NHANES Actigraphy Singular Spectrum and Regression Analysis

Analyzing human activity is a complex problem for researchers to decipher. Many researchers in clinical informatics have recently been using actigraphy monitoring, which is a non-intrusive and continuous measurement of human activity and movement, to evaluate sleeping patterns, sleeping quality and health problems related to sleep.[1] This paper aims to analyze how the demographics and risk factors of an individual can impact the actigraphy. In particular, since age and gender are significant demographic factors and BMI and total cholesterol are pivotal risk factors, we include these variables into our model to analyze the actigraphy. There are five sub-datasets from the NHANES 2003-2004 data used in this study: actigraphy data, demographic data, blood pressure data, total cholesterol data and body measures data.

To pre-process the data, we first subsetted the focus on the first 12 hours of the actigraphy data for each of the 7160 patients. Subsequently, we performed singular spectrum analysis (SSA), in which we derived a Toeplitz matrix from the autocorrelation for each patient and performed eigendecomposition which places the 720 eigenvalues of actigraphy in decreasing order. Finally, we merged the eigenvalues with the other four datasets with sequence number (SEQN) being the unique identifier.

To understand the decreasing trend of sorted eigenvalues for each patient, we plotted the first 30 SSA eigenvalues for 250 randomly chosen patients in Figure 1. We can see that the first eigenvalue differs tremendously among the patients; the value ranges between approximately 0 to more than 500. Since the SSA eigenvalue shows how chaotic and heterogeneous is the patient's activity, the results suggest that some patients have extremely regular daily activity (low eigenvalue), while some patients have very chaotic activity (high eigenvalue). On the other hand, after the 15th eigenvalue, we see that all 250 patients have stabilized at low eigenvalues. This result indicates that patients tend not to persistently take part in chaotic activities.

To quantitatively model the relationship between the heterogeneity of patient's demographics and the SSA eigenvalue, we first perform five linear regressions, with age and gender being the predictors and the third to seventh eigenvalues, namely, s_3, s_4, s_5, s_6, s_7, being the response. It is because Figure 1 shows that the first two eigenvalues are heavily right-skewed, making it plausible that using respectively the third to seventh eigenvalues as response to the demographic variables will give more informative findings. Therefore, for $i = 3, 4, 5, 6, 7$, the linear regression models are as follows:

$$s_i = \beta_0 + \beta_1(age) + \beta_2(gender)$$

Since the regression coefficients of all five models are statistically significant ($p < 0.01$), Table 1 should demonstrate meaningful regression results. Specifically, both age and gender have negative coefficients, indicating that people tend to follow more regular activity schedules as age increases, and females are less diverse in daily activity on average. However, since the R^2 and adjusted R^2 are smaller than 0.03 in all five models, the demographic data do not effectively capture the variability of the SSA eigenvalues appeared in the model. Therefore, it is reasonable to also include the risk factors of patients. We particularly chose total cholesterol, body mass index (BMI) and systolic blood pressure (SBP) since they represent partially the health status of patients. We refit the linear regression models with total cholesterol, BMI and SBP included:

$$s_i = \beta_0 + \beta_1(age) + \beta_2(gender) + \beta_3(Cholesterol) + \beta_4(BMI) + \beta_5(SBP)$$

Including the risk factors into the linear model yields more interesting results. Unlike the previous models that all predictors are statistically significant, only age, gender and BMI are significantly influencing the actigraphy eigenvalues when including the risk factors. Since BMI is defined as $BMI = \frac{Weight}{Height^2}$, the negative regression coefficient of BMI in turn suggests that individual's body fatness, rather than the level of blood pressure, is influential in reducing the diversity of individual's activity on average. One possible interpretation for this situation is that when patients suffer from obesity, they are less likely to involve in vigorous physical activities such as basketball game, tennis singles and hiking.[2]

In terms of goodness of fit, it is unfortunate that including the risk factors only slightly increased the R^2 from roughly 0.03 to 0.05, which is still unreasonable to say that linear regression can effectively model the relationship between SSA eigenvalues and patient's demographic data and risk factors. We also checked the variance inflation factor (VIF) for the five predictors. The results show that the VIF for all variables are less than 2, which suggest that there is no indication of multicollinearity within the variables.

In conclusion, linear models have difficulty to model the variability of actigraphy data based only on the demographic data and the risk factors. To capture the potentially non-linear relationship between the predictors and SSA eigenvalues, there are future opportunities to build complex models such as b-spline and generalized additive model with specification of smoothing methods. For the predictors, it is also conducive to consider patients' demographics such as race, education level and residential community and nutrient information because these factors may also be significant to represent the large spectrum of the U.S. population.

