Health Label and Behavioral Feature Prediction Using Bayesian Hierarchical Vector Autoregression Models

Ethan N. Lyon[1], **Luis H. Victor**[2], **Akane Sano**[1]
Rice University
[1] Electrical and Computer Engineering
[2]Applied Physics
enl1,lhv1,akane.sano@rice.edu

***Abstract*— The rising availability and accessibility of data from wearable devices and ubiquitous sensors allow the leveraging of computational methods to address human health and behavioral challenges. In particular, recent works have created time series, interpretable, and generalizable models for predicting patient healthcare outcomes from multidimensional data including expensive self-reported patient data, clinical data, and data from mobile and wearable devices. In this work, we used a Bayesian Hierarchical Vector Autoregression (BHVAR) model to predict behavioral and self-reported health outcomes on college student participants from passively collected data from their smartphones, wearable devices, and environment, as well as their self-reports. We also evaluated how the model performed being trained on 3, 7, 11, and 13 different features including some actionable and modifiable behavioral features. Then, we showed the value of augmenting self-reported datasets with many different types of data by demonstrating that additional inferences can be made with no significant toll on accuracy in comparison to using only self-reported features. Our models proved to be robust despite the greatly increased variable count as the reduced mean squared error (RMSE) of BHVAR over the patient-specific, maximum likelihood estimate (MLE) model was 10.5%, 14.9%, 26.6%, 39.6% in the 3, 7, 11, and 13 variable models respectively. We also obtained patient-level insights from clustering analysis of patient-level coefficients.**

I. Introduction

A. Motivation and Related Work

Recent advances in big data analytics and the availability of user data have prompted researchers to develop methods for understanding human health and behavior [6]. For example, the ability to predict and draw inferences for both individual and population-level outcomes is of great interest to the medical profession. Multivariate, time-series data available from surveys, and wearable or ubiquitous sensors in particular can be used to construct and train predictive health models. Discovering the relationships between these passively collected features and self-reported data is critical for researchers and clinical practitioners to identify the best strategies for adjusting human behavior for improved health and well-being whether through visualization tools, nudges, or other methodologies [3]. Various previous works have leveraged longitudinal health data and machine learning models for predicting patient-level outcomes [7, 11, 8]. However, these works have relied on expensive feature acquisition, as in [8], which limit the practical deployment of these models.

Stress and mood label predictions from wearable, smartphone and environmental data shows promising predictive power. Bogomolov et al. [1] utilized said data, in tandem with taking into consideration a participant's baseline personality data, and achieved reasonable accuracy in predicting mood labels using a Random Forest model. Then, Taylor et al. [12] integrated physiological, survey, and smartphone data to achieve an accuracy of 70.17% on a daily, two-class, happy-sad prediction task using neural networks. Both these works demonstrate the potential of inexpensive feature collection for healthcare models, however lack interpretability and the opportunities for inference present in Bayesian models. Some of these works have also attempted to develop a personalized model to predict health conditions; however, understanding the relationships between input data and output labels has been understudied. In addition, personalizing general models to individuals is still challenging. In this work, we contribute to the development of an accessible and low-cost methodology for improving health and behavior outcomes by evaluating the performance of a Bayesian Hierarchical Vector Autoregressive (BHVAR) model on making individual and population-level predictions from passively collected and self-reported data.

II. Methods

A. Data

Our dataset was collected from 243 students in a US university, and included wearable, smartphone, survey and weather data (acceleration, screen and call time, GPS, daily activities, self-reported wellbeing 0-100, weather). Two hundreds and twenty eight students contributed 30 days of consecutive data and the remaining 15 contributed 90 days of consecutive data. A total of 15 daily features were computed, including 10 modifiable behavioral features such as exercise duration (Table I) [10].

B. Models

The Bayesian hierarchical, vector autoregressive (BHVAR) model [5] was designed to address the lack of sparsity in maximum likelihood models [13] and the constraining sparsity

This work is supported by NSF #2047296 and #1840167.

