Functional Analysis of Wearable Data

Lilia Chang[1] and Bora Uyumazturk[2]

[1] Institute of Computational and Mathematical Engineering, Stanford University, lilia.chang@stanford.edu
[2] Department of Computer Science, Stanford University buyumazturk@stanford.edu

Abstract. In recent years wearable device data have become more prominent, with numerous studies illustrating their ability to continuously and reliably track various metrics relevant to human health [5]. However, methods of analyzing the continuous data still rely on primitive summary statistics. Such methods can fail to accurately model important biomarkers such as average resting heart rate [4][7] . We attempt to illustrate the promise of functional data analysis techniques for wearable data by using them to improve predictive performance for resting heart rate. We apply various methods, namely, k-means, principal components analysis, non-negative matrix factorization, and the Discrete Fourier Transform on the daily and weekly patterns in the data. The DFT coefficients outperform the other methods as linear predictors of resting heart rate, as measured by mean absolute error estimated via leave-one-out cross validation. We also find that a weekly scale is more effective for modeling variation in resting heart rate than a daily one.

1 Introduction

The measurement of basic physiological parameters requires health care visits for the majority of the population. These visits are mostly infrequent, and so by the time a significant change in health is detected, it may be months after the condition's conception. Biosensor wearable devices have emerged in the last decade as a cheaper, more efficient alternative to in-person health care for basic tasks such as achieving accurate, basic metrics. Additionally, continuous monitoring through wearables provides the opportunity to better explain differences in individual health outcomes.

However analysis of wearable data thus far mainly relies on simple summary statistics like the mean and standard deviation. For our project we explore whether the continuous features derived from wearable sensors can be used to explain variation in individual health outcomes, specifically average resting heart rate: a clinically relevant biomarker, the variation of which is poorly explained by raw summary statistics. Li, et. al. [7] show that there is loose correlation between the average number of steps per day and the average heart-rate, with R2 of 0.12.

Using the dataset provided by Li, et. al. we try various ways of capturing the functional characteristics of the activity levels. To validate the methods we use them in a simple linear model to better predict resting heart rate. Out of the numerous methods we tested, linear models fit on the magnitudes of the Fourier coefficients outperform other more adaptive matrix factorization techniques, achieving a lower mean absolute error (MAE) in leave-one-out cross validation. As a secondary observation, we find that analyzing the data at a weekly scale is more effective for predicting resting heart rate than the daily scale.

2 Data

We rely on the dataset provided by Li, et. al, that includes per-second accelerometer measurements of 43 individuals. The subjects are observed over varying time windows using a Basis wearable device. The individuals are aged between 35 and 70 years old and are observed between a few weeks and 2 years each. The per-second measurements on each person include the `accel_magnitude` and heart rate; the static data include the persons average resting heart rate, age, gender, and other basic demographic info. See Table 1 for summary statistics.

Table 1: Summary statistics of data.

	Age	BMI	Resting HR	Skin temp.	Steps per day	Days observed
Number of records	38	36	43	43	43	43
Average	55.76	29.36	72.1	89.19	5226.99	152.3
Standard Deviation	10.02	4.83	6.75	1.88	2435.66	117.53

The number of observations we have per person is on the order of millions and so it is necessary to aggregate the data prior to doing analysis. Li, et. al. deal with this issue by taking the average activity (`accel_magnitude`) per hour per day.

However there is good reason to believe that taking the average per hour per day removes useful variance within people's schedules. We demonstrate this observation in Figure 1. We also show the range in daily average activity levels and the median weekly activity levels in Figures 2. As such we work with the time-series data of the observed acceleration (also referred to as "activity levels"), averaged over both per hour, per day and per hour, per week.

