Statistical Independence and Correlation Analysis among MBTI Personality Dimensions

Chia-Wen Lu (A69044492) and Jia-Ming Lin (A69044401)

Department of Electrical and Computer Engineering, University of California San Diego

1 Introduction

The Myers-Briggs Type Indicator (MBTI) is a widely utilized psychological instrument; however, the statistical independence and correlation among its four primary dimensions, as well as their relationship with various demographic factors, requires comprehensive analysis. To address this problem, this project applies statistical analysis to rigorously examine the dependence structure among MBTI dimensions, assess the validity of the independence assumption, and explore whether these relationships are consistent across demographic groups.

While prior studies have examined the psychometric properties of MBTI instruments, relatively few have directly investigated the interdependencies among the MBTI dimensions themselves using modern statistical techniques. In addition, demographic variables, including age, gender, and education, are known to be associated with personality-related measures, raising the possibility that observed dependencies among MBTI dimensions may reflect shared demographic influences rather than intrinsic relationships.

In this project, we employ a set of complementary statistical methods to capture different forms of dependence. Pearson correlation is used to assess linear relationships, Mutual Information to detect nonlinear dependencies, partial correlation to control for confounding variables, and Graphical Lasso to infer conditional independence structures. In addition, exploratory analyses are conducted to characterize the relationships between MBTI dimensions and demographic variables.

By integrating these approaches, this project aims to systematically evaluate the independence assumption underlying the MBTI framework and to examine how demographic factors influence observed personality patterns. All code is available at https://github.com/chiawen0104/mbti_stat_analysis.

2 Dataset Description

The Kaggle dataset "Predict People Personality Types" contains 43,744 observations with nine variables, including age, gender, education, introversion score (I/E), sensing score (S/N), thinking score (T/F), judging score (J/P), interest, and personality type.

Education is coded as a binary variable, where a value of 1 indicates individuals with a graduate-level education or higher. Higher scores on the introversion scale indicate a greater tendency toward extraversion, while higher scores on the sensing, thinking, and judging scales indicate stronger preferences for sensing, thinking, and judging, respectively. Due to missing or unspecified values in the interest variable, this column was excluded from the analysis.

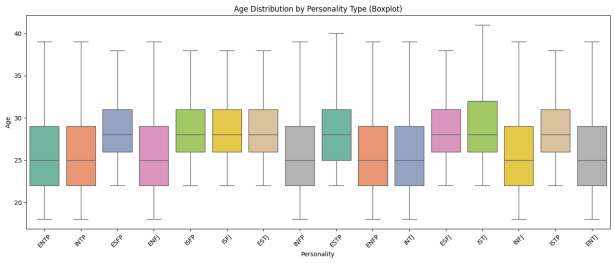

Figure 1: Age Distribution by Personality Type

3 Data Exploration

We conducted an Analysis of Variance (ANOVA) to examine the relationship between age and personality. The results ($F = 284.06, p < 0.001$) reveal highly significant differences in mean age across personality types. The effect size ($\eta^2 = 0.089$) indicates that approximately 8.9% of the variance in age is explained by personality type, representing a medium-to-large effect in behavioral science research. Additionally, Figure 1 illustrates a demographic skew toward participants in their 20s and early 30s, likely reflecting a higher propensity among younger individuals to engage with personality assessments for self-exploration.

To investigate the dependency between gender and personality, we performed a Chi-square test. The results ($\chi^2 = 910.57$, $p < 0.001$) allow us to reject the null hypothesis, confirming a significant association between the two variables. As shown in Figure 2, this association is characterized by a distinct gender split between Extroversion and Introversion. Specifically, females are overrepresented in several Extroverted types (ENTJ, ENTP, ESTJ), whereas males show a higher proportional dominance in Introverted types (INFJ, INFP, ISFP).

Finally, we examined the association between educa-

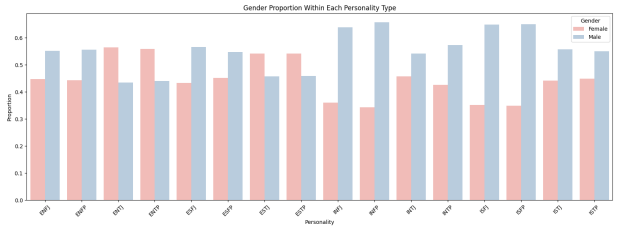

Figure 2: Gender Proportion Within Each Personality Type

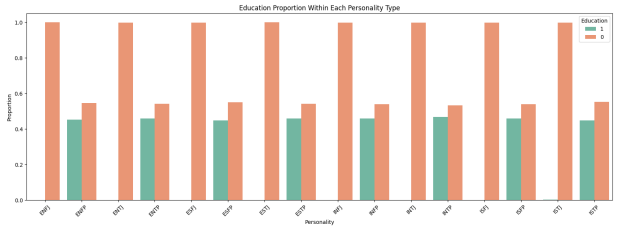

Figure 3: Education Proportion Within Each Personality Type

tion level and personality type using a chi-square test of independence. The test yielded a large chi-square statistic ($\chi^2 = 12,874.24$, $p < 0.001$), indicating a statistically significant association between the two variables. Given the large sample size, this result reflects substantial differences in the distribution of education levels across personality types rather than a deterministic relationship.

