How Many Personality Types Are There?

Edric Svarte
University of Waterloo - December 5th, 2023
esvarteb@uwaterloo.ca

Abstract

We apply unsupervised learning techniques to psychometric data in search of personality trait clusters. We do not find convincing evidence for the existence of distinct personality types. However, clustering results suggest two or three potential groups, differing principally in neuroticism. Further investigation is required.

1 Introduction

Do personality *types* exist? Personality *traits* surely do. For instance, the Big Five taxonomy proposes five personality dimensions: openness, conscientiousness, extroversion, agreeableness and neuroticism (Widiger 2017, 11-32). On the other hand, the existence of distinct personality types remains contested (Gerlach et al. 2018, 735). Nonetheless, it seems reasonable to assume that stable[1] traits may constitute a "personality" or "character." A personality type would thus be a collection of correlated traits.

Interest in grouping individuals into neat, platonic categories dates back at least as far as the Greeks. Indeed, Hippocrates based a crude theory of medicine upon his belief in four personalities (Merenda 1987, 367). Currently, the Myers-Briggs system (MBTI) proposes sixteen personalities, though the evidence for this model is sparse (Gerras and Wong 2016, 55). Indeed, critics blame the MBTI for leveraging the so-called Barnum effect; the tendency to interpret vague descriptions as highly personal and precise (Pittenger 1993, 6).

A sound personality theory would improve our understanding of others' needs, preferences, and motivations. Moreover, such a model provides predictive power; certain types may have particular tendencies. This information is valuable, and hence, this problem is worthwhile.
We contribute by applying machine learning (ML) techniques to personality test data. In particular, we rely on clustering algorithms that group "similar" observations. However, more work in this area is required since we fail to find reliable evidence for the existence of distinct types.

2 Related Work

Gerlach et al. (2018) use Gaussian Mixture Models (GMMs) to identify personality trait clusters. In particular, the researchers apply factor analysis to personality test answers. This technique scores participants on five personality domains resembling the Big Five traits. However, their method provides continuous rather than discrete scores.
The researchers find initial evidence suggesting thirteen clusters. Upon investigation, however, this solution was found to overfit the data. In fact, only four clusters were meaningful. These results imply four personality types, which Gerlach et al. (2018) label "average," "self-centred," "reserved," and "role-model." For instance, the self-centred individual is highly extroverted, though low in openness. Finally, the authors apply the same approach to three other data sets. Impressively, they obtain nearly identical results. These findings are compelling since all four data sets contain more than one hundred thousand samples. However, survey data remain prone to various biases.

[1] Meaning these traits change little over time.

On the other hand, Sava and Popa (2011) use K-Means Clustering to identify distinct personality groups. The researchers find evidence for two solutions, three or five clusters, both revealed to be stable via cross-validation. Crucially, response bias may explain these results. Sava and Popa (2011) cite related studies which find only three personality types when the data are self-reported. On the other hand, observational data reveal additional types. Furthermore, clustering results are sensitive to the data's country of origin. For instance, clusters may be less obvious or fewer in number when the surveyed population is "homogeneous." Finally, the authors propose the five-cluster solution, which better explains observed behaviours (e.g., smoking habits). Unfortunately, Sava and Popa's (2011) sample comprises only 1039 participants (Sava and Popa 2011, 366). Given the relatively low quality of self-report data, a larger sample size may be preferred.

Mount et al. (2005) first attempt to define the concept of personality. Interestingly, their definition echoes our intuition. In particular, the notion of a personality is meaningful if stable characteristics influence actions (Mount et al. 2005, 448).
The researchers show that personality traits and vocational interests are distinct. More specifically, hierarchical clustering reveals three initial interest clusters: enterprising-conventional, realistic-investigative, and artistic-social. Similarly, two initial personality trait clusters form: open-extroverted and conscientious-stable. However, interests and personality traits merge only during the final step, implying that both are largely unrelated. Lastly, their model suggests a three-faceted individual determined by social and vocational interests, as well as degree of achievement striving.

3 Data

3.1 Overview

The data were obtained from *Open Psychometrics*, an online psychology data repository. The fifty variables (columns) of interest correspond to fifty personality test questions. Study subjects indicated their degree of agreement (1-5) with fifty statements about themselves. For instance, "I am the life of the party" (EXT-1). The questionnaire comprises five sections of ten questions each. Each portion assesses the strength of a Big Five trait in a participant.