2.1 Preprocessing

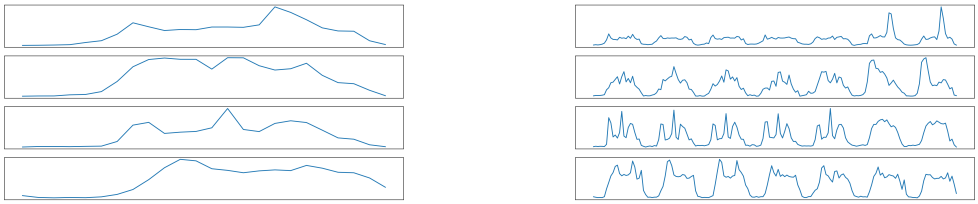
Due to the limited observation time for some individuals, there were a significant amount of missing values after taking the mean value per hour of the week. We elected to do complete case analysis, since simple imputation schemes such as forward filling, backward filling, or mean imputation would likely interfere significantly interfere with the shape of the activity curves, corrupting our results. Complete case analysis reduced our effective dataset size to 32 individuals for at the weekly scale (at the daily scale we had observations for all 43 individuals).

For the methods which take place in the time-domain, alignment of each person's time series can significantly impact results. For example, two individuals whose routines are identical might seem very different (if one were to take the $L2$ distance) simply because one tends to wake up later. To mitigate this, we aligned the time series by the effective start of the day for each individual. To do this, we first approximated the second derivative of their activity using finite differences, and then found the first hour after midnight whose second derivative exceeded a threshold which chose manually.

3 Methods

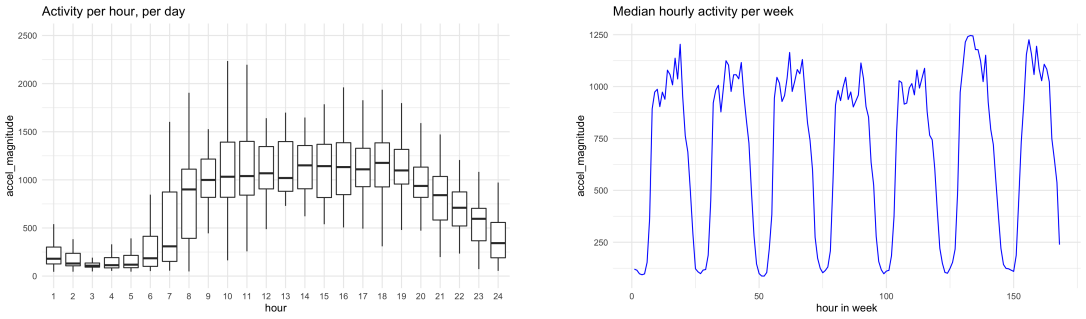
The framework for our analysis is as follows.

1. Consider the activity-level matrix A, for which A_{ij} has the ith person's activity levels at time j.
2. Factorize A to get personalized coefficients of "basis" activities for each person. That is, if $f_i(t)$ is person i's activity at time t, we estimate $c_{i,j}$'s such that,
$$f_i(t) = c_{i,1}\phi_1(t) + ... + c_{i,k}\phi_k(t) \tag{1}$$
3. Regress average heart rate on coefficients in addition to s_i, the average steps per day for person i (heart rate: Y, regress Y on c_i's)
$$y_i \sim s_i + c_{i,1}\beta_1 + c_{i,2}\beta_2 + ... + c_{i,k}\beta_k \tag{2}$$
4. Interpret results.



(a) Daily averages (b) Weekly averages

Fig. 1: Average daily and weekly activity levels for the same four individuals. Notice the granular periodicity in the weekly average curves in contrast to the flattened daily curves.



(a) Daily averages (b) Weekly median

Fig. 2: Box plot of daily hourly activity and median weekly, hourly activity.

We discuss the various factorization methods we use.

3.1 Functional factorization methods

Discrete Fourier Transform (DFT) The Discrete Fourier Transform (DFT) represents a signal $f(n)$ of length N as a linear combination of complex exponentials. Formally, it decomposes f as follows:

$$f(n) = \sum_{k=0}^{N-1} \langle f, w_k \rangle w_k(n),$$

where $w_k(n) = e^{2\pi \frac{ik}{N} n}$ and $\langle f, g \rangle = \sum_n f(n)\overline{g(n)}$ [1]. The scalar values $F_k = \langle f, w_k \rangle$ are called *Fourier coefficients*. They capture the presence of various frequencies in f. Note that w_k has period equal to N/k, a fact that will assist with interpretation in the discussion.

