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. 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. (See Appendix@ref(EDA country)) 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 key motivation behind our adoption of this model is that Dirichlet Process Gauassian mixture is non-parametric, in that it assumes a nonfixed number of clusters K. It has the following advantages: 1) it learns the number of clusters K from the data. It is extremely convenient especially provided that we do not have strong belief of exact number of personality types underlying this population. 2) it eliminates the necessity of the model selection procedure 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 vaired K using criterion like BIC.[3] With DPGMM, the model itself returns the optimal "posterior" size.

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 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 $\boldsymbol{\theta_i}$.

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 the postive support matches that of α, and $Gamma(2, 4)$ gives us an expected value of $alpha = 0.5$. Literature has shown that the prior expeceted number of clusters can be expressed using concentration paramter $alpah$ as follows: $\alpha log(N) = 0.5\, timeslog(10000) = 4.6$.[6]This matches the reporting of a meaingful cluster size of 4 on the same personality data we used in a recent study,[3] so it makes sense for us to set the $Gamma(2, 4)$ as prior for α. 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). [7] This scheme is used in r-package $DirichletProcess$ [8]. 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. The chain included 1000 iterations. We provided the summary statistics from the chain as follows. For each iteration, we estimated the number of clusters through 1) use posterior draw of α to perform stick breaking process in getting exact number of clusters (they can be quite large, i.e. 300-400 total clusters).(we used $PosteriorClusters$ method) [8] 2) since we are only interested the major personality types, we choose to only retain clusters with size proportion (out of 10,000) greater than $\epsilon = 0.1$. This truncated the number of clusters to a magnitude of 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.) The mean estimator is $rmean_estimator$. We rounded up to $K = 5$.We picked the last iteration for our posterior estimate of cluster paramters $\boldsymbol{\theta_i}$ due to the complication of taking the average over cluster paramters of different size.

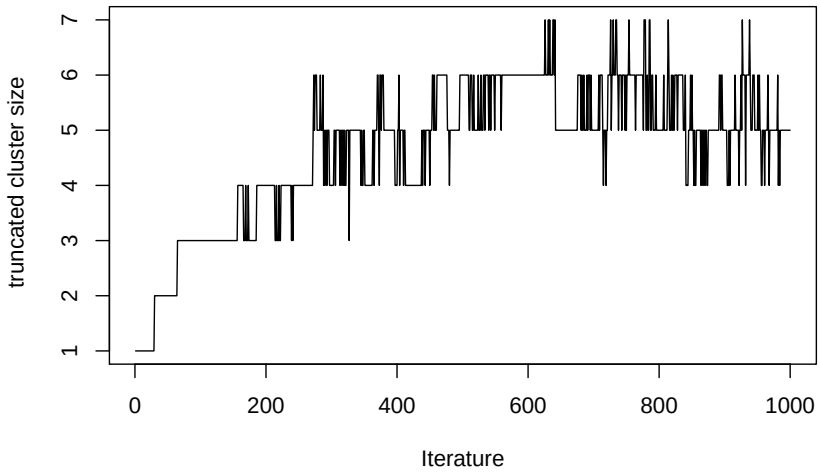

Figure 1: Traceplot for number of cluster

3.2 Gaussian mixture Model

To allow our model to capture more data in population, we will use the posterior five clusters to specify a prior to the Gaussian mixture model with fixed cluster size K and run a paramteric Gaussian mixture model on 200,000 individuals. The motivation is that since this process is less costly, we can run on more data to further update the "belief" on personality types and thus achieve more model generatlization. The key assumption we implicitly made here is that DPGMM on random 10,000 individuals is representative of the whole population so that at least $K = 5$ is fairly accurate assumption to feed into paramteric Gaussian mixture. Note that it does not have the same "accuracy" requirment on cluster parameters since Bayesian learning will keep updating them using 200,000 data while K is fixed throughout this phase. This process is justified since it 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:

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

We assigned the following prior:

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

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

We inference on the posterior cluster paramters by MCMC sampling, using the r package mixAK.[9], which does exert some additional prior constraints. We plug in the values of $(\boldsymbol{\mu_j}, \Sigma_j)$ using cluster paramters 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. mixAK also allows a uniform parameter V and places a hyper Gamma priors on V. (See our elicitation of hyper prior in Appendix) Our MCMC estimator is obtained by simply taking the cluster parameters' mean across all iterations.

