Vector Autoregression (VAR) of Longitudinal Sleep and Self-report Mood Data: an N-of-1 Study

Jeff Brozena

May 3, 2023

Penn State University

Goals

1. Motivate the problem
2. Define the problem and approach
3. Demonstrate findings
4. Close with ways forward and limitations

Introduction

Motivation

Long-term self-management of chronic illnesses such as bipolar disorder require persistent awareness of illness state[1] over long periods of time and at varying time scales[2]

In the context of this specific illness, prior work has demonstrated the vital role of sleep in order to promote mood stability and prevent symptomatic episodes[3]

[1] 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," *Journal of the American Medical Informatics Association* 23, no. 3 (May 1, 2016): 477–84, https://doi.org/10.1093/jamia/ocv165.

[2] 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," in *Designing Interactive Systems Conference* (DIS '22: Designing Interactive Systems Conference, Virtual Event Australia: ACM, 2022), 1907–20, https://doi.org/10.1145/3532106.3533531.

[3] Greg Murray and Allison Harvey, "Circadian Rhythms and Sleep in Bipolar Disorder," *Bipolar Disorders* 12, no. 5 (2010): 459–72, https://doi.org/10.1111/j.1399-5618.2010.00843.x.

Motivation

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

Motivation

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.[4]

[4] Massimiliano de Zambotti et al., "The Sleep of the Ring: Comparison of the ŌURA Sleep Tracker Against Polysomnography," *Behavioral Sleep Medicine* 17, no. 2 (March 4, 2019): 124–36, https://doi.org/10.1080/15402002.2017.1300587.

Motivation

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.

Motivation

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.

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.

Problem setup

Dataset description

The sleep score dataset was created using the second- and third-generation Oura Ring and contains 1,455 nights of sleep bout data occurring between February, 2019 and March, 2023.

Dataset description

Table 1: Descriptive statistics of Oura Ring sleep score data

Descriptor	Value
Total nights	1455
Missing nights	1
Mean	73.82
SD	12.36
Max	97
Min	30

Dataset description

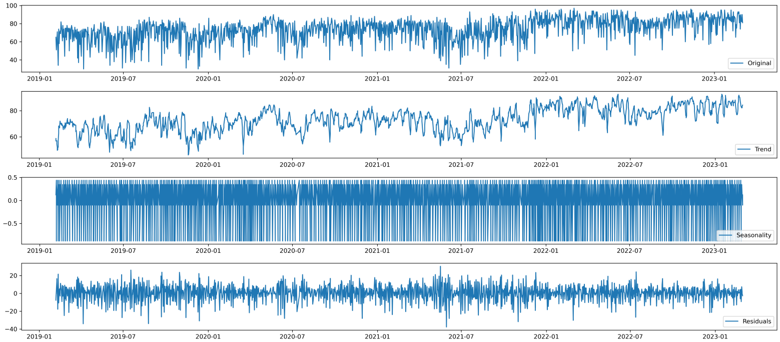

Figure 1: Decomposition of sleep time series

Dataset description

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.”[5]

[5] “eMoods” (eMoods, 2023), https://emoodtracker.com.

Dataset description

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.”[5]

The version used through this period contains preset mood categories (depressed, irritable, anxious, and elevated) and logs 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.

[5] “eMoods” (eMoods, 2023), https://emoodtracker.com.

Dataset description

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

State	Count
irritable	100
anxious	88
depressed	103
elevated	48

Overview of analysis steps

1. Performed Augmented Dickey-Fuller tests on all variables to **test stationarity**

Overview of analysis steps

1. Performed Augmented Dickey-Fuller tests on all variables to **test stationarity**
2. Selected **optimal lag order** using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Hannan-Quinn Information Criterion (HQIC), and final prediction error (FPE)

Overview of analysis steps

1. Performed Augmented Dickey-Fuller tests on all variables to **test stationarity**

2. Selected **optimal lag order** using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Hannan-Quinn Information Criterion (HQIC), and final prediction error (FPE)

3. Fit a **VAR(2) model** on the multiple time series data

Overview of analysis steps

1. Performed Augmented Dickey-Fuller tests on all variables to **test stationarity**
2. Selected **optimal lag order** using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Hannan-Quinn Information Criterion (HQIC), and final prediction error (FPE)
3. Fit a **VAR(2) model** on the multiple time series data
4. Performed **Granger causality** test in order to assess predictive relationships between variables

Overview of analysis steps

1. Performed Augmented Dickey-Fuller tests on all variables to **test stationarity**
2. Selected **optimal lag order** using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Hannan-Quinn Information Criterion (HQIC), and final prediction error (FPE)
3. Fit a **VAR(2) model** on the multiple time series data
4. Performed **Granger causality** test in order to assess predictive relationships between variables
5. Plotted **impulse response function** to further explore the temporal relationships between variables

On stationarity

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. The Dickey-Fuller test is one mechanism to determine whether a time series is stationary.

On stationarity

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. The Dickey-Fuller test is one mechanism to determine whether a time series is stationary.

