CS-C4100
Digital Health and Human Behavior
Fall 2024

Interaction Analysis with Apple Watch Data

Table of contents

1 Introduction

Achieving sufficient, high-quality sleep can affect both physical and mental health. Adequate sleep has been proven to improve areas such as cardiovascular health, hormone regulation, and immunity. [1] Consequently, the motivation to study the quality of sleep is evident, due to major effects in multiple functions of the human body.

Human sleep consists of five stages, which include being wake, and a sleeping cycle. The sleeping cycle can be divided into two different stages: rapid eye movement (REM), and non-rapid eye movement (NREM), that can be further divided into three phases: N1, N2 and N3. The time spent in each cycle varies between individuals, and is altered by factors such as age, circadian rhythms, medication or depression. For a healthy individual, sleep of optimal quality is composed of approximately 75% of NREM-sleep, divided across N1, N2, and N3 phases as 5%, 45% and 25% respectively, and 25% of REM-sleep. On average, the cycle repeats 4 to 6 times per night. [2] Figure 1 visualizes the transitions between the five stages during sleep.

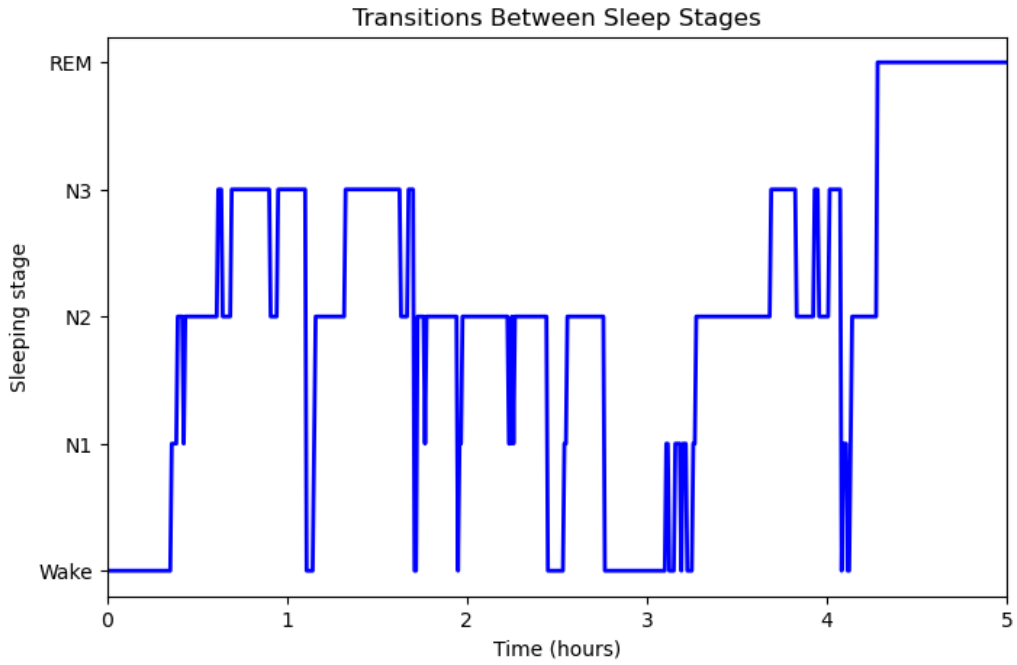

Figure 1. Transitions between sleep stages of an individual from hour 0 to hour 5 of a sleeping opportunity. The data is obtained from [3].
The data corresponds to measurements for participant id 2638030.