4 Results

5 Conclusion

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

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

: 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

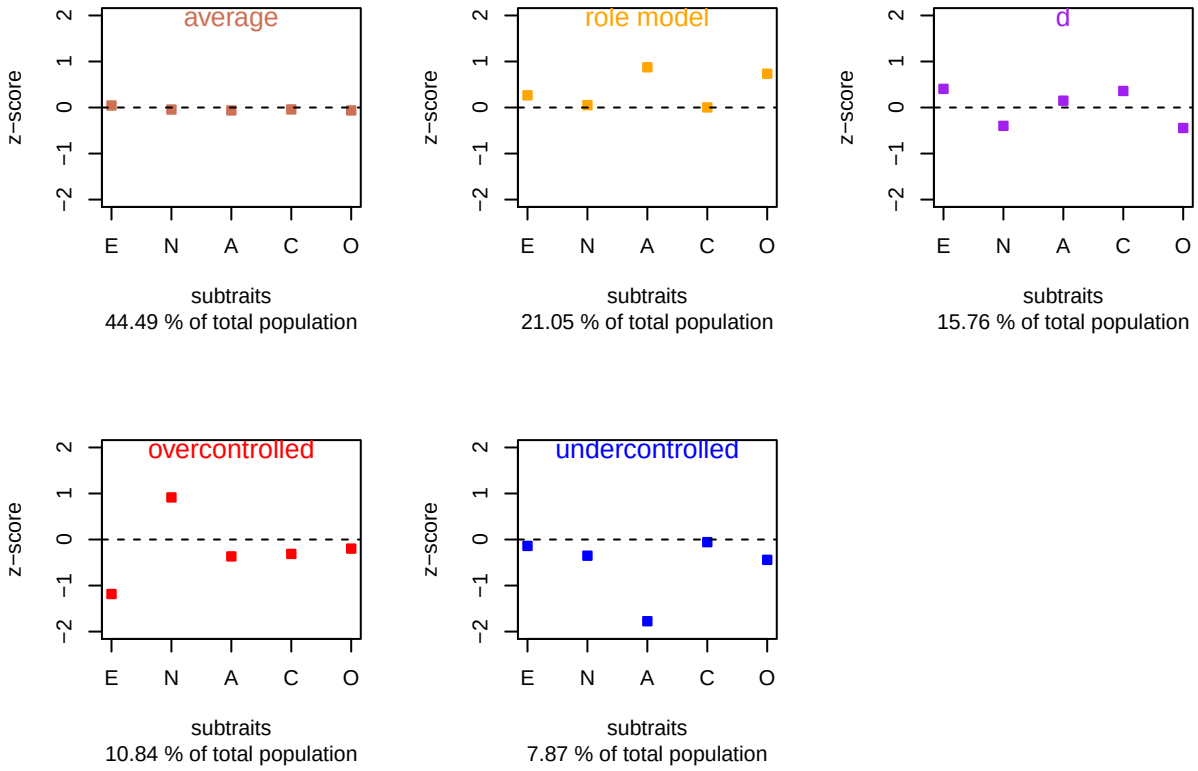

Figure 2: Mean vectors of 5 identified clusters of personality types

[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] 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.

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

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

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

[10] 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.

[11] 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/.

[12] 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.

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

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

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

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.[9] 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, $\gamma_j^{-1} \sim Gamma(g_j, h_j)$.