We hypothesize that the frequency information in the DFT can be helpful for understanding continuous activity data. However, in general F_k is complex valued, so in the final regression we use the magnitude of the F_k to get a single real number for each k. This has other advantages as well. Because the time shifting only effects the phase in the frequency domain, using the magnitudes of the DFT components make the representation invariant to time shifts. This resolves the issue of misalignment between time series, which can have a large effect on techniques which take place in the time-domain. Figure 3 shows the transformation of a single individual's average weekly schedule. Note that because the input is real value, the magnitude is symmetric, so in practice we use only the positive frequency components in our representations. We also note that since the 0-frequency component of the DFT is just the average value of the signal over the entire time domain, we omit that from our representation as well. We include boxplots of the Fourier coefficients of the examples in the dataset for both daily and weekly timescales in Figure 4.

3.2 Baseline methods

K-Means We run K-Means directly on the persons' observed, average hourly activity levels (over both days and weeks). That is, each row of the aforementioned matrix A is a data point. The regression of the heart rate is then done on the one-hot encoded vectors, with indices corresponding to each of the k classes. This baseline method was attempted by Li, et. al., who find that optimal k on the daily averages is 4. We do a more robust analysis using both the daily and weekly averages.

Principal Component Analysis (PCA) We run PCA on the persons' observed average activity levels as another baseline method. We take the principal k eigenvectors as the basis vectors and the coefficients on each of the eigenvectors as the input to the regression. Given that the size of the input data after taking averages is manageable, PCA makes sense as another baseline method to attempt.

Non-negative Matrix Factorization (NMF) NMF takes an input matrix and factorizes it as the dot product between two matrices W in $\mathbb{R}_+^{m \times k}$ and H in $\mathbb{R}_+^{k \times n}$ as shown in 5. The non-negative constraint allows for better interpretability for our setting and in general sometimes is more efficient [6][3]. Then w_i^T, the ith row of W, are the coefficients of person i on the components in H. We take W as the input matrix to the regression.

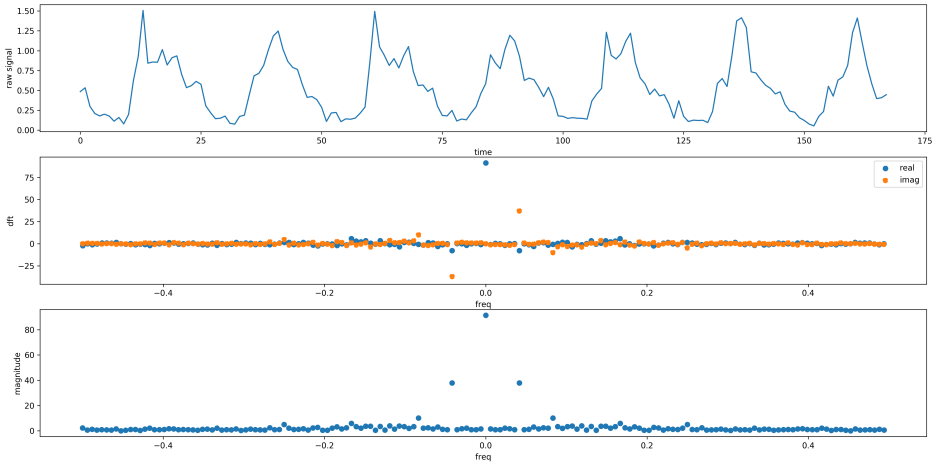
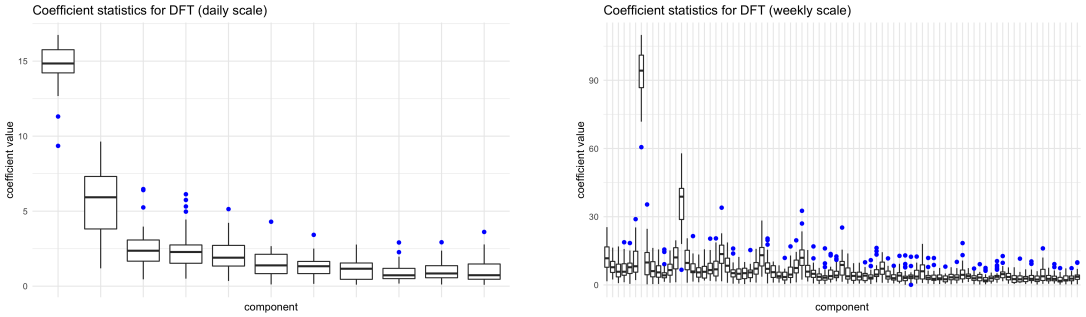


Fig. 3: A subject's raw weekly average signal and their DFT. We plot both real and imaginary coefficients in the second plot and plot the magnitude of the coefficients in the last plot.



