BIOS611: Introduction to Data Science **(Time: 12/05/21)**

Final Report of BIOS 611

Instructor: Prof. Toups
Name: Yixiang Qu

Contents

1 Introduction

With progress on both the theoretical and the computational fronts, the use of spline modelling has become an established tool in statistical regression analysis. In particular, splines are regularly used for building explanatory models in clinical research, since irregular time series often occurs in biomedical researches[1]. Indeed, many new methodological developments in modern biostatistics make use of splines to model smooth functions of interest. However, when it comes to very complicated model structure, traditional splines methods may not work well and may not help to make use of all the information. In my first year of my PhD, I made a deep learning tool for better interpolate microbiome data. In order to compare its performance with traditional spline methods, I made this project. It will use Rshiny to show the difference of the interpolation results of my deep learning method and the B-spline method.

2 The complicated data structure of microbiome time series data

The time series data two dimensions originally. However, since there are multiply observations of microbiome species at one time point, the dataset has three dimensions, called a multivariate time series dataset.

Consider a N-subject time series dataset with V-dimensional covariates. We can use $j \in \{1, 2, \ldots, N\}$ to denote the j-th subject. For the j-th subject with L_j samples, we use $\mathbf{T_j} = (t_{j,1}, t_{j,2}, \ldots, t_{j,L_j})$ to denote the observation time points. And for the i-th ($i \in \{1, 2, \ldots, L_j\}$) time point, the observation values $\mathbf{y_{j,i}}$ will be a V-dimensional vector. Therefore, for the j-th subject, we can denote the observation values as $\mathbf{Y_j} \in \mathbb{R}^{L_j \times V}$. And if there are missing values in the j-th subject, we can use mask matrix $\mathbf{M_j} \in \{0, 1\}^{L_j \times V}$ to indicate the missing information, where 0 indicated that the information is missing at the specific time point for the certain subject.

Table 1: A example data format of multivariate microbiome time series dataset

ID	Time	Value_1	Value_2	...	Value_V	Mask_1	Mask_2	...	Mask_V
1	$t_{1,1}$	$y_{1,1,1}$	$y_{1,1,2}$		$y_{1,1,V}$	$m_{1,1,1}$	$m_{1,1,2}$		$m_{1,1,V}$
1	$t_{1,2}$	$y_{1,2,1}$	$y_{1,2,2}$		$y_{1,2,V}$	$m_{1,2,1}$	$m_{1,2,2}$		$m_{1,2,V}$
1	$t_{1,3}$	$y_{1,3,1}$	$y_{1,3,2}$		$y_{1,3,V}$	$m_{1,3,1}$	$m_{1,3,2}$		$m_{1,3,V}$
2	$t_{2,1}$	$y_{2,1,1}$	$y_{2,1,2}$		$y_{2,1,V}$	$m_{2,1,1}$	$m_{2,1,2}$		$m_{2,1,V}$
2	$t_{2,2}$	$y_{2,2,1}$	$y_{2,2,2}$		$y_{2,2,V}$	$m_{2,2,1}$	$m_{2,2,2}$		$m_{2,2,V}$
...									
j	$t_{j,1}$	$y_{j,1,1}$	$y_{j,1,2}$		$y_{j,1,V}$	$m_{j,1,1}$	$m_{j,1,2}$		$m_{j,1,V}$
j	$t_{j,2}$	$y_{j,2,1}$	$y_{j,2,2}$		$y_{j,2,V}$	$m_{j,2,1}$	$m_{j,2,2}$		$m_{j,2,V}$
...									
j	t_{j,L_j}	$y_{j,L_j,1}$	$y_{j,L_j,2}$		$y_{j,L_j,V}$	$m_{j,L_j,1}$	$m_{j,L_j,2}$		$m_{j,L_j,V}$
...									
N	$t_{N,1}$	$y_{N,1,1}$	$y_{N,1,2}$		$y_{N,1,V}$	$m_{N,1,1}$	$m_{N,1,2}$		$m_{N,1,V}$
N	$t_{N,2}$	$y_{N,2,1}$	$y_{N,2,2}$		$y_{N,2,V}$	$m_{N,2,1}$	$m_{N,2,2}$		$m_{N,2,V}$
N	$t_{N,3}$	$y_{N,3,1}$	$y_{N,3,2}$		$y_{N,3,V}$	$m_{N,3,1}$	$m_{N,3,2}$		$m_{N,3,V}$

3 Interpolation results of different spline methods

3.1 Deep learning interpolation

The model to interpolate the time series data is related to my paper, which is under preperation, so I apologize that I cannot write too much details in the report. A concise diagram is shown as Fig. 1.