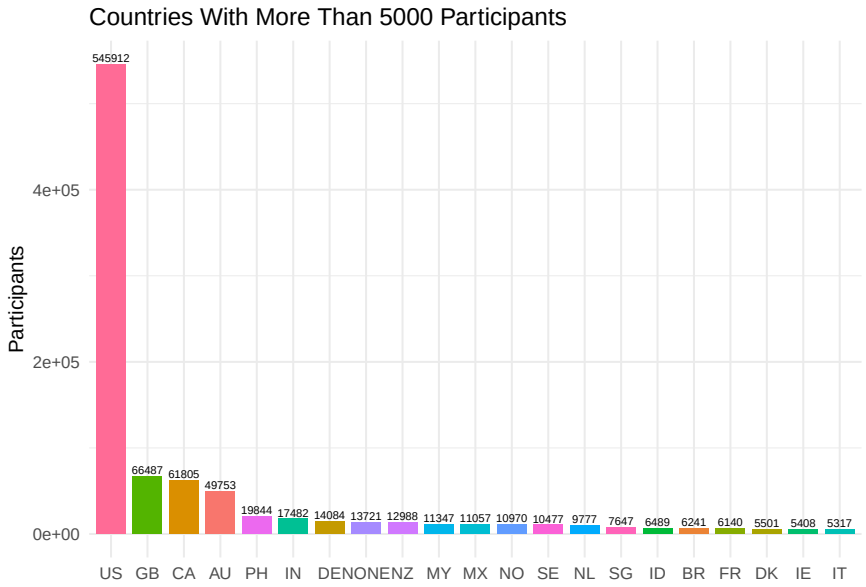

Figure 3: Number of participants in countries

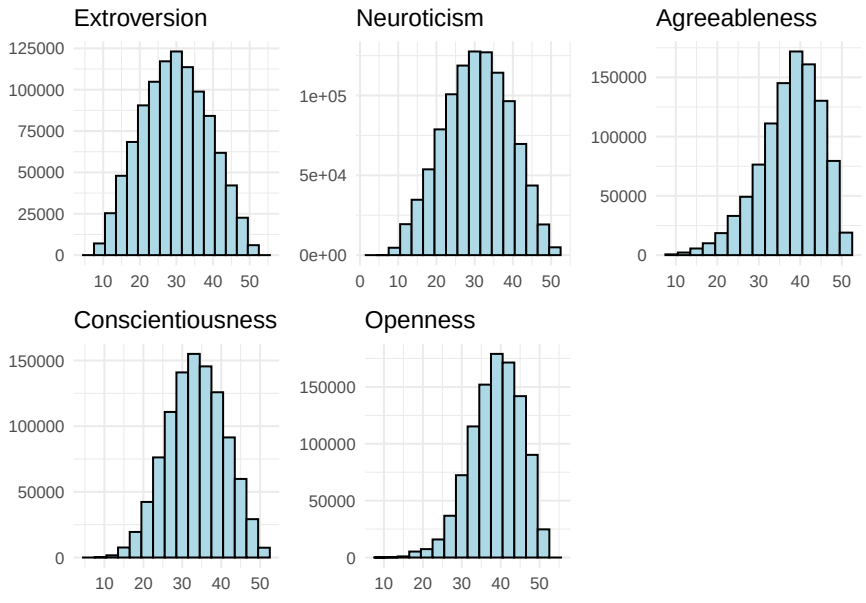

Figure 4: Normalcy check

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 $fitdistrplus$. [10] [11] 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

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.