TABLE I

Surveys:	Morning Happiness 3, Time in Bed* 3, Academic Duration* 7, Sleep Time* 7, Awakening Duration* 7, Exercise Duration* 13
Phone/Wearable:	Call Duration* 7 Screen Time* 11, Total Distance travelled* 11, Regions of Interest Visited* 13, Step Count* 3
Environment:	Avg. Cloud Cover 11, Temperature 11

Features used in the models. The number indicates the inclusion into the 3, 7, 11, and 13 variable models. * indicates actionable and modifiable behavioral features.

of the regularized linear model [5, 2] by implementing an elastic net prior for coefficient estimation. Bayesian Hierarchical (BH) modeling allows the integration of population and patient-level observations to obtain more accurate predictions and more informative inferences. Vector autoregression (VAR) is a mainstay of multivariate time-series analysis and captures the endogenous interdependencies of the variables in the data. In this work, we evaluate the architecture of the BHVAR model to an arbitrary number of variables from the dataset described above.

The vector auto regression model is given by:

$$\mathbf{y}_{nt} = \sum_{i=1}^{p} A_{ni}\mathbf{y}_{n,t-i} + \boldsymbol{\epsilon}_{nt}, t = p+1, \cdots, T_n \qquad (1)$$

Where $\mathbf{y}_{nt}$ is a column vector of R features for time $t = 1, \cdots, T_n$ for patient $n = 1, \cdots, N$. A_{ni} is an $R \times R$ matrix that represents the lag-i coefficients for time lags $i = 1, \cdots, p$ in a p-lag VAR model, VAR(p). We adopted a VAR(1) model, in which one previous day's data is used as predictors for the next day. $\boldsymbol{\epsilon}_{nt}$ is a multivariate normal (MVN) error term where $\boldsymbol{\epsilon}_{nt} \sim MVN(0, \boldsymbol{\Lambda}^{-1})$. $\boldsymbol{\Lambda}$ is the precision matrix and is the same for all patients. The coefficient matrices for patient n is represented by $\mathbf{w}_n$: $\mathbf{w}_n = vec([A_{n1}, \cdots, A_{np}])$ It is given by: $\mathbf{w}_n = \mathbf{w} + \mathbf{v}_n$, where $\mathbf{w}$ are the population-level coefficients common to all patients, and $\mathbf{v}_n$ are the patient-level coefficients for patient n.

To generate the posterior distributions in the BH model, Markov Chain Monte Carlo simulations (MCMC) is used. An elastic net prior is used to estimate the coefficients and introduce sparsity. The VAR coefficients are obtained from the posterior modes (in this case the maximum likelihood of the posterior distributions).

The coefficient for the maximum likelihood estimate (MLE) model is obtained via the maximum likelihood weights for each individual.

I.e. solve for $\mathbf{w}_n$ in $\mathbf{H}_n\mathbf{w}_n = \mathbf{y}_n$ for each individual, $\mathbf{n}$.

Where $\mathbf{H}_n = \begin{pmatrix} \mathbf{y}_{np} & \cdots & \mathbf{y}_{n,T_n-1} \\ \vdots & \ddots & \vdots \\ \mathbf{y}_{n1} & \cdots & \mathbf{y}_{n,T_n-p} \end{pmatrix}$

C. Experiment

Each variable was transformed to log scale, demeaned, scaled to unit variance, and detrended. Four simulation chains of MCMC were created and combined to obtain the posterior distributions and coefficients. To assess the prediction accuracy, all but the last day of data for every patient was used to train the model. Predictions on the last day are given by: $\hat{\mathbf{y}}_{n,T_n+1} = \sum_{i=1}^{p} \hat{A}_{ni}\mathbf{y}_{n,T_n} + \hat{\boldsymbol{\epsilon}}_{n,T_n+1}$. The prediction was made for one time-step ahead, T_{n+1}, into the future for participant n. The mean-squared error (MSE) per variable between the predictions, $\hat{\mathbf{y}}_{n,T_n+1}$, and test cases $\mathbf{y}_{n,T_n+1}$, for all participants $n = 1, \cdots, N$ was obtained via: $\frac{1}{N}\sum_{n=1}^{N}\left(\hat{\mathbf{y}}_{n,T_n+1} - \mathbf{y}_{n,T_n+1}\right)^2$

The overall MSE for the model was calculated as the average MSE per variable. The reduced mean square error (RMSE) was calculated as the percent reduction in MSE from the MLE to the BHVAR model.