As discussed in article [2], the quality of human sleep can be studied in multiple ways. For clinical purposes, the most applied method is polysomnography (PSG), which involves monitoring multiple physiological signals of a sleeping individual. Clinical observation of the output of PSG allows for the diagnosis of conditions such as various sleep-related breathing disorders and narcolepsy. In addition, sleep difficulties have been negatively associated with well-being and quality of life. Consequently, the article emphasizes on the clinical significance of understanding the physiology and pathophysiology of sleep. [2]

The development of integrated health monitors in smart devices, such as smart watches and phones offer a new direction for studying sleep. As a result, a study [3]

was conducted to investigate the possibilities of the metrics obtained from smart devices. The goal was to evaluate the descriptiveness and prediction quality of health data recorded by a smart watch. In the study, physiological health data was observed from 39 individuals participating in the study. The model evaluation was conducted for multiple models, with neural nets producing the best results. The mentioned method attained an accuracy of 90% for sleep-wake classification, and 72% for classifying between wake, NREM and REM. The results were further generalized by evaluating the accuracy of the models on unseen data from a different study, Multi-ethnic Study of Atherosclerosis (MESA). For the MESA data set, the model was able to produce results comparable to the test set from the original smart watch data. [3]

Naturally, the next step proposed in the study [3], is to expand the model for multiple sleep stages. While the accuracy of 90% for sleep-wake predictions by the model in presented [3] is promising, the accuracy drops noticeably dividing sleep to NREM and REM stages. As suggested, our attempt in this project was to build a model for predicting up to 5 stages of sleep. The model was trained and evaluated on the same data set as in [3]. Sleep, as a highly complex phenomenon, may require more complex models to accurately depict sleep and predict all the possible stages. A proposed solution by [3, 4] is to expand the models to incorporate the probabilities of sleep stage transitions.

In paper [5], hypnograms were analyzed using a seven-state continuous-time Markov model (CTMM). The aim was to study the complex behavior of sleep stage switching, which common analysis methods regularly fail to model. The purpose of the CTMM was to model the time-dependent sleep stage transitions and distinguish three types of wake stages. The model fitted on MESA data set with the CTMM incorporated allowed the differentiation of sleep architectures in the data set population. Further implying the complexity of sleep, physiological factors such as age and sex showed distinctive patterns on the sleep stage transitions. Sleeping disorders, such as restless sleep and insomnia also proposed varying sleeping patterns. [5]

This project work was directly motivated by the analysis conducted in [3-6]. Additionally, the importance of accurate sleep stage modeling emphasized on [2] further drives motivates to conduct predictive analysis on sleep. As a result, a regular logistic regression model, and custom model with incorporated transition probabilities were constructed. The goal was to determine the capability of logistic regression in predicting up to 5 distinct stages. The custom model was built to evaluate whether a simple Markov chain, determined from the general sleep patterns of individuals, could improve the results. Additionally, the correlations between physiological measures were investigated. The physiological measures consisted of heart rate, motion and sleep stage recordings, obtained from [3, 4].

After training and fitting the model on the data set, the logistic regression model had an accuracy of 49%, 63%, 70% and 92% for predicting 2, 3, 4 and 5 stages respectively. The custom model accuracies of 48%, 60%, 70% and 92%. The 4- and 5-stage results for both models were almost identical to those obtained in the original study [3]. From correlation analysis, the average heart rate variation had a correlation of 0.20 with sleep score. Average activity had a near zero to minimal negative correlation.

2 Problem Formulation

The aim of this project is to build a model for predicting wake, NREM-phases and REM sleep stage from physiological factors. As suggested in study [3], the predicted labels will include all 5 distinct sleep stages. The prediction will be based on heart rate variation and motion during sleep. In addition, to make the results comparable to those in the original study [3], the model will also be applied on predicting wake-sleep, and combined sleep stage settings. Consequently, the data set is the same as in the original study, with the exclusion of the separate testing data set.

In addition to the predictability of physiological variables, the direct correlation with sleep quality is of interest. In a study of the restless legs syndrome [6], leg movements during wakefulness were associated with increase in heart rate. The results determined a possible linkage between sympathetic activation and increased risk of cardiovascular complications. [6] As a result, the correlation of heart rate and motion measurements from [3, 4] with sleep quality is investigated.The sleep quality measure will be further discussed in section 3, Dataset Description.

The switching of sleep stages is commonly observed to follow a pattern, and to be highly time-dependent. [4, 5]. The pattern of sleep stage switching can also be observed in figure 1. However, sleep disorders or neuropathology can result in an irregular sleeping pattern. Proper modeling of the sleep stage switching and the underlying patterns can be advantageous for neuroscience, and dosing of sleeping medicine [2, 5]. As a result, and as further suggested in [3], the attempt is to model the transitions between stages by introducing a transition probability matrix as priori, data-driven knowledge on a custom model.

