Bayesian Unsupervised Clustering Method For Uncovering Latent Personality Types

Boxuan Li, Nianli Peng, Danny Luo

4/24/2021

1 Introduction

The Five Factor Model (FFM) of personality is a model for personality assessment that has been widely studied and applied in the field of Psychology. [1] It proposes 5 domains across which one's personality could be characterized. They are Openness to Experience, Conscientiousness, Extraversion, Agreeableness and Neuroticism (or in abbreviation, OCEAN) respectively.

While FFM presents a viable framework to evaluate individual personality's scores on those five traits, it does not identify any personality type by itself. A fuller ultilization of FFM data usually means analyzing in depths the interaction between each dimension or moving a step further in classifying individuals into homogenous personality profiles that could be interpretable under FFM.[2] Identifying those latent personality types will be of trememdous psychometric values. It will not only reveal correlations between each dimension of personality traits, but will also present us a fuller picture of compositions of human personalities. An ideal latent personality classification would also yield a simple and univariate measure of individual personality, that could be used in causal inference and prediction widely in the field of psychology and behavioral science.

Recent literatures have attempted with various techniques to approach this clustering tasks to identify personality types from FFM, including Latent Profile Analysis, Gaussian Mixture Model combined with Factor Analysis. [2][3] This project will add on to this line of inquiry with two main objectives: 1) to develop an efficient Bayesian method to uncover latent personality in the Big Five dataset and 2) to compare cluster findings with existing literature on personality types.

We propose an Bayesian unsupervised clustering algorithm that leverages a two-fold modeling structure:

- A non-parametric Dirichlet process Gaussian mixture model to decide the optimal number of clusters and their respective subpopulation parameters using a small portion of data
- Feed the above result as a prior into Gaussian mixture model with fixed cluster size, utilizing a greater proportion of the dataset

We adopt this two phase modeling due to expensive computational cost given the gigantic dataset and the costly nature of Dirichlet process model. The final output would yield a clustering of all individuals into different latent personalities type that is highly interpretable using FFM framework.

2 Data

The dataset contains $1,015,342$ questionnaire answers collected through an interactive online personality test by Open Psychometrics from 2016 to 2018. The personality test was constructed with the "Big-Five Factor Markers" from the International Personality Item Pool, developed by Goldberg (1992).[4][5] It consists of

fifty items that the respondent must rate on how true they are about him/her on a five point scale from "Very Inaccurate", "Moderately Inaccurate", "Neither Inaccurate nor Accurate", "Moderately Accurate", and "Very Accurate". Responses to this test was recorded anonymously. (For a detailed list of individual questions, see Appendix 5.5)

After eliminating 89150 missing values, we have a total of 1013558 valid observations. We see that the vast majority of the participants are from the U.S. (See Appendix @ref(EDA country)) This means that the cross-cultural generalizability of our results might be limited.

Among the 50 items in the survey, some questions are positively-keyed (e.g. "I am the life of the party") while some are negatively-keyed (e.g. "I don't talk a lot"). For + keyed items, the response "Very Inaccurate" is assigned a value of 1, "Moderately Inaccurate" a value of 2, "Neither Inaccurate nor Accurate" a 3, "Moderately Accurate" a 4, and "Very Accurate" a value of 5. For − keyed items, the response "Very Inaccurate" is assigned a value of 5, "Moderately Inaccurate" a value of 4, "Neither Inaccurate nor Accurate" a 3, "Moderately Accurate" a 2, and "Very Accurate" a value of 1.

Once numbers are assigned for all of the items in the scale, we will sum all the values to obtain a total scale score for each of the five personality traits. The distribution of "Extroversion"', "Neuroticism", and "Conscientiousness" looks pretty symmetric, but that of "Agreeableness" and "Openness" looks left-skewed. Since we will be approximating the distribution of trait scores as normal distributions, we should proceed with caution when analyzing these two traits.

3 Model

3.1 Nonparametric Learning for cluster size K

We assume the score vector for each individual in the survey comes from a mixture Gaussian distribution with unknown number of components K in the mixture. The normal assumption could be largely justified by the symmetric bell shape distribution of 3 dimensions in the personality data as shown in Appendix 5.2.

To find the unknown number of components or cluster number K, we adopt a Dirichlet process Gaussian mixture model(DPGMM) that is a widely used clustering tool documented in literature.[6] The key motivation to adopt this model is that Dirichlet Process Gaussian mixture is non-parametric, by which it assumes a unfixed number of clusters K. It has the following advantages: 1) it determines the number of clusters K from the data, which is conveniently flexible especially provided that we do not have a strong belief of exact number of personality types underlying this population 2) it eliminates the necessity of model selection procedures if we were to use parametric models. If a parametric model is adopted, optimal number of clusters would have to be tested via different runs of model with varied K using criterion like BIC.[3] With DPGMM, we can infer the optimal cluster number from the posterior.

The sampling model of DPGMM has the form below:

$$\begin{aligned} y_i &\sim N(y|\boldsymbol{\theta_i}), \\ \boldsymbol{\theta_i} = \{\boldsymbol{\mu_i}, \Sigma_i\} &\sim G, \\ G &\sim DP(\alpha, G_0) \end{aligned}$$

To give a brief overview, the process first draws a distribution G from Dirichlet Process DP with concentration parameter α and a base distribution of G_0. G_0 is a joint distribution of Gaussian parameters $\boldsymbol{\mu}, \Sigma$, from which we assume all Gaussian mixture parameters are drawn. The hierarchical process first draws a distribution G from the DP, where $G = \sum_{K=0}^{\infty} \pi_k \delta_{\boldsymbol{\theta_k}}$. That is, we can understand G as $K \to \infty$ random discrete probability measure, where $\delta_{\boldsymbol{\theta_k}}$ is a point mass centered on $\boldsymbol{\theta_k}$ [7]. A "stick-breaking" construction of the DP process suggests that most probability mass is concentrated on a few values so that when $\boldsymbol{\theta_i}$ is being simulated from G, it will mostly likely take on only a finite discrete values given appropriate concentration value α and those few values become our cluster parameters $\boldsymbol{\theta_i}$.

