Data Driven Dynamic Discovery of the PPG

Jacob Sindorf

***Abstract*—Data driven dynamic discovery has become a major focus in research based on unknown and complicated systems. Specifically time series datasets from bio-sensors due to their time evolving states, and importance in medical interpretation. Understanding the underlying dynamics of these time series signals would provide essential information in a wide range of applications. One particular bio-signal is the photoplethysmography (PPG) signal. PPG, an optical signal commonly used for heart rate monitoring, has a periodic like pattern. This research applies data driven dynamic discover to the PPG signal with the intent to provide accurate descriptions of the underlying dynamics. Two main methods are explored, being a linear system identification and sparse Identification of nonlinear dynamics (SINDy). SINDy provided promising results and through reinitializing can provide a reconstructed signal capturing some of the PPGs dynamics. Future work to build on the foundations described in this work would provide an accurate PPG reconstruction. These directions include using real and noisy data as well as complicating the SINDy library.**

I. Introduction

Modeling dynamical systems remains an important challenge in research. By obtaining accurate dynamical descriptions of systems, it is possible to study new scenarios and discover even more about the system being observed than before. Highly complicated mathematical models can be created, or estimates and guesses can be formulated. All of which may or may not provide an accurate description of the systems dynamics. One particular group of dynamic discovery is heavily tied to a systems measured data. This grounds the research in a field based on observed data leading to the field of data driven discovery. Data driven discovery however relies only on the data presented and thus it could be highly dependant on having some previous knowledge of the system that produced the data. Understanding or even visualizing data may not always be a possibility bringing up the need for data driven modeling.

Data driven science can be applied to many things and forms the basis for things such as machine learning or as mentioned, discovery of a systems dynamics. An application of this would be for dynamical discovery in bio-signals. Bio-signals here can be generalized as signals received from a sensor placed on a living thing. In this case, human bio-signals include signals received from sources such as accelerometers, electrocardigrams, or photoplethysmography to name a few. Using data driven dynamical discovery on human bio-signals would allow for a greater understanding of the signal received. The uses for better understanding a signals underlying dynamics are endless. With the dynamics described, it could have implications in denoising time series data for better health screening, improving machine learning models to identify specific signal dynamics, or even improving bio-signal simulations given different real parameters. Here a specific bio-signal named photoplethysmography (PPG) is explored in a way to form the basis of not only PPG dynamical discovery, but bio-signal dynamic discovery.

II. Background

A. The PPG Signal

Photoplethysmography (PPG) describes a bio-signal that measures the absortivity of light passed through the skin to detect blood volume changes [1]. Driven by the cardiac cycle, a light sensor, consisting normally of a red and infrared LED, passes light through the skin and the unabsorbed light returned to a photo-diode can be interpolated into the PPG waveform. An example PPG waveform is shown in fig.1. Here it is possible to see the main ideal features where the higher peak describes the systolic peak, and the smaller peak describes the diastolic peak. The time series signal remains relatively periodic in time without the corruption from noise sources. Free of noise, a PPG signal is usually used to measure heart rate and blood oxygen levels or SPO2. With a clean signal it is also possible to compute the derivatives which are labeled velocity PPG (VPG) for the first derivative, and acceleration PPG (APG) for the second derivatives. These values more accurately recognize the inflection points of the signal and allow for more interpretation [2]. A visual representation of these values can also be seen in fig. 1 as approximate derivatives of the simulated PPG signal.

As a cheap and low complexity signal, PPGs have become a rising subject in applications of machine learning, activity recognition, and health screening research projects. However, as mentioned, PPG signals suffer from many noise sources, preventing clean and interpretable signals in many instances. The noise sources are mainly derived from motion artifact (MA) but can also come from power-line interference, drift, or high frequency interruption. These sources of motion change the overall dynamics of the received signal changing the time series dataset. Exploration of a PPG's dynamics through data driven modeling as well as the effects of noise on the data driven dynamics remains as an important area to fully understand the PPG waveform.