3 Dataset Description

The data set originates from a study conducted in [3, 4]. 39 subjects from the University of Michigan participated in the study. The mean age of the participants was 29.4 years, with a standard deviation of 8.52 years. Before the measuring phase took place, the participants were screened through a questionnaire, which ensured the absence of a diagnosed sleep-affective disorders, such as insomnia, cardiovascular diseases, restless legs syndrome, or sleep-related breathing disorders. Individuals that had traveled across more than two time zones within the last month, or worked night shift regularly, were also excluded. [3]

The study began with an ambulatory recording period, which lasted for 7 to 14 days. In the beginning of the period, the participants were given an Apple Watch (Apple inc.), which was used to measure the physiological data from the participant's wrist. The 7- to 14-day ambulatory period was used in advance to build the circadian rhythm of each individual and was afterwards used as a clock proxy feature in the original study. The ambulatory period was followed by an 8-hour sleep opportunity, in which the subjects underwent polysomnography (PSG) in a sleep lab, while simultaneously wearing the Apple Watch. During the opportunity, the PSG recorded the sleep stages while the Apple Watch tracked motion and heart rate in real-time. [3]

The motion was measured as vectors of acceleration in the x, y and z directions by accelerometers in units of g (9.81 m/s^2). The heart rate was measured by the Apple watch using photoplethysmography (PPG) in beats per minute. The sleep stages are obtained from PSG, which integrates brain, physiological and physical activity to discrete stage labels [5]. The measurements were conducted on a total of 39 individuals. Due to sleep affecting conditions, an individual's measurements were discarded due to an indication of REM-sleep behavior disorder, and further 3 due to signs of obstructive sleep apnea. In addition, 4 more subjects were excluded due to incomplete data. In case if the battery ran out during the PSG, the data was cropped to only the valid data points, which resulted in varying sample sizes for some individuals. [3]

The features were extracted from the measurements by applying a pre-processing script to the raw data, obtainable from [3, 4]. With the pre-processing script, the raw data is transformed to features representing 30-second epochs, with the 30-second timestamps saved in feature "time". For each scored epoch, a local window of 10 minutes is cropped around the epoch. The acceleration vectors are transformed into activity counts, indicating motion in a time window. For heart rate, the samples were interpolated, smoothed, and finally filtered to amplify periods of high change. Finally, the measurements are normalized, and the standard deviation in the 10-minute window around the epoch was used as the sample, which indicates the variation in heart rate. The normalization was done by dividing the values with the 90th percentile of the absolute difference between consecutive measurements and the mean heart rate across the whole sleeping opportunity. [3]

After applying the pre-processing script to the features, the data set consisted of 4 features: "hr", "activity count", "time", and "cosine" the label "psg labels" for each individual. The feature "cosine" represents the clock proxy feature, which was applied in the original study as prior information of an individual's circadian rhythm. [3] The clock proxy feature was excluded, due to the inclusion of data-driven prior information in the form of probability transition matrix. Consequently, the effect of pre-processing script on clock proxy feature is not further elaborated.

The variable "psg labels" contained values from 0 to 5, instead of 0 to 5 with 4

excluded. The PSG measurement of “4” indicates the NREM-phase N4, which is commonly included in the stage N3 [7]. Similarly to the “psg_labels” variable in the original study, we only applied analysis on the values 0 to 5, with 4 excluded. Therefore, the PSG labels representing NREM-phase N4, included in N3, and consequently transformed to value 3.

Additionally, a new variable, “sleep_score”, was introduced for correlation analysis. The sleep quality score was based on the recommended amounts of N1-, N2-, N3-phases and REM-sleep, proposed in [2]. For each participant in the study, a sleep quality score of 0 – 100% was computed individually. The score was constructed as the sum of the percentage value of each stage sleep divided by the respective guidelines. For excess sleep of certain stage, the optimal value was chosen instead.