We place the following priors on the paramters α and G_0:

$$\alpha \sim Gamma(a = 2, b = 4)$$
$$G_0(\boldsymbol{\mu}, \Sigma) \sim N(\boldsymbol{\mu}|\boldsymbol{\mu_0} = 0, \Sigma_0)IW(\nu_0, \Phi_0)$$

Our prior choices are justified as follows. We chose a Gamma prior since it has a positive support that matches α. We determined the parameters $Gamma(2, 4)$ so we will get an expected value of $\alpha = 0.5$. Literature has shown that the prior expected number of clusters can be expressed using concentration parameter α as $\alpha log(N)$[8]. In our case, it evaluates as 4.6051702, which roughly matches the conclusion of a meaningful cluster size of 4 on the same personality data we used in a recent study [3]. Lastly we place a semi-conjugate Normal-InverseWishart prior on G_0 since we do not assume additional constraints on any dependency between mean and variance. This model choice is apparent since a draw from G_0 is a cluster parameter $\{\boldsymbol{\mu_i}, \Sigma_i\}$. We set the prior parameter μ_0 to 0, ν_0 to be 1 for the base distribution as the data is scaled, and $\Phi_0 = I$ to represent a non-informative prior belief.

For posterior sampling of *alpha*, we adopted the MCMC sampling scheme as described by West (1992) [9], and uses r-package $DirichletProcess$ [10] that applies this scheme. Due to the complexity of the full sampling scheme and the output specification of the package $DirichletProcess$, the full details and posteriors will not be discussed here at length.

Since the process is very computationally costly, we chose to run the model with a random sample of 10,000 individuals out of over 1,000,000 data in total to obtain an optimal cluster number and the corresponding cluster parameters, while later using less costly Gaussian mixture model with fixed K to incorporate more data. The MCMC chain included 1000 iterations. To obtain an MCMC estimator of K from the chain, we proceed as follows: - first use posterior draw of α to perform stick breaking process in getting exact number of clusters (which can be quite large, i.e. 300-400 total clusters), details of implementation could be found in the R documentation of $PosteriorClusters$ method [10] - Since we are only interested the major personality types, we choose to only retain clusters with size proportion (out of 10,000) greater than $\epsilon = 0.1$.

After removing the clusters with a few elements, We truncated the number of clusters to less than 10. We used this truncated "number of clusters" to derive a mean estimate of cluster number. (The traceplot of truncated number of clusters is in Fig 1) As it appears to converge after 400 iteratoins, we used a burnin period of 400 iterations and then calculated the mean estimator to be $rmean_estimator$. We rounded to $K = 5$. We picked the last iteration for our posterior estimate of cluster parameters $\boldsymbol{\theta_i}$ due to the complication of taking the average over cluster parameters of different size.

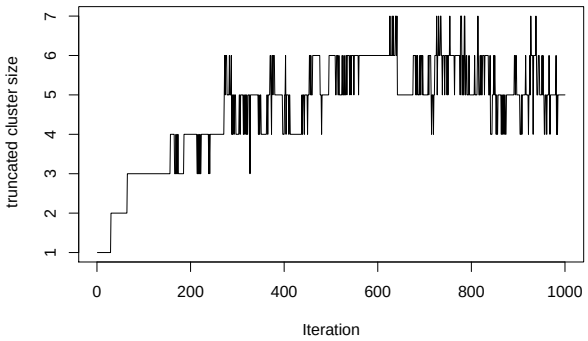

Figure 1: Traceplot of truncated cluster number

3.2 Gaussian mixture Model

We then use the optimal cluster number $K = 5$ and their corresponding parameters as the prior to feed the Gaussian mixture model with fixed cluster size $K = 5$ and run a parametric Gaussian mixture model on 200,000 samples, with the assumption that DPGMM on random 10,000 individuals is large enough to represent the population data, thus guaranteeing the value of K is reasonable. As the MCMC process with know number of clusters is less costly, we could calculate the posterior based on a much larger set of data to update our "belief" on personality types. Hence, the posterior will allow for better inference on the population. This process inherently uses the "Bayesian" philosophy of using new information to update prior belief coming from limited data.

The fixed K mixture model can be described by the following:

$$y_i|z_i = j \sim N(\boldsymbol{\mu_j}, \Sigma_j),$$
$$P(z_i = j) = p_j,$$

We assigned the following prior:

$$(\boldsymbol{\mu_j}, \Sigma_j) \sim N(\boldsymbol{\mu_{0j}}, \boldsymbol{\Phi_j}) \times Wishart(n, V) \forall j = 1, ..., K,$$
$$\boldsymbol{p} \sim Dirichlet(\boldsymbol{\alpha}),$$

where for each individual i, z_i is a latent unobserved component membership variable indicating which component in the mixture it belongs to.

We inference on the posterior cluster parameters by MCMC sampling, using the r package mixAK [11], which does exert some additional prior constraints. We plug in the values of $(\boldsymbol{\mu_j}, \Sigma_j)$ using cluster parameters we derive from DPGMM. However, α vector is required to be uniform. Thus we place a small $\boldsymbol{\alpha} = \mathbf{1}$, representing weak prior belief, allowing the model to learn the posterior weighted towards data itself. Also, the MixAK only allows a uniform parameter V, and gives the option to specify V only through specifying the hyperprior parameters of a Gamma distribution.[11] According to the documentation page of the package, matrix V is assumed to be diagonal with $\gamma_1, \gamma_2, ..., \gamma_p$ on the diagonal, and for each γ_j, $\gamma_j^{-1} \sim Gamma(g_j, h_j)$.

