Pre-processing and model identification from sparse non-uniform data

General definitions for the proposed problems

With abundance of data, data-driven model identification has become an important problem in dynamical systems and its allied fields. In health-care related fields, the ever-increasing reliance on patient-specific data is becoming common place. This is even more so for diseases with complex pathology, such as Early Stage Renal Disease (ESRD). The availability of abundant data and use of mechanistic physiology-based models to provide treatments is becoming acceptable. However, in some renal pathophysiologies, the underlying mechanisms exhibiting the dynamical behavior of measured clinical variables are obscured. Therefore, data-driven model identification techniques must be used. However, there are a few issues that hinder the successful use of these techniques. The multiscale nature of these pathologies, and unequal sampling frequencies of the observed variables makes the application of data-driven model derivation difficult.

Let $\vec{f}: \mathrm{R}^N \mapsto \mathrm{R}^N$ be a function that defines a dynamical system of interest, such as:

$$\dot{\vec{x}} = \vec{f}(\vec{x},\vec{\alpha})$$

where $\vec{x}(t)$ is a N-dimensional vector that fully defines the state of the system, and $\vec{\alpha}$ is a m-dimensional vector of parameters. Let $\vec{x}(t)$ represent a trajectory of the dynamical system, given the initial condition $\vec{x}_0 = \vec{x}(t = 0)$. The goal of this approach is to use observational patient data to infer an appropriate form of the dynamical system f (see references a-c below), along with the parameters $\vec{\alpha}$.

Let X be a set of this observational data that represents the time series for $\vec{x}(t)$. However, this representation is not resolved uniformly in time and each variable's data varies in sparsity and noise amplitude. More generally, one can say that:

$$X_{ij} = x_i\left(t_{ij}\right) + \mathrm{A}_i\mathcal{G}(\mu = 0, \sigma^2 = \sigma_i^2) \qquad (1)$$

where A_J is the noise amplitude, and $\mathcal{G}$ is a Gaussian distribution. Important to note is that for each variable x_i there is an associated variance σ_i^2, and that different variables have a different rate of missing values (or are sampled at different rates). The timestamps t_{ij} is a jagged array: the number of columns within each row is different, and the sampling rate within each row is nonuniform. For realistic data sets, there is a tradeoff: variables with a higher sampling rate (sampled two to three values a week) tend to have a very high variance, and variables with low variance have a very low sampling rate (one value every three months).

While the process of finding f from X is well established for well-resolved data, extending these ideas to sparse data sets requires a better quantitative understanding of the given data. These characteristics are needed to reliably modify the method described in the references in order to apply it to the many clinical problems where the available data are obscured by the differential sampling rates among clinical variables. The tasks required for this problem are based on the references below, but alternate approaches are welcome that account for differentially sampled datasets of a multiscale physiological system.

Pre-Processing

1. Given the data set X, estimate the underlying variance to each different variable, i.e., estimate the values σ_i^2. A robust estimate of the data variance σ_i^2 is needed to inform Gaussian process interpolation of the time series $\{X_{ij}\}$.
2. Given the data set X, recover an approximation for the original time series $x_i(t)$ for $i=1,\ldots,N$, i.e., find an approximation for the underlying value of the state variable $x_i(t)$ (eq. (1)) for specific values of t, and the derivative $\dot{x}_\iota(t)$.

Inferring the Dynamical System

Given the data set X, estimate at specific values of t an expected value and variance for the variables $x_i(t)$ and their derivative $\dot{x}_i(t)$, for every i. The values of t can be around positions where there are no missing values in the original data set. Knowing the variance and the derivatives of the processed data are close to that of the real data points, we can develop a more robust method to find a reliable approximation of f for the type of dataset we have, where data are sampled at different frequencies and the system being described is multi-scale.

References

a. Brunton, S. L. *et al.* (2016) 'Discovering governing equations from data by sparse identification of nonlinear dynamical systems', *Proceedings of the National Academy of Sciences of the United States of America*, 113(15), pp. 3932–3937.
doi: 10.1073/pnas.1517384113

b. Lagergren, J. H. *et al.* (2020) 'Learning partial differential equations for biological transport models from noisy spatio-temporal data', *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2234), p. 20190800.
doi: 10.1098/rspa.2019.0800.

c. Mangan, N. M. *et al.* (2016) ‘Inferring Biological Networks by Sparse Identification of Nonlinear Dynamics’, *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, 2(1), pp. 52–63. doi: 10.1109/TMBMC.2016.2633265.