We evaluated the accuracy of the model for different numbers of variables. Datasets of sizes 3, 7, 11, and 13 variables were used in evaluating the model. Participants missing more than 20% of data for any variable were excluded. Any remaining missing data was imputed using the median value for each variable, for each participant.

Since our model produces patient-level coefficients, the ability to model the differences between them can further provide valuable inference into behavioral patterns. To this end, we performed a K-means clustering analysis on these coefficients. The optimal number of clusters was determined via the graphing of Silhouette score across multiple values of K from $K = 2, ..., 30$. It is important to note that clustering results were obtained using the median of posterior distributions instead of the mode. This is because where as using the mode is important for inducing sparsity and thus interpretability, using the median captures a more robust metric that is better for heterogenous comparisons.

III. Results

TABLE II

# of Features	3	7	11	13
MLE	1.37	1.28	1.50	1.67
BHVAR	1.23	1.09	1.10	1.01
RMSE	**10.5%**	**14.9%**	**26.6%**	**39.6%**

Mean-squared error of the Maximum Likelihood Estimate (MLE) and Bayesian Hierarchical Vector Autoregression (BHVAR) models with the reduced-MSE (RMSE) per number of features

Benchmarking MSE: Our BHVAR model achieved an MSE of 1.23, reduced from an MSE of 1.37 from the MLE model, when using the three variables 'Morning Happiness', 'Time in Bed', and 'Step Count.' Our results shows the improved performance of the BHVAR model when compared to a patient-level, MLE model. Specifically, the BHVAR model achieved a RMSE over the MLE model for all models as seen in Table II.

The BHVAR model performed robustly compared to the MLE model in the presence of many features. Not only was the MSE better for each model, the average MSE per variable was the lowest in the 13 variable model whereas the MLE model showed its maximum MSE per variable. It is also interesting to note that while 'Step Count' had a considerably high MSE when being predicted next day, when

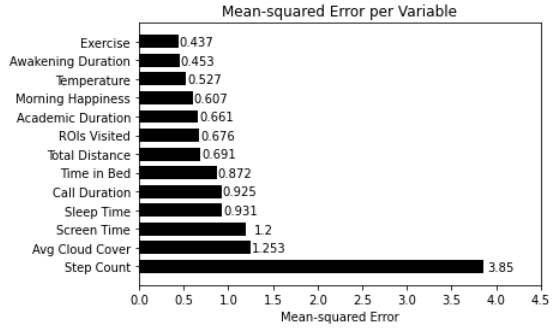

Fig. 1: Mean-squared error per variable for the 13-variable model

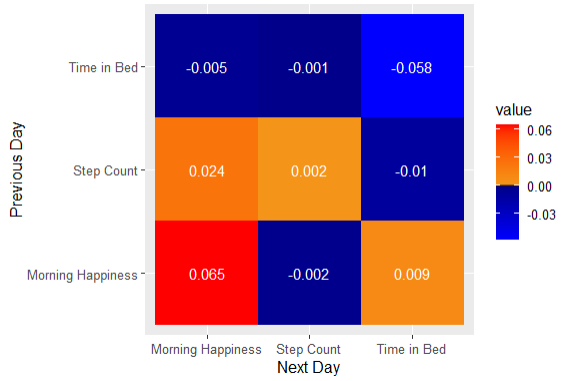

Fig. 2: Population level coefficients for the 3 variable model. Coefficients represent the mode of the posterior distribution in this case

used to predict next day 'Morning Happiness', we observe a significant relationship (zero not contained within the 90% confidence interval for the posterior distribution) in the three variable model. This is an example of the benefit the model gains from including a feature which may be hard to predict, but is itself still useful for prediction of labels. MSE was calculated per variable in **Figure 1**. An interesting result is the relatively low MSE of the self-reported mood label, 'Morning Happiness'. The MSE is comparable 'Temperature' - indicating that this abstract emotional label can be predicted as reliably next-day as temperature change in the Fall and Spring (the seasons in which the data was collected) using our model. We also note that the RMSE for 'Morning Happiness' is 9.67% for the 3 variable model, and 48.56% for the 13 variable model. This result showed that with proper surveys and labeling, it is possible to utilize patient diary entries to create robust predictions using the BHVAR model.

