Figure 1: Heat map of entire data set

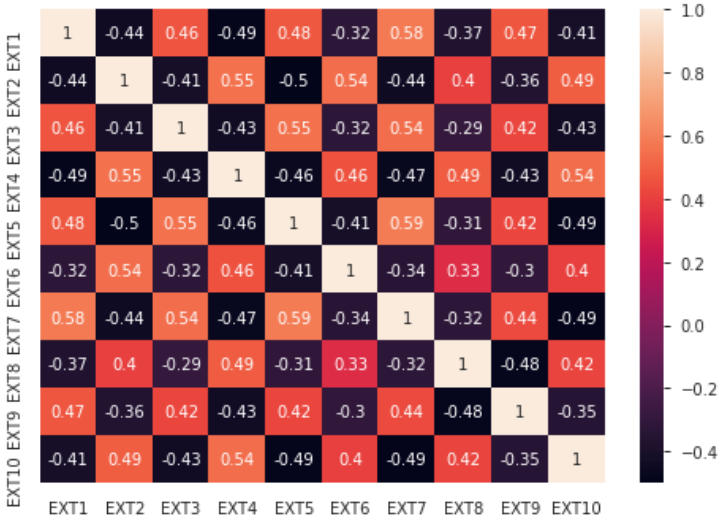

Figure 2: Heat map of surgency or extraversion

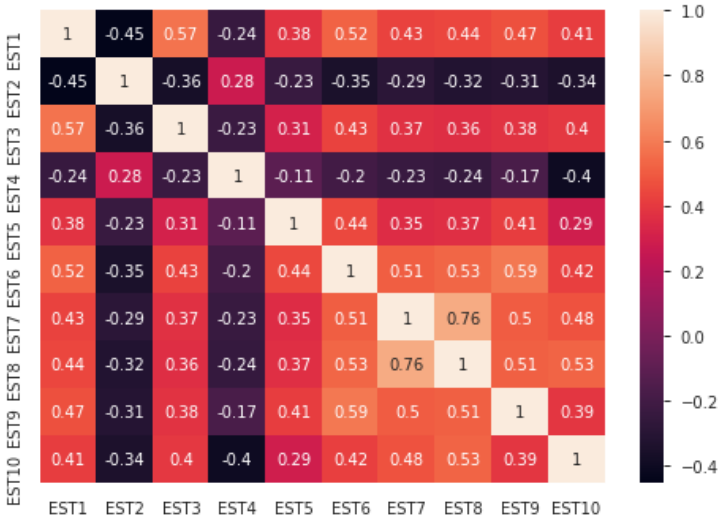

Figure 3: Heat map of emotional stablity

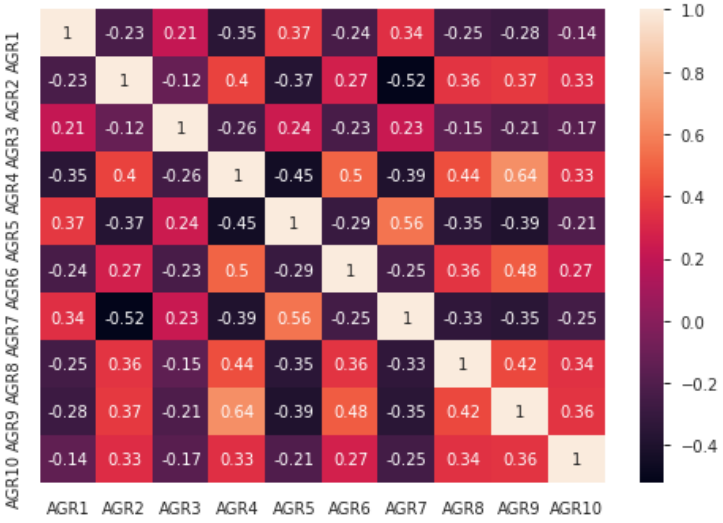

Figure 4: Heat map of agreeableness

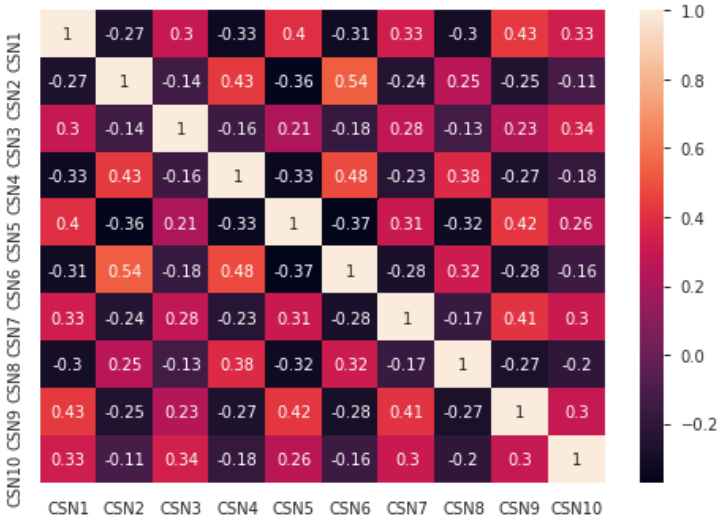

Figure 5: Heat map of conscientiousness

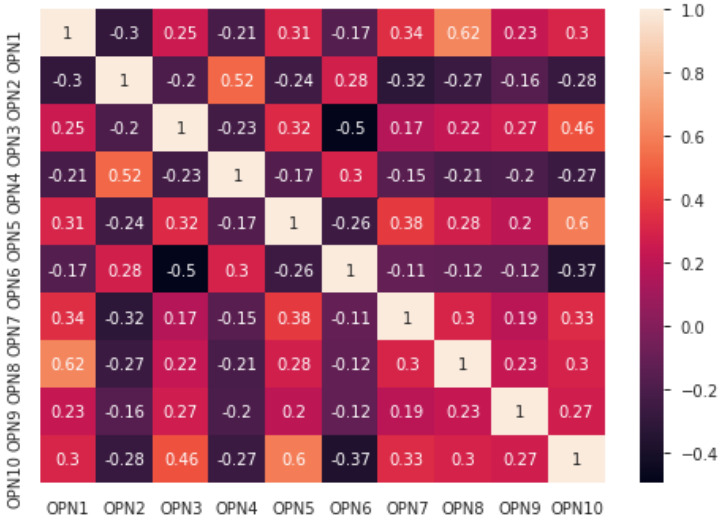

Figure 6: Heat map of intellect or imagination

Next important step during data analise was generating histograms for each question. Thanks to that, we could check how exactly were distributed answers in our data set with 1 000 000 samples. The exemplary plots for questions are presented below. We can think about these plots, as approximation of probability for grades in categorical distribution for all 50 questions. Most of these histograms would be also well described by Gaussian distribution.