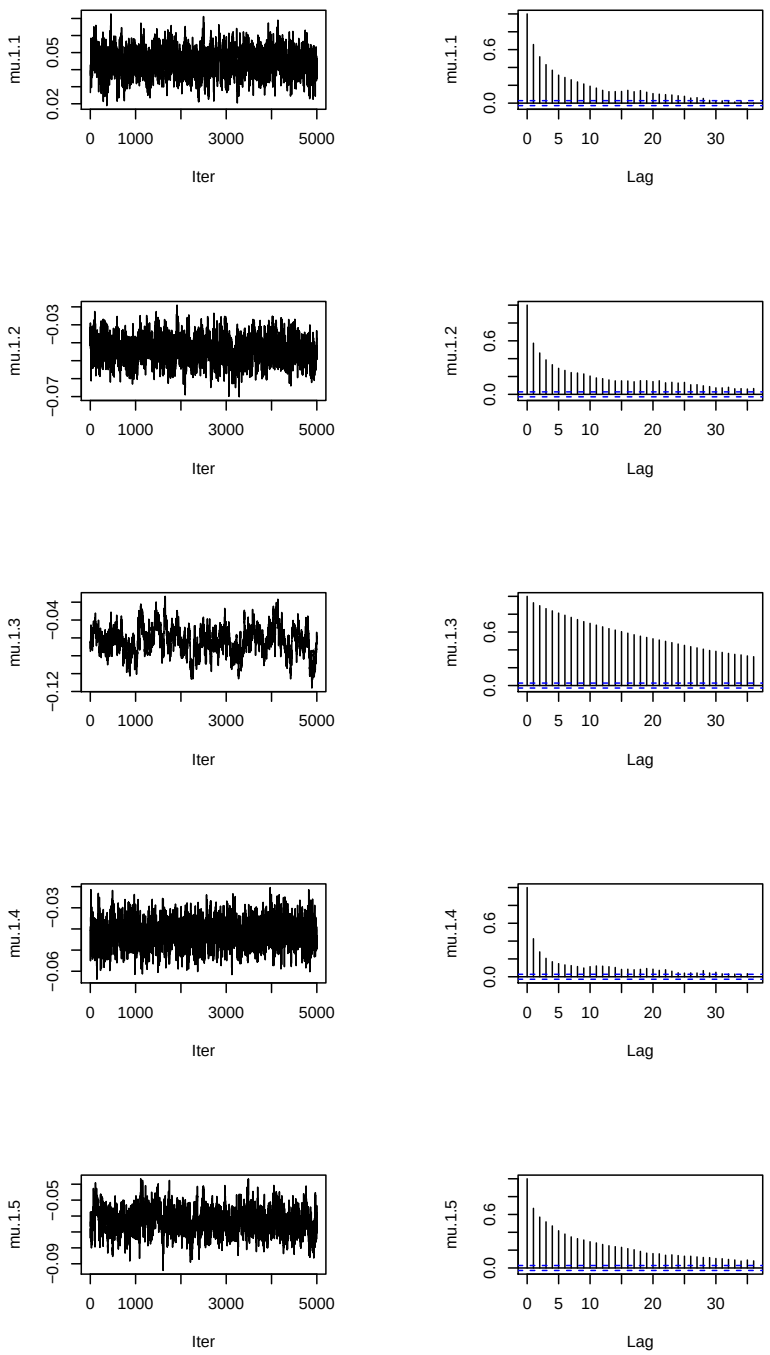

Figure 5: Traceplots and Autocorrelation plots for cluster means

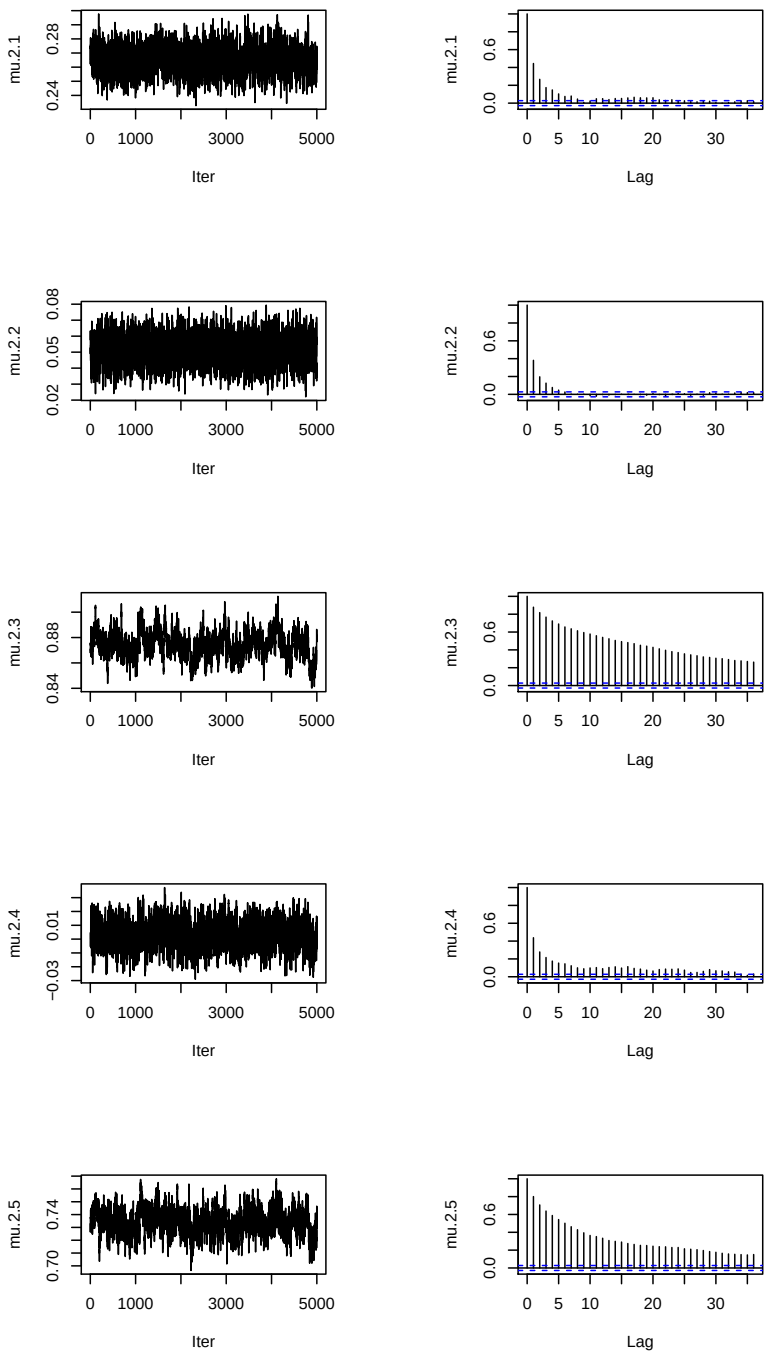

