Vector Autoregression (VAR) of Longitudinal Sleep and Self-report Mood Data

Jeff Brozena

College of Information Sciences and Technology
Penn State University, United States
brozena@psu.edu

May 2, 2023

Abstract

Self-tracking is one of many behaviors involved in the long-term self-management of chronic illnesses. As consumer-grade wearable sensors have made the collection of health-related behaviors commonplace, the quality, volume, and availability of such data has dramatically improved. This exploratory longitudinal N-of-1 study quantitatively assesses four years of sleep data captured via the Oura Ring, a consumer-grade sleep tracking device, along with self-reported mood data logged using eMood Tracker for iOS. After assessing the data for stationarity and computing the appropriate lag-length selection, a vector autoregressive (VAR) model was fit along with Granger causality tests to assess causal mechanisms within this multivariate time series. Oura's nightly sleep quality score was shown to Granger-cause presence of depressed and anxious moods using a VAR(2) model.

1 Introduction

Long-term self-management of chronic illnesses such as bipolar disorder require persistent awareness of illness state over long periods of time and at varying time scales [11, 9, 8]. Remaining consistently aware of key indicators signalling the onset of a chronic condition allow individuals a chance at early intervention to reduce the severity of a givene episode. For example, an individual may modify behavior, engage their health practitioners, or adjust medication dosage. However, bipolar disorder is an illness that often degrades an individual's self-awareness and capacity for self-monitoring during symptomatic periods.

In the context of this specific illness, a volume of prior work has demonstrated the vital role of sleep in order to promote mood stability and prevent symptomatic episodes [5, 12, 4]. Although the particulars of this topic fall beyond the scope of this paper, these nuanced relationships may in fact be self-reinforcing and bidirectional — poor sleep may lead to episodic onset, which may also lead to worsening (or shortening) sleep bouts.

Given the importance of sleep in the ongoing management of this illness, accurate consumer-grade alternatives to polysomnography (considered the gold standard of sleep tracking) have emerged over the last few years. Indeed, comparatively inexpensive sleep tracking technologies like the Oura Ring have dramatically improved the quality of information that can be used to augment and inform these self-monitoring activities. Objective sensor-based tracking technology can be complemented with subjective self-report measures in order to form a more complete picture of physical and mental health across time. Given the aforementioned interplay of sleep and mood, this combination of subjective and objective tracking creates the possibility of longitudinal analysis — and potentially deepens one's capacity for self-awareness.

Following four years of consistent sleep and mood tracking, I sought to more formally interpret the data I had collected to quantify what I had previously intuited: that certain mood states could be understood (and potentially even predicted) by recent sleep trends. Indeed, this intuition has been demonstrated quantitatively in existing literature [1, 10, 6]. As this work also demonstrates, combining data from consumer wearable technology and subjective self-report logs allows for a more comprehensive picture of health.

I will first describe the vector autoregression (VAR) method and subsequent tests, namely the Granger causality test and an impulse response analysis, that were performed to achieve these goals.

First, I will describe the methods used to achieve these goals, providing a brief overview of vector autoregression, Granger causality, and impulse response functions. Next, I will detail the findings of these methods on the dataset. This work concludes with a discussion of the methods and their potential applications in future work.

2 Problem setup

A multivariate time series analysis was performed using a vector autoregressive (VAR) model fit using ordinary least squares. An optimal lag order was first obtained using a combination of Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Hannan-Quinn Information Criterion (HQIC), and final prediction error (FPE). After fitting a VAR(2) model on the multiple time series data (outlined below), a Granger causality test was performed in order to assess the predictive relationships between variables. Finally, an impulse response analysis was plotted to further explore the temporal relationships between variables, specifically between sleep and self-reported mood.

2.1 Vector Autoregression

A VAR(p) model for a multivariate time series is a regression model for outcomes at time t and time lagged predictors, with p indicating the lag. Given $p = 1$, the model would be concerned with one observation prior to t. As noted by Lütkepohl [7] (as cited in [13]), a $T \times K$ multivariate time series (where T is the number of observations and K is the number of variables) can be modeled using a p-lag VAR model, notated as

$$\begin{aligned} Y_t &= \nu + A_1 Y_{t-1} + \ldots + A_p Y_{t-p} + u_t \\ u_t &\sim \mathsf{Normal}(0, \Sigma_u) \end{aligned} \tag{1}$$

where A_i is a $K \times K$ coefficient matrix.

Intercept terms are included in ν and regression coefficients are included as the subscripted A values. This equation is solved using ordinary least squares (OLS) estimation. The vector autoregressive (VAR) model is a flexible method for the analysis of causality in this setting.