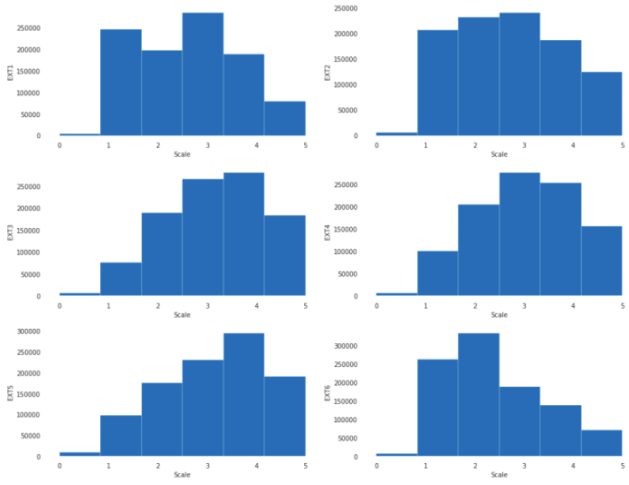

Figure 7: Histograms of answers for questions

The expressly interesting aspect of this data set is the influence of test par-

ticipant nationality on answers. Below there is one more histogram, which represents the nationality distribution of test participants. The biggest bar shows the number of USA representations, which dominated our dataset.

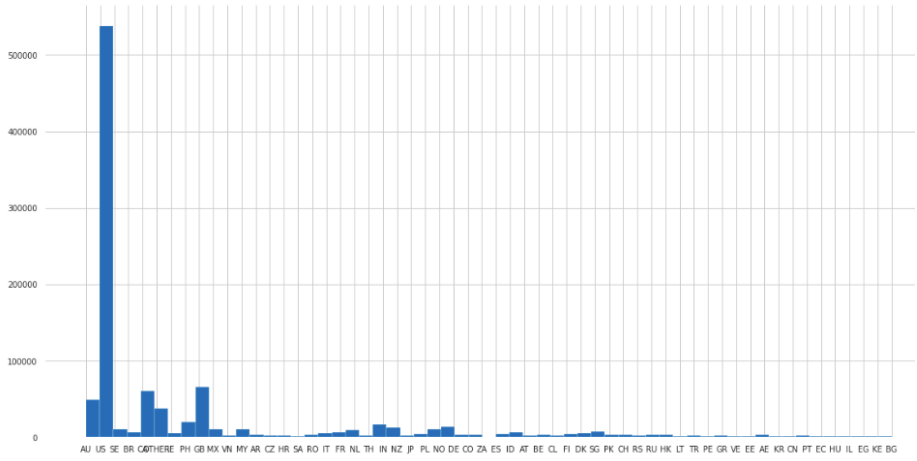

Figure 8: nationality histogram of test's participants

WWe prepared a quick comparison of answers between American and polish users, by plotting for each nation histograms of answers for every question. By analyzing this data, we got the following

- Americans seem to be much more open than Polish people
- Polish people seem to be more emotional and to feel sad more frequently than Americans
- USA people looks to be more emphatic than Polish people
- USA people looks to be a little better organised than Polish people
- Polish people seem to be more intellectual and they use their imagination more willingly

We want to mention, that there is a big difference in data sizes. There are 537755 tests from the USA and only 4586 from Poland. A smaller amount of samples from Poland may influence results. There is a lot of researchers, how country influence on personalities of it citizens. Though we didn't focus on this aspect in our researchers and modeling, we want to mention that it is an interesting part of this data set and it may be interesting to develop this in the future.

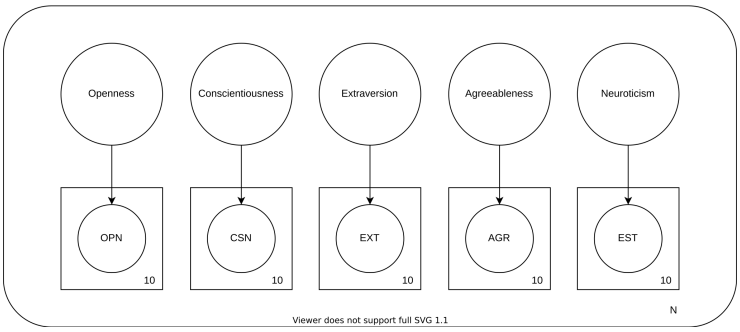

Figure 9: Bayesian net modeling questionare anwsers based on 5 traits. Note that there are no trainable parameters outside od parameters

2 Bayesian network models

2.1 Bayesian interpretation of Big Five model

2.1.1 Overview

In the Big 5 Personality traits model, we assume that, that user personality can be described by 5 dimensional, orthogonal vector, where each dimension is orthogonal to each other (at least in respect to quiz questions). This model can be quite explicitly represented by the Bayesian network.