As we checked out the covariance matrix of the traits, we find that most of them are diagonally dominant, in which case the diagonal entries (variances) are much larger than the off-diagonal entries (covariances), mostly by one or two orders of magnitude. Hence, it is reasonable to simplify V to a diagonal matrix to save the computational cost. Next, we checked if the variances of each trait follow an inverse-gamma distribution.Since matrix V is the scale matrix parameters of the Wishart prior placed on cluster covariance matrix Σ_j and that it does not change with different subpopulations, we performed the following elicitation:

- Draw 1000 random samples of size 1000 from the dataset and calulated the 5 by 5 covariance matrix C_i for $i = 1, 2, ...1000$.
- For j =1,2,..,5, extract samples $L_j = \{C_i[j, j]\}_{i=1}^{1000}$.
- Forj =1,2,..,5, fit a gamma distribution using L_j, using the r package $fitdistrplus$. [12] [13] and using the estimation from the output to be our $\{g_j, h_j\}$

For MCMC diagnostics, we examined the traceplots for the parameters we are interested in (μ and weights) to check if the MCMC generates results representative of the real distribution.(See Appendix **??**) We observed that the autocorralaion of the parameters vary–the autocorrelation of weights are generally later than that of μ. The traceplots show that the sampling from the distribution is in a fairly consistent way and not meandering around the sample space, which indicates chain mixes well.

4 Results

Our analysis yields five clusters each with distinct and interpretable personality types, and we plotted the mean vector of the Gaussian distribution of each type. (See Figure 2). The first and most heavily weighted

(with a weight of 44.49%) cluster has the personality score very close to average in all subtraits category, thus representing an average type of personality.

The rest of the clusters follows more interesting patterns. We denote the personality type represented by the yellow color as "role model" since it has the highest score of "Agreeableness" and "Openness" and above average in the rest of three subtraits, all socially desirable traits except for Neuroticism (it is only slightly above average so it still supports our classification nicely).

In fact, three replicable personality types has been under consistent research focus and appeared widely in Literature since 2004, and they are "Resisilient","Overcontrolled",and "Undercontrolled", also known as ARC-type classification. [14][3] A documented association of those three personality types with Big Five models are summarized by Donnellan and Robins in 2010; we listed their summary in Table 1. [14]

Comparing the finding and our cluster model, high identifiability of our clusters with minor refinement could be observed. The purple, red and blue cluster in Figure 2 can be nicely identified with "Resislient","Overcontrolled" and "Undercontrolled" respectively by matching the score distribution with documented characteristics in Table 1. The match is not exact, however. As we can see in the "Undercontrolled" group, Conscientiousness is not significantly low below the average. In addition, "Extroversion" and "Conscientiouness" are only slightly above average.

The reasonable consistency of our clustering results with existing literature indicates validity of our modeling approach while minor deviations reveal the unique characteristics of this dataset at hands and contribute to the continued debate over replicability and robustness of the ARC-type classification.[14] Our results demonstrate that ARC-type classification might be only a minimal set of the general topology of personality types as we introduced two new types, "average" and "role model", a result that is also obtained by Northwestern study in 2018.[3] Those two pesonality types could be analyzed more extensively in the future and cross-validated from other datasets; and subsequent research could be to investigate the predictive validity of personal developmental outcome on those two types.

Table 1: Literature findings of Big Five trait profiles correlation with the three replicable personality types

	Resilient	Overcontrolled	UnderControlled
Extroversion	high	low	–
Agreeableness	–	–	Low
Conscientiousness	high	–	Low
Neuroticism	low	high	–
Openness	–	–	–

5 Conclusion

This paper provides an alternative clustering methodology to uncover latent personality types using Big Five personality data. Nonparametric learning using Dirichlet Process Gaussian Mixture model was used to decide the optimal cluster size while subsequent Gaussian mixture model was used to update and refine the cluster parameters. It provides a valuable methodology to conduct Bayesian unsupervised clustering on Big Five personality data, and the model strength are validated by existing literature as some consistency is observed.

Furthermore, the results validated the existing "ARC" types of personality: "UnderControlled","Overcontrolled" and "Resislient", but also showed the existance of two new types, "Role model" and "Average", which shares some consistency with another recent study.

Future work should be directed towards applying the algorithm to more cross-cultural dataset and to eliminate more modeling constraints as we currently are subject to. The classification results of this study should also be critically assessed and built upon in future research.