2.2 Granger Causality Testing

Granger causality tests were performed in order to better assess the predictive capacity of the Oura sleep score on self-reported mood states. Granger causality defines one type of relationship between time series [3] and states that a variable *Granger causes* another variable if "the prediction of one time series is improved by incorporating the knowledge of a second time series" [1].

Two autoregressive models are fit to the first time series, once with and once without the inclusion of the second time series. The improvement of the prediction is measured as the ratio of variance of the error terms. The null hypothesis states that the first variable *does not* Granger cause the second variable and is rejected if the coefficients for the lagged values of the first variable are significant.

Granger causation tests were applied using sleep scores as a single predictor and each mood state as outcome variables.

2.3 Impulse Response Function Visualization

An impulse response function (IRF) is the "reaction of a dynamic system in response to an external change" [17]. Plotting an IRF allows for the interpretation of the impulse of a predictor on other variables on subsequent periods. Given sleep scoring as a predictor, an IRF visualization was created to better understand its impact on mood state. Figure 3 displays the results of this analysis over a 10-day period.

3 Experimental Results

3.1 Dataset Description

The sleep score dataset was created using the second- and third-generation Oura Ring. The proprietary Oura sleep score is on a scale of 1 to 100 and incorporates a variety of sensor-based measures (i.e., heart rate variability, resting heart rate, body temperature) across time. Although the specifics of this algorithm are not public, the Oura Ring has been found to produce accurate measures of sleep timing and heart rate variability when compared against polysomnography [18]. As detailed in Table 1, my use of the Oura Ring was consistent across time. The dataset contains 1,455 nights of sleep bout data occurring between February, 2019 and March, 2023.

	Value
Total nights	1455
Missing nights	1
Mean	73.82
SD	12.36
Max	97.00
Min	30.00

Table 1: Descriptive statistics of Oura Ring sleep score data

Each day at 4:30pm I received a notification prompting me to log my subjective state in eMood Tracker, a mobile application for iOS. eMood Tracker is "recommended by psychologists, therapists, and social workers" and is intended to "track symptom data relating to Bipolar I and II disorders" [2]. The version used through this period contains preset mood categories (depressed, irritable, anxious, and elevated) and allow users to log the presence and intensity on a scale of 0 to 3, where 0 is "not present" and 3 is "severe". The resulting dataset contains the most severe mood state per day. The contents of this dataset are outlined in Table 2.

3.2 Data Analysis

All analysis were performed in Python version 3.11.0 [16] using `pandas` 1.5.3 [15] for data preprocessing and `statsmodels` 0.13.5 [13] for modeling. Dickey-Fuller tests of stationarity were performed using `statsmodels` and the `pymdarima` library [14], a clone of R's `auto.arima`.

3.3 Stationarity, Decomposition, and Autocorrelation

A stationary time series contains no periodic fluctuation ("trend"). Without stationarity, the means and correlations given by a model will not accurately describe a time series' true signal [1]. If a time series is found not to be stationary, an approach known as differencing can be applied to achieve stationarity. The Dickey-Fuller test is one mechanism to determine whether a time series is stationary.

Two Dickey-Fuller tests were performed on each time series, first via `statsmodels` and then, additionally, using `pymdarima`'s `should_diff()` function to assess the need for differencing. The `statsmodels` approach, an Augmented Dickey-Fuller test (ADF), yielded a significant p-value of .001 indicating support for the null hypothesis that the time series is not stationary. However, the ADF performed via the `pymdarima` approach

EMA Categories	**Count**
irritable	100
anxious	88
depressed	103
elevated	48

Table 2: Count of days where EMA item contains a non-zero value

using an alpha value of 0.05 yielded a non-significant p-value of 0.01 indicating that no differencing was required in order to produce a stationary time series. For the purposes of this study, I followed the results of the `pymdarima` library and assumed stationarity.

An exploratory time series decomposition visualization was created to better understand the presence of trend in the sleep score dataset. Figure 1 contains these results. Additionally, a partial autocorrelation function was plotted using `statsmodels`, displayed in Figure 2. Notably, partial autocorrelation appears to drop to zero for lag values greater than 2.