3.2 Preprocessing and Cleaning

The survey's structure provides a natural way to reduce the data's dimensionality from fifty to five. In particular, we sum individuals' scores from each section; implicitly assuming the validity of the widely accepted Big Five taxonomy (Allik and McCrae 2002, 1-3). In truth, fifty dimensions may offer a more accurate representation of the data. However, our algorithms may become impractically slow due to the sparsity of high-dimensional neighbourhoods (Hastie et al. 2009, 23). Also note that the data were not normalized or standardized since all variables lie on the same scale.

The data set contains 1,015,341 observations (rows). Unfortunately, 1783 observations were missing all entries (columns). These examples were removed since imputation is not possible. In fact, mean imputation is possible, though unwise, since this procedure distorts the empirical CDF (Houari et al. 2014, 100).
Those who recorded the data added an "IPC" variable, indicating the number of records from each user's IP address. For "maximum cleanliness," the codebook indicated that only entries with IPC values of one should be used. Consequently, we removed all other entries, implying 695,704 remaining data points. Additionally, 1696 observations had zeros in at least one column. Recall that all questions are scored from one to five. Accordingly, the minimum value for each column should be ten. Upon inspection, we discovered that many of these examples contained several zero entries and low scores in all others. We removed these samples since it was unclear what imputation would entail. Afterwards, we found 685 entries with at least one score lower than ten. We chose to keep these samples since they were not obviously corrupted. However, an argument could be made for removing these entries as well. Finally, we proceed with 694,008 observations.

Lastly, the data type is ordinal for all columns. We assume equal intervals; distances between categories are equal. This assumption appears reasonable, though we mention it in the discussion.

3.3 Exploratory Analysis

Here we display interesting data characteristics, and perform initial investigations. First, we obtained a random subsample[2] of $5 \cdot 10^4$ variates from each trait (each column). Each plot in the matrix below provides three density estimates: a histogram, a Gaussian, and a kernel density estimate (KDE). The histogram details relative frequencies for the subsample, while the Gaussian was fit via maximum likelihood estimation (MLE) over the entire sample. On the other hand, the kernel density was fit to the subsample, and the bandwidth determined via Silverman's heuristic.

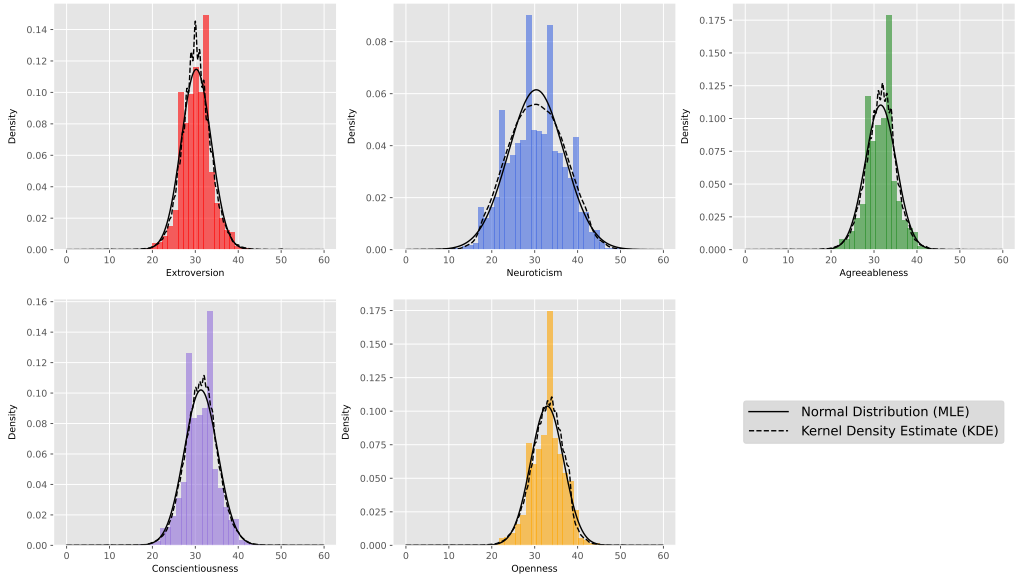

Figure 1: Distributions of Estimated Big Five Trait Scores

The normality assumption seems reasonable for all five variables. However, all five distributions appear lighter-tailed than Gaussian ones. Indeed, this finding suggests that very low or high scores are less common than expected from Gaussian data.