(a) Daily timescale (b) Weekly timescale

Fig. 4: Coefficient statistics for DFT on both timescales.

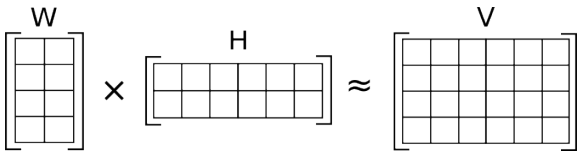
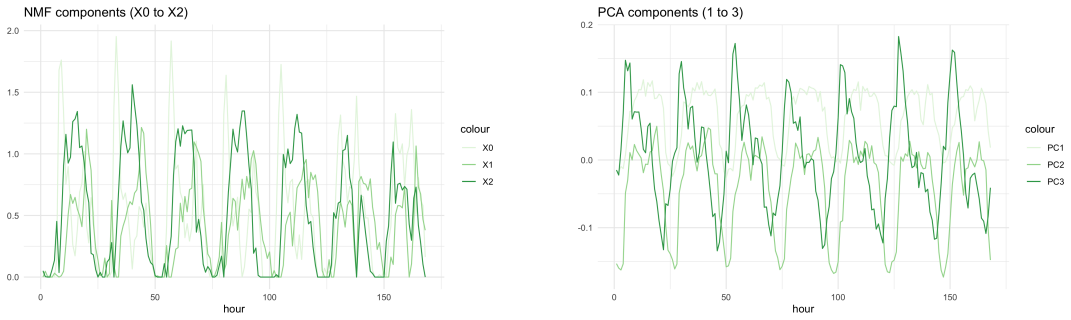


Fig. 5: NMF factorizes matrix V into the product of matrices W and H. In our setting, V=A.



(a) NMF (b) PCA

Fig. 6: Three most significant components from NMF and PCA on weekly timescales.

3.3 Predicting resting heart rate

To assess the value of the features extracted by the different methods, we include them in a regression on resting heart rate. In order to reduce confounding for overall activity, we include average amount of steps per day (following [7]) as a feature in the regression. We also normalize all time series to $L2$ norm 1 before applying matrix factorization to hide any magnitude information. This way we hope to assess the ability of the various techniques to extract valuable information about people's activity patterns apart from simply the amount of activity they experience on average.

3.4 Model Evaluation

In order to assess the quality of each factorization, we fit linear models to predict average resting heart rate and compared their performance using mean absolute error. We prefer MAE in this setting for ease of interpretability and robustness to outliers. R2 is less relevant in this context since we are interested in the generalization of the extracted features, and not goodness-of-fit to the training set. Estimating generalization error is difficult due to the limited number of example. The dataset contains records for only 43 individuals, some of whom are only recorded very briefly. As mentioned above, we applied complete case analysis, which made the final dataset sizes 42 for the daily scale and 32 for the weekly scale.

In order to compensate for the limited amount of data, we used a nested leave-one-out cross validation approach [2]. The hyperparameter we tuned was the number of components or clusters to use. The final performance of each method was estimated as follows: for each example, we created a development set consisting of all of the other examples. Then a hyperparameter was chosen using leave one out cross validation on this development set. Finally, the model was retrained on the entire development set and used to predict the original held out example. We repeated this for each example to get our predictions. In our results section we report the mean absolute error as well as 95% confidence intervals using standard errors and the t-distribution approximation. These results are presented in Table 2 and depicted visually in Figure 7.