2.1.2 Distributions

Each trait will be modeled as a Beta distribution, with α and β conditioned on country. ie:

$$Openness \sim Beta(\alpha = \alpha_{openness}, \beta = \beta_{openness}) \tag{1}$$

At last, we assume that each statement agreement comes from a binomial distribution, where n=4, and p=Trait, eg.

$$OPN_i^* \sim Binomial(n = 4, p = Openness_c) + 1 \tag{2}$$

We do it for every question, for every sample. This way, we can interpret trait as a measure of likelihood, that someone would agree with the statement (ie. we transform a 5-scale problem into a binary problem so that the trait can be interpreted as the probability of someone agreeing with the question. The +1 comes from the fact, that opinions range from 1 - 5, rather than 0 - 4. The star in the notation means a standardized answer. Some statements are reverse (ie. I don't like people in extraversion trait). By standardized statement we mean, that for such cases we flip the question into agreeing form (I like people) and change the answer as following

$$OPN_i^* = 6 - OPN_i \tag{3}$$

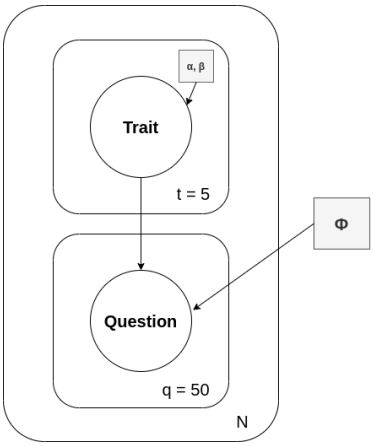

Figure 10: Bayesian net modeling questionare anwsers based on 5 traits that will be created based on data.

This model will serve as comparision for a more general model presented in next subsection.

2.2 Bayesian network model with generalized categorical probabilities

2.2.1 Overview

In reference Big Five model we assumed that anwser were sampled from repeated Bernoulli process. This assumption allowed for clear probabilistical interpretation of traits. Second assumption was that each trait only affets staments from the group that was desiged to extract this trait. In generalized model (later referenced as proposed model) we remove these assumptions.

2.2.2 Distributions

Trait, similary to first model, will be drawn from Beta distribution.

$$Trait_i \sim Beta(\alpha = \alpha_i, \beta = \beta_i) \tag{4}$$

The difference between this model and fixed model presented before, is that Questions will be drawn from categorical distribution. Probs of that distribution will be a deterministic function of Traits

$$Question_i \sim Categorical(f(Traits; \Phi)) + 1 \tag{5}$$

Where $f(Traits; \Phi)$ will be a general function parametrized by Φ. In our project, we represented $f(Traits; \Phi)$ as neural network with one hidden layer, where Φ are the weights of the neural network. As previously, map 0-4 range into desired 1-to-5 range by adding one to sampled values.

3 Dirichlet Process Mixture Models

In purpose to find main types of personalities in our Big five psychological test data set we used Dirichlet Process Mixture Models, that solves this clustering problem. Dirichlet Process Mixture Models is example of Bayesian non parametric model.

Bayesian non parametric models are models where the number of parameters grow freely with the amount of data provided. Big advantage of DPMM, in comparison to other clustering models, is that there is no need to pass desired number of cluster — model determinate it itself. This is very useful in case of our task, where we don't know how many different personalities can hide in data set.

3.1 The Dirichlet Process

Dirichlet's processes are a family of probability distributions over discrete probability distributions. Dirichlet's process (DP) is specified by some base probability distribution $G_0 : \Omega \to \mathbb{R}$ and a positive, real, scaling parameter commonly denoted as α. A sample G from a Dirichlet process is itself a distribution over . For any disjoint partition $\Omega_1, ..., \Omega_k$ of Ω, and any sample $G \sim DP(G_0, \alpha)$, we have:

$$(G(\Omega_1), ..., G(\Omega_k)) \sim \text{Dir}(\alpha G_0(\Omega_1), ..., \alpha G_0(\Omega_k))$$

This is taking a discrete partition of our sample space Ω and subsequently constructing a discrete distribution over it, using the base distribution G_0. Now, as we are familiar with Dirichlet Process, we can move to similar term — Chinese Restaurant Process.

3.2 The Chinese Restaurant Process

Imagine a restaurant with infinite tables, that accepts customers one at a time. The n-th customer chooses their seat according to the following probabilities:

- with probability $\frac{n_t}{\alpha+n|1}$, sit at table t, where nt is the number of people at table t
- with probability $\frac{\alpha}{\alpha+n|1}$, sit at an empty table

If we associate each table with draw from distribution G_0 and unnormalized probability mass n_t to that draw, the resulting distribution is equivalent to draw from Dirichlet process. We can use this process to define our non parametric mixture model. For each row of data (customer), we will assign it to a cluster (table), given by categorical distribution parameters, which were drown from categorical conjugate — Dirichlet distribution (G_0). This process is overall describe of what we are doing. To be more particular, in implementation is used Stick-Breaking Method.

3.3 The Stick-Breaking Method

This process, used for clasterisation with DPMM, proceeds as follows:

- draw $\beta_i \sim \text{Beta}(1, \alpha)$ for $i \in \mathbb{N}$
- draw parameters for i-th distribution $\theta_i \sim G_0$ for $i \in \mathbb{N}$
- construct the mixture weights π by taking $\pi_i(\beta_{1:\infty}) = \beta_i \prod_{j<i}(1 - \beta_j)$
- for each observation $n \in \{1, ..., N\}$, draw "table" $z_n \sim \pi(\beta_{1:\infty})$, and then draw $x_n \sim f(\theta_{z_n})$

$$z_n \sim \pi(\beta_{1:\infty}),\ \theta_{z_n} \sim Dirichlet(\tau),\ \theta_{z_n} = [p_{grade1}, p_{grade_2}, p_{grade_3}, p_{grade_4}, p_{grade_5}]^{50},$$
$$x_n \sim Categorical(\theta_{z_n}),\ x_n \in \{grade_1, grade_2, grade_3, grade_4, grade_5\}^{50}$$

In theory, there may be infinite amount of tables and corresponding to them weights/parameters, but in implementation we guess value T, which constraints upper limit of possible clusters. In our task, we want to find main types of personalities. In particular, we are interested in finding parameters of each categorical distributions, which are creating these clusters (personalities). Parameters of each categorical distribution are five probabilities of every possible answer for question (1,2,3,4,5) for all 50 questions from the test. As prior of categorical distribution parameters, we use it's conjugate — Dirichlet distribution.