B. SINDy

Sparse Identification of nonlinear dynamics (SINDy) is a dynamical system discovery method based on data driven modeling. It was first proposed in [4] and included in [5] and has become a growing method in defining a systems dynamics through just data alone. SINDy can be summarized from [4] as follows:

SINDy itself uses sparse regression in order to find the most relevant terms needed to define a systems dynamics. This therefore makes the governing equations sparse in high dimensional nonlinear function space. In order to use SINDy, data for a system in time $x(t)$ is required, and in the most simple case, it is assumed that the derivative exists or it is possible to derive it numerically from $x(t)$. This time data can be arranged into two separate matrices, labeled $\mathbf{X}$ and $\dot{\mathbf{X}}$. A library $\Theta(\mathbf{X})$ can be created consiting of potential nonlinear functions of $\mathbf{X}$. These usually include the first to the fifth order polynomials, to even trigonometric terms. A sparse regression problem can then be created in order to determine the sparse vector coefficients stored in a matrix Ξ. The equation can be written as $\dot{\mathbf{X}} = \Theta(\mathbf{X})\Xi$, and one Ξ has been solved, each row can have a model constructed of them.

Solving a SINDy problem effectivley allows one to discover the dynamics of a system in which only the data is known. Not only are the dynamics found, but they are found sparsley, meaning only the most relevant terms are kept, keeping the overall system less complex and a lower order if possible. This strategy works well with time series data and brings up the potential to discover new information from signals in which the dynamics are not fully known.In this work, the discovery of the dynamics of a PPG signal are started and explored. Using simulated PPG signals, a basis for PPG dynamics can be created for future dynamic discovery. Starting from the simplest linear model to SINDy.

III. Modeling

A. Linear System Identification

Discovery of a linear relationship of a dynamical system can be simplified by starting with the expression $dx/dt = \phi x$. The linear slope formula can be brought to matrix form to represent the time series dataset and its derivative. To do so the time series dataset can be described as x and its first $\dot{x}$. Taking the first expression, it is possible to write it as $\dot{\mathbf{X}} = \phi\mathbf{X}$ with $\mathbf{X} = \begin{Bmatrix} x \\ \dot{x} \end{Bmatrix}$. This creates a linear relationship between the time series data and its relationship based on the matrix ϕ. Using Matlab's backslash function, ϕ can be solved by $\phi = \mathbf{X}\backslash\dot{\mathbf{X}}$. With the solution to ϕ, the equation can then be numerically solved through Matlab's built in ode45 solver, and a time series estimation can be created and graphed based on the linear system identification.

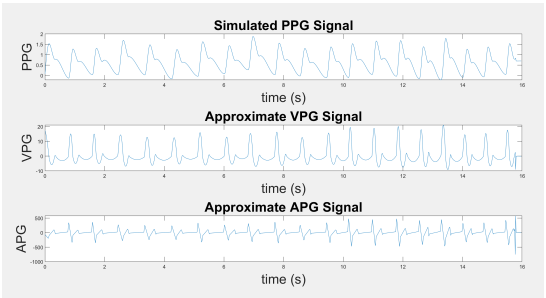

Fig. 1. Simulated PPG waveform from neurokit2 [3] with a heart rate of 75 bpm, and a sampling rate of 64 Hz.

In order to fit the problem, it is best to discretize $\dot{\mathbf{X}} = \phi\mathbf{X}$ by taking $\dot{\mathbf{X}}$ as a set of time series data from $2 \dots n$ and $\mathbf{X}$ from $1 \dots n-1$. This can then be rewritten as the matrix:

$$\begin{bmatrix} x_{n+1} & x_{n+2} & \dots & x_{n+m} \\ \dot{x}_{n+1} & \dot{x}_{n+2} & \dots & \dot{x}_{n+m} \end{bmatrix} = [\phi] \begin{bmatrix} x_n & x_{n+1} & \dots & x_{n+m-1} \\ \dot{x}_n & \dot{x}_{n+1} & \dots & \dot{x}_{n+m-1} \end{bmatrix}$$

With this framework this problem can be further expanded upon. As PPG data is a time series dataset, it can be seen as a projection of some unobserved internal elements of a dynamical system onto a real domain [**?**]. It is important to find someway to correlate how much of the signal would be needed to accurately capture the signals overall behavior. This starts with the method of average mutual information (AMI). Running AMI reveals some steep drop off on a chart that indicates the appropriate number, 'm', to be considered. The methods of this algorithm are explored in a MATLAB code from [6]. Here m can also be seen in the matrix representation as described previously. This allows for a mathematical way to determine the size of the matrix that would be able to capture the systems dynamics.