Population-Level Coefficient Matrix: We further examined the population level coefficient in the 3-variable model in **Figure 2**, inferences about the relationship between previous day's variables to current day variables. Coefficient matrices like these were also available at the patient-level as well, making possible both individualized inference and heterogenous comparisons between patients - the ability to make such comparisons is evaluated in the next section on

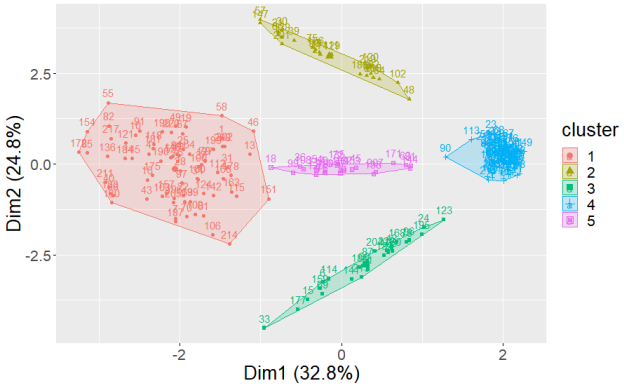

Fig. 3: K-means clustering results for the 3 variable model with 'Time in Bed', 'Morning Happiness', and 'Step Count'

clustering.

We observed most of the strong relationships are variables with themselves. Coefficient matrices can be useful in determining day-to-day, modifiable patterns which affect wellbeing labels. A significantly positive relationship was found between previous day 'Step Count' and next day 'Morning Happiness'. Finding these strong relationships can help patients become more aware of how their day-to-day behavior is affecting their personal health and mood.

Clustering Analysis: **Figure 3** shows the results of K-means clustering performed on the 3-variable model which includes the features Step Count, Morning Happiness, and Time in Bed. The dimension of these coefficients are $R * R = 9$. These 5 clusters were deemed to be optimal via the Silhouette score. The clustering results in 9 dimensions were projected down to two dimensions using partial component analysis, where the two most significant components onto which the coefficients were project on to explain 32.8% and 24.8% of the variance respectively. The structured nature of these clusters indicate that useful insights can be obtained using this sort of analysis. By subtracting the mean of the patient level coefficients from each cluster's center, five interpretable patterns of behavior emerge. Cluster 4 represented patients with significantly lower valued coefficients, while cluster 5 represented patients those with higher valued coefficients. Clusters 1, 2, and 3 represented patients that have lower coefficients in predicting next day 'Time in Bed', 'Morning Happiness', and 'Step Count' respectively. Being able to assess how exactly different patients responded to modifiable behaviors this way show potential in personalizing treatment.

IV. Discussions

An advantage a statistical model such as BHVAR has over traditional machine learning models is interpretability. Coefficient distributions and the information that can be gained from examining them are a clear advantage BHVAR has over more obtuse machine learning frameworks such as neural networks. When dealing with prediction for issues as serious as healthcare outcomes, being able to 'look under the

hood' of a model via the particular properties of the posterier distribution, as we obtain via MCMC, may be valuable.

Predictive models such as BHVAR potentially have value as inexpensive diagnostics and as an aid in just-in-time intervention. Consulting a healthcare professional can be costly and take a considerable amount of time before an assessment can be made. Instead, this model may be able to swiftly react to a decline in health and forecast the likely state of a person the next day. Despite improvements in prediction accuracy, further work exploring the efficacy of the BHVAR model aiding in these clinical or behavorial interventions is necessary before being deployed for such purposes. Clustering analysis of patient-level coefficients for our three variable model was insightful into providing five clearly separate groups with interpretable coefficient patterns. While this level of structure in clustering is not a guarantee, useful inference into treatments and how specific patients respond to daily activities may be able to be gained. The BHVAR model shows promise in modeling the relationships between self-reported labels and features in a way that is more easily deployed in low resource settings. Previous work done predicting mood labels from inexpensive smartphone and environmental features [1, 4, 9] suggests that including these types of features alongside patient personality data in the model can lead to reasonable predictive power.