Table 1: Estimated Means and Standard Deviations for Big Five Trait Scores

Variable	$\hat{\mu}$	$\hat{I}^{\mu}_{95}$	$\hat{\sigma}$	$\hat{I}^{\sigma}_{95}$
Extroversion (E)	30.2	[30.2, 30.2]	3.5	[3.5, 3.5]
Neuroticism (N)	30.4	[30.4, 30.4]	6.5	[6.5, 6.5]
Agreeableness (A)	31.6	[31.6, 31.6]	3.6	[3.6, 3.6]
Conscientiousness (C)	31.3	[31.3, 31.3]	3.9	[3.9, 3.9]
Openness (O)	32.9	[32.9, 32.9]	3.8	[3.8, 3.9]

There is little uncertainty surrounding population parameter estimates. Interestingly, neuroticism scores exhibit the highest variance. This finding is significant, though this may be due to the large sample size. On the other hand, extroversion exhibits the lowest variance of all trait scores.

We now apply principal component analysis (PCA) to better visualize the data, and identify potential clusters. We obtained the principal components using a (standardized) random subsample of 10^5 data points, and then projected these data onto the resulting eigenvectors. The first two and three principal components explain 55%, and 73% of the subsample variance, respectively.

[2]Subsampling ensures a scalable approach. Under mild assumptions, sampling from the empirical CDF, $\hat{F}_X$, is roughly equivalent to sampling from the true, unknown CDF, F_X, by the Gilvenko-Cantelli theorem.

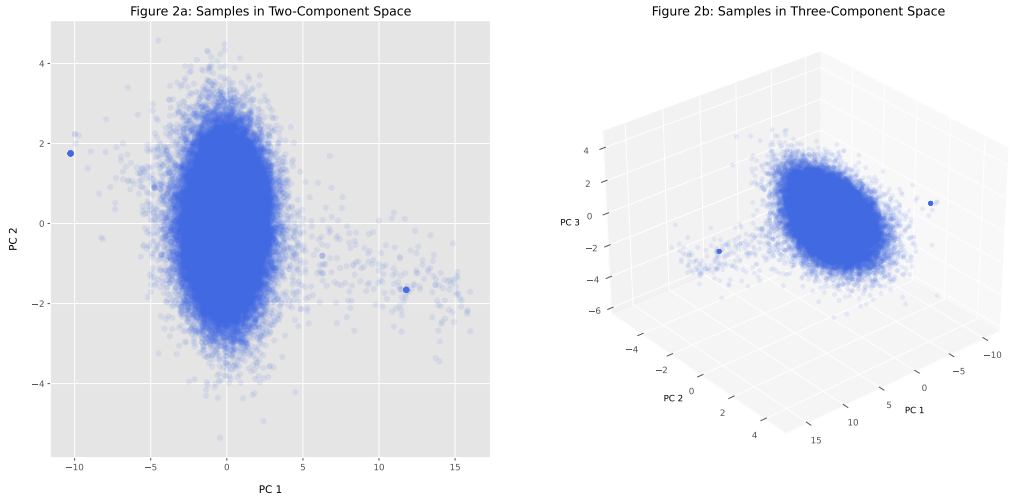

Figure 2: PCA Results

There are no obvious clusters; there is only one high-density region in both spaces. Interestingly, we notice two groups of (potential) outliers near the boundaries of the data. However, these groups contain relatively few samples compared to the main group. Accordingly, we choose to run clustering algorithms on the untransformed data.

4 Analysis and Results

4.1 K-Means Clustering

The K-Means Clustering algorithm partitions n data points into K mutually exclusive clusters. Model hyperparameters include the number of clusters, K, and the distance function. The former can be determined via elbow and silhouette methods. On the other hand, we use the Euclidean distance in this analysis. Indeed, this function provides a desirable statistical interpretation: the algorithm minimizes within-cluster variances (Hastie et al. 2009, 510).

We restrict our search for K to the set $[1,\ 20] \subset \mathbb{N}$. We use four different initial cluster assignments (seeds) for each value of K. This technique might clarify whether solutions are spurious and correspond to local minima.