4 Experiment results

4.1 Simple bayesian interpretation

4.2 Big Five model as generator

Most obvious thing is to treat the model as a generative one. Then we can measure how probable is that the data came from the posterior. To quantify this measure, we can use a negative log of probability (to prevent underflow from product of small numbers). We split the data on test and training set. In a Five Trait model no training data is used for interfence, as model does not have any parameter outside of sample plate. In theory, we could fit priors on training data, however then we would assume knowledge that original model does not specify. On testing set, the model got value of

$$-log(\mathcal{L}(x)) = 34379.34$$

Its hard to meaningfully interpret this value, however it will be used as a benchmark for second model.

4.2.1 Reference model as classificator

The beauty of fully probabilistical models is that we can easyli use the model for classification of missing data. To setup experiment, we randomly and uniformly removed some percent of data from test dataset. Then we conditioned model on the data we observed, and predict values of unobserved variables. To quantify performanced, we calulcated f1 scores on most probable missing values predicted by the model.

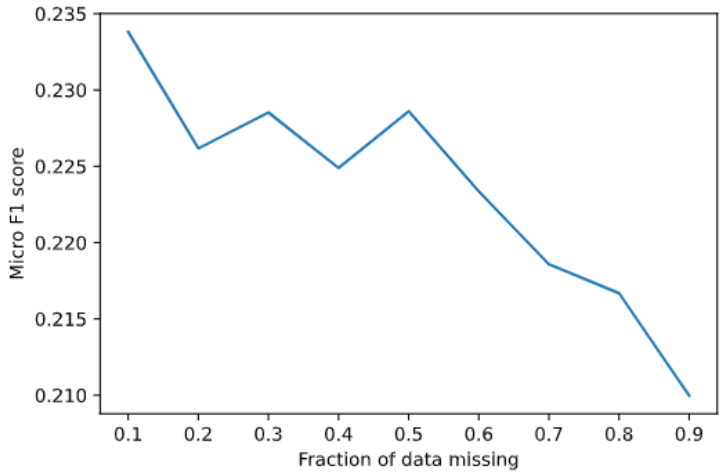

Figure 11: Micro f1 score of classification based on how much data was missing.

4.2.2 Our proposed model as generator

In this model, we can no longer (at least apriory) assing meaning to traits. Model should learn such transformation from traits to categorical probs, that would end up in minimizing ELBO (ie. describe observed data the best it can). Actually those paramters would also end up describing prior belief (figure 12).

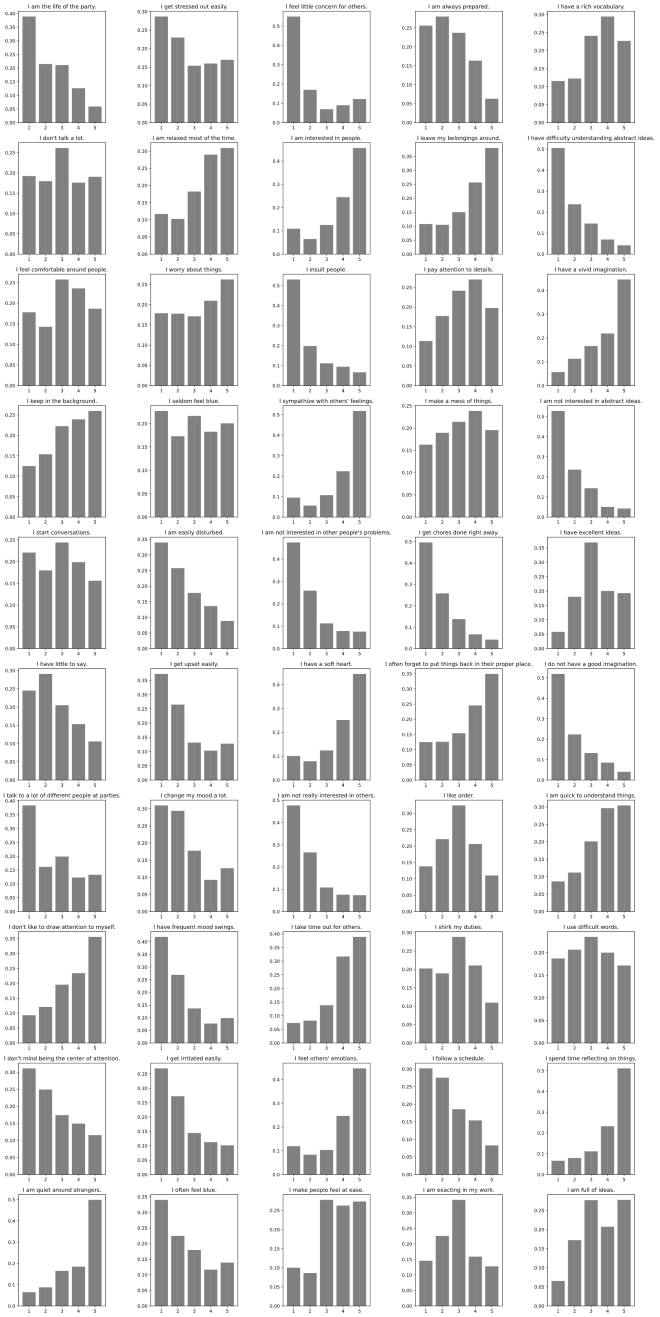

Figure 12: Estimated probability distributions of predicted anwsers without any conditioning. Refer to the notebook for high-resolution version.

To discribe quality of generative model we use a negative log prob metric (however prior is uniform, so defacto we are calulating log likelihood).

$$-log(\mathcal{L}(x;\Phi)) = 27318.283$$

That is a substantial improvement (over 25% in base-e logarithmic scale) over naive reference model. The cost of that improvement is larger complexity and lower interpretability. However, as we will see in folowing subsection, intepretability is still quite high (at least on intuitive level). It turns out that those latens have quite clear interpretation

4.2.3 Interpreting semantical meaning of the latent variables

In this model we can't just use the structure of the model do infer the meaning of the variables (like we did in 5 Traits Model). To learn about the meaning of those latents, we conditioned on them having their maximum value, and checked what distributions changed the most. To quantify the maning "changed the most" we used a KL divergence as a measure of how much posterior diverged from the prior. Then, to infer the meaning, we looked at 5 questions which diverged the most on observing a single trait at the time (and only this trait). Results (presented on figure 13) are astonishingly easy to intepretet.