Figure 6: Traceplots and Autocorrelation plots for cluster means

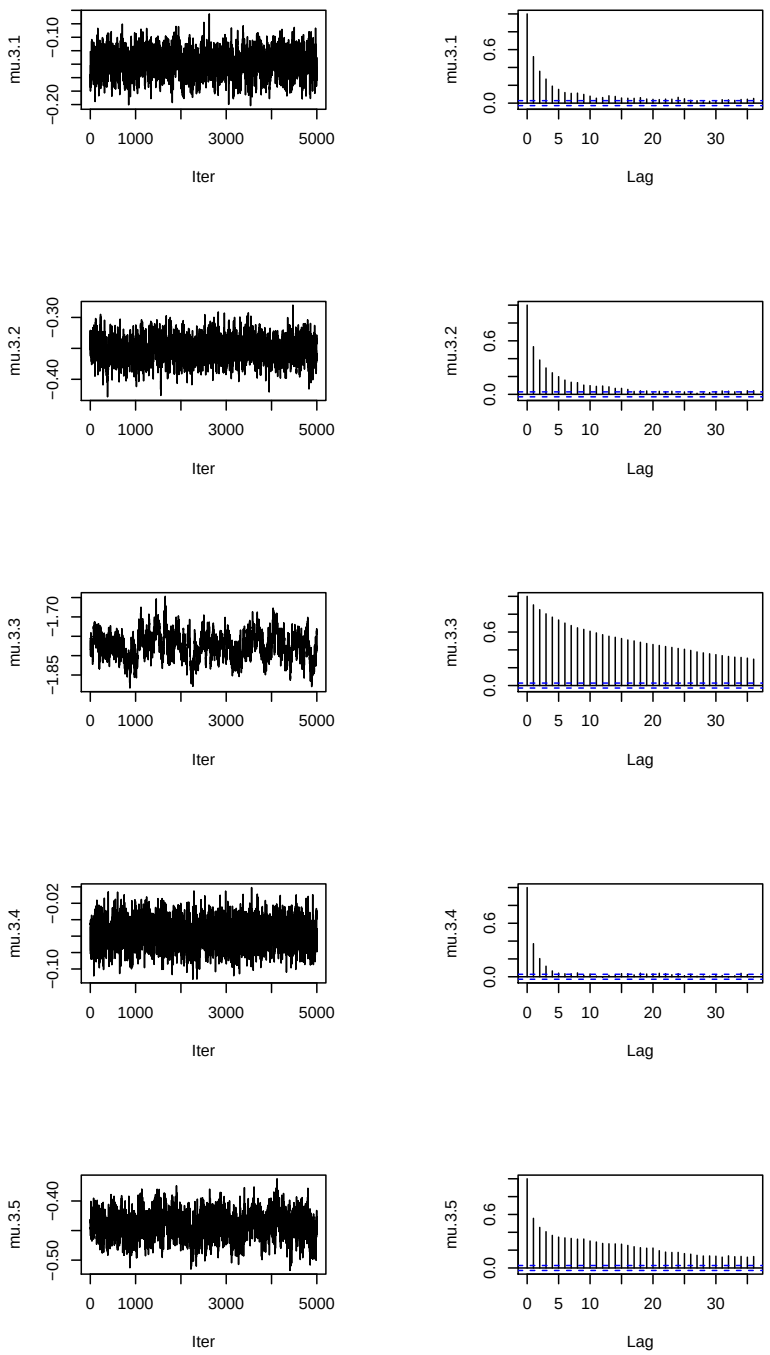

Figure 7: Traceplots