For interpretation purposes, we also report the leave-one-out (unnested) MAE for each number of components for each model to see how generalization error changes as we add further flexibility to the model. We use a range of 0-12 components on the daily scale (since the DFT only has 12 components

to use), and a range of 0-24 components for the weekly scale to avoid over-parametrization issues (since there are only 32 examples). The results of this are shown in Figure 8. Finally, we fit a least angle regression (LAR) on the residuals of the regression of heart rate on average steps per day using the 24 DFT components at the weekly scale. In Figure 9 we plot the sample paths for the components to get a sense of which frequency ranges are most predictive.

Table 2: MAE for various methods and timescales

	Steps		**DFT**		**NMF**		**PCA**		**K-Means**	
daily	5.12	(4.00, 6.24)	5.48	(4.39, 6.58)	4.98	(3.77, 6.18)	6.13	(4.73, 7.52)	5.66	(4.57, 6.76)
weekly	5.28	(3.86, 6.69)	4.03	(2.62, 5.44)	6.52	(4.78, 8.26)	6.29	(4.62, 7.96)	6.47	(4.98, 7.95)

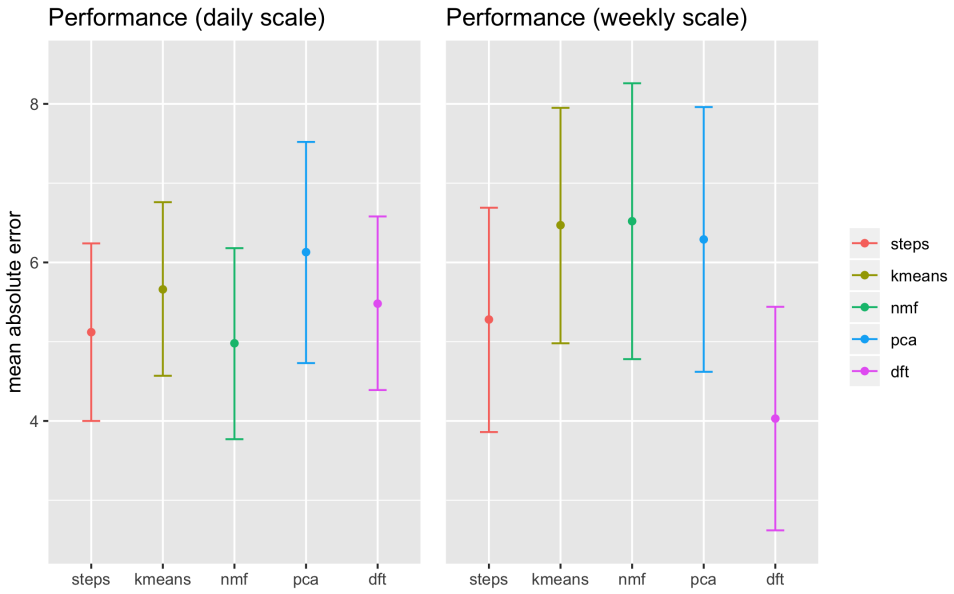


Fig. 7: MAE on the hold-out test set (steps refers to baseline with no matrix factorization features)

4 Results

While the limited sample size makes drawing clear conclusions difficult, DFT on the weekly scale has the best nominal performance (MAE 4.03 (2.62, 5.44)). It also the only method which seems to substantively improve upon the baseline regression using only average steps (MAE 5.12 (4.00,

6.24)). While these results seem promising, to confirm our hypothesis that DFT is better suited than other methods to extract functional characteristics of activity data we would need a larger sample size.

In line with our hypothesis that averaging over all days would remove useful variance, no method clearly outperforms the baseline in the daily scale. Interestingly, all comparison methods perform worse on the weekly scale except for DFT (MAE 4.03 (weekly) vs. 5.48 (daily)). See Figure 7 and Table 2 to view the MAE and confidence intervals.

In Figure 8 we see the variation of the MAE as we add more components. While the curves generally overlap on the weekly scale, supporting the hypothesis that performance is essentially random, the linear regression clearly improves as lower DFT components (1-7) and higher DFT components (14-20) are added. Components 8-13, on the other hand, seem to increase generalization error.

The least angle regression paths in Figure 9 suggest that the low and high components are negatively correlated with resting heart rate, while the middle components are positively correlated with resting heart rate.