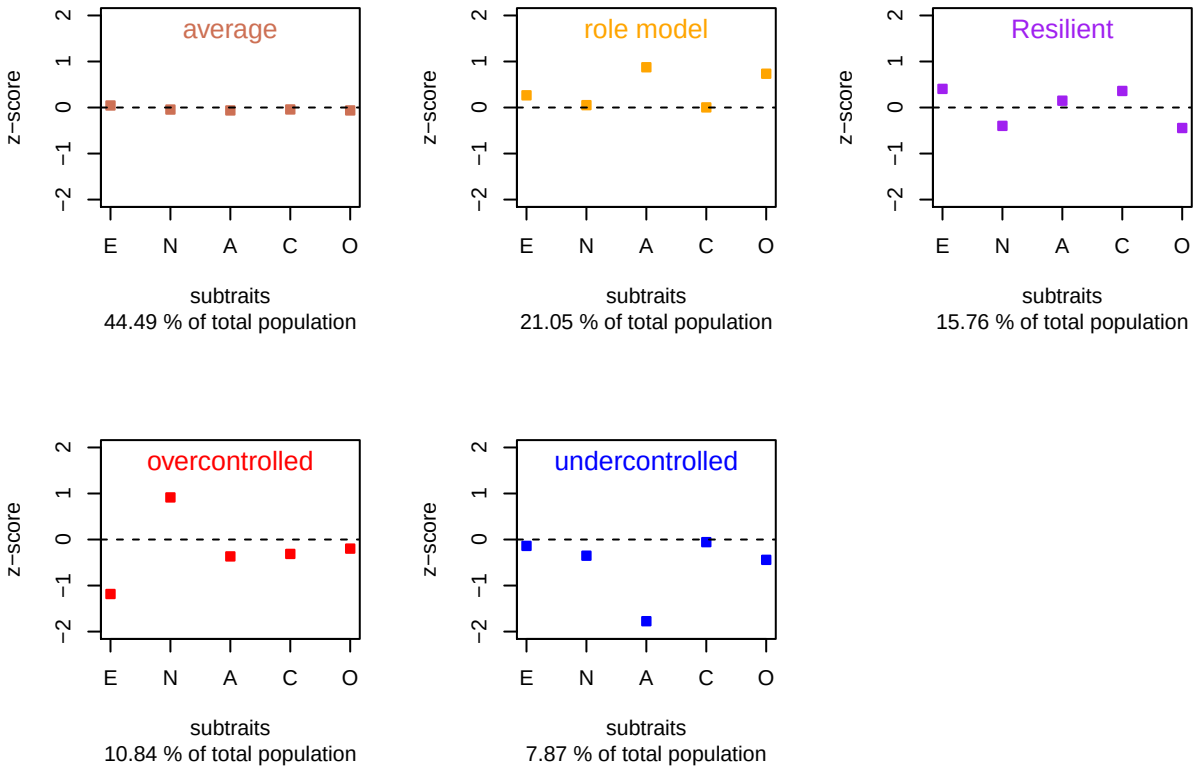

Figure 2: Mean vectors of 5 identified clusters of personality types

References

[1] N. Taylor and G. P. de Bruin, "The basic traits inventory," in *Psychological assessment in south africa: Research and applications*, Wits University Press, 2013, pp. 232–243.

[2] E. L. Merz and S. C. Roesch, "A latent profile analysis of the five factor model of personality: Modeling trait interactions," *Pers Individ Dif*, vol. 51, no. 8, pp. 915–919, Dec. 2011, doi: 10.1016/j.paid.2011.07.022.

[3] M. Gerlach, B. Farb, W. Revelle, and L. A. Nunes Amaral, "A robust data-driven approach identifies four personality types across four large data sets," *Nat Hum Behav*, vol. 2, no. 10, pp. 735–742, Oct. 2018, doi: 10.1038/s41562-018-0419-z.

[4] "Big five personality test. Open-source psychometrics project," Aug. 02, 2019. https://openpsychometrics.org/tests/IPIP-BFFM/.

[5] "Big five personality test kaggle." https://www.kaggle.com/tunguz/big-five-personality-test (accessed Apr. 24, 2021).

[6] D. Görür and C. Edward Rasmussen, "Dirichlet process gaussian mixture models: Choice of the base distribution," *J. Comput. Sci. Technol.*, vol. 25, no. 4, pp. 653–664, Jul. 2010, doi: 10.1007/s11390-010-9355-8.

[7] Y. W. Teh, "Dirichlet process," in *Encyclopedia of machine learning*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2010, pp. 280–287.

[8] Y. P. Raykov, A. Boukouvalas, M. A. Little, and others, "Simple approximate map inference for dirichlet processes mixtures," *Electronic Journal of Statistics*, vol. 10, no. 2, pp. 3548–3578, 2016.

[9] M. West, "Hyperparameter estimation in dirichlet process mixture models," *Duke University ISDS Discussion Paper\# 92-A03*, p. 6, 1992.

[10] K. M. Gordon J. Ross Dean Markwick, *Dirichletprocess: Build dirichlet process objects for bayesian modelling*. 2020.

[11] A. Komárek, *MixAK: Multivariate normal mixture models and mixtures of generalized linear mixed models including model based clustering*. 2020.

[12] R. P. [. Marie-Laure Delignette-Muller [aut] Christophe Dutang [aut], *Fitdistrplus: Help to fit of a parametric distribution to non-censored or censored data*. 2020.

[13] M. L. Delignette-Muller and C. Dutang, "fitdistrplus: An R package for fitting distributions," *Journal of Statistical Software*, vol. 64, no. 4, pp. 1–34, 2015, [Online]. Available: https://www.jstatsoft.org/v64/i04/.

[14] M. B. Donnellan and R. W. Robins, "Resilient, overcontrolled, and undercontrolled personality types: Issues and controversies," *Social and Personality Psychology Compass*, vol. 4, no. 11, pp. 1070–1083, 2010.

Appendix

5.1 EDA country

5.2 EDA Normalcy

5.3 Hyper Prior elicitation for V for package mixAK

MixAK only allows a uniform parameter V, and gives the option to specify V only through specifying the hyperprior parameters of a Gamma distribution.[11] According to the documentation page of the package, matrix V is assumed to be diagonal with $\gamma_1, \gamma_2, ..., \gamma_p$ on the diagonal, and for each γ_j, $\gamma_j^{-1} \sim Gamma(g_j, h_j)$.