and Autocorrelation plots for cluster means

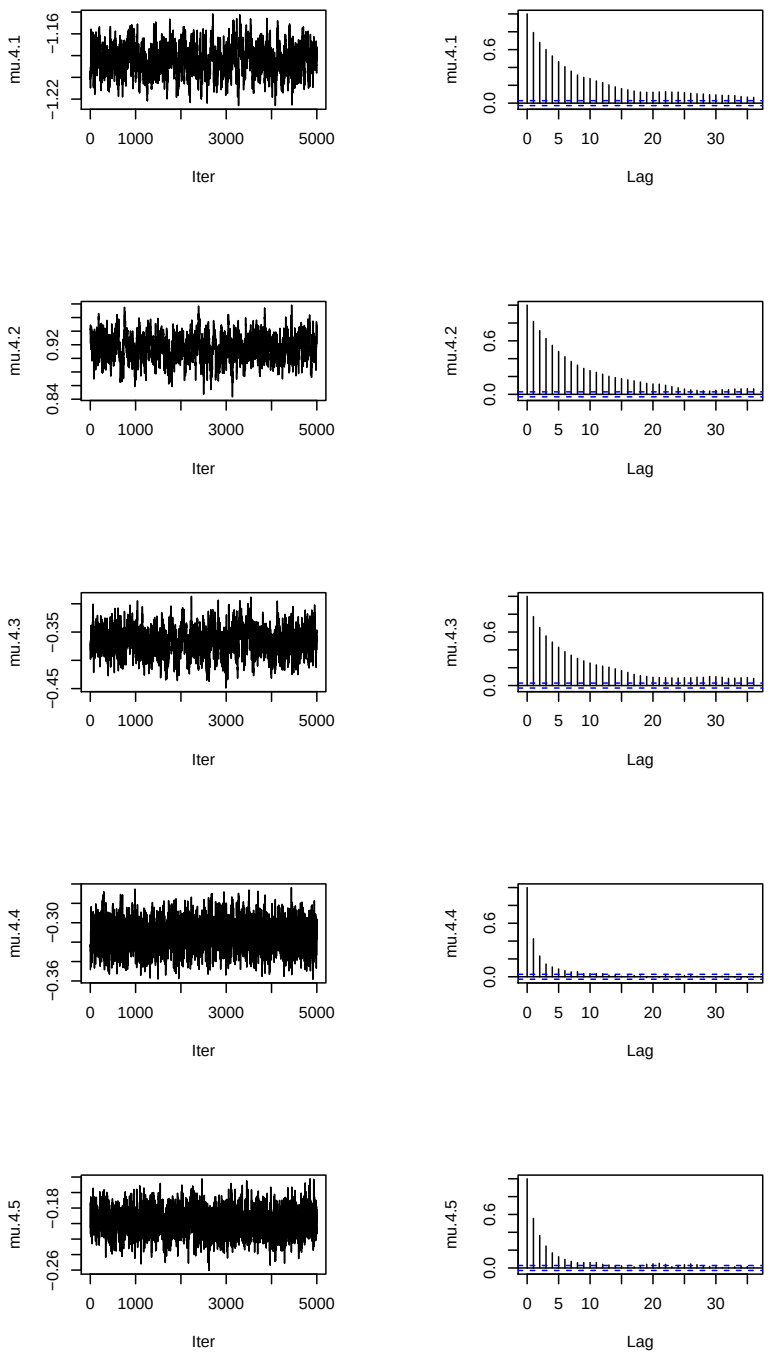

Figure 8: Traceplots and Autocorrelation plots for cluster means