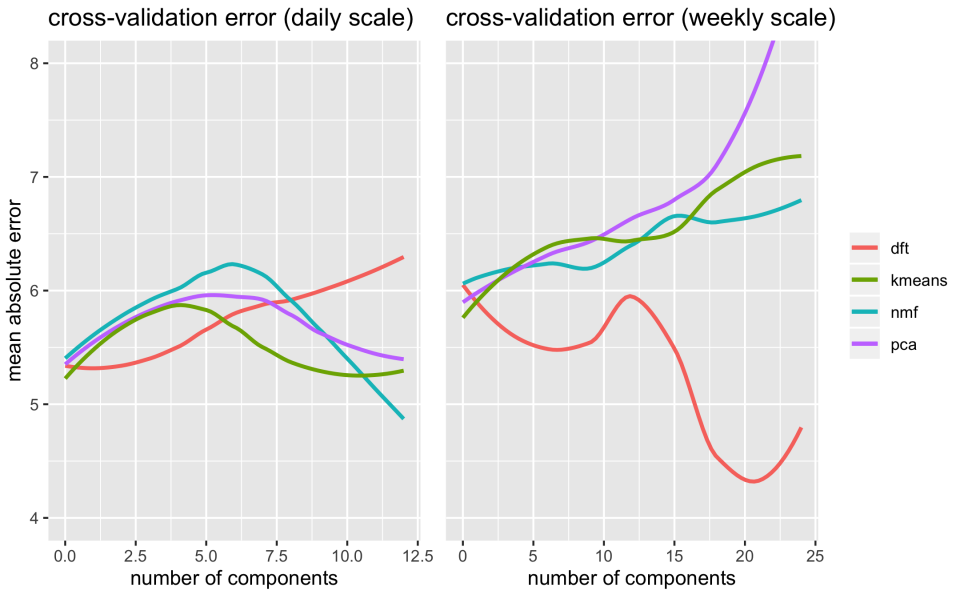


Fig. 8: MAE as number of components increases

5 Discussion

There is good reason to believe that DFT should perform better than the baseline models on the weekly timescales: as mentioned in §3.3.1, the "basis" components in DFT are periodic functions themselves. We demonstrate in Figures 1 and 2 the periodicity of subjects' activity levels, especially

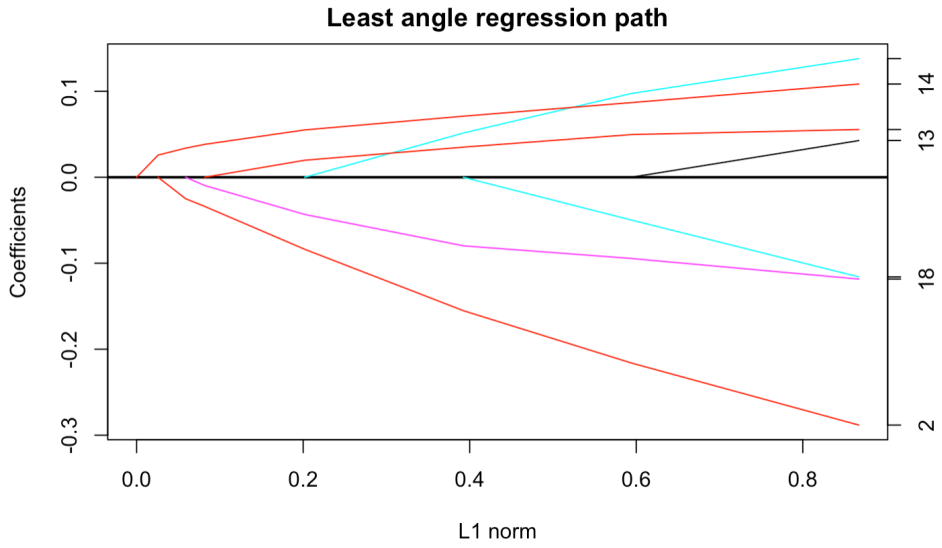


Fig. 9: Least angle regression path on DFT weekly components (first 8 steps)

in the weekly averages, and thus DFT seems to be an appropriate factorization method for this setting.

