Bayesian Unsupervised Clustering Method For Uncovering Latent Personality Types

Boxuan Li, Nianli Peng, Danny Luo

4/24/2021

Contents

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. To fully extract the value from FFM data usually means analyzing in depths the interaction between each dimensions 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]

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

- A non-parametric Dirichlet process Gaussian mixture model to estimate size of clusters and their respective subpopulation parameters using a small portion of data
- Feed the above result as prior into Gaussian mixture model with fixed cluster size, utilizing the rest of our data

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

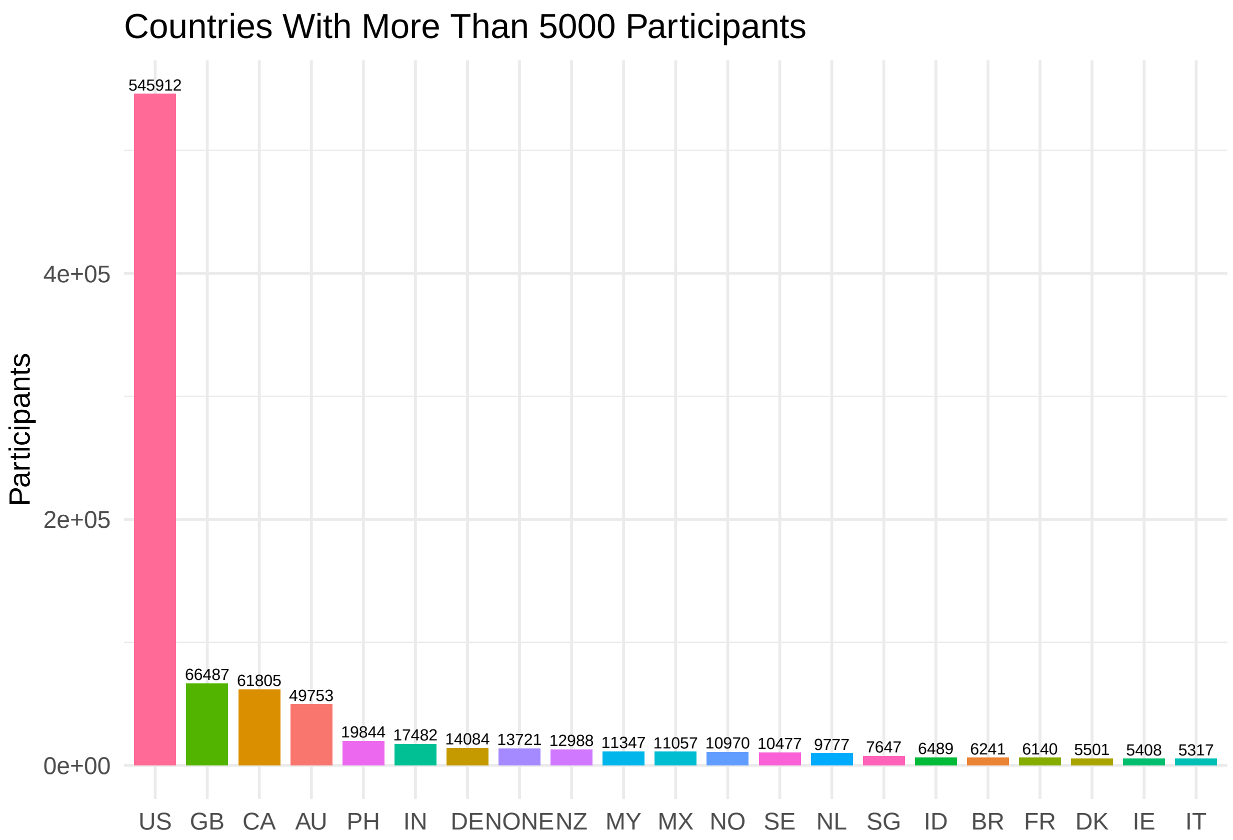
2 Data

This 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). 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. More information about each question is included in the appendix.

In this study we will analyze the data set and use a Bayesian unsupervised learning algorithm for clustering the participants.

It looks like there are 89150 missing values.

After eliminating missing values, we have 1013558 valid observations.



We see that the vast majority of the participants are from the U.S. We might be exposed to selection bias.

Among the 50 items in the survey, some are positive (e.g. "I am the life of the party") while some are negative (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 EDA.