I used `statsmodels` and `pymdarima` in python to determine that these were stationary datasets.

Introducing 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.

Introducing Vector Autoregression

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

$$Y_t = \nu + A_1 Y_{t-1} + ... + A_p Y_{t-p} + u_t$$

$$u_t \sim \text{Normal}(0, \Sigma_u)$$

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

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

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 — AIC, BIC, FPE, HQ

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 — AIC, BIC, FPE, HQ

This selection process yielded a tie between lag-1 and lag-2, with each labeled as the minimum across these criteria

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 — AIC, BIC, FPE, HQ

This selection process yielded a tie between lag-1 and lag-2, with each labeled as the minimum across 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

Lag order selection

	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

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

Introducing Granger causality

Granger causality defines one type of relationship between time series[6] 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"[7]

[6] C. W. J. Granger, "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," *Econometrica : Journal of the Econometric Society* 37, no. 3 (1969): 424–38, http://www.jstor.org/stable/1912791.

[7] 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," *Nursing Research* 66, no. 1 (2017): 12–19, https://doi.org/10.1097/NNR.0000000000000193.

Introducing Granger causality

Granger causality defines one type of relationship between time series[6] 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”[7]

Two autoregressive models are fit to the first time series — once with and once without the inclusion of the second time series

[6] C. W. J. Granger, “Investigating Causal Relations by Econometric Models and Cross-Spectral Methods,” *Econometrica : Journal of the Econometric Society* 37, no. 3 (1969): 424–38, http://www.jstor.org/stable/1912791.

[7] 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,” *Nursing Research* 66, no. 1 (2017): 12–19, https://doi.org/10.1097/NNR.0000000000000193.

Introducing Granger causality

Granger causality defines one type of relationship between time series[6] 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”[7]

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

[6] C. W. J. Granger, “Investigating Causal Relations by Econometric Models and Cross-Spectral Methods,” *Econometrica : Journal of the Econometric Society* 37, no. 3 (1969): 424–38, http://www.jstor.org/stable/1912791.

[7] 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,” *Nursing Research* 66, no. 1 (2017): 12–19, https://doi.org/10.1097/NNR.0000000000000193.

Introducing Granger causality

Granger causality defines one type of relationship between time series[6] 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”[7]

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

[6] C. W. J. Granger, “Investigating Causal Relations by Econometric Models and Cross-Spectral Methods,” *Econometrica : Journal of the Econometric Society* 37, no. 3 (1969): 424–38, http://www.jstor.org/stable/1912791.

[7] 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,” *Nursing Research* 66, no. 1 (2017): 12–19, https://doi.org/10.1097/NNR.0000000000000193.

Introducing IRF analysis

An impulse response function (IRF) is the “reaction of a dynamic system in response to an external change”[8]

[8] 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,” *Sensors* 23, no. 1, 1 (January 2023): 332, https://doi.org/10.3390/s23010332.

Introducing IRF analysis

An impulse response function (IRF) is the “reaction of a dynamic system in response to an external change”[8]

Plotting an IRF allows for the interpretation of the impulse of a predictor on other variables on subsequent days

[8] 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,” *Sensors* 23, no. 1, 1 (January 2023): 332, https://doi.org/10.3390/s23010332.

Introducing IRF analysis

An impulse response function (IRF) is the “reaction of a dynamic system in response to an external change”[8]

Plotting an IRF allows for the interpretation of the impulse of a predictor on other variables on subsequent days

Given sleep scoring as a predictor, an IRF visualization was created to better understand its impact on mood state

[8] 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,” *Sensors* 23, no. 1, 1 (January 2023): 332, https://doi.org/10.3390/s23010332.

Results

VAR(2)

Sleep score **was found** to be a significant positive predictor of depression, also confirmed via Granger causality tests. 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

Granger causality

Sleep score **was shown** to Granger-cause both depressed and anxious mood.

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.

Impulse response function

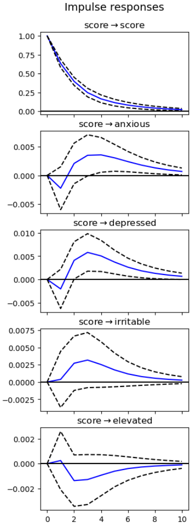

Figure 2: Plot of Impulse Response Function, Lag 0 to 10. Standard errors are plotted at the 95% significance level. The effect of an increase to sleep score on depressed and anxious moods appear to be most felt only after several days of its impact, peaking at roughly 3 days and then gradually decaying.

Summary

Summary

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.

Summary

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.

Limitations: My reliance on algorithmic ADF tests to assess stationarity (rather than directly assessing the data myself) leaves room for error. An incorrect assessment of stationarity risks the accuracy of the remainder of the analysis

Summary

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.

Limitations: My reliance on algorithmic ADF tests to assess stationarity (rather than directly assessing the data myself) leaves room for error. An incorrect assessment of stationarity risks the accuracy of the remainder of the analysis

Future directions: 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