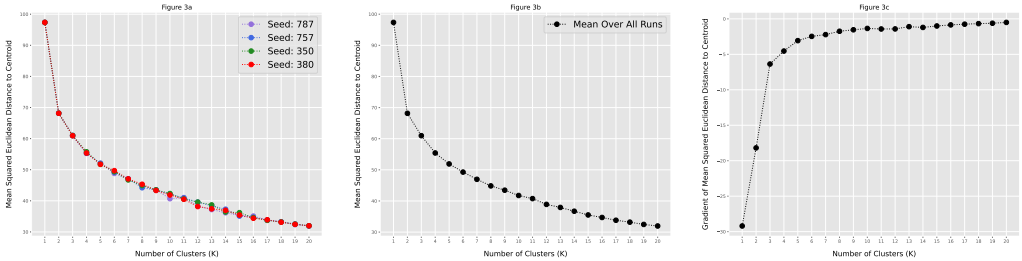

Figure 3: Squared L_2 Distance to Centroid as a Function of K

Figure 3a details the average squared distance between points and their corresponding centroids. The elbow method suggests choosing the "elbow" of the curve as K. In our case, this technique yields largely inconclusive results. Indeed, the curves in figures 3a and 3b lack obvious elbows. On the other hand, figure 3c details the gradient[3] of the curve. This plot is more informative; notice that the gradient appears to "flatten" noticeably for $K \geq 7$.

[3]The gradient is calculated via second differences for interior points, and first differences at the boundaries. The documentation for the NumPy *gradient()* function cites Fornberg (1988).

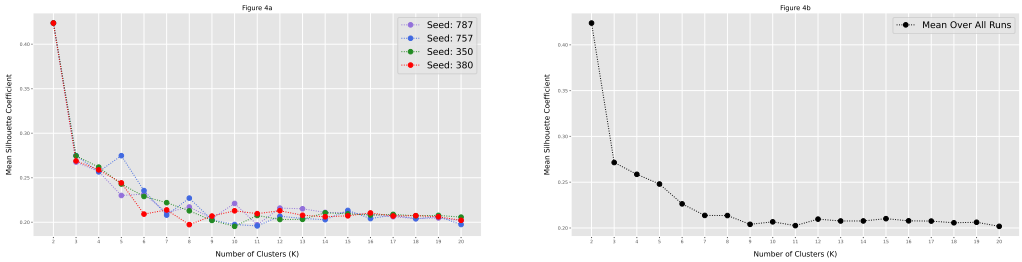

Figure 4: Mean Silhouette Coefficient as a Function of K

The silhouette coefficient measures the degree of cohesion within each cluster, and varies from -1 (worst) to 1 (optimal). In general, this method is not ideal for massive data given its $O(n^2)$ time complexity (Petrovic 2006, 10). Notice that the silhouette coefficient drops markedly at $K = 2$, and once more at $K = 5$.

Overall, $K = 2$ and $K = 3$ appear to be sensible choices for K. We now examine the predicted clusters for $K = 2$ and $K = 3$ from the first of the four runs. Results are illustrated below.

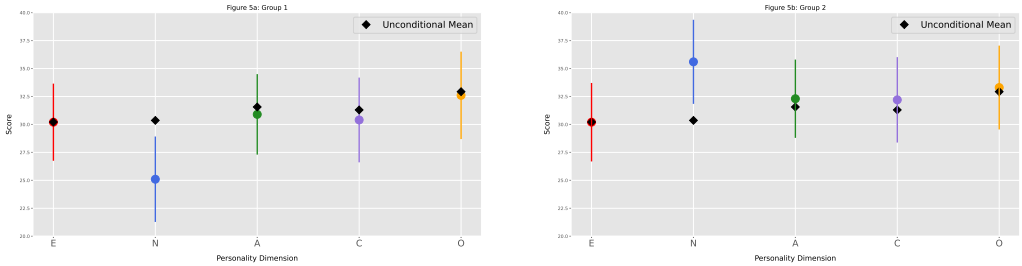

Figure 5: Mean and Standard Deviation Estimates for $K = 2$

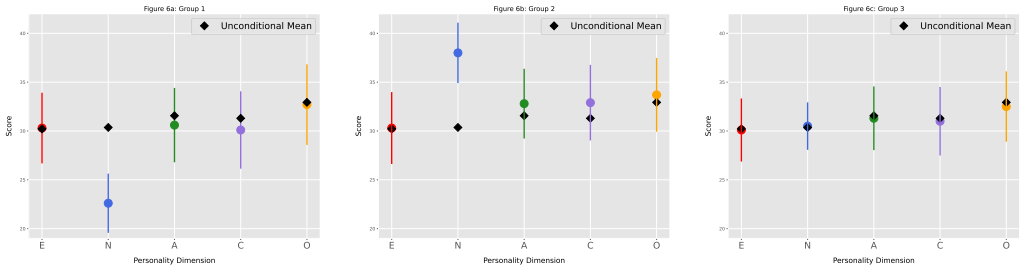

Figure 6: Mean and Standard Deviation Estimates for $K = 3$

The table below details estimated centroids.