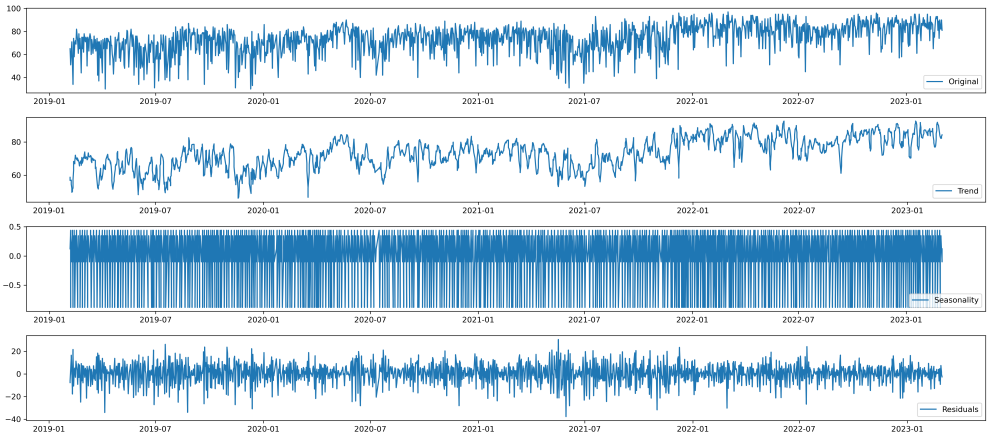

Figure 1: Decomposition of sleep time series

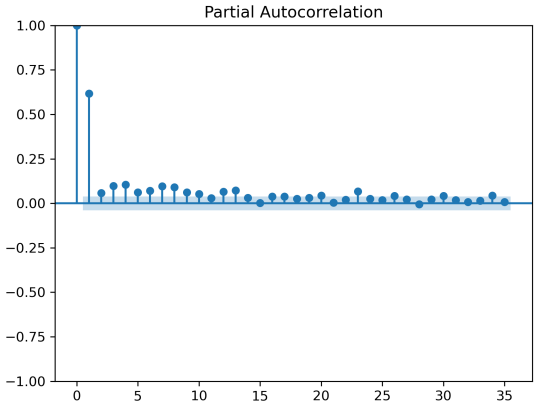

Figure 2: Partial autocorrelation of sleep time series

3.4 Lag Order Selection

The number of lags amount to the number of preceding days included as predictor values in the model. The `statsmodels` function `select_order()` was used to assess an optimal lag value with possibilities between 0 and 15. The optimal lag value was determined using four information criteria — Akaike Information Criteria (AIC), Bayes Information Criterion (BIC), Final Prediction Error (FPE), and Hannan-Quinn (HQ) criterion. As detailed in Table 3, this selection process yielded a tie with lag-1 and lag-2 each labeled as the minimum on these criteria. Rather than taking further quantitative approaches, lag-2 was selected based on prior knowledge of sleep quality and the onset of mood states.

	AIC	BIC	FPE	HQIC
0	1.688	1.708	5.408	1.695
1	-0.02482	0.09412*	0.9755	0.01979*
2	-0.03545*	0.1826	0.9652*	0.04635
3	-0.03417	0.2830	0.9664	0.08481
4	-0.02907	0.3872	0.9714	0.1271
5	-0.02537	0.4900	0.9750	0.1680
6	-0.01701	0.5975	0.9832	0.2135
7	-0.01940	0.6943	0.9809	0.2483
8	-0.01014	0.8026	0.9900	0.2948
9	-0.0008049	0.9111	0.9993	0.3413
10	0.01201	1.023	1.012	0.3913
11	0.02510	1.135	1.026	0.4415
12	0.03723	1.246	1.038	0.4909
13	0.04867	1.357	1.050	0.5395
14	0.06022	1.468	1.063	0.5882
15	0.07076	1.577	1.074	0.6359

Table 3: VAR Order Selection (* highlights the minimum)

3.5 Vector Autoregression Model

All time series under analysis were found to be stationary (ADF test $p < .05$). The results of the VAR(2) model predicting mood states using sleep score are shown in Table 4. Sleep score was found to be a significant positive predictor of depression, also confirmed via Granger causation tests. Oura's sleep score did not positively or negatively predict other mood states in this model.

	coefficient	std. error	t-stat	prob
L1.score	0.633262	0.027574	22.966	0.000
L1.anxious	0.153275	0.446110	0.344	0.731
L1.depressed	0.477164	0.409130	1.166	0.243
L1.irritable	-0.282988	0.412509	-0.686	0.493
L1.elevated	-0.220198	0.655784	-0.336	0.737
L2.score	-0.003080	0.027452	-0.112	0.911
L2.anxious	0.353528	0.445359	0.794	0.427
L2.depressed	1.241873	0.409667	3.031	0.002
L2.irritable	-0.080069	0.412341	-0.194	0.846
L2.elevated	-0.499540	0.657230	-0.760	0.447

Table 4: VAR results for equation score

3.6 Granger Causality

The results of the Granger causation tests are shown in Table 5. Sleep score was shown to Granger-cause both depressed and anxious mood.