Lastly, there are some limitations in this work. First, the assumptions made about the prior distributions are MVN. In a more carefully designed model, the prior distribution of the variables would better reflect the appropriate distribution of the data. Second, all evaluation performed both in the paper only used VAR models with a time lag of $p = 1$. Optimal time-lag selection can be done using the Akaike Information Criterion (AIC). However, calculating AIC can be computationally expensive as it requires computing the BHVAR model for each time lag. Since the amount of coefficients in the model is $R^2 * p$, where R is the number of variables and p the time lag, including more variables is more computationally expensive than increasing the time lag. Thus, evaluating the model on datasets with high dimensionality may reduce the feasibility of training the model with larger lags. While not explored here, occluding the effects of variables with themselves can lead to an understanding of which proxy variables have the most effect in accurately predicting valuable next-day emotion labels. Significance of predictions can be made via confidence intervals derived from the posteriors, however this still does not give adequate insight into how we can create a model for next day emotional states via proxies. The easiest way to examine the effect of a proxy with the current model is observing the reduction of MSE when comparing a model trained with only the emotional label, and then examining the reduction of MSE with a model trained with that label and the desired proxies.

V. Conclusion

We used a BHVAR model to predict behavioral and self-reported health outcomes using passively collected data from smartphones, wearable devices, and environment, as well as their self-reports. We compared our models trained on 3, 7, 11, and 13 different features including some actionable behavioral features. Our models were robust despite the greatly increased variable count, reducing the MSE of BHVAR over the patient-specific MLE model. We also analyzed patient-level insights from clustering analysis of patient-level coefficients.

References

[1] Andrey Bogomolov et al. "Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits". In: ACM International Conf. on Multimedia (2014).

[2] Jerome Friedman. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: Journal of Statistical Software (2010).

[3] Ralph Hertwig and Till Grüne-Yanoff. "Nudging and boosting: Steering or empowering good decisions". In: Perspectives on Psychological Science (2017).

[4] Natasha Jaques et al. "Predicting students' happiness from physiology, phone, mobility, and behavioral data". In: ACII (2015).

[5] Feihan Lu et al. "Bayesian hierarchical vector autoregressive models for patient-level predictive modeling". In: Plos One (2018).

[6] Ahmed A Moustafa et al. "Applying big data methods to understanding human behavior and health". In: Frontiers in computational neuroscience (2018).

[7] Jenna M Reps et al. "Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data". In: JAMIA (2018).

[8] Daniel Sanchez-Morillo and Fernandez-Granero. "Use of predictive algorithms in-home monitoring of chronic obstructive pulmonary disease and asthma". In: Chronic Respiratory Disease (2016).

[9] Akane Sano and Rosalind W. Picard. "Stress Recognition Using Wearable Sensors and Mobile Phones". In: Humaine Association Conference on ACII (2013).

[10] Akane Sano et al. "Identifying Objective Physiological Markers and Modifiable Behaviors for Self-Reported Stress and Mental Health Status Using Wearable Sensors and Mobile Phones: Observational Study". In: Journal of Medical Internet Research (2018).

[11] Zach Shahn, Patrick Ryan, and David Madigan. "Predicting health outcomes from high-dimensional longitudinal health histories using relational random forests". In: The ASA Data Science Journal (2015).

[12] Sara Taylor et al. "Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health". In: IEEE/Transactions on Affective Computing (2017).

[13] Yao Zheng et al. "An Idiographic Examination of Day-to-Day Patterns of Substance Use Craving, Negative Affect, and Tobacco Use Among Young Adults in Recovery". In: Multivariate Behavioral Research (2013).