# Results of SINDy

Kevin Cui

July 10, 2019

## 1 Processing of data

Using Xingjia's concern level data, I separated out all the participants who take more than 3 tests (2581 people), and created three csv files to store these matrix, (Y1, R1, R2 respectively). Since the number of people who take four tests are much less than that of people who take three, I choose three snapshots for SINDy algorithm for accuracy, but the result might be inaccurate due to the lack of time frames (usually a minimum one thousand snapshots is required for performing SINDy). I can later perform the SINDy algorithm on four snapshots to see if the results are different.

## 2 Issue of data on performing SINDy

In the morning, Jerry said he tried the SINDy algorithm on our data, and he tested the $48 * 10$ data from our data, which is 48 people multiply by 10 questions. When I talked with him about this,he said that he just wanted to test is the algorithm can actually work. According to the paper by Steven L. Bruntona, to perform the SINDy algorithm, the matrix that we input should have each row representing the state of different variables at each snapshot, and each column representing the state of one variable at each snapshot. As represented by the Figure 1:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}^T(t_1) \\ \mathbf{x}^T(t_2) \\ \vdots \\ \mathbf{x}^T(t_m) \end{bmatrix} = \overset{\text{state}}{\overrightarrow{\begin{bmatrix} x_1(t_1) & x_2(t_1) & \cdots & x_n(t_1) \\ x_1(t_2) & x_2(t_2) & \cdots & x_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(t_m) & x_2(t_m) & \cdots & x_n(t_m) \end{bmatrix}}} \downarrow \text{time}$$

Figure 1: The input matrix of SINDy algorithm

In DMD, we could input the entire matrix with column representing number of questions and row representing each person to get the transition matrix $A$ since we can find different linear transformations for different stages (i.e. A for Y1 to R1, and R1 to R2 are different). However, the SINDy algorithm does not allow us to do this, since if we still input the matrix of number of questions multiply by each person into the SINDy algorithm, we will get the change of 8 concern scores over the sequence of person, (i.e. the derivative means the concern level change from person labelled $i$ to $i + 1$, rather than the time derivative we want). Furthermore, if we do this, we cannot relate each snapshot to each other in time series.

# 3 How I deal with the issue

To deal with the conflict of our data with the SINDy algorithm, we need to input $8 * 3$ matrix with 8 representing 8 variables (8 concern levels) and 3 representing the number of snapshots. To reduce the loss of information by combining the concern level of 2581 people into one single value, I perform the SINDy algorithm on each individual and iterated is 2581 times, and then take the average of these coefficients to get the final coefficient matrix. Admittedly, taking the average of these coefficients still results in the loss of information, it traces the individual change and gives a better estimation for the overall dynamic system. Some other better techniques of determining the coefficients by eliminating the influence of outliers may perform better than simply averaging the coefficients, and we may talk about that in later meeting.