After finding the AMI, the system can be solved for ϕ through MATLAB's built in backslash function as mentioned. Here ϕ should create a set of solutions for ϕ. The 4 sets of values can then be brought to MATLAB's curve fitter as a polynomial and fit to find a descriptive equation. Altering the degree and robust setting allow for optimization of the r^2 value giving the final constants and equation. For example, a 1 degree equation would be displayed as p1x + p2. Once the 4 value sets have been solved into equations, they can be numerically solved to reconstruct a PPG signal.

B. Nonlinear System Identification with SINDy

Expanding the equation from the linear system identification would be a nonlinearity in the form $f(x)$, giving the equation $\dot{\mathbf{X}} = A\mathbf{X} + f(x)$. Reducing $f(x)$ to a state to simplify the equation to a linear problem is not as trivial as $f(t)$. Here $f(x)$ represents nonlinear functions such as polynomials or trigonometric ones. for example, in a discretized matrix form, consider $f(x) = x^2$, the discretized form can be written as $x_{n+1} = Ax_n + ax_n^2$, then the next time step would be $x_{n+2} = Ax_{n+1} + ax_{n+1}^2$. Through a sparse regression, it would be possible to solve the constants of a. This is where the method of SINDy becomes useful. SINDy can take the $f(x)$ term similarly to the library $\Theta(\mathbf{X})$. The library holds nonlinear functions that may represent the dynamics of the system . Most commonly, it uses the first to the fifth order polynomial, and or trigonometric functions. An application of SINDy has also been developed as a series of examples from [4] that take a time series data set x, obtain the derivatives, build the library, then solve with Matlab's ode45. The overall framework of SINDy is then easily applied to many different time series applications for sparse dynamic discovery.

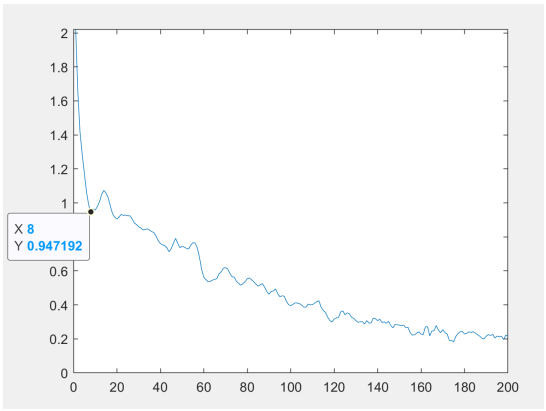

Fig. 2. AMI results chart displaying point of deepest descent.

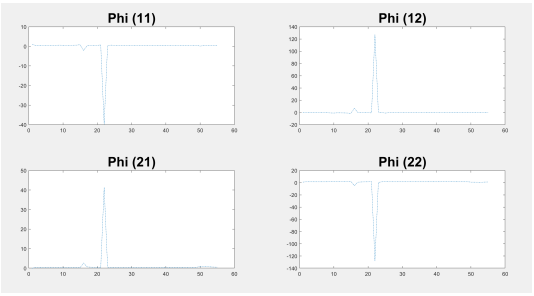

Fig. 3. ϕ solutions displayed as the ϕ matrix.

IV. Results

Overall results for this paper assist in creating the foundation for future work and studies involved in dynamic system discovery of the PPG signal. Here the preliminary results of linear system identification and SINDy are explored. These methods can then be compared and contrasted, providing a basis for future discovery.

A. Linear System Identification

In order to solve the linear system identification problem the data had to be organized and the AMI found. Fig.2 displays the results for the AMI indicating that an m=8 captures the majority of the dynamics needed.