After the pre-processing steps, one more individual was removed due to only 118 recordings. After pruning, the final data set consisted of 24948 rows. The rows were formed of recordings from 30 individuals. The minimum number of recordings for an individual was 415, and maximum 978, with an average of 845 and a median of 941 recordings. The variables consisted of 3 features: “hr”, “activity_count” and “time”, a label “psg_labels” and the “sleep_score” for correlation analysis. There were no missing features or label values in the data.

4 Methods

The methods -section introduces the chosen machine learning model. The section also investigates the construction of the transition probability matrix, and the procedure of nested k-fold cross-validation. Finally, the section will shortly present correlation as a simple statistical analysis tool.

4.1 Multinomial Logistic Regression

Logistic regression is a common classification technique in machine learning. The basis of the method is the logarithm of the odds ratio, which refers to the odds of two events occurring relative to each other. The odds ratio can be transformed to probabilities with the logit function. The output class of logistic regression is the class with higher probability. [8]

For the purpose of this project to predict multiple classes, we will be applying multinomial logistic regression. Similarly to binary logistic regression, the multinomial version observes the probabilities of the output being a certain category. In the multinomial model, relative risk ratios are applied over odds ratios. [8]

Multinomial logistic regression was applied with class LogisticRegression from skicit-learn [9]. The class LogisticRegression applies a regularized logistic regression, with the one-versus-rest (OvR) scheme for multiclass classification. With OvR, each label class is in turn treated as the correct class, while the other classes are treated as false. The prediction of the method is based on the probabilities of different classes, computed with the class function “predict_proba”. The predict_proba function returns the probabilities of observing each class, for all rows of features. The prediction is based on the highest probability class. [9]

In addition to the traditional multinomial logistic regression, a customized version was also constructed. As the response of the traditional model is unordered, the model must be extended to model the pattern and sequential behavior of stage-switches. As a solution, transition probabilities were incorporated in the prediction phase of the logistic regression model.

The customized model was built by modifying functions “fit” and “predict” of the LogisticRegression class from skicit-learn. In addition, a new function for computing the transition matrix was also added. The custom “predict” function was modified to predict the labels sequentially according to the predicted probabilities. These predicted probabilities were adjusted by probabilities from transition probability matrix in each time bin. This resulted in a combined prediction by the prediction from the multinomial logistic regression, combined with the transition probabilities of the time-interval.

Logistic regression was chosen for this project, as the predictions are based on probabilities. Consequently, the inclusion of adjustive prior probabilities is achievable. In addition, the original logistic regression is computationally relatively easy, and fast to hyperparameter tune.

4.2 Transition matrix

To include the pattern, time-dependent nature of switching between sleep stages in the model, transition matrix is applied in the prediction phase of the custom logistic regression model. As described in [5], the stage-switches are modeled as a Markov chain. The Markov chain represents the probabilities of transitions from the current state [10]. Since the data is discretized into time intervals, we can apply the discrete form of the Markov chain.

A transition matrix is used to represent the Markov chain transition probabilities. The transition matrix will be constructed using information from the PSG measurements. With five possible sleeping stages: wake, REM, and NREM-phases, the transition matrix will be of the form 5x5. The rows of the matrix will represent the probabilities of transition from the current stage to the other stages. The elements of the matrix P[i, j], for i-th row, j-th column, represent the probability of transition from stage i to stage j. [10]

The transition probabilities will be directly assessed by the normalized relative frequencies of stage switches in the training data. For each sleep stage, the frequency of observing the next state will be treated as the probability measure. To further emphasis the time-dependency in the matrix, the "time" variable of the training data is split into T intervals. The transition matrix will be computed individually for each interval, as the transition probabilities should vary with respect to time, due to patterns of the transitions [5]. The transition probabilities will be computed for each individual separately. For each time interval, the transition matrices of individuals will be summarized and finally normalized. The accuracy of the custom model is evaluated with varying values of T.

4.3 Cross-validation

For a model to predict with high accuracy on unseen data, the model should be trained with a large and diverse enough data set. By increasing the variability in the training data set, we decrease the chance of overfitting on training data, increase robustness, and decrease the generalization error. [11] Due to small amount participants in the study, we will be applying nested n-fold cross-validation on the data set to increase its relative size. Notably, the splitting of the data set will be done with respect to the participants.

