[Note: All submissions must be in PDF format. Failure adhere to the format requirements may result in rejection of your submission without review]

Name:	Hossein Soleimani
Affiliation:	Johns Hopkins University
Email:	Hossein.soleimanib@gmail.com
Presentation type (select one):	Poster

Multivariate Longitudinal Models for Electronic Health Records

Hossein Soleimani[1], Wenbo Pan[1], James Hensman[2], Suchi Saria[1]
[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD; [2]Division of Medicine, Lancaster University, Lancaster, UK

Abstract

We propose multivariate longitudinal models for electronic health record (EHR) data. EHR typically consists of data variables measured with different frequencies; i.e. at any time point, measurements for some variables are available while the others are missing. A standard approach for handling missing data in multivariate analysis is imputation in which each missing data point in one variable is estimated based on the observed values of other variables. This approach, however, may introduce bias and change the representation of the data. Our model, in contrast, has a statistically principled mechanism for handling temporally misaligned data. Using only observed data from different variables, which may, in general, be measured with different frequencies, our model learns a linear decomposition of the data into a set of latent functions. Further, to account for heterogeneity across patients, our model specifies two types of latent functions: functions which are shared across patients and those which are patient-specific. Results of our experiments show that our model achieves promising performance even in cases where significant portions of the measurements from some EHR variables are missing. Our longitudinal model can also very naturally be used in other important healthcare applications such as early warning system and survival analysis.

Introduction

Analyzing electronic health record (EHR) data can lead to significant improvement in various aspects of healthcare. Precision medicine, early warning systems, and disease trajectory prediction are a few example fields which can benefit from effective computational models developed on EHR data. In this work, we focus on modeling longitudinal EHR data, a central task of computational health data analysis.

Missing data and measurements with different frequencies are common problems in health data which should be addressed in developing longitudinal models of health data. EHR data for each individual patient typically consists of several variables which are measured with different frequencies. In particular, at any given time point, we may only have observations for *some* of the variables while the measurements for the rest of variables are missing; i.e. the data is temporally misaligned.

Standard multivariate longitudinal models do not naturally handle temporally misaligned data. In standard multivariate analysis, data is first completed using an imputation process in which, typically, each missing data for every variable is estimated based on the observations from other variables. This simple process, however, can introduce bias and may significantly affect our analysis.

In this work, we develop a multiple-output Gaussian process[1] for multivariate longitudinal modeling of EHR data, which can naturally handle temporally misaligned data. Our model, which follows ideas similar to linear models of coregionalization, uses only observed data from different EHR variables to discover a set of statistically independent latent functions underlying the observations. More specifically, for each EHR variable, our model discovers a linear decomposition into the set of latent functions. These latent functions, which are shared across all variables of each patient, capture all dependencies and correlations between EHR variables. Our model reconstructs and predicts the missing portions of EHR variables using linear combinations of the latent functions. Our model also provides a full distribution on all EHR variables, hence propagating uncertainty in the input, which is due to missing data and measurement irregularities, to the output in a statistically principled fashion. We can use the information from the full distribution to further propagate the uncertainty in observations to any downstream quantity of interest such as hazard functions or early warning predictors.

Figure 1. shows the predicted values (black solid line) and the discovered latent functions (dashed lines) by our model for four EHR variables of one patient. For each variable, we only observe the red points. We also hold out the black points for performance validation. Note that the measurements for different variables are highly misaligned. Nevertheless, our model can still make relatively good predictions. The shaded area around the prediction line denotes confidence interval of prediction.

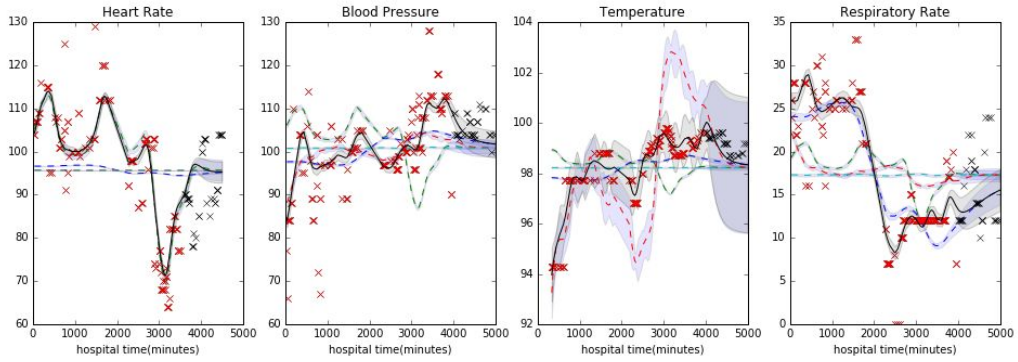

Figure 1. Prediction (solid black line) and latent functions (dashed lines) on four EHR quantities of one patient. Red points and black points are respectively training and test data points.

In addition to handling heterogeneity across EHR variables for one patient, our model also accounts for heterogeneity across patients in the population. More specifically, some latent functions in our model are shared across all patients while others are patient-specific. The former type of latent functions capture the patterns which are common in all individuals in the population. The latter type, however, allows us to model specific patterns exhibited by each patient.

We also develop a scalable learning algorithm for our model which allows us to process EHR data for millions of patients. In future work, we plan to use our longitudinal model in some important healthcare applications such as early warning systems and time-to-event analysis.

References

1. Alvarez MA, Lawrence ND. Computationally efficient convolved multiple output Gaussian processes. *Journal of Machine Learning Research 2011;*12:1459-1500.