Table 2: Centroids for $K = 2$ and $K = 3$

	$K = 2$		$K = 3$		
Variable	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$
Extroversion (E)	30.2	30.2	30.3	30.3	30.1
Neuroticism (N)	25.1	35.6	22.6	38.0	30.5
Agreeableness (A)	30.9	32.3	30.6	32.8	31.3
Conscientiousness (C)	30.4	32.2	30.1	32.9	31.0
Openness (O)	32.6	33.3	32.7	33.7	32.5
Mixing proportion	0.50	0.50	0.29	0.29	0.42

First consider the $K = 2$ solution (left). The algorithm has separated the sample in half. We notice that group two is higher in neuroticism, agreeableness and conscientiousness than group one. On the other hand, both groups are equal in extroversion, and similar in openness.
Now consider the $K = 3$ solution (right). Once again, the clusters appear balanced in terms of mixing proportions. Once more, we notice one group (group two) which dominates others in terms of neuroticism, agreeableness and conscientiousness. Extroversion does not seem to vary greatly.

Overall, the difference in neuroticism scores is most salient. Interestingly, we noticed that this variable had the largest variance of all features during the exploratory analysis. Strangely, we do not recover any of the types proposed by Gerlach et al. (2018) or Sava and Popa (2011). Moreover, we cannot test for significant differences in means since this amounts to data snooping; we are more likely to find significant differences. However, it may still be useful to analyze various sums of squares. In particular, the unconditional variance in the j-th variable can be decomposed as follows:

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i^{(j)} - \hat{\mu}^{(j)})^2}_{TSS^{(j)}} \equiv \underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i^{(j)} - \hat{\mu}_{C_{(i)}}^{(j)})^2}_{WSS^{(j)}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_{C_{(i)}}^{(j)} - \hat{\mu}^{(j)})^2}_{BSS^{(j)}}$$

The proof is given in the appendix. Here, $\hat{\mu}_{C_{(i)}}^{(j)}$ denotes the estimated mean of the j-th variable in the cluster corresponding to observation i, $C_{(i)}$. Let $R^{(j)} = WSS^{(j)}/TSS^{(j)}$ and denote by $\bar{R}$ the average of $R^{(j)}$ over all five variables. If $R^{(j)}$ is large,[4] then the clusters do not account for a significant portion of the variance in the j-th variable. On the other hand, $\bar{R}$ illustrates how much variance is accounted for by the clusters, on average. Crucially, we find that $\bar{R}_{K=2} = 0.85$, and $\bar{R}_{K=3} = 0.81$. As suspected, the clusters do not account for much variation in the observed data.

In the next section, we employ a probabilistic, parametric clustering algorithm. Unlike the non-parametric technique explored above, the following algorithm may provide greater interpretability.

4.2 Gaussian Mixture Model

A K-component Gaussian Mixture Model (GMM) assumes the data originate from a mixture of K Gaussian distributions. Expectation Maximization (EM) is a common method to estimate these models. Crucially, EM is a "soft" K-Means algorithm, which calculates posterior, class membership probabilities for each data point (Hastie et al. 2009, 512). The hyperparameter K can be selected via the Bayesian Information Criterion (BIC), the formula for which is given below.

$$BIC \equiv p \cdot \ln(n) - 2 \cdot l(\theta^*)$$

Here, $l(\theta^*)$ denotes the value of the log-likelihood function at the optimal parameter vector θ^*, and $p,\ n$ denote the number of parameters and samples, respectively. Lower BIC values indicate a parsimonious fit. Notice that the model is increasingly penalized for large p as n increases.
We return to our search for K in the set $[1,\ 20] \subset \mathbb{N}$. Once again, we use four different random starting points for each K. Results are found below.

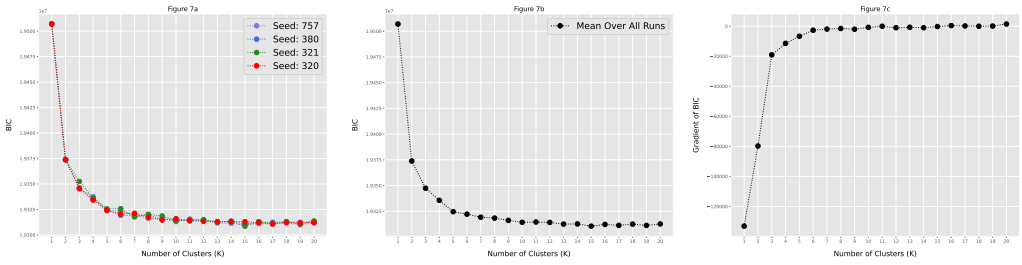

Figure 7: BIC as a Function of K

[4]The goal is not to minimize this quantity. Indeed, the ratio becomes zero if we create a cluster for each data point.