Reference:

1. Peters, B. (n.d.). Fitness Trackers: Discover How Actigraphy Detects Sleep-Wake Patterns. Retrieved December 17, 2018, from https://www.verywellhealth.com/what-is-actigraphy-3015130
2. Examples of Moderate and Vigorous Physical Activity. (2017, May 08). Retrieved December 19, 2018, from https://www.hsph.harvard.edu/obesity-prevention-source/moderate-and-vigorous-physical-activity/

Appendix:

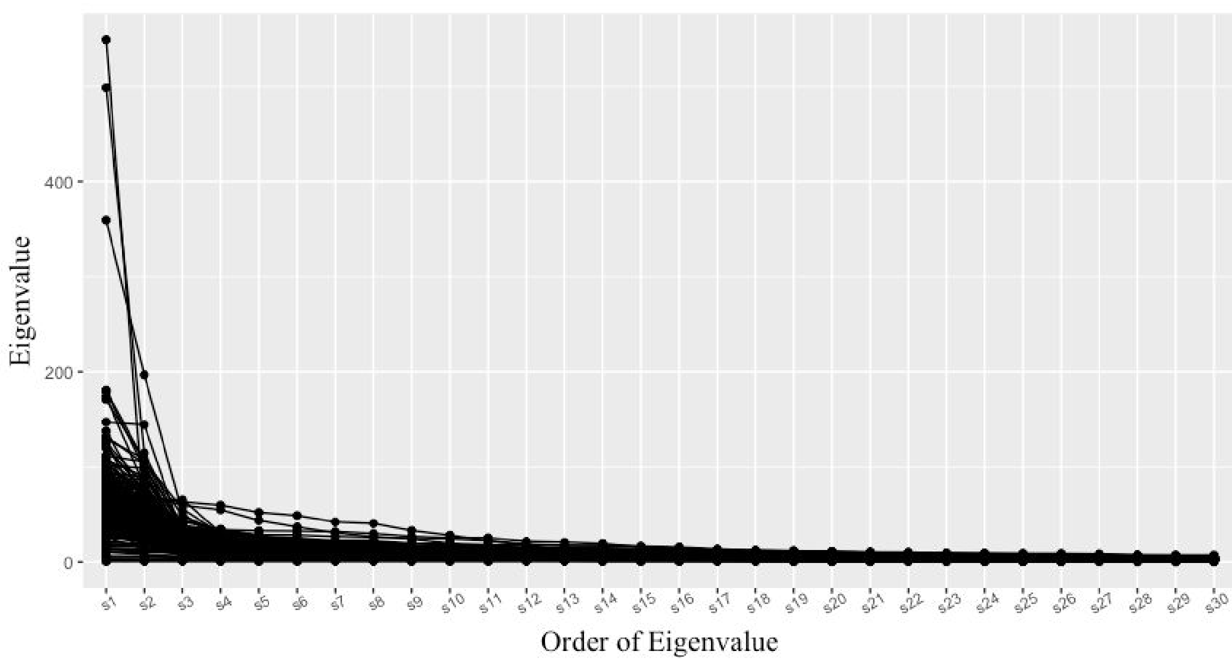

Figure 1: The decreasing trend of top 30 eigenvalues for randomly chosen 250 out of 7160 actigraphy patients

	Linear Model s3	Linear Model s4	Linear Model s5	Linear Model s6	Linear Model s7
(Intercept)	25.2188	20.2868	16.1851	14.0731	12.3446
RIDAGEYR	−0.0506	−0.0355	−0.0288	−0.0232	−0.0198
RIAGENDR	−2.3213	−1.9028	−1.4634	−1.2137	−1.0400

Table 1: Outputs of simple linear regression model with patient's age and gender as predictors and respectively the third, fourth, fifth, sixth and seventh actigraphy eigenvalue as response

	Linear Model s3	Linear Model s4	Linear Model s5	Linear Model s6	Linear Model s7
(Intercept)	28.3173	24.5215	19.8415	17.6084	15.5145
RIDAGEYR	−0.0327	−0.0139	−0.0133	−0.0088	−0.0067
RIAGENDR	−2.1474	−1.6919	−1.2937	−1.0152	−0.8689
BPXSY1	−0.0157	−0.0149	−0.0093	−0.0070	−0.0060
BMXBMI	−0.1187	−0.1492	−0.1300	−0.1300	−0.1149
LBXTC	0.0052	0.0022	0.0005	−0.0003	−0.0007

Table 2: Outputs of simple linear regression model with patient's age and gender, systolic blood pressure, body mass index and total cholesterol as predictors and respectively the third, fourth, fifth, sixth and seventh actigraphy eigenvalue as response