As shown in Figure 3, the association is primarily driven by differences along the Judging–Perceiving (J–P) dimension. Personality types characterized by Judging (e.g., ENFJ, ISTJ) are predominantly associated with Education Level 0, whereas Perceiving types (e.g., ENFP, ISTP) display a more heterogeneous distribution, including a higher proportion of Education Level 1.

Although Judging (J) types are often assumed to be associated with academic discipline and achievement, the observed distribution suggests a different pattern, with Education Level 1 being more prevalent among Perceiving (P) types. Rather than reflecting differences in ability, this pattern may be related to variation in educational trajectories or timing. For example, individuals with Judging preferences may be more likely to complete formal education earlier and transition into the workforce, whereas those with Perceiving preferences may be more inclined to pursue extended, non-linear, or exploratory educational pathways. This interpretation remains speculative and should be examined in future research using measures that directly assess educational motivation, career timing, and academic engagement.

4 Literature Review

Previous research indicates that the four dimensions of the MBTI are not entirely independent of one another. Johnson and Saunders [1] found that some MBTI subscales simultaneously reflect multiple dimensions, suggesting overlap among dimensions rather than complete separation. In contrast, Francis and Village [2] showed that, despite clearer measurement of individual dimensions, these dimensions are better viewed as related psychological traits rather than fully independent categories.

In addition, Furnham's work [3] has shown that the Thinking–Feeling dimension of the MBTI is strongly associated with Agreeableness in the Big Five, Judging–Perceiving with Conscientiousness, Extraversion–Introversion with Extraversion, and Sensing–Intuition with Openness. By contrast, Neuroticism shows inconsistent associations with multiple MBTI dimensions and is not explicitly represented within the MBTI framework. Overall, existing studies consistently suggest that the four MBTI dimensions exhibit systematic relationships and overlap, rather than strict independence.

However, much of the existing literature focuses on psychometric validity or cross-instrument comparisons, with limited direct testing of whether the four MBTI dimensions are empirically independent. To address this gap, our study applies statistical analyses to examine the dependence structure among MBTI dimensions, assess the validity of the independence assumption, and explore whether these relationships vary across demographic groups.

5 Correlation and Independence

To analyze the dependence structure among MBTI personality dimensions, we apply four statistical methods, each capturing a different aspect of variable relationships. **Pearson correlation** provides a baseline assessment of linear associations between MBTI scores and demographic factors. **Mutual Information** extends this analysis by detecting nonlinear dependencies that may not be reflected in linear measures. To distinguish genuine relationships from those driven by other variables, we compute **Partial Correlations**, which evaluate the direct association between each pair of dimensions while controlling for all others. Finally, we use **Graphical Lasso** to infer the overall conditional independence network, identifying which pairs of variables remain directly connected after accounting for the full multivariate structure. Together, these methods allow us to assess linear, nonlinear, and conditional relationships, forming a comprehensive framework for evaluating independence among the MBTI dimensions.

5.1 Pearson Correlation

The sample Pearson correlation coefficient is defined as

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ denote the sample means of X and Y, respectively. We compute a Pearson correlation matrix including the four MBTI dimension scores as well as age, gender, and education. As shown in Figure 4, the MBTI dimensions (Introversion, Sensing, Thinking, and Judging Scores) exhibit minimal linear correlation. Among demographic variables, age shows a weak positive correlation with the Sensing score ($r = 0.17$), while education