Notice that the curves in figures 7a and 7b lack evident elbows. Figure 7c details the BIC gradient; the change in BIC becomes insignificant for $K \geq 7$. Once again, $K = 2$ and $K = 3$ are arguably the most reasonable choices for K.

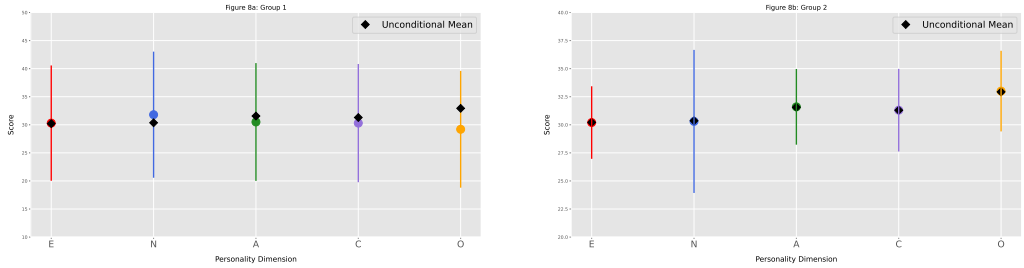

Figure 8: Mean and Standard Deviation Estimates for $K = 2$

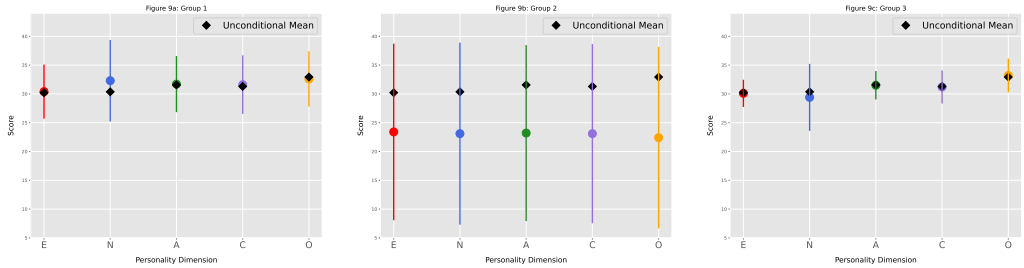

Figure 9: Mean and Standard Deviation Estimates for $K = 3$

The parameter estimates are given in the table below.

Table 3: Parameter Estimates for $K = 2$ and $K = 3$

	$K = 2$		$K = 3$		
Variable	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$
Extroversion (E)	30.3	30.2	30.4	23.4	30.1
Neuroticism (N)	31.8	30.3	32.3	23.1	29.4
Agreeableness (A)	30.5	31.6	31.7	23.2	31.2
Conscientiousness (C)	30.3	31.3	31.6	23.1	31.2
Openness (O)	29.2	33.0	32.6	22.4	33.2
Mixing proportion	0.02	0.98	0.34	≈ 0.00	0.66

Consider the $K = 2$ solution. Unlike the K-Means clusters, these classes are severely imbalanced. In particular, only 2% of the sample was assigned to group one. Notice that the between-cluster variance in personality trait scores is marginal. Unfortunately, there is no obvious interpretation for these findings.

Crucially, convergence to a local minimum is unlikely to be the cause of these results. Indeed, we repeated the analysis for $K = 2$ ten times using different starting points. All runs yielded nearly identical estimates.

The $K = 3$ solution is likewise incredible. Notice that nearly 0% of the sample was assigned to group two. This group comprises 2480 members, whose trait scores are far below average (on average). Moreover, trait scores appear highly variable across individuals within this class.

Here we find that $\bar{R}_{K=2} \approx 1$, and $\bar{R}_{K=3} = 0.97$. As suspected, the clusters do not account for much variation in the data. It remains unclear why we obtained these results. GMMs are relatively flexible since clusters may be "stretched" or "compressed" by the variance-covariance matrix. We expected GMMs to yield better results than the K-Means algorithm. In fact, the opposite appears to be true for these data.

5 Discussion

We begin by discussing the data and potential sources of bias. Recall that the data were obtained from an online self-report survey. First, individuals uninterested in online surveys are excluded from this sample (selection bias). This bias may lead us to underestimate the number of clusters, since certain personality types may avoid surveys.
Next, participants may only provide socially acceptable answers, or ones that support their self-images (response/self-serving bias). For instance, individuals may not want to admit that they "shirk their duties" (CSN-8).
Finally, participants may only give extreme or neutral responses. In the former case, responses become caricatures of participants. In the latter case, data become equally distorted. This bias may inflate or decrease variances, depending on the tendencies of surveyed individuals. For example, we would underestimate the variance of all variables if most individuals provided deceivingly neutral responses.