After finding this value, it is possible to solve for ϕ, which yielded the following solutions found in fig.3. Then taking each value of the solution set, the final representative equation for ϕ can be found. Through a polynomial curve fitting and a degree of 1, each of the 4 equations achieved an r^2 value of 0.93. The final results of the system can then be displayed in the following matrix: ϕ =

$$\begin{bmatrix} -0.01255t + 0.5211 & 0.004627t + 0.4854 \\ 0.05845t + 0.543 & -0.01533t + 1.535 \end{bmatrix}$$

Reconstructing the PPG signal numerically would then be possible given the values found in phi however this is further elaborated in the future work section.

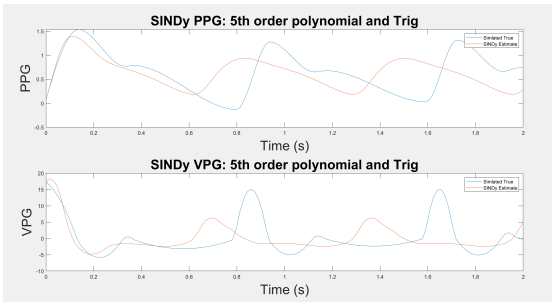

Fig. 4. Preliminary SINDy results on a PPG signal. The top chart zooms in on the first 2 seconds of results of the PPG. The bottom chart zooms in on the first 2 seconds of results of the VPG.

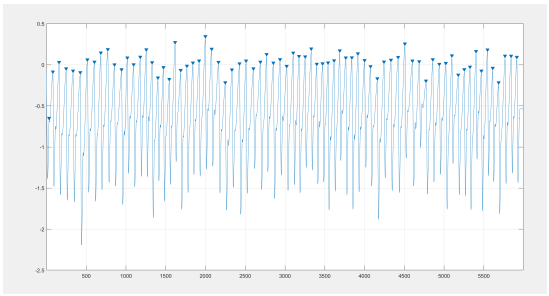

Fig. 5. Full inversed PPG signal (-PPG) and the found peaks needed to separate by period.

B. SINDy

Through SINDy, dynamic discovery given the sparse regression has shown promising results. By applying the SINDy formulation to the PPG signal, then reconstructing a signal based on the results, it is possible to see the promise of the SINDy method in discovering the time series dynamics. Fig.4 shows some preliminary results using the SINDy method. A zoomed in version of this is shown to highlight the first waveform of the PPG. Here the estimated SINDy result was reconstructed and plotted and performed with some decent accuracy at capturing the PPG dynamics. However a major issue arises as the system does not reinitialize over time, and ends up repeating the same periodic wave.

To solve this, it would be best to run the SINDy algorithm on each individual period, or pulse, of the PPG. To do so, fig.5 depicts a method in Matlab using peak finder. By inversing the PPG signal,$-$signal, and finding the peaks, it is possible to identify each individual pulse. The SINDy algorithm can then be applied to each period separately, allowing it to reinitialize over time. Overall, two methods were explored to test SINDy's ability to discover a systems dynamics based on data.

The first would be to just consider a 1D system, or just x, the full PPG signal. Using x and its derivative dx, the Θ library can be built, and by sparsifying, a set of values X_i can be found. Using the first value of each pulse as the initial value, the SINDy algorithm can be finalized and an ode can be solved to reconstruct a PPG signal. Fig.6 shows the results for SINDy and just using the PPG signal. Overall the results struggle to capture the dynamics of the system.

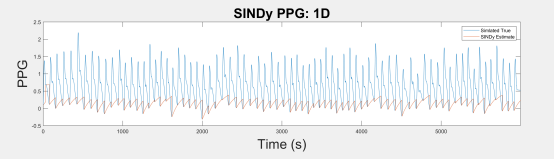

Fig. 6. Results for a 1D SINDy application comparing the PPG signal to the SINDy reconstruction. Uses 5th order polynomial and trigonometric terms.