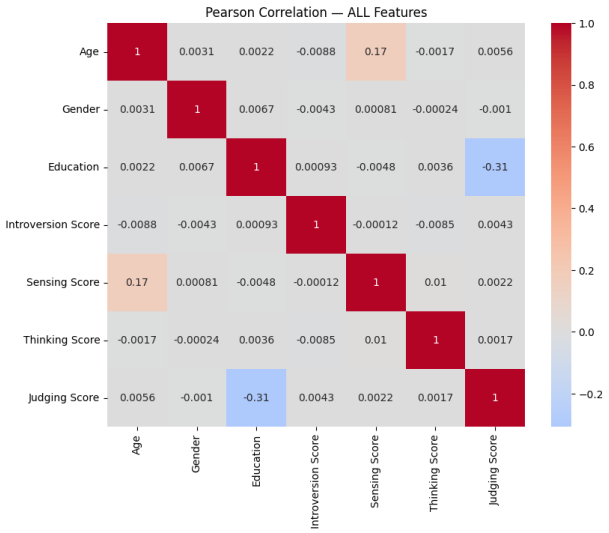

Figure 4: Pearson Correlation Matrix of All Features

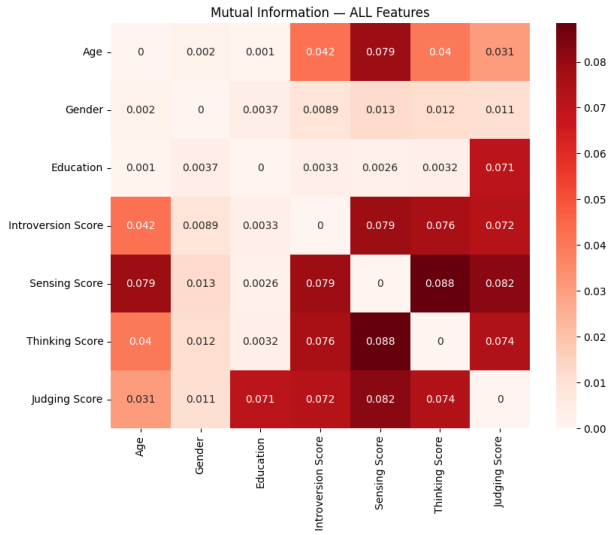

Figure 5: Mutual Information Matrix of All Features

exhibits a moderate negative correlation with the Judging score (r = -0.31).

5.2 Mutual Information

We next examine nonlinear dependencies using Mutual Information (MI), which measures general statistical dependence without assuming linearity. In contrast to Pearson correlation, MI reveals small but nonzero dependencies among the MBTI dimensions. For two variables X and Y, MI is defined as

$$I(X;Y) = \int \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \, dx \, dy.$$

where $p(x, y)$ denotes the joint probability density function of X and Y, and $p(x)$ and $p(y)$ denote their marginal distributions. Because the dataset contains both continuous and categorical variables, MI is estimated using different estimators. For continuous variables, we employ a k-nearest neighbor–based regression estimator, while for categorical variables, a classification-based estimator is used, following the implementations in `scikit-learn` (`mutual_info_regression` and `mutual_info_classif`).

Figure 5 reports the Mutual Information (MI) matrix across MBTI dimension scores and demographic variables. Compared to the Pearson results, MI reveals small but nonzero nonlinear dependencies among the four MBTI dimensions. In particular, the Sensing dimension exhibits the largest MI values with the other MBTI dimensions, including Thinking (MI = 0.088), Judging (MI = 0.082), and Introversion (MI = 0.079), suggesting that Sensing may capture limited nonlinear structure not reflected in linear correlations. Nevertheless, the absolute magnitudes of these MI values remain small, indicating weak nonlinear dependence overall.

Regarding demographic variables, age shows weak nonlinear associations with Sensing (MI = 0.079) and Introversion (MI = 0.042), while education exhibits a weak association with Judging (MI = 0.071). Gender shows consistently low MI values with all MBTI dimensions.

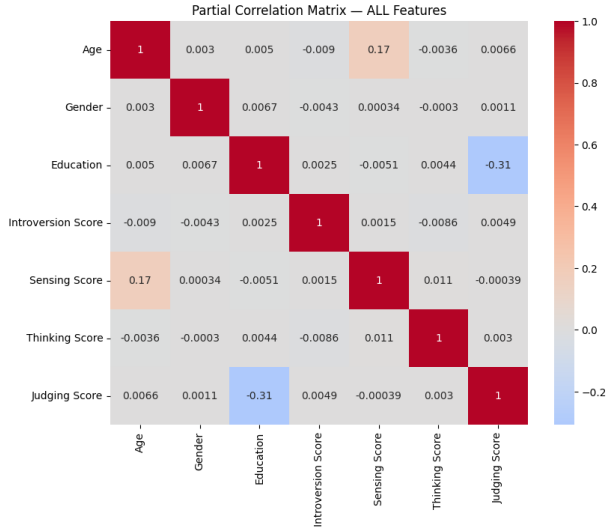

Figure 6: Partial Correlation Matrix of All Features

Overall, these results suggest that MBTI dimensions are approximately independent after accounting for age, gender, and education.

5.3 Partial Correlations

To assess direct linear associations while controlling for confounding effects, we compute the partial correlation (PC) matrix across all variables. We follow the standard definition described in Wikipedia and estimate it using the `pingouin` package.

