STA 440: Case 1 Report #2

Jeffrey Ho, Amanda, Levenberg, Lucy Lu

September 14, 2016

Background

This analysis aims to study the relationship between brain fiber connections and the commonly studied Big Five personality traits, with particular focus on openness. Each personality trait is measured on a spectrum between two extremes, and openness of individuals are quantified as integer values whose low extremes correspond to consistent/cautious and high extremes to inventive/curious. Through network modeling and analysis, this study aims to elucidate whether a subject's brain connectivity is associated with their level of openness, while controlling for confounding variables such as the age and sex of the subject.

Modeling

Our network model estimates the probability of connection between any two brain regions of a particular subject based on the subject's personal information (denoted by x_i in Equation 1) as well as the 2-D coordinates of the regions (denoted by Z_u, Z_v in Equation 1). A logit link function is used to map outputs of the linear predictor $\eta_{u,v}$ (Equation 2) to valid probabilities within the range [0, 1].

$$Pr(y_{i,u,v} = 1 | x_i, Z_u, Z_v) = \text{logit}(\eta_{i,u,v}), u, v \in \{1, ..., 68\}, u \leq v \tag{1}$$

$$\eta_{i,u,v} = \alpha^T x_i + \lambda ||Z_u - Z_v||_2 \tag{2}$$

The three models we parametrized have increasing levels of complexity, as we are interested in whether additional estimators, especially subject-specific openness, are useful in modeling subject-specific brain connectivity.

$$\begin{aligned}
&\text{Model 1} : x_i = (1)^T \\
&\text{Model 2} : x_i = (1, \text{Sex}_i, \text{Age}_i)^T \\
&\text{Model 3} : x_i = (1, \text{Sex}_i, \text{Age}_i, \text{Openness}_i)^T
\end{aligned}$$

Parameters are estimated via Maximum Likelihood Estimation (MLE) (Likelihood is shown in Equation 3).

$$L = \prod_{i=1}^{n} \prod_{u=1}^{68} \prod_{v=u+1} y_{i,u,v}^{\text{logit}(\eta_{i,u,v})} (1 - y_{i,u,v})^{1-\text{logit}(\eta_{i,u,v})} \tag{3}$$

Results show that the Euclidean distance between regions remain a significant estimator throughout the models. This is reasonable since we would expect that the closer the regions are, the more likely they are connected. An interesting observation is that subject's age turns out to be a significant estimator – given two regions, the older a subject is, the more likely there is a connection. The ages of subjects in this study range from 18 to 29, and the observation is supported by a study that shows one's brain continues to mature with incresing degree of connectivity in one's mid-20s. Last but not the least, openness is found to be significant. A possible explanation could be that curiosity acts as a stimulus to neuro connection. Detailed summary of parameter estimations can be found in Appendix Table 1.

Assessment

The Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC) serve as promary measures of fitness. Both criteria describe a model's fitness in terms of the compromise between its accuracy and complexity. In particular, the BIC punishes free parameters more seriously than the AIC. As indicated by the criteria values, the third model has the minimal AIC and BIC, suggesting that it is the best fit among the three. Detailed summary of AICs and BICs can be found in Appendix Table 2.

We also used 10-fold cross-validation to assess performance of the models. Subjects are divided into training set, which contains approximately $\frac{9}{10}$ of all subjects, and test set, which contains the rest $\frac{1}{10}$. Paramaters are estimated using training set, and the models are then fit to test set. Area Under Curve(AUC)s are recorded and used as the standard for comparison, since it is invariant against class skew. Results show that all three models have similar performance with average training AUC at 0.83 and average test AUC at 0.82, suggesting fair accuracy. Detailed summary of AUCs can be found in Appendix Table 3.

Discussion

Several limitations exist in our model that could potentially be improved upon. For instance, distance between brain regions could be measured in latent space instead of Euclidean space. Indeed, brain regions in close proximity but in different brain hemispheres may vary to a greater degree than those a similar distance apart in the same hemisphere. Moreover, the analysis does not take into account the number of connections between regions - simply their existence. A more extensive analysis may attempt to weight connections based on the relative number of connections available.

Secondly, an alternative Bayesian analysis may provide a more appropriate fit. The current model with Openness, though more appropriate than the other models examined, still retains a fairly high AIC and BIC. Through the use of other models and methods, it may be possible to improve upon fitness to better model the data at hand.

Appendix

Table 1: Parameter Estimation

Model	Parameter	Estimate	SE	p-value
1	intercept	2.39	0.01	< 0.001
	distance	-0.36	0.00	< 0.001
2	intercept	2.23	0.04	< 0.001
	distance	-0.36	0.00	< 0.001
	sex	0.02	0.02	0.06
	age	0.01	0.01	< 0.001
3	intercept	2.05	0.05	< 0.001
	distance	-0.36	0.00	< 0.001
	sex	0.01	0.01	0.16
	age	0.01	0.00	< 0.05

Table 2: AIC and BIC

Model	AIC	BIC
1	217266	217287
2	217255	217296
3	217207	217259

Table 3: Average AUC of Training and Test Sets

Model	training	test
1	0.8281	0.8310
2	0.8281	0.8309
3	0.8282	0.8308