3.7 Impulse Response Analysis

As shown in Figure 3, the impact of sleep score on the four self-reported mood states varies over a 10-day period. Standard errors are plotted at the 95% significance level. The effect of an increase to sleep score on depressed and anxious moods appear to be most felt only after several days of its impact, peaking at roughly 3 days and then gradually decaying.

Causal Variable	Variable	Test statistic	Critical value	p-value	df
sleepscore	**depressed**	**5.384**	2.997	**0.005**	(2, 6535)
sleepscore	**anxious**	**3.294**	2.997	**0.037**	(2, 6535)
sleepscore	irritable	1.347	2.997	0.260	(2, 6535)
sleepscore	elevated	1.203	2.997	0.500	(2, 6535)

Table 5: Granger Causality Tests for Sleep Score on Self-reported Mood States

4 Discussion

This exploratory study affirms the role that self-tracking technologies can play in the ongoing management of affective disorders. This type of N-of-1 analysis would be impossible without inexpensive wearable sensors, and the quality of this dataset is directly related to how non-invasive this particular wearable is.

This work is not without limitation. My reliance on an algorithmic ADF test to assess stationarity (rather than directly assessing the data myself) could leave room for error. In the context of this work, an incorrect assessment of stationarity risks the accuracy of the remainder of the analysis. Additionally, this work only assesses the influence of the Oura sleep score on mood. In reality, this is likely closer to a bidirectional influence and this should be reflected properly in the analysis.

In future work, I hope to incorporate machine learning techniques for time series segment annotation in order to explore the possibility of automatic labeling of time periods where signals indicate the presence of an oncoming episode.

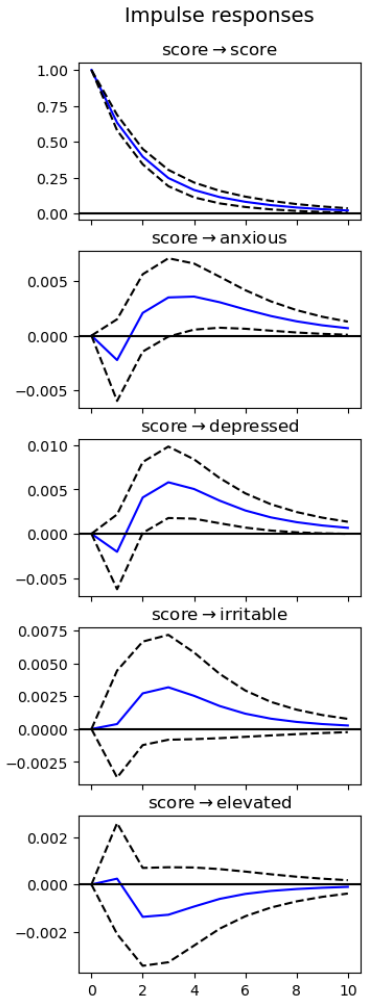

Figure 3: Plot of Impulse Response Function, Lag 0 to 10

References

[1] Eliezer Bose, Marilyn Hravnak, and Susan M. Sereika. "Vector Autoregressive (VAR) Models and Granger Causality in Time Series Analysis in Nursing Research: Dynamic Changes Among Vital Signs Prior to Cardiorespiratory Instability Events as an Example". In: *Nursing research* 66.1 (2017), pp. 12–19. ISSN: 0029-6562. DOI: 10.1097/NNR.0000000000000193. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5161241/ (visited on 04/25/2023).

[2] *eMoods*. en. 2023. URL: https://emoodtracker.com (visited on 04/29/2023).

[3] C. W. J. Granger. "Investigating causal relations by econometric models and cross-spectral methods". In: *Econometrica : journal of the Econometric Society* 37.3 (1969). Publisher: [Wiley, Econometric Society], pp. 424–438. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1912791 (visited on 05/01/2023).

[4] June Gruber et al. "Sleep matters: Sleep functioning and course of illness in bipolar disorder". en. In: *Journal of Affective Disorders* 134.1 (Nov. 2011), pp. 416–420. ISSN: 0165-0327. DOI: 10.1016/j.jad.2011.05.016. URL: https://www.sciencedirect.com/science/article/pii/S016503271100262X (visited on 04/29/2023).