Fortunately, the sample size is large. It seems reasonable to assume that a considerable portion of the data is representative of the global population. Indeed, individuals from 223 countries participated in the survey. The global nature of the survey may lessen the effect of culture-specific biases. Moreover, this fact somewhat mitigates the effect of culture on the number of clusters (Section 2).

We now discuss our methods. First, recall that we assumed equal intervals; we treated ordinal data as continuous. The aforementioned extreme response bias affects the soundness of this assumption. For instance, a subject may interpret scores 2 and 3 as "closer" than 1 and 2. If true, the intervals are no longer equal, and the assumption is unwarranted. However, it is impossible to know whether participants understood the survey this way. On the other hand, models exist to handle ordinal data; applying these techniques to this data set constitutes a possible extension.
Next, we only considered two algorithms: K-Means Clustering and GMMs. Other algorithms and techniques exist, for instance, hierarchical or density-based clustering. However, we sought to investigate the existence of distinct personality types. Indeed, partitional clustering is entirely appropriate for this task.

6 Conclusion

We find mixed results; most of our procedures to determine K were inconclusive. However, $K = 2$ and $K = 3$ appeared to be the most reasonable choices for the number of clusters. Unfortunately, the clusters for $K = 2$ and $K = 3$ proved unconvincing upon inspection. Indeed, these clusters did not account for a significant portion of the variation in the observed data. As noted in the discussion, this may suggest biased data rather than the non-existence of personality types.

Unlike Gerlach et al. (2018), we obtained especially poor results from GMMs. Indeed, the K-Means algorithm provided more interpretable results. In general, the latter technique is more suitable for high-dimensional, large data since GMMs involve $O(d^3)$ matrix inversion (Pinto and Engel 2015), where d denotes the number of variables.

In summary, we fail to find convincing evidence for the existence of distinct personality types. However, absence of evidence should not be confused with evidence of absence. In particular, more research is required to draw a definite conclusion. Possible extensions include repeating the analysis on other data sets, and using observational rather than self-report data. As mentioned in the discussion, exploring other clustering algorithms may also be worthwhile.

References

Allik, J. and R. R. McCrae (2002). A five-factor theory perspective. *The five-factor model of personality across cultures*, 303–322.

Fornberg, B. (1988). Generation of finite difference formulas on arbitrarily spaced grids. *Mathematics of computation 51*(184), 699–706.

Gerlach, M., B. Farb, W. Revelle, and L. A. Nunes Amaral (2018). A robust data-driven approach identifies four personality types across four large data sets. *Nature human behaviour 2*(10), 735–742.

Gerras, S. J. and L. Wong (2016). Moving beyond the mbti. *Military review 96*(2), 54–57.

Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*, Volume 2. Springer.

Houari, R., A. Bounceur, A. K. Tari, and M. T. Kecha (2014). Handling missing data problems with sampling methods. In *2014 International Conference on Advanced Networking Distributed Systems and Applications*, pp. 99–104.

Merenda, P. F. (1987). Toward a four-factor theory of temperament and/or personality. *Journal of personality assessment 51*(3), 367–374.

Mount, M. K., M. R. Barrick, S. M. Scullen, and J. Rounds (2005). Higher-order dimensions of the big five personality traits and the big six vocational interest types. *Personnel psychology 58*(2), 447–478.

Petrovic, S. (2006). A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. In *Proceedings of the 11th Nordic workshop of secure IT systems*, Volume 2006, pp. 53–64. Citeseer.

Pinto, R. C. and P. M. Engel (2015). A fast incremental gaussian mixture model. *PloS one 10*(10), e0139931.

Pittenger, D. J. (1993). Measuring the mbti... and coming up short. *Journal of Career Planning and Employment 54*(1), 48–52.

Sava, F. A. and R. I. Popa (2011). Personality types based on the big five model. a cluster analysis over the romanian population. *Cognitie, Creier, Comportament/Cognition, Brain, Behavior 15*(3).

Widiger, T. A. (2017). *The Oxford handbook of the five factor model*. Oxford University Press.