# 4 Result of SINDy

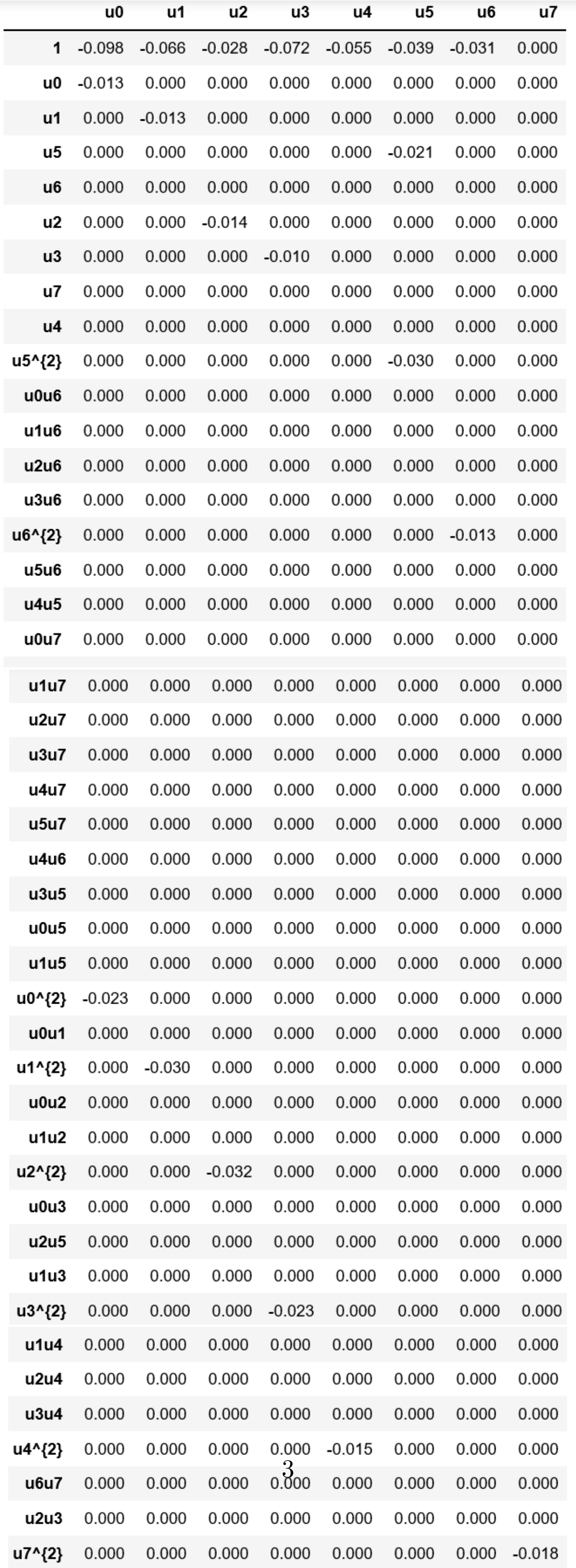
## 4.1 Coefficient matrix

In the iterations, I set the threshold cutoff value for sparsity to 0.01 (I tried other values, but 0.01 performs the best in terms of sparsity and accuracy) to perform the SINDy algorithm on each participants, and average the coefficient matrix and set the threshold to 0.01 again to ensure sparsity.

- All the time derivative of concern levels only depend on themselves, rather than the concern levels in other sections, which does not violate the common sense. With these autonomous differential equations, we can easily solve the system by hand.

- All the coefficients are negative, meaning that the concern level decreases with time in general trend, and it demonstrates the effect of program, and also obeys our intuitive understanding of concern level.

Figure 2 is the output of SINDy algorithm on the eight concern level factors, $u_0$ represents section $a$, $u_1$ represents $b$, etc.



| | u0 | u1 | u2 | u3 | u4 | u5 | u6 | u7 |
|---|---|---|---|---|---|---|---|---|
| **1** | -0.098 | -0.066 | -0.028 | -0.072 | -0.055 | -0.039 | -0.031 | 0.000 |
| **u0** | -0.013 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u1** | 0.000 | -0.013 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u5** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | -0.021 | 0.000 | 0.000 |
| **u6** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u2** | 0.000 | 0.000 | -0.014 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u3** | 0.000 | 0.000 | 0.000 | -0.010 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u7** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u4** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u5^{2}** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | -0.030 | 0.000 | 0.000 |
| **u0u6** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u1u6** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u2u6** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u3u6** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u6^{2}** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | -0.013 | 0.000 |
| **u5u6** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u4u5** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u0u7** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u1u7** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u2u7** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u3u7** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u4u7** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u5u7** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u4u6** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u3u5** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u0u5** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u1u5** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u0^{2}** | -0.023 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u0u1** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u1^{2}** | 0.000 | -0.030 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u0u2** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u1u2** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u2^{2}** | 0.000 | 0.000 | -0.032 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u0u3** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u2u5** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u1u3** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u3^{2}** | 0.000 | 0.000 | 0.000 | -0.023 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u1u4** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u2u4** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u3u4** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u4^{2}** | 0.000 | 0.000 | 0.000 | 0.000 | -0.015 | 0.000 | 0.000 | 0.000 |
| **u6u7** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u2u3** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **u7^{2}** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | -0.018 |