[5] Allison G. Harvey, Lisa S. Talbot, and Anda Gershon. "Sleep Disturbance in Bipolar Disorder Across the Lifespan". en. In: *Clinical Psychology: Science and Practice* 16.2 (2009), pp. 256–277. ISSN: 1468-

2850. DOI: 10.1111/j.1468-2850.2009.01164.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-2850.2009.01164.x (visited on 04/29/2023).

[6] Salar Jafarlou et al. "Objective Prediction of Next-Day's Affect Using Multimodal Physiological and Behavioral Data: Algorithm Development and Validation Study". EN. In: *JMIR Formative Research* 7.1 (Mar. 2023). Company: JMIR Formative Research Distributor: JMIR Formative Research Institution: JMIR Formative Research Label: JMIR Formative Research Publisher: JMIR Publications Inc., Toronto, Canada, e39425. DOI: 10.2196/39425. URL: https://formative.jmir.org/2023/1/e39425 (visited on 04/12/2023).

[7] Helmut Lütkepohl. *New introduction to multiple time series analysis*. en. OCLC: ocm61028971. Berlin: New York : Springer, 2005. ISBN: 978-3-540-40172-8.

[8] Shazmin Majid et al. "Exploring self-tracking practices for those with lived experience of bipolar disorder: Learning from combined principles of Patient and Public Involvement and HCI". en. In: *Designing Interactive Systems Conference*. Virtual Event Australia: ACM, June 2022, pp. 1907–1920. ISBN: 978-1-4503-9358-4. DOI: 10.1145/3532106.3533531. URL: https://dl.acm.org/doi/10.1145/3532106.3533531 (visited on 07/01/2022).

[9] Emma Morton et al. "'Taking back the reins' – A qualitative study of the meaning and experience of self-management in bipolar disorder". en. In: *Journal of Affective Disorders* 228 (Mar. 2018), pp. 160–165. ISSN: 0165-0327. DOI: 10.1016/j.jad.2017.12.018. URL: https://www.sciencedirect.com/science/article/pii/S0165032717317913 (visited on 10/02/2022).

[10] Isaac Moshe et al. "Predicting Symptoms of Depression and Anxiety Using Smartphone and Wearable Data". In: *Frontiers in Psychiatry* 12 (2021). ISSN: 1664-0640. URL: https://www.frontiersin.org/articles/10.3389/fpsyt.2021.625247 (visited on 04/12/2023).

[11] Elizabeth L Murnane et al. "Self-monitoring practices, attitudes, and needs of individuals with bipolar disorder: implications for the design of technologies to manage mental health". In: *Journal of the American Medical Informatics Association* 23.3 (May 2016), pp. 477–484. ISSN: 1067-5027. DOI: 10.1093/jamia/ocv165. URL: https://doi.org/10.1093/jamia/ocv165 (visited on 01/12/2022).

[12] Greg Murray and Allison Harvey. "Circadian rhythms and sleep in bipolar disorder". en. In: *Bipolar Disorders* 12.5 (2010), pp. 459–472. ISSN: 1399-5618. DOI: 10.1111/j.1399-5618.2010.00843.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1399-5618.2010.00843.x (visited on 04/29/2023).

[13] Skipper Seabold and Josef Perktold. "statsmodels: Econometric and statistical modeling with python". In: *9th python in science conference*. 2010.

[14] Taylor G. Smith et al. *pmdarima: ARIMA estimators for Python*. 2017. URL: http://www.alkaline-ml.com/pmdarima.

[15] The pandas development team. *pandas-dev/pandas: Pandas*. Feb. 2020. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[16] *The Python Language Reference*. URL: https://docs.python.org/3/reference/index.html (visited on 05/01/2023).

[17] Herman J. de Vries et al. "Wearable-Measured Sleep and Resting Heart Rate Variability as an Outcome of and Predictor for Subjective Stress Measures: A Multiple N-of-1 Observational Study". en. In: *Sensors* 23.1 (Jan. 2023). Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, p. 332. ISSN: 1424-8220. DOI: 10.3390/s23010332. URL: https://www.mdpi.com/1424-8220/23/1/332 (visited on 04/12/2023).

[18] Massimiliano de Zambotti et al. "The Sleep of the Ring: Comparison of the ŌURA Sleep Tracker Against Polysomnography". In: *Behavioral Sleep Medicine* 17.2 (Mar. 2019). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/15402002.2017.1300587, pp. 124–136. ISSN: 1540-2002. DOI: 10.1080/15402002.2017.1300587. URL: https://doi.org/10.1080/15402002.2017.1300587 (visited on 04/29/2023).