In nested n-fold cross-validation, the original data is split into n-folds. For each fold in is used as the test set, and n-1 folds are used as the training set. In nested cross-validation, the training set is further split into k-folds. From the k-folds, one set in turn is used as a validation set to determine the optimal hyperparameters, whereas the k-1 folds are used in training with varying hyperparameter(s). The outer test set is finally evaluated with a model trained on the outer training set, and the best hyperparameter value(s) obtained from the inner loop. In both the outer and inner loops, the accuracy of the model is computed as the average accuracy across the folds.

4.4 Correlation

In addition to the advanced machine learning method, correlation was applied as a statistical method to study subject-level dependencies. Precisely, the correlation method applied was Pearson's linear correlation, which measures the linear correlation between variables. Pearson's correlation was chosen due to being easily interpretable, all thought incapable of capturing non-linear dependency. [12]

The linear correlation between variables is measured by a correlation coefficient. The coefficient is a value between -1 and 1, where the value of 0 indicates no linear correlation, and absolute values close to 1 indicating a strong linear correlation. The empirical correlation matrix was calculated using a function "corrcoef" from NumPy [13]. The correlation matrix returned by the corrcoef -function is according to the equation

$$R_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}, \tag{1}$$

where C denotes covariance, and R_{ij} is the Pearson sample correlation between variables x_i and x_j. [13]

5 Results

Both logistic regression models were hyperparameter tuned in a nested k-fold cross-validation loop, with n = 5 outer splits, and k = 4 inner splits. The inverse of regularization strength parameter "C" in skicit's LogisticRegression was tuned with values 100, 10, and 1. The penalty of "l2" was applied across all the models, while other parameters were kept as default. [9] When applying the custom logistic regression model, the time binning was done with T bins, with T varying from 2 to 16. In addition, the model was also applied on the combined sleep stage settings. Figure 2 shows the relation of testing accuracy of the custom model with respect to the amount of bins T. Table 1 summarizes the accuracies for the regular and best-performing custom logistic regression models, across varying amounts of unique sleep stages.

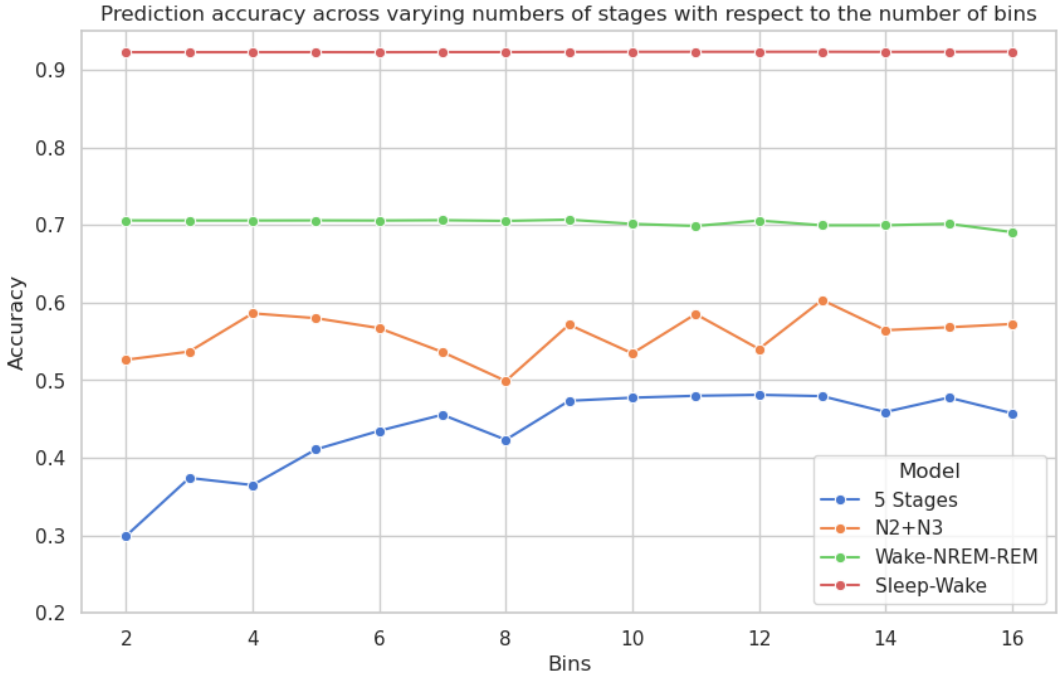

Figure 2. Custom logistic regression model test accuracy with varying number of time-interval bins.