We cleary see that most diverged questions for each of the traits share a similar semantical meaning. We propose following interpretations:

- Shyness
- Apathy
- Emotional instability
- Professorism
- Leadership

Most of the proposed traits are reasonably self-explanatory. However, for the lack of the better word, we did introduce a neologism "Professorism". This is trait that we cleary recognaize, that is prevelant in the academic and artistic environment. It describes peoples who read and think a lot, however easily loose touch with reality.

4.2.4 Proposed model as classifier

We setup experiment in a similar way we did in reference mode. Portion of data was removed. We see that micro f1 score is nearly twice as as large as the one in reference model (figure 14).

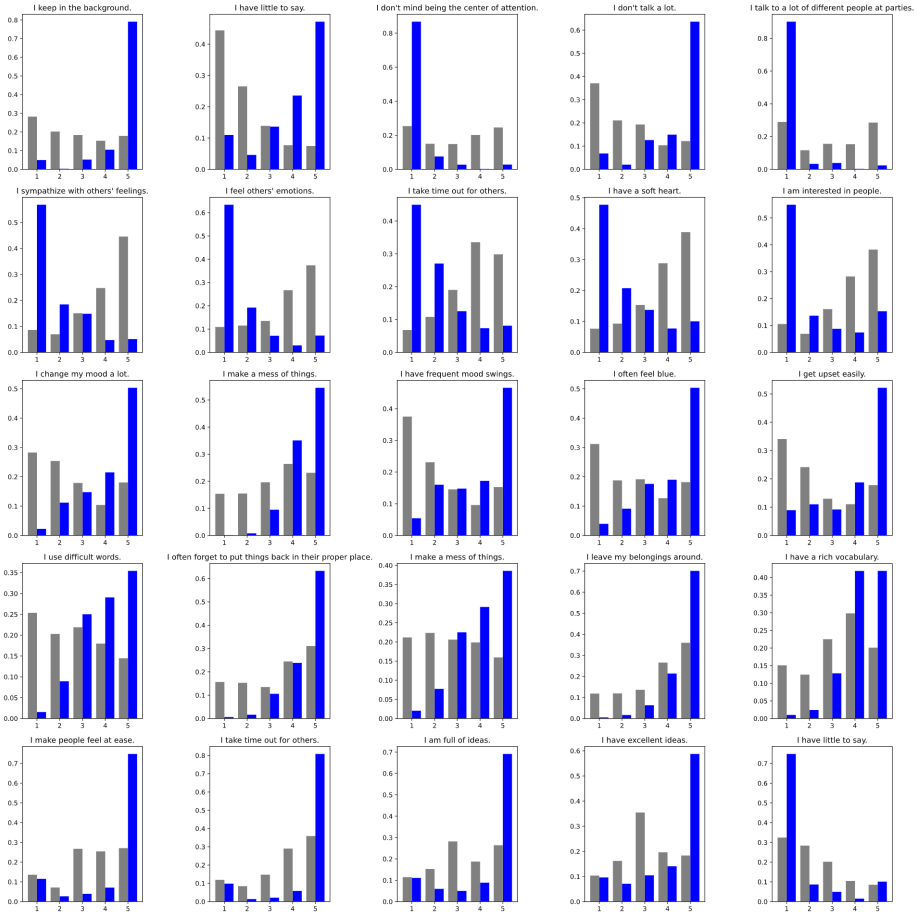

Figure 13: Comparison of posterior (blue) and prior (gray) probability distributions. Each row contains 5 questions with highest divergence from the prior.

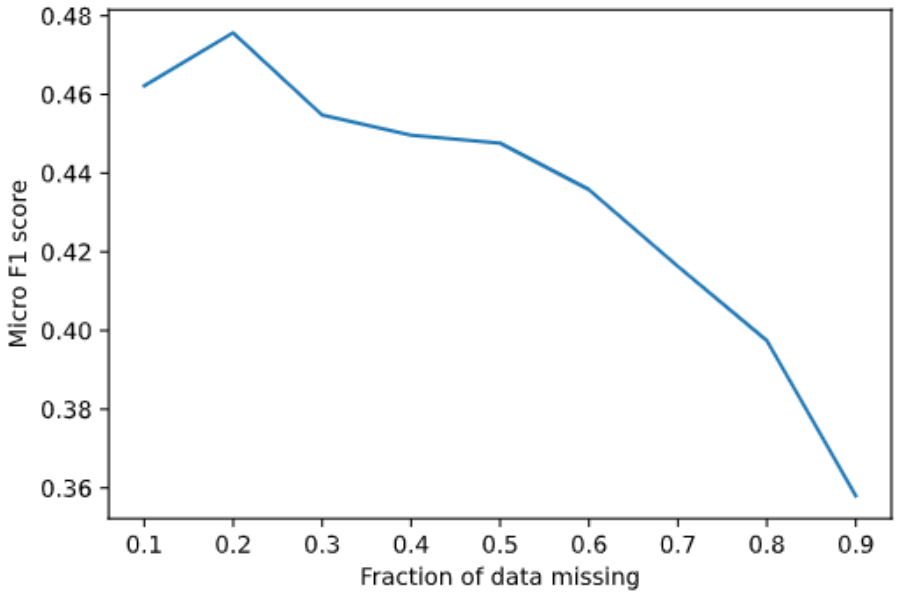

Figure 14: Comparison of posterior (blue) and prior (gray) probability distributions. Each row contains 5 questions with highest divergence from the prior.

4.3 Dirichlet Process Mixture Models

We used DPMM for clusterisation task. Our goal was to find some main distributions, which would produce different personalities that are present in data set. As described in section "Models", there is necessity to set parameter T - which constraints cluster's number, and α - which controls if single observation should join to the existing cluster, or if it should create a new one. Moreover, it would be good to choose good learning rate.