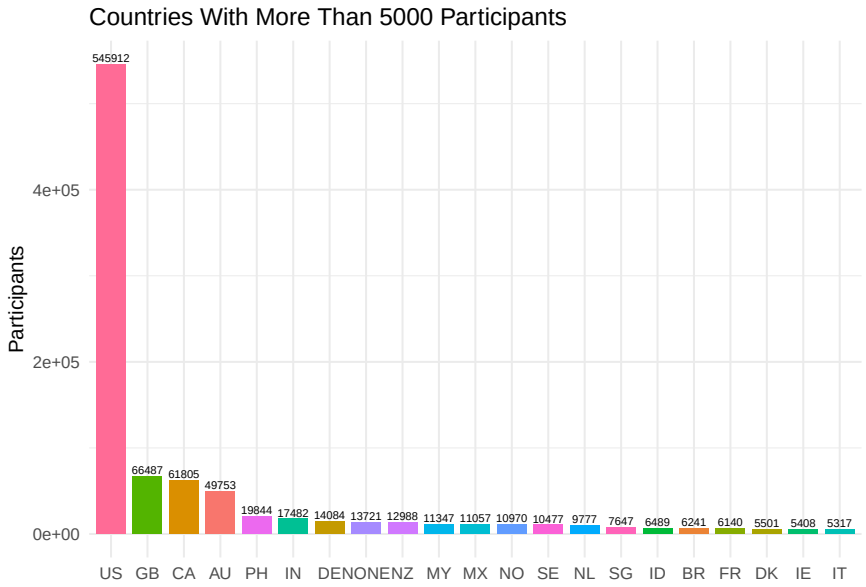

Figure 3: Number of participants in countries

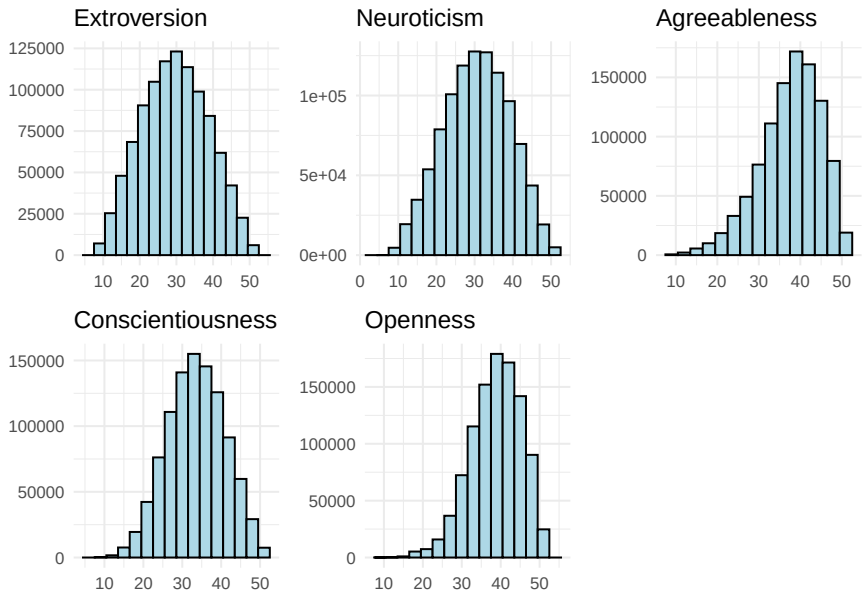

Figure 4: Normalcy check

We checked that the covariance matrix of some ramdom subset of the population are approximatley diagonal. Since matrix V is the scale matrix oarameters of the Wishart prior placed on cluster covariance matrix Σ_j and that it does not change with different groups, we performed the following elicitation:.

- Draw 1000 random samples of size 1000 from the dataset and calulated the 5 by 5 covariance matrix C_i for $i = 1, 2, ...1000$.
- For j =1,2,..,5, extract samples $L_j = \{C_i[j, j]\}_{i=1}^{1000}$.
- Forj =1,2,..,5, fit a gamma distribution using L_j, using the r package $fitdistrplus$. [12] [13] and using the estimation from the output to be our $\{g_j, h_j\}$

The fitting was highly accurate as we checked the Q-Q plot to be almost matching.

5.4 MCMC Diagnostics for cluster means and weights #{diagnostics}

5.5 Personality Test Questions / Comprehensive Data description.

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.

- 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.