Table 1: Test accuracies across varying amounts of labels. Only the best custom model is included.

Model	Stages (labels)	Accuracy
Regular logistic regression	wake, N1, N2, N3, REM	0.49
Custom logistic regression (T = 13)	wake, N1, N2, N3, REM	0.48
Regular logistic regression	wake, N1, N2 + N3, REM	0.63
Custom logistic regression (T = 13)	wake, N1, N2 + N3, REM	0.60
Regular logistic regression	wake, N1 + N2 + N3, REM	0.70
Custom logistic regression (T = 13)	wake, N1 + N2 + N3, REM	0.70
Regular logistic regression	wake, sleep	0.92
Custom logistic regression (T = 13)	wake, sleep	0.92

Pearson's correlation was computed for variables "hr", "activity count" and "sleep_score". For each individual, the mean heart rate variation and activity count were computed. Therefore, Pearson's correlation was evaluated between the sleep scores and the means of the variables. Figure 3 visualizes the heatmap of correlations between the 3 variables.

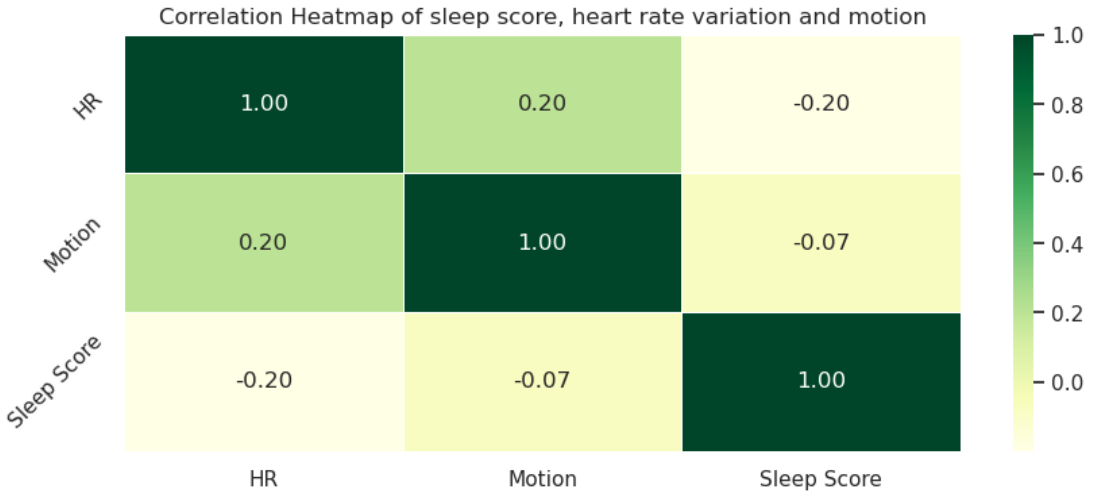

Figure 3. Heatmap of correlations between variables.

6 Conclusion & Discussion

In the original study, the neural nets had an accuracy of 90% for sleep-wake classification, and 72% for 3-stage-classification [3]. These results are almost identical to the 92% and 70% accuracies obtained from the 2 models. However, for 4- and 5-stage classification, the accuracy of both models dropped significantly. The lower accuracy for multiple classes could be explained by the thin border between the different NREM-phases. Applying only two features, heart rate and motion might be incapable of capturing the precise physiological differences of NREM-phases.