To evaluate models quality, we used ELBO and following measures:

- silhouette index - maximized measure. It contrasts the average distance of elements in individual clusters with the average distance of elements in other clusters.For the average distance of objects inside the x_i cluster and the average distance of objects to other clusters y_i

$$SC = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)/max(x_i, y_i)$$

- davies bouldin index -minimized measure. It checks the relationship between the relative distance between clusters and the distance of elements within clusters. The measure is given by the formula for n clusters, the expected value c_x and for the average distance σ_x of elements x to the

expected value c_x:

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{j\neq i}(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)})$$

. We also prepared visualizations, which compares expected values and probabilities of each cluster.

Now we are going to describe experiments we conducted in order to choose these hyperparameters. Firstly, we tried to find out optimal learning rate. Below there are results of experiment, where we evaluated models with different pair of parameters (learning rate, T). We checked all combinations for learning rate $\in \{0.1, 0.01, 0.05, 0, 001, 0, 0001\}$ and $T \in \{2, 5, 10, 100\}$. Test was conducted on 10000 sample sized data set, 200 model iterations and $\alpha = 0.15$.

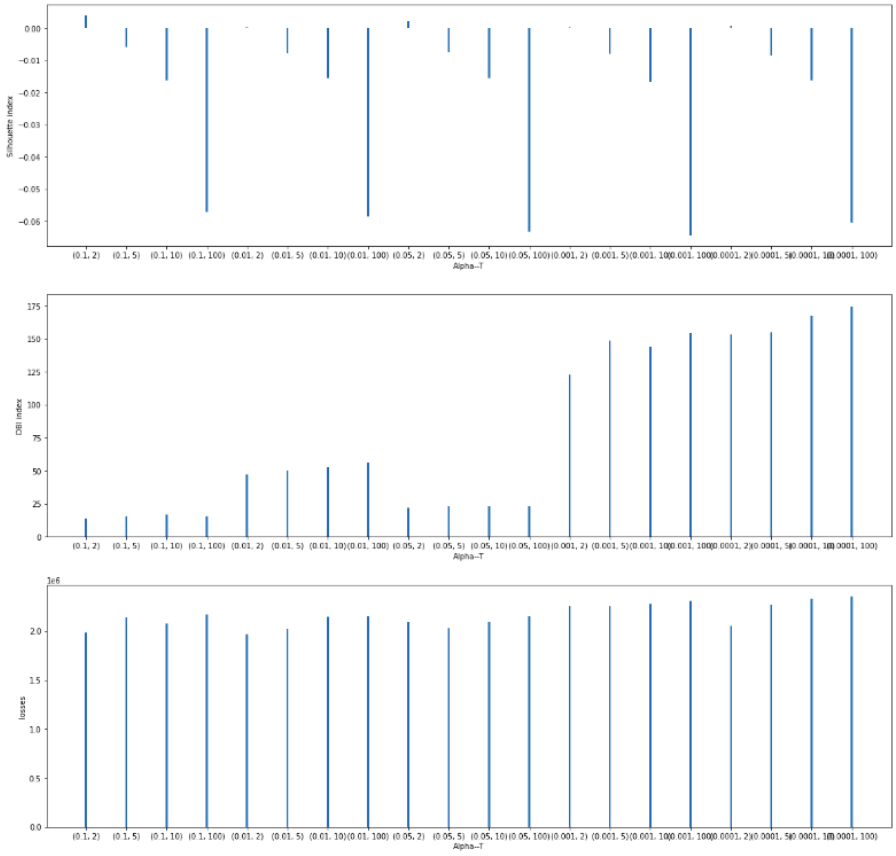

Figure 15: The upper chart plots silhouette index, middle one plots DBI, and third plots ELBO. In X axis there is pair (learning rate, parameter T)

After seeing these results, we decided to operate on learning rate 0.05 during

further experiments. Next step was to evaluate models according to pair of parameters (α, T), $\alpha \in \{0.001, 0.1, 10, 100\}$ and $T \in \{2, 50, 100\}$. This experiment was conducted on 1000 sample sized data set, 200 model iterations and learning rate = 0.05.

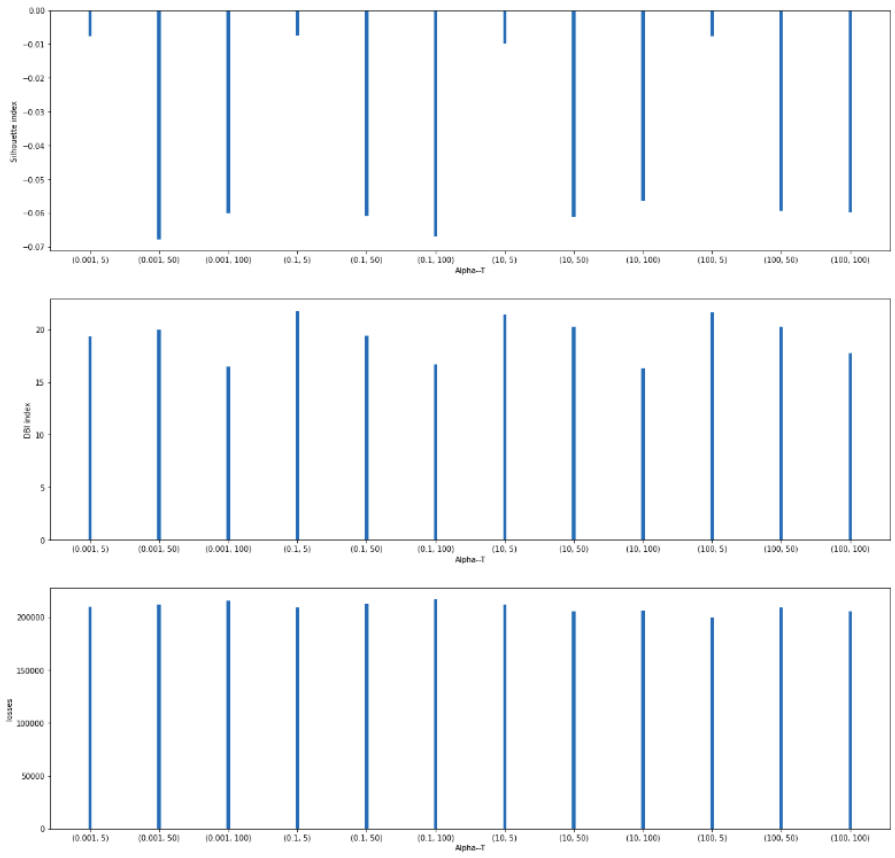

Figure 16: The upper chart plots silhouette index, middle one plots DBI, and third plots ELBO. In X axis there is pair (alpha, parameter T)

We can see, that silhouette index is better for 5 clusters ($T = 5$), but generally models seems to not react much on parameters (alpha, parameter T) changes, what is disappointing. Moreover, silhouette index is near 0, what means, that clusters covers similar data area.