Figure 2: Coefficient Matrix of SINDy

## 4.2 Plot of trend of 8 concern level scores in 6-questionnaires period

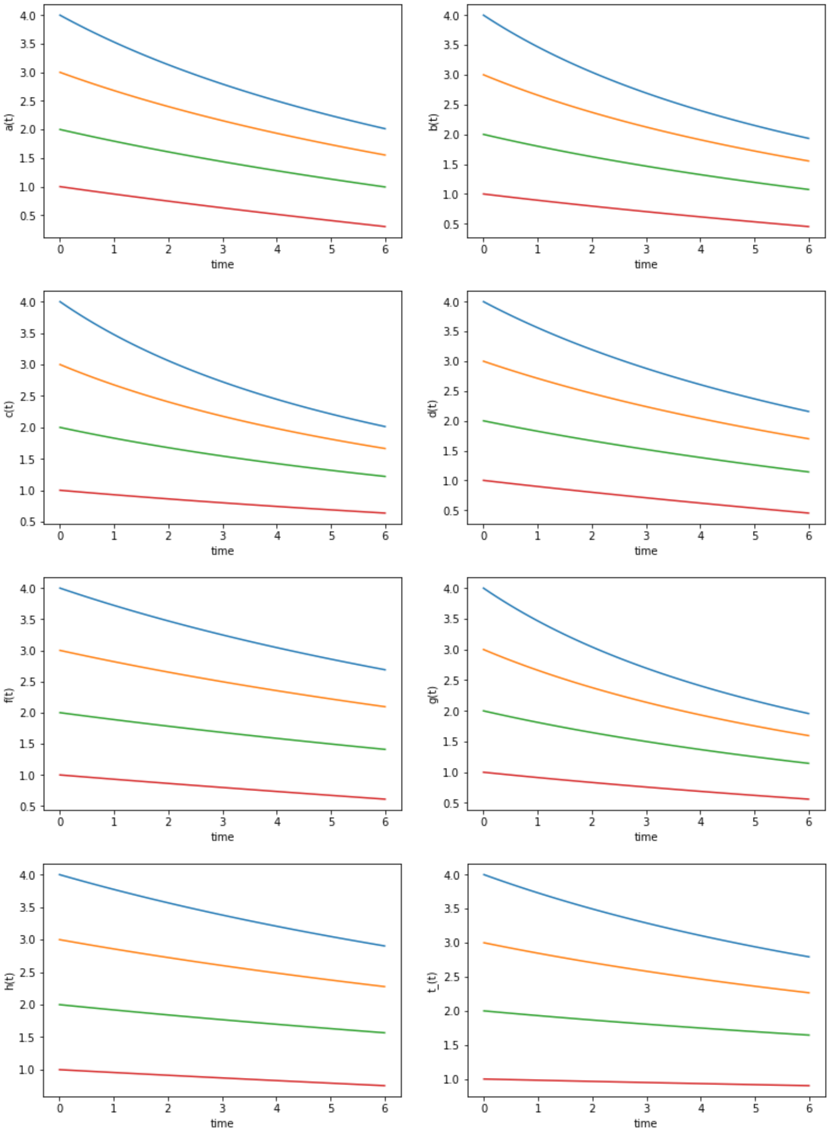


Figure 3: Trend of 8 concern level scores at different initial value

# 5 The analysis of the result

## 5.1 The accuracy of the system

The inaccuracy arise mainly from two aspects:

### 5.1.1 Lack of snapshots (main issue )

### 5.1.2 Setting the threshold

### 5.1.3 Averaging the coefficients

## 5.2 Next steps to make prediction

### 5.2.1 Try four snapshots rather than three

### 5.2.2 Try some other ways of calculating the final coefficients other than averaging

### 5.2.3 Predicting the trend using the system

The prediction might differ hugely from the actual trend due to the issues stated in the section 4, and the accuracy cannot outperform the ensemble methods (including the bagging method we just applied), but it is easier to interpret.