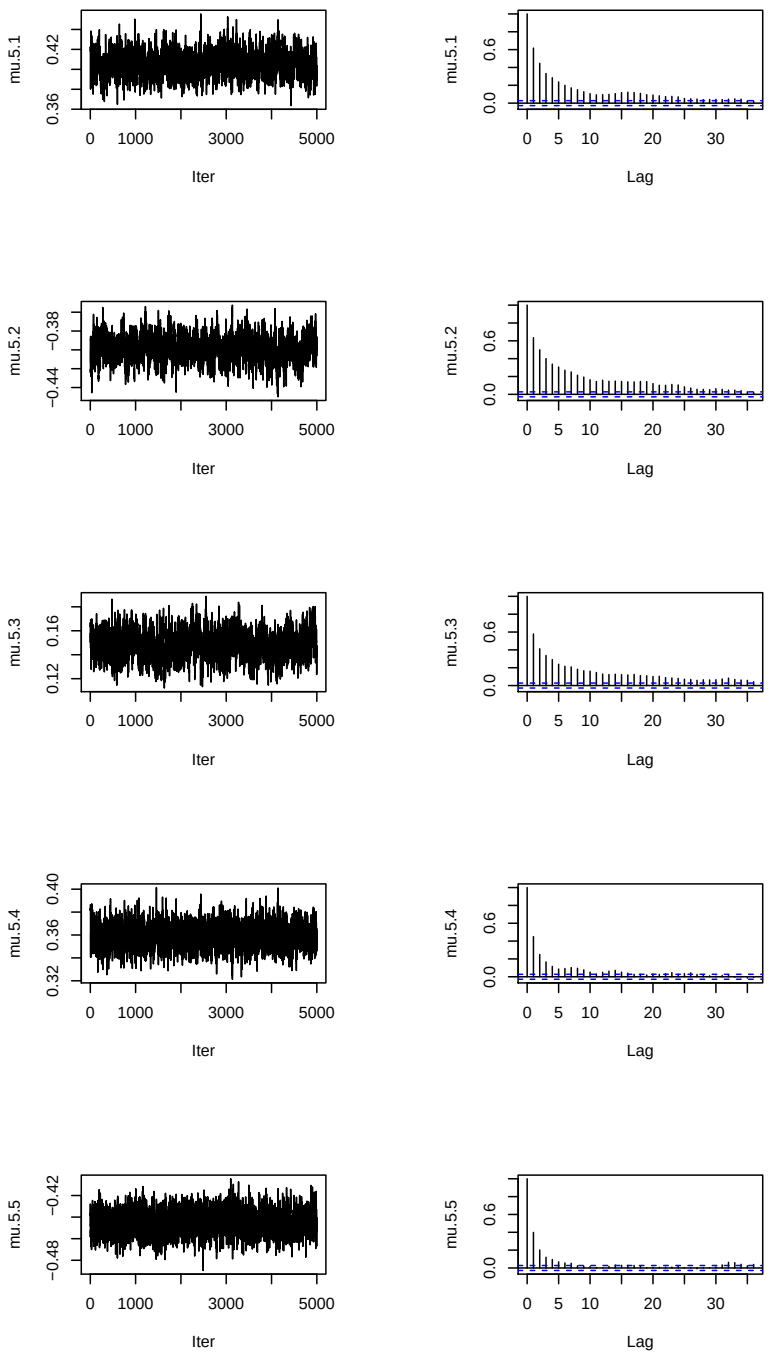

Figure 9: Traceplots and Autocorrelation plots for cluster means