To better evaluate, how model is working, we prepared some visualizations, which shows how much categorical distributions from different clusters are similar. We will analyze model, with $T = 10, \alpha = 0.1$ and learning rate=0.05 with 500 iterations. There were created 10 clusters with corresponding weights: $[0.1036, 0.1014, 0.1030, 0.0992, 0.0982, 0.0988, 0.0979, 0.0975, 0.0983, 0.1019]$. Below there is presented 10 bar plots, where each plot is corresponding to cluster.

On each plot we can see expected grade value for every 50 questions.

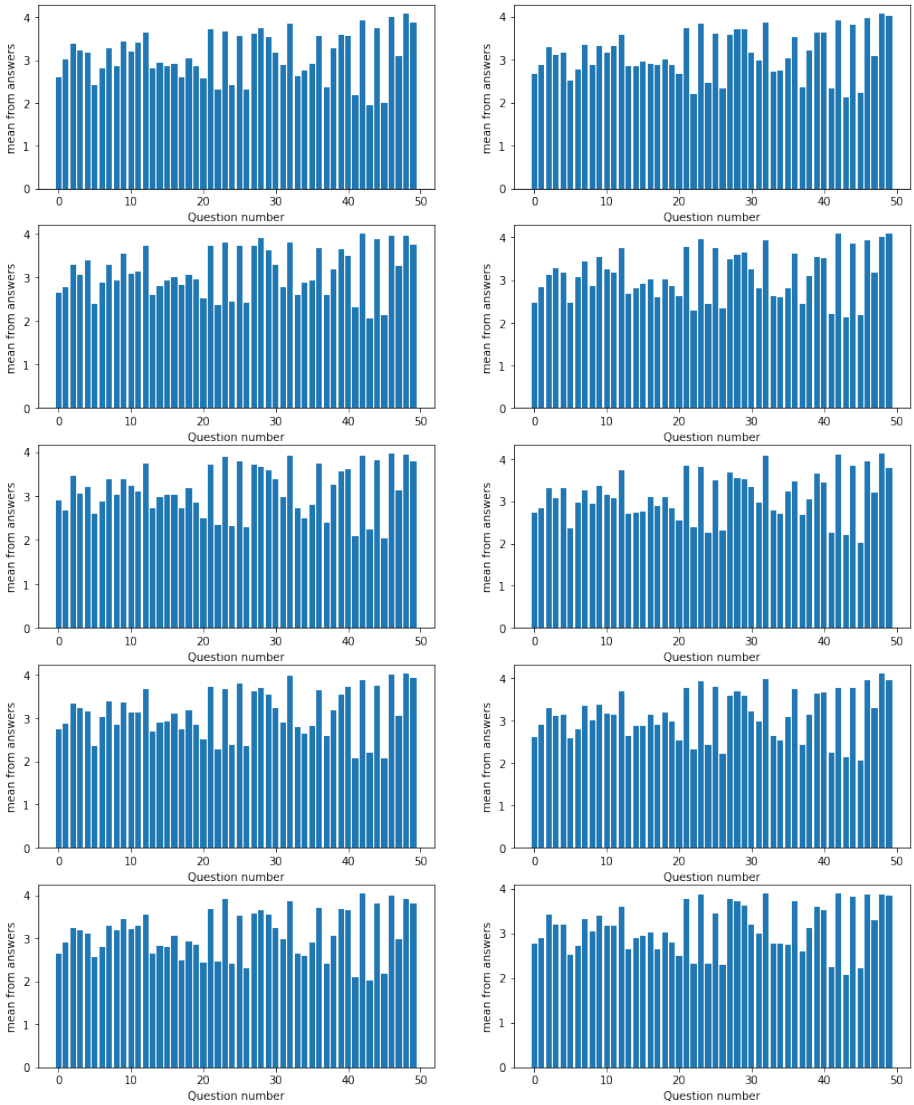

Figure 17: Expected grade values for each questions, 10 plots - 10 clusters

As we can see, these plots are quite similar, what isn't desirable. We wish clusters would be distanced to each other. Anyway, even if expected values are similar, there is chance that probabilities in categorical distributions are different, so it is worth to check it. Below there are plots, that compares probabilities, for each question, between 5 clusters, that have been randomly chosen from entire set of 10 clusters.