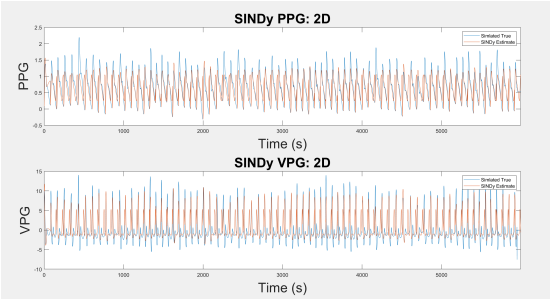

Fig. 7. Results for a 2D SINDy application comparing the PPG signal to the SINDy reconstruction (top), and VPG to reconstruction (bottom). Uses 5th order polynomial and trigonometric terms.

Moving on from 1D leads to a 2D approach. Here instead of x being just the PPG signal, x now contains the VPG as well. As discussed in the methodology, we can represent x as [PPG; VPG], then its derivative, dx, would be [VPG;APG]. Here initial values would be the first value of each pulse given the PPG and VPG, otherwise the methods would be the same as the 1D case. The results are shown in fig.7 where we can now reconstruct both the PPG and VPG signal. Overall the SINDy algorithm begins to capture some of the dynamics.

V. Conclusion

Data driven modeling to discovery a systems underlying dynamics remains a major research direction in time series data. This is especially beneficial in bio-sensor data such as PPG, as an understanding of the dynamics can lead to better simulated signals or abnormalities. This research provides a basis for future work to continue fine tuning and discovering the dynamics of a PPG signal. This includes the PPG signal and its derivatives, VPG and APG. Linear system identification remains the simplest form of dynamical discovery given the linear framework it can be modeled after. The results here are not fully finished as reconstruction of the signal was not performed. Reconstruction of the signal after solving for ϕ remains a challenge even with MATLAB's built in ODE solvers. However, the results can be interpreted in a way to highlight areas of improvement. The first here would be to look into a system that is time variant. The current system would be a time invariant problem, meaning it would only work well on a small set of time series data, which contradicts. This is because time series data depends on time and shows an evolution in time. The second main concern would be to fit the linear model to the periodic like pattern a PPG follows. The benefit to fitting and exploring a linear model still remains relevant due to the lower computational cost of most linear systems.

Perhaps the most insightful results arise from the SINDy algorithm. SINDy has shown great promise in dynamical discovery given a set of data. Here the simplest form of SINDy is explored using both a 1D and a 2D model. It is apparent that the signal must be reinitialized, and that a 2D system resembles the true signal more closely. Given just a 5th order polynomial and a sine and cosine value to the Θ library still yielded a reconstructed signal. With some future exploration, SINDy can be expanded to better fit the PPG signal and allow for a smoother reconstruction.

VI. Future Work

This research has provided groundwork for many future directions. The first few most impactfull directions lie with improving and testing the methods mentioned with both more data and real data. The method for linear system identification requires a reconstruction and to further explore it, a forcing term may be added. A forcing term adds an f(t) to the equation and allows for more to be captured. As for SINDy, given the current results, it would require exploring adding more terms to the Θ library. Specifically more trigonometric terms.

After finalizing the methods, more simulated data can be added and used. This would allow for more cases and situations to be covered in the system. Then given the simulated results, real PPG data can then be used and compared. After that, the next most important direction would be to focus on noisy PPG data. As mentioned, noisy PPG data has become a prevalent issue in PPG interpretation. By exploring the differences in dynamical discovery of simulated, real, and noisy datasets, a basis for future comparisons has been solidified.

With that, the next step would then to be to explore other methods and determine their performance. This would include machine learning as well as specific use cases for the dynamics of the PPG signal.

References

[1] P. A. Kyriacou and S. Chatterjee, "2 - the origin of photoplethysmography," in *Photoplethysmography*, J. Allen and P. Kyriacou, Eds. Academic Press, 2022, pp. 17–43. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128233740000049

[2] M. Elgendi, *PPG Signal Analysis: An Introduction Using MATLAB*. CRC Press, 2020.

[3] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, feb 2021. [Online]. Available: https://doi.org/10.37582Fs13428-020-01516-y

[4] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 113, no. 15, pp. 3932–3937, 2016. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1517384113

[5] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019.

[6] "Mutual average information." [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/880-mutual-average-information