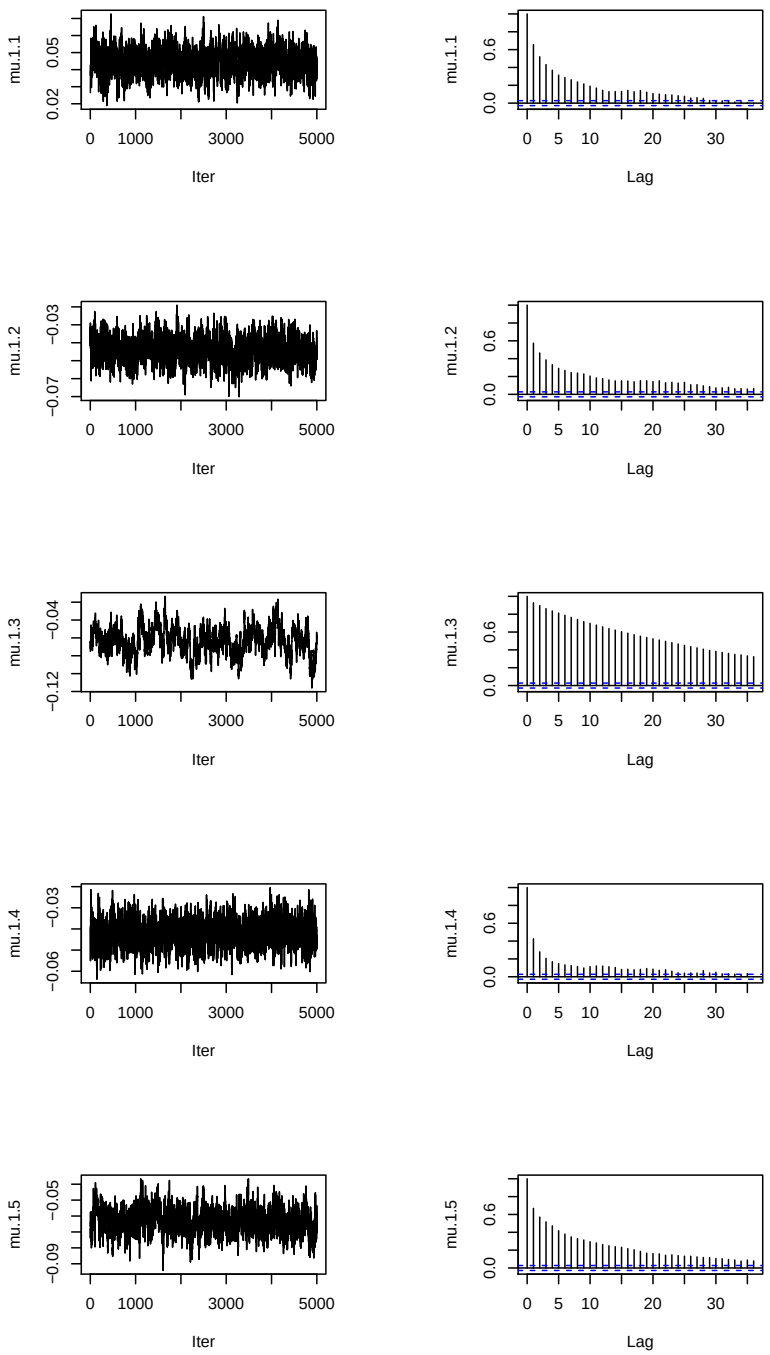

Figure 5: Traceplots and Autocorrelation plots for cluster means

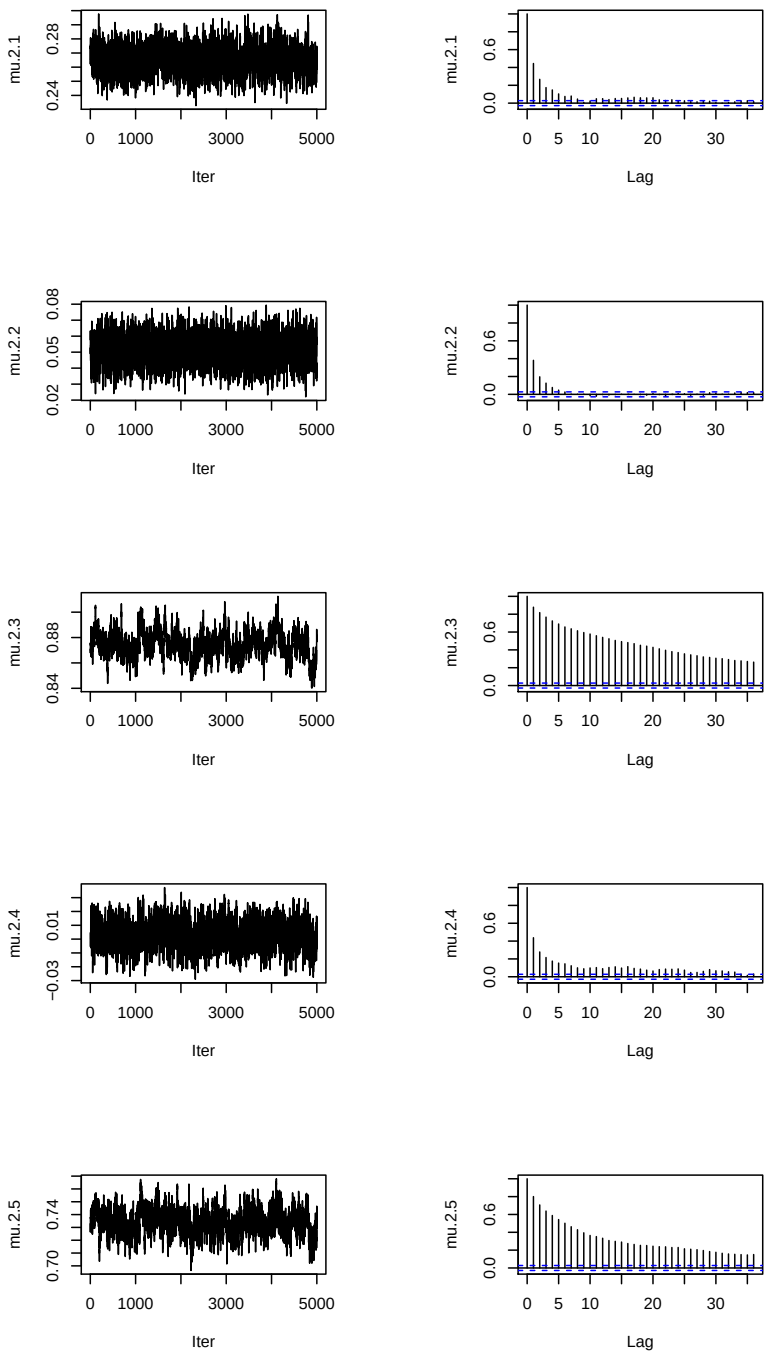

Figure 6: Traceplots and Autocorrelation plots for cluster means