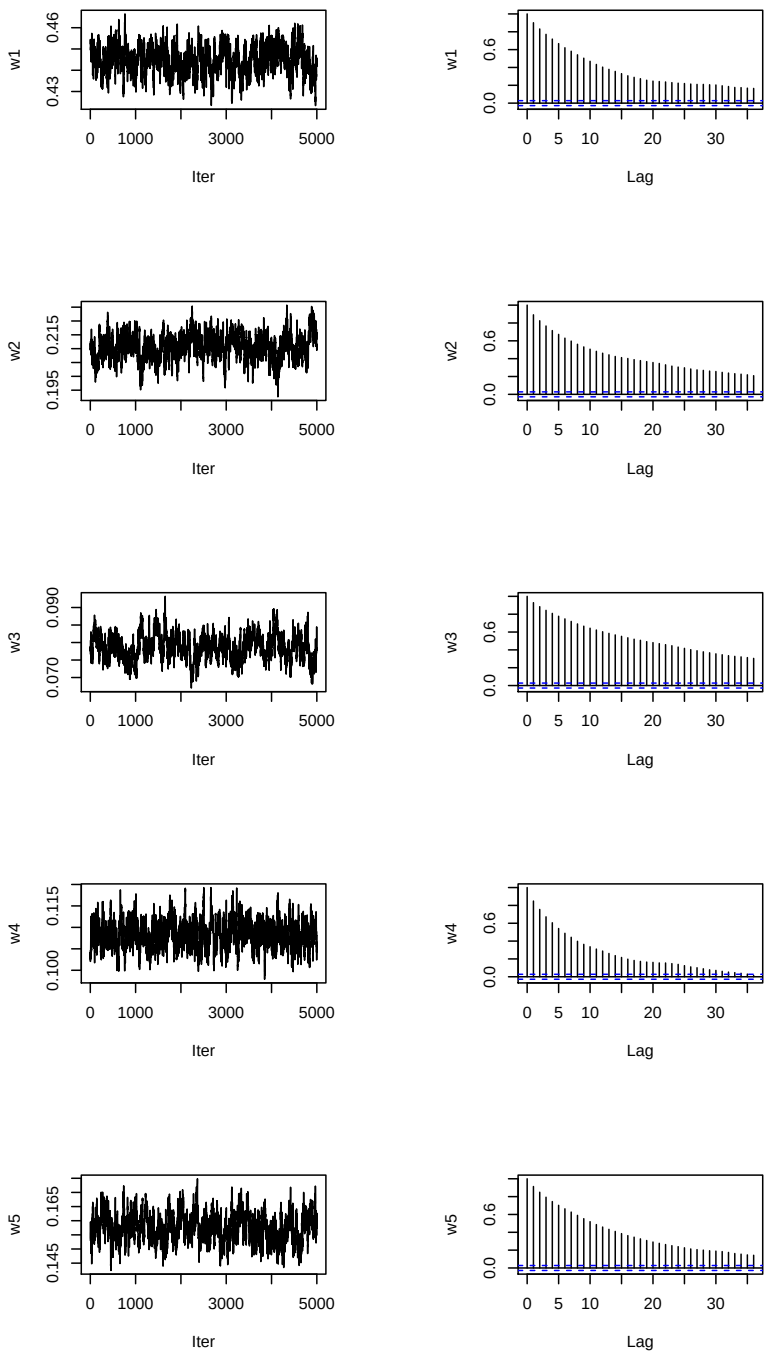

Figure 10: Traceplots and Autocorrelation plots for weights

- 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.