To find the unknown number of components or cluster number K, we adopt a Dirichlet process Gaussian mixture model(DPGMM).It is a widely used clustering tool documented in literature.[4] The sampling model

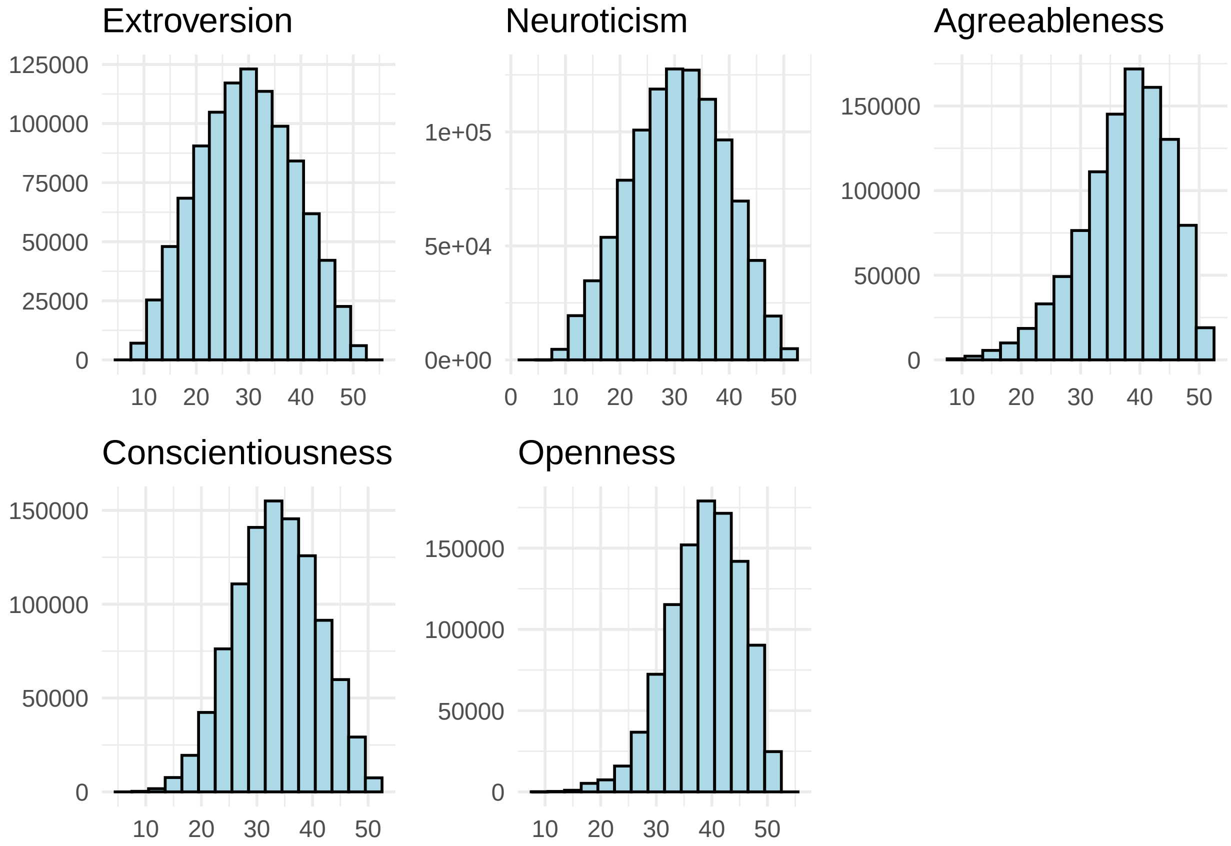


Figure 1: Normalcy check

of DPGMM has the form below:

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

The process works by first drawing a distribution G from Dirichlet Process DP with concentration paramter α and a base distirbution of G_0. G_0 is a joint distribution of Gaussian paramters $\boldsymbol{\mu}, \Sigma$, which we assume all Gaussian mixture paramters come from. The hierachial process first draw 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 $\boldsymbol{\theta_k}$.[5]. A stick-breaking property construction of the DP process suggests that most probability mass is concentrated on a few values, that is, when $\boldsymbol{\theta_i}$ is being simulated from G, it will mostly likely take on only a few discrete values given appropriate concentration value α and those few values become our cluster parameters $\boldsymbol{\theta_i}$. This process is non-parametric, providing the key advantage of nonfixed cluster size. Thus the cluster size is fully learned from the data itself. We place the following priors on the paramters α and G_0:

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

We chose a Gamma prior since the postive support matches that of α, and $Gamma(2, 4)$ gives us an expected value of $alpha = 0.5$. A small α generally yields more concentrated distribution, thus less clusters as we would expect. Literature has reported a meaingful cluster size of 4 on the same personality data we used.[3]. For the base distribution, since the data is scaled, we will set the prior paramter μ_0 to 0, ν_0 to be 1 and $\Phi_0 = I$ to represent a non-informative prior belief. Semi-conjugacy and full conditionals can be established for G_0 since we do not assume any dependency between mean and variance. For posterior sampling of $alpha$, we adopted the MCMC sampling scheme as described by West (1992). [6] This scheme is used in r-package *DirichletProcess* [7]. Due to the complexity of the full sampling scheme, the full details and posteriors will not be discussed here at length. MCMC sampling shcemes return the following clusters.

3.2 Iterative Gaussian mixture Model

We will use the posterior clusters as our prior to the Gaussian mixture model with fixed cluster size.The motivation is that since the DPGMM process is very costly, we will then "learn" the rest of the data using the fixed number of cluster size in DPGMM. This should be appropriate since we used 10k random samples that could be representative of the whole population profiles. The fixed K mixture model can be described by the following:

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

We assigned the following prior:

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

where for each individual i, z_i is a latent unobserved component membership variable indicating which component in the mixture it belongs to. We plug in the values of $(\boldsymbol{\mu_j}, \Sigma_j)$ using cluster paramters we derive from DPGMM. α vector is also derived from the weights from DPGMM.

We inference on the posterior cluster paramters by MCMC, using the r package mixAK.[8] We initiate the chain using the same paramters we got from DPGMM in the last step for faster convergence.

Since the methods error out when running larger set of data, we break down the rest of our data into chunks and iteratively ran the MCMC procedure, feeding the Maximum-A-Posterior estimation of last the MCMC chain into the prior of the next one. In that way, we are iteratively "learning" the population patterns.

4 Results

Posterior analyses from the fitted Bayesian model, and a translation of such findings into meaningful & understandable conclusions for the target audience (e.g., engineers, business managers, policy-makers, etc). See project rubric for details.

5 Conclusion

A summary of key findings and potential impacts of your project.

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 EDA.

To find the unknown number of components or cluster number K, we adopt a Dirichlet process Gaussian mixture model(DPGMM).It is a widely used clustering tool documented in literature.[4] It has the form below:

$$
\begin{aligned}
y_i &\sim N(y|\theta_{\mathbf{i}}) \\
\theta_i &\sim N()
\end{aligned}
$$

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$f\left(k\right) = \binom{n}{k} p^k \left(1-p\right)^{n-k} \tag{1}$$

: Discussion & justification of the proposed Bayesian model framework (prior and sampling model). This discussion should elicit prior information on the problem, the data sources available, and relevant project goals. Any “downstream” uses of the model (e.g., for prediction, optimization, ranking) should be discussed in detail here. See project rubric for details.

This is test for referencing equation. See equation (1)

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] 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.

[5] 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.

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

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

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

[9] M. A. Hayat, J. Wu, and Y. Cao, "Unsupervised bayesian learning for rice panicle segmentation with UAV images," *Plant Methods*, vol. 16, no. 1, p. 18, Feb. 2020, doi: 10.1186/s13007-020-00567-8.

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

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

[12] G. J. Ross and D. Markwick, "Dirichletprocess: An r package for fitting complex bayesian nonparametric models," p. 42.

Appendix

5.1 1 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.
- 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.

The time spent on each question is also recorded in milliseconds. These are the variables ending in _*E*. This was calculated by taking the time when the button for the question was clicked minus the time of the most recent other button click.

- dateload The timestamp when the survey was started.
- screenw The width the of user's screen in pixels
- screenh The height of the user's screen in pixels
- introelapse The time in seconds spent on the landing / intro page
- testelapse The time in seconds spent on the page with the survey questions
- endelapse The time in seconds spent on the finalization page (where the user was asked to indicate if they has answered accurately and their answers could be stored and used for research. Again: this dataset only includes users who answered "Yes" to this question, users were free to answer no and could still view their results either way)
- IPC The number of records from the user's IP address in the dataset. For max cleanliness, only use records where this value is 1. High values can be because of shared networks (e.g. entire universities) or multiple submissions
- country The country, determined by technical information (NOT ASKED AS A QUESTION)
- lat_appx_lots_of_err approximate latitude of user. determined by technical information, THIS IS NOT VERY ACCURATE.
- long_appx_lots_of_err approximate longitude of user