Appendix

Variance Decomposition Proof

Consider a random variable $X^{(j)}$ with corresponding distribution function $F^{(j)}$. Let $x_i^{(j)}$ denote the i-th realization of $X^{(j)}$. Let $\hat{\mu}^{(j)}$ denote the sample mean of variates drawn from $F^{(j)}$. Suppose that x_i is assigned to cluster $C_{(i)}$, and let $\hat{\mu}_{C_{(i)}}^{(j)}$ denote the mean of points in $C_{(i)}$. We claim that the following equation balances:

$$\underbrace{\sum_{i=1}^{n}(x_i^{(j)} - \hat{\mu}^{(j)})^2}_{TSS^{(j)}} \equiv \underbrace{\sum_{i=1}^{n}(x_i^{(j)} - \hat{\mu}_{C_{(i)}}^{(j)})^2}_{WSS^{(j)}} + \underbrace{\sum_{i=1}^{n}(\hat{\mu}_{C_{(i)}}^{(j)} - \hat{\mu}^{(j)})^2}_{BSS^{(j)}}$$

Proof. Consider $TSS^{(j)}$:

$$\sum_{i=1}^{n}(x_i^{(j)} - \hat{\mu}^{(j)})^2 = \sum_{i=1}^{n}(x_i^{(j)} - \hat{\mu}_{C_{(i)}}^{(j)} + \hat{\mu}_{C_{(i)}}^{(j)} - \hat{\mu}^{(j)})^2 = \sum_{i=1}^{n}((x_i^{(j)} - \hat{\mu}_{C_{(i)}}^{(j)}) + (\hat{\mu}_{C_{(i)}}^{(j)} - \hat{\mu}^{(j)}))^2$$

Expanding the quadratic, we obtain:

$$= \sum_{i=1}^{n}(x_i^{(j)} - \hat{\mu}_{C_{(i)}}^{(j)})^2 + \sum_{i=1}^{n}(\hat{\mu}_{C_{(i)}}^{(j)} - \hat{\mu}^{(j)})^2 + 2\sum_{i=1}^{n}(x_i^{(j)} - \hat{\mu}_{C_{(i)}}^{(j)})(\hat{\mu}_{C_{(i)}}^{(j)} - \hat{\mu}^{(j)})$$

We switch to more convenient notation. Suppose that $C_{(k)}$ denotes the k-th cluster, where $k \in [1,\ K] \subseteq \mathbb{N}$. A point x_i is assigned to cluster k if and only if $i \in C_{(k)}$.

$$= \sum_{i=1}^{n}(x_i^{(j)} - \hat{\mu}_{C_{(i)}}^{(j)})^2 + \sum_{i=1}^{n}(\hat{\mu}_{C_{(i)}}^{(j)} - \hat{\mu}^{(j)})^2 + 2\sum_{k=1}^{K}(\hat{\mu}_{C_{(k)}}^{(j)} - \hat{\mu}^{(j)}) \sum_{i \in C_{(k)}}(x_i^{(j)} - \hat{\mu}_{C_{(k)}}^{(j)})$$

Rewriting the rightmost term:

$$= \sum_{i=1}^{n}(x_i^{(j)} - \hat{\mu}_{C_{(i)}}^{(j)})^2 + \sum_{i=1}^{n}(\hat{\mu}_{C_{(i)}}^{(j)} - \hat{\mu}^{(j)})^2 + 2\sum_{k=1}^{K}(\hat{\mu}_{C_{(k)}}^{(j)} - \hat{\mu}^{(j)})[\sum_{i \in C_{(k)}}(x_i^{(j)}) - |C_{(k)}| \cdot \hat{\mu}_{C_{(k)}}^{(j)}]$$

$$= \sum_{i=1}^{n}(x_i^{(j)} - \hat{\mu}_{C_{(i)}}^{(j)})^2 + \sum_{i=1}^{n}(\hat{\mu}_{C_{(i)}}^{(j)} - \hat{\mu}^{(j)})^2 + 2\sum_{k=1}^{K}(\hat{\mu}_{C_{(k)}}^{(j)} - \hat{\mu}^{(j)})(|C_{(k)}| \cdot \hat{\mu}_{C_{(k)}}^{(j)} - |C_{(k)}| \cdot \hat{\mu}_{C_{(k)}}^{(j)})$$

Indeed, the rightmost term vanishes:

$$= \sum_{i=1}^{n}(x_i^{(j)} - \hat{\mu}_{C_{(i)}}^{(j)})^2 + \sum_{i=1}^{n}(\hat{\mu}_{C(i)}^{(j)} - \hat{\mu}^{(j)})^2$$

□

Variance decompositions like the one above are common in statistics. This demonstration is nothing novel.