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 online by Open Psychometrics. The test uses the Big-Five Factor Markers from the International Personality Item Pool, developed by Goldberg (1992). The test consists of fifty items that the respondent must rate on how true they are about him/her on a five point scale where 1 = Disagree, 3 = Neutral and 5 = Agree. Responses to this test was recorded anonymously.

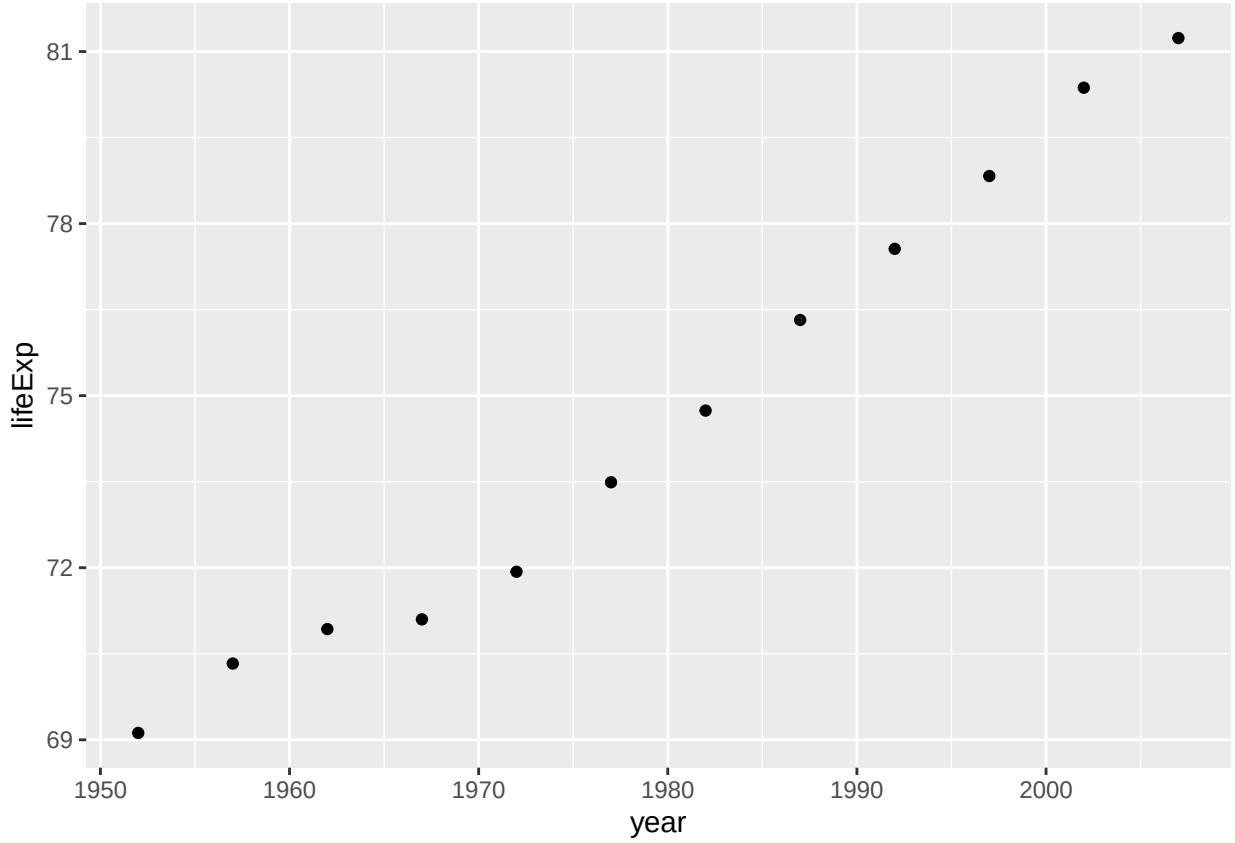

Figure 1: Life expectancy from 1952 - 2007 for Australia. Life expentancy increases steadily except from 1962 to 1969. We can safely say that our life expectancy is higher than it has ever been!

3 Model

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

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

: 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)

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.

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] 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.

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

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

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

[8] 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.

Appendix