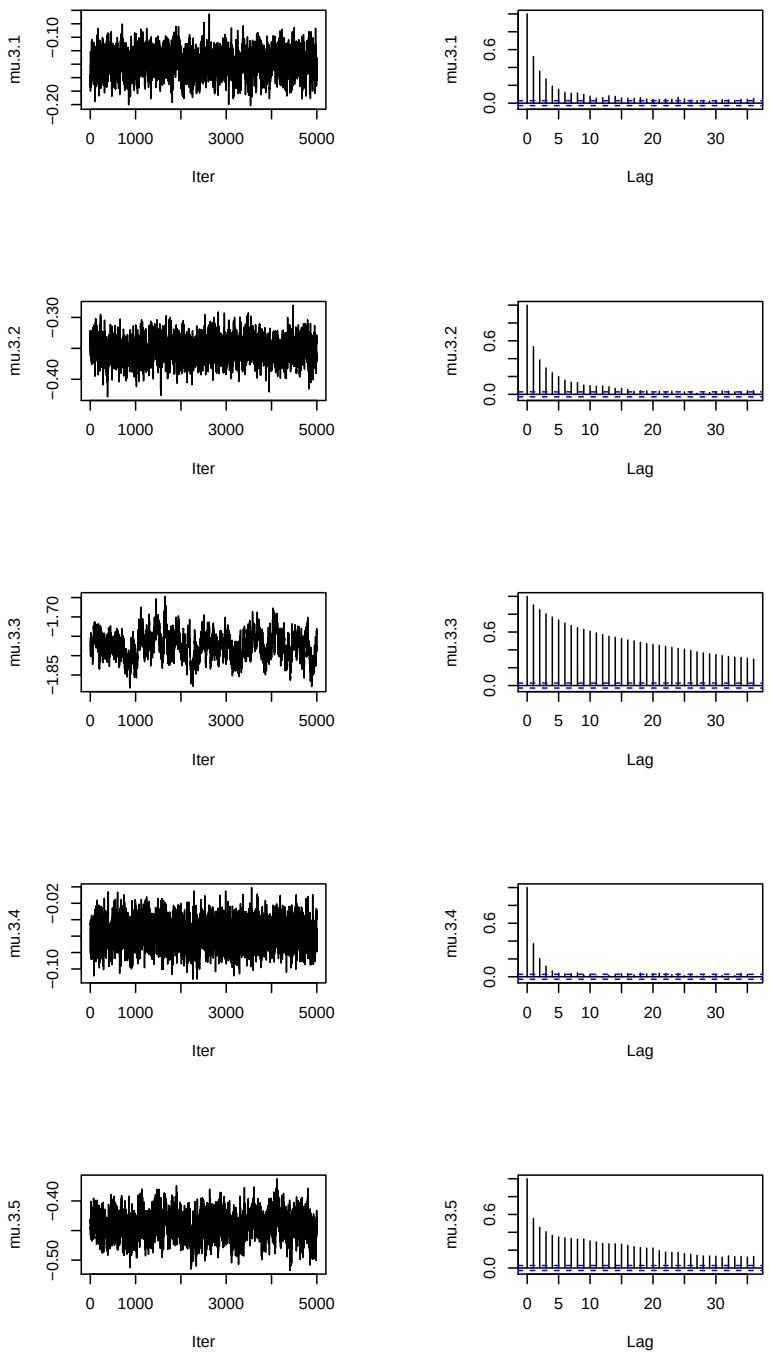

Figure 7: Traceplots and Autocorrelation plots for cluster means

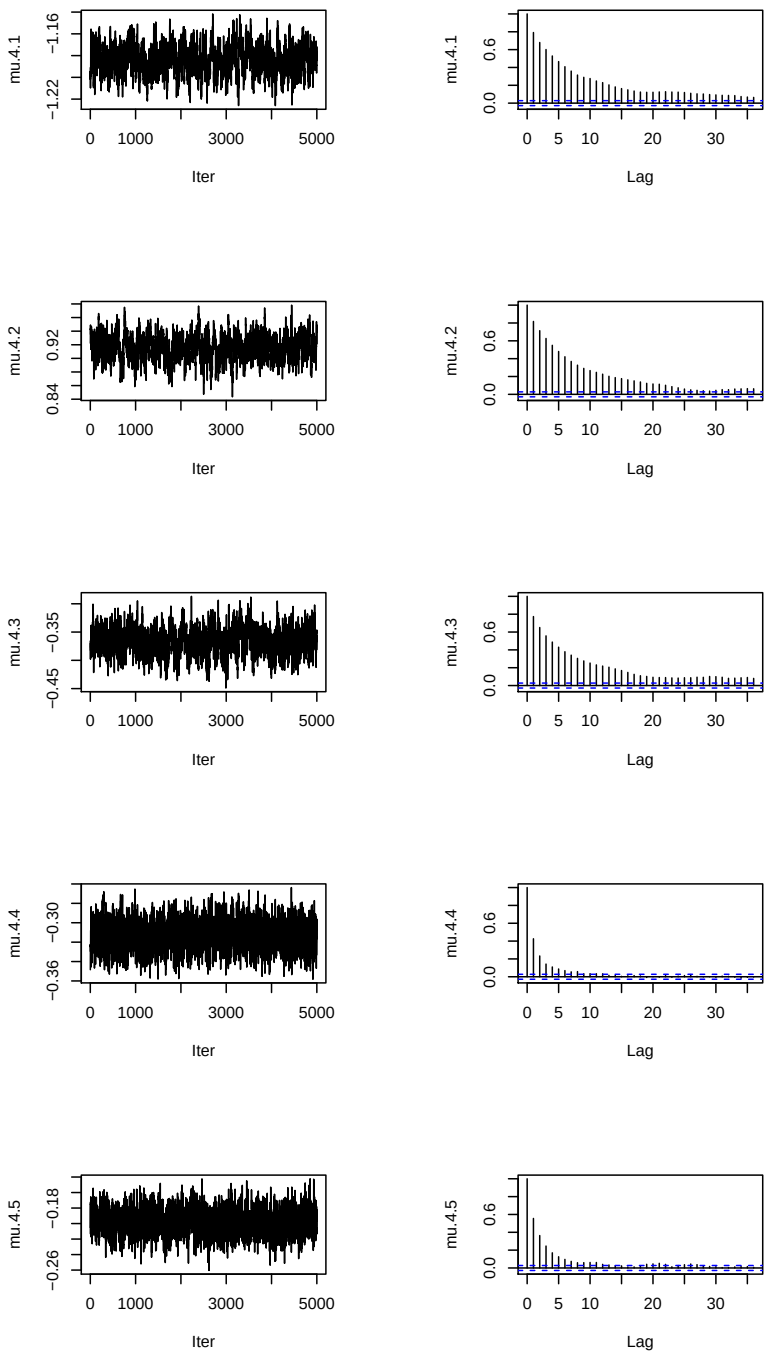

Figure 8: Traceplots and Autocorrelation plots for cluster means

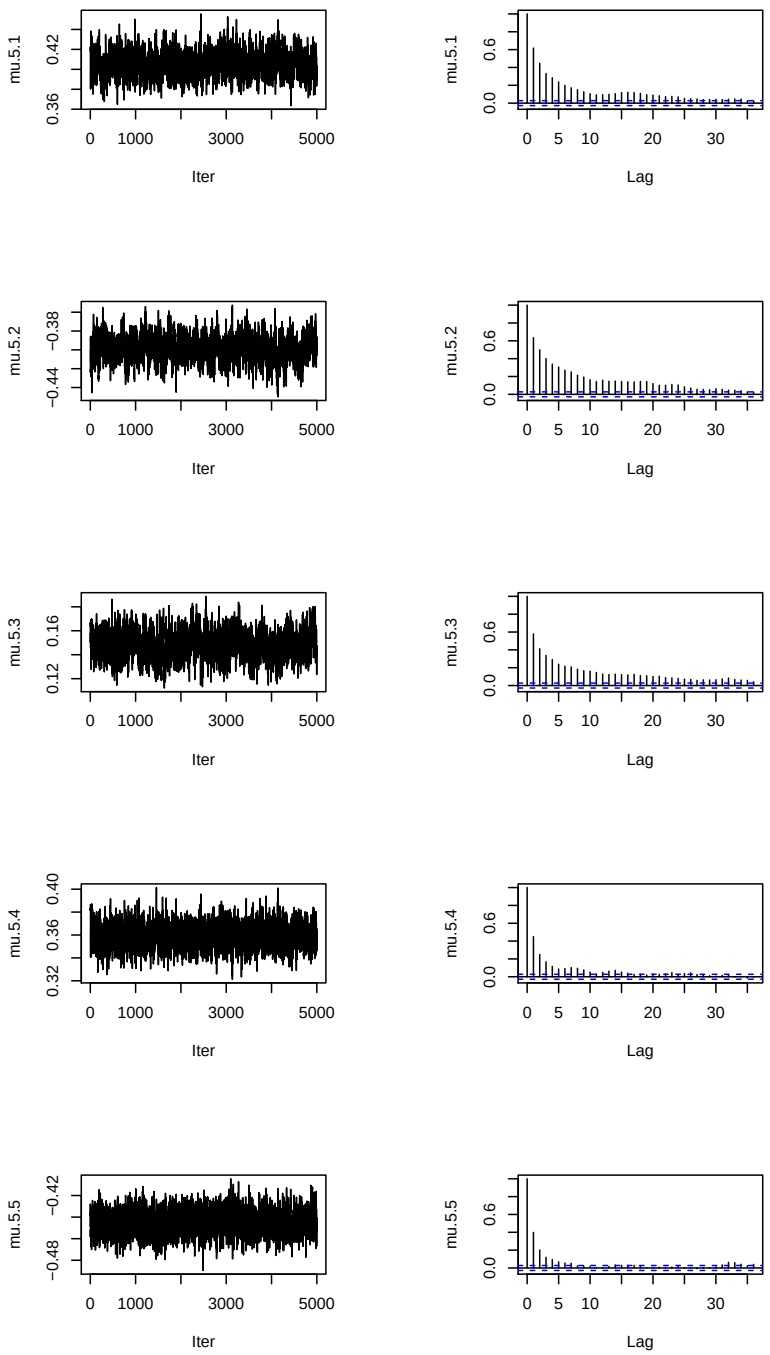