As shown in Figure 6, PCs among the four MBTI dimensions remain close to zero, consistent with the Pearson correlation results 4. This indicates that no substantial direct linear associations exist between MBTI dimensions after controlling for other personality dimensions and demographic variables. In addition, the associations between age and the Sensing dimension, as well as between education and the Judging dimension, remain largely unchanged compared to the Pearson analysis, suggesting that these relationships are not driven by confounding effects. In summary, the PC analysis

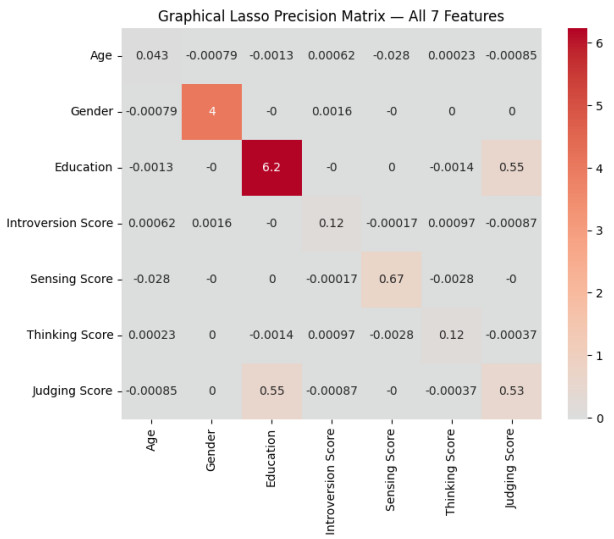

Figure 7: Graphical Lasso Precision Matrix of All Features

strengthens the conclusion of approximate linear independence among the MBTI personality dimensions.

5.4 Graphical Lasso

Although Pearson, MI, and partial correlation capture marginal linear, general nonlinear, and conditional linear relationships, respectively, they do not explicitly characterize the global conditional independence structure among variables. To further identify direct dependencies and conditional independence relationships, we apply Graphical Lasso (GL) to estimate a sparse precision matrix, which encodes the conditional independence graph.

GL estimates a sparse precision matrix under a multivariate Gaussian assumption. Let S denote the sample covariance matrix and let $\Theta = \Sigma^{-1}$ denote the precision matrix. The GL estimator is defined as the solution to the following ℓ_1-penalized Gaussian log-likelihood optimization problem:

$$\hat{\Theta} = \arg\max_{\Theta \succ 0} \left(\log\det(\Theta) - \mathrm{tr}(S\Theta) - \lambda \sum_{i \neq j} |\Theta_{ij}| \right),$$

where λ is a regularization parameter controlling sparsity.

Figure 7 shows the GL precision matrix, where non-zero off-diagonal entries indicate conditional dependence. The estimated matrix is highly sparse and consistent with the Pearson and partial correlation results. In particular, all off-diagonal entries among the four MBTI dimensions are close to zero, indicating no direct conditional dependencies after controlling for other variables. In contrast, a non-zero off-diagonal entry between education and the Judging dimension indicates that these variables are not conditionally independent, consistent with earlier analyses. Collectively, the GL results provide strong evidence that the MBTI personality dimensions are approximately independent under the Gaussian graphical model.

Conclusions

This study examined the dependence structure among MBTI personality dimensions and their relationships with demographic variables using multiple statistical approaches, including Pearson correlation, Mutual Information, partial correlation, and Graphical Lasso. Across all methods, the four MBTI dimensions exhibit consistently weak associations, providing convergent evidence for their approximate independence. Linear and nonlinear analyses reveal only minimal dependencies, primarily involving the Sensing dimension, while Graphical Lasso yields a sparse precision matrix with no conditional dependencies among the MBTI dimensions.

In contrast, demographic variables show stable and interpretable associations with specific MBTI dimensions. Age is most strongly related to the Sensing dimension, education to the Judging–Perceiving dimension, and gender exhibits significant distributional differences across personality types. Overall, these findings support the conceptualization of MBTI dimensions as largely independent constructs, while highlighting the importance of demographic context in shaping observed personality distributions.

References

[1] David R. Saunders Donald A. Johnson. Confirmatory factor analysis of the myers-briggs type indicator—expanded analysis report. *Educational and Psychological Measurement*, 50, 1990.

[2] Village Andrew Francis, Leslie J. The francis psychological type scales (fpts): factor structure, internal consistency reliability, and concurrent validity with the mbti. *Mental Health, Religion & Culture*, 25(9):931–951, 2022.

[3] Adrian Furnham. The big five versus the big four: the relationship between the myers-briggs type indicator (mbti) and neo-pi five factor model of personality. *Personality and Individual Differences*, 21:303–307, 1996.