The other methods do seem to pick up on the periodic nature of the activity levels as shown by Figure 6. However they seem to pick up on spurious patterns and each component has one peak per day – the components differ in their scales and their shift, but their overall shapes are similar to one another. They lack flexibility in the frequency domain. In contrast, each of DFT's components have a different frequency: the kth component has period $168/k$ hours. Our results then seem to indicate that the frequency of activity per week matters more than shifts in activity.

Not all frequencies seem to have the same impact on resting heart rate, however. The plot of MAE against the number of components (Figure 8) and the LAR paths (Figure 9) both suggest bimodal behavior; while contributions from low (up to component 7) and high (greater than 18) seem to decrease resting heart rate, middle frequency components seem to be associated with increased resting heart rate.

Translating frequencies these back into hours gives a clue to what may be happening. The components up to component 7 have periods span multiple days, while the components past the 18th component have periods that are contained within a single half-day. We should expect these to be the dominant periods in person's schedule. The middle components (10-17), on the other hand, have periods between 12 and 17 hours long. We posit that having high magnitude on these frequencies indicates very sporadic, unpredictable behavior which cannot be captures by functions with periods greater than one day or less than eight hours. The boxplot in Figure 4 shows that these middle coefficients have significant variance, consistent with the biomodal hypothesis. However, more work must be done to further explore the association between these components and resting heart rate.

6 Conclusion

In this paper we aimed to go beyond basic summary statistics when analyzing continuous streams of activity data. We hypothesized that the DFT would be an effective method for processing such data, and that a weekly scale, rather than a daily scale, would improve predictive power. We validated this approach by predicting resting heart rate. Our results suggest that DFT outperforms other baselines in explaining residual variation in resting heart rate and that the weekly scale is indeed more fruitful, however a larger dataset would be required to draw such conclusions with confidence.

Additionally, from the cross-validation curves and least angle regression paths we observed a negative association between very low and very high frequency Fourier coefficients and resting heart rate, while components with frequency between those extremes (period approximately 12-17 hours long) seem to be positively associated with resting heart rate. We posit an explanation for this phenomenon and suggest directions for future work.

7 Contributions

Lilia dealt with most of the data pre-processing, explored the data and gathered summary statistics, and handled the k-means, PCA, and NMF methods. Bora helped optimize the pre-processing code, handled the DFT method, and implemented leave-one-out cross validation.

References

1. Bracewell, R.N.: Fourier transform and its applications (1999)
2. Cawley, G.C., Talbot, N.L.: On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. **11**, 2079–2107 (Aug 2010)
3. CICHOCKI, A., PHAN, A.H.: Fast local algorithms for large scale nonnegative matrix and tensor factorizations. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences **E92.A**(3), 708–721 (2009). https://doi.org/10.1587/transfun.E92.A.708
4. Jensen, M.T., Suadicani, P., Hein, H.O., Gyntelberg, F.: Elevated resting heart rate, physical fitness and all-cause mortality: a 16-year follow-up in the copenhagen male study. Heart **99**(12), 882–887 (2013). https://doi.org/10.1136/heartjnl-2012-303375, https://heart.bmj.com/content/99/12/882
5. Kim, J., Campbell, A.S., de Ávila, B.E.F., Wang, J.: Wearable biosensors for healthcare monitoring. Nature Biotechnology **37**(4), 389–406 (2019). https://doi.org/10.1038/s41587-019-0045-y, https://doi.org/10.1038/s41587-019-0045-y
6. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Leen, T.K., Dietterich, T.G., Tresp, V. (eds.) Advances in Neural Information Processing Systems 13, pp. 556–562. MIT Press (2001), http://papers.nips.cc/paper/1861-algorithms-for-non-negative-matrix-factorization.pdf
7. Li, X., Dunn, J., Salins, D., Zhou, G., Zhou, W., Schüssler-Fiorenza Rose, S.M., Perelman, D., Colbert, E., Runge, R., Rego, S., Sonecha, R., Datta, S., McLaughlin, T., Snyder, M.P.: Digital health: Tracking physiomes and activity using wearable biosensors reveals useful health-related information. PLOS Biology **15**(1), 1–30 (01 2017). https://doi.org/10.1371/journal.pbio.2001402, https://doi.org/10.1371/journal.pbio.2001402