Figure 9: Traceplots and Autocorrelation plots for cluster means

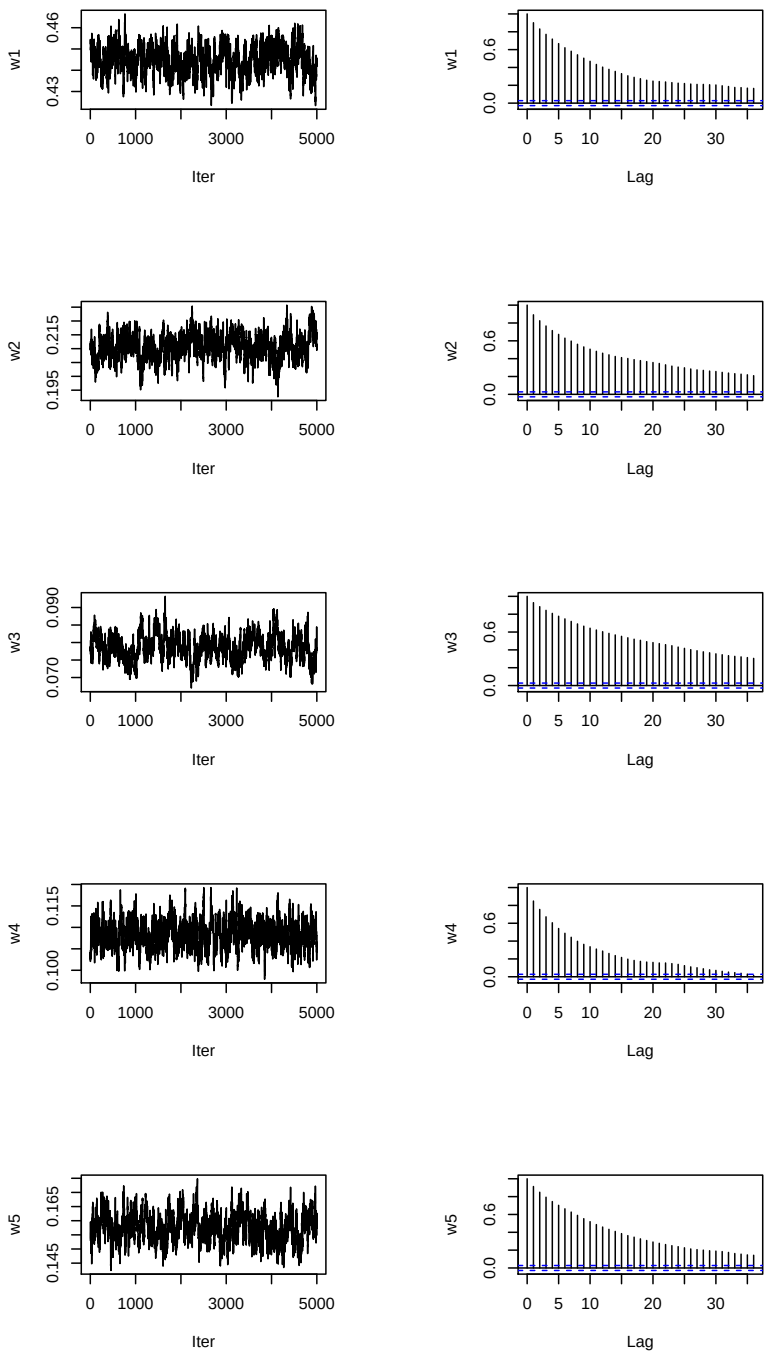

Figure 10: Traceplots and Autocorrelation plots for weights

- 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.