By comparing the results between models from table 1, it can be determined that the best-performing custom model produced almost identical results to regular logistic regression. A possible cause could be that the adjusted probabilities in the custom model changed only slightly compared to the original model. For further testing, including transition probabilities from a larger population could produce better results. Additionally, the application of a more complex form of Markov chain could be beneficial. For the custom model, the accuracy of 4- and 5-stage classifying task increased slightly with respect to increasing number of bins. However, for classification of fewer stages the influence of the bin size on performance of the model decreased.

In the hyperparameter tuning stage, the pool of possible parameters was relatively small. During implementation of the custom model, varying parameters for “penalty”, “solver”, and larger range for “C” of the class LogisticRegression from skicit were investigated. However, as the implementation of the custom model was computationally heavy to evaluate, some hyperparameters were set as constant. For similar reasons, the original choice of applying support vector machines was abandoned.

The results for the correlation analysis were -0.20 for heart rate and sleep score, -0.07 for motion and sleep score, and 0.20 for heart rate and motion. As intuitively predicted, the average heart rate and motion show linear correlation across the 30 individuals. Additionally, heart rate has a negative correlation of -0.20 with sleep score. The negative correlation suggests that subjects with greater heart rate variation may experience suboptimal amounts of sleep. Correlation value of -0.07 between average motion and sleep score imply no linear dependency between motion and sleep score.

The implementation of the sleep score could result in improper correlation analysis, as the score was constructed by comparing the frequencies of sleep stages to the recommended guidelines. With this implementation, the excess time spent sleeping in each stage was discarded. While the different sleep stages have varying effect on the human body, [2], a better solution would be weight the excess time, rather than discarding the excess time. This results in inaccuracies of the measure, especially in situations where the PSG labels fluctuate quickly.

For future work, the generalization of the model on data such as the MESA data set is of interest. Also, the addition of more explanatory variables could improve the distinction of stages such as N2 and N3. Due to time-related issues, a more complex hidden Markov model (HMM) was discarded. Implementing a more complex Markov method to model time-dependence could produce better results. In addition, predictions of the sequential sleep metrics could be improved by applying a more complex, time-series based method.

References

[1] Baranwal, N., P.K. Yu, and N.S. Siegel, *Sleep physiology, pathophysiology, and sleep hygiene.* Progress in Cardiovascular Diseases, 2023. 77: p. 59-69.

[2] Patel AK, R.V., Shumway KR, et al., *Physiology, Sleep Stages.* StatPearls, 2024.

[3]. Walch, O., Y. Huang, D. Forger, and C. Goldstein, *Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device.* Sleep, 2019. 42(12): p. zsz180.

[4] Walch, O., *Motion and heart rate from a wrist-worn wearable and labeled sleep from polysomnography.* PhysioNet, 2019.

[5] Jacobs, J., Caitlin E. Martin, Bernard Fuemmeler, and Shanshan Chen., *Profiling the Sleep Architecture of Ageing Adults Using a Seven-State Continuous-Time Markov Model.* Journal of Sleep Research n/a, no. n/a (n.d.): e14331., 2024.

[6] Benbir Senel, G. and D. Karadeniz, *Spectral electroencephalographic changes and heart rate variability accompanying leg movements during suggested immobilization test in patients with Restless Legs Syndrome.* Sleep Medicine, 2022. 100: p. S233.

[7] Gardiner, C., et al., *The effect of caffeine on subsequent sleep: A systematic review and meta-analysis.* Sleep Medicine Reviews, 2023. 69: p. 101764.

[8] Hilbe, J.M., *Logistic regression models*. Chapman & Hall/CRC texts in Statistical Science Series. 2009, Boca Raton, Florida ;: CRC Press.

[9] Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 2011. 12: p. 2825--2830.

[10] Brooks, S. and ProQuest, *Handbook for Markov chain Monte Carlo*. Chapman & Hall/CRC handbooks of modern statistical methods. 2011, Boca Raton, Fla: Taylor & Francis.

[11] Mohri, M., A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning, Second Edition*. 2018, Cambridge, UNITED STATES: MIT Press.

[12] Murphy, K.P., *Machine learning: a probabilistic perspective*. 2012: The MIT Press.

[13] Harris, C.R., et al., *Array programming with NumPy.* Nature, 2020. 585(7825): p. 357-362.