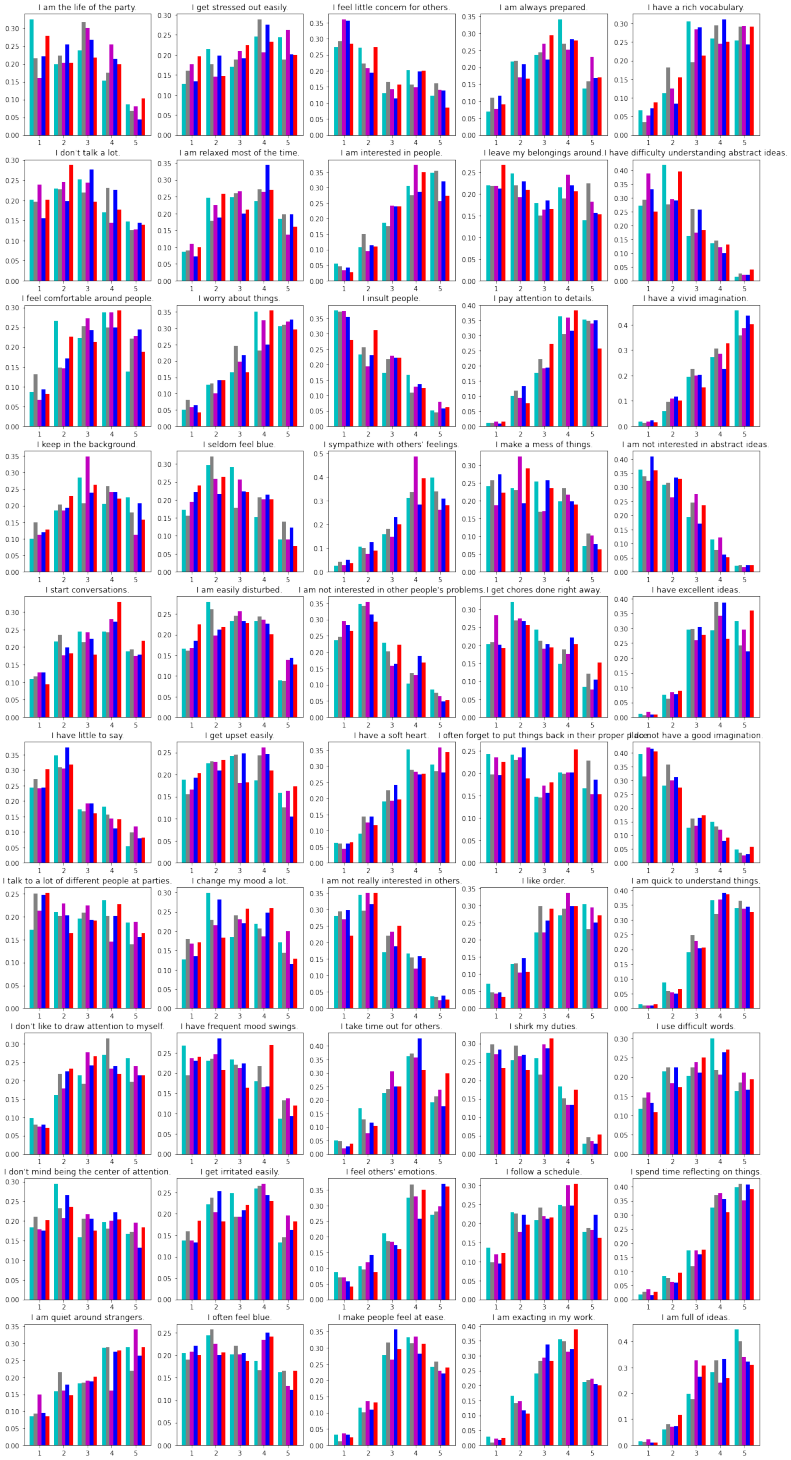

Figure 18: Probability distribution comparsion between 5 clusters

We can see, that there are certain differences in distributions, but still all plots are similar. This may be caused by the fact, that clusters must group samples in 50 dimensions (50 questions for one test), therefore one cluster has mixed 50 categorical distributions with 5 probabilities each.

Below there is one more plot, that shows how ELBO changes during next model iterations. We can see here, that after 100 iterations ELBO stops rapidly decrease (so 200 iterations number, assumed in previous experiments, was safety option).

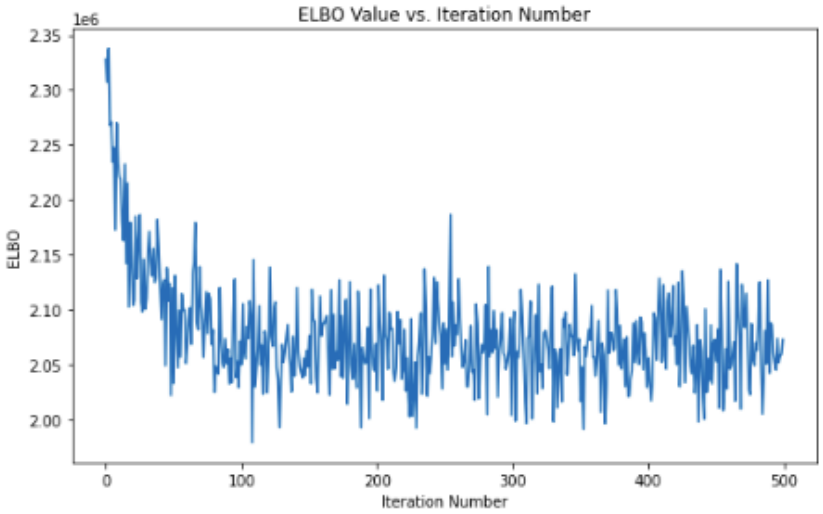

Figure 19: ELBO value vs iteration number

Moreover, silhouette score in this experiment is -0.019 (what's mean, that clusters are very close to each other), and davies bouldin score is 23.111, (where score 0 is the best for this index).

In conclusion, although DPMM implementation was successful, the resulting clusters aren't satisfying, and probably using another clustering model would be good idea. However, this data set is complex in a way that it has 50 dimension, which makes clustering a hard problem.

5 Summary

After taking several approaches into this data set, we think that Bayesian Network with learnable trait-to-questions-probs transfer function was the most successful. Visualizations, based on largest KL-divergance gave us a resonable suggestions of new traits that could make a base for personality tests. Also, thanks to probabilistic approach, we were able to create a model that can infere knowlede for just partial information. This way we can predict missing user statements (and probability distributions of those anwsers), just based on partial informatin, real-time (in the meaning of online inference, not CPU time).

Both the likelihood and f1 scores of proposed model were substantionaly better than nonparametric, naive Five Traits model.

Dimensionality and nolinearity made clustering problematic and although we thought that Dirichlet Process Mixture Model will give better results, we think that we successfully analyzed this data set. We were not sucessfull in finding a disting clusters of personality.

6 Continuation

There are some aspects, where our work could be extended and developed, for example with usage of data about nationality to enhance the model. Morover the personality triats we discovered are particulary intresting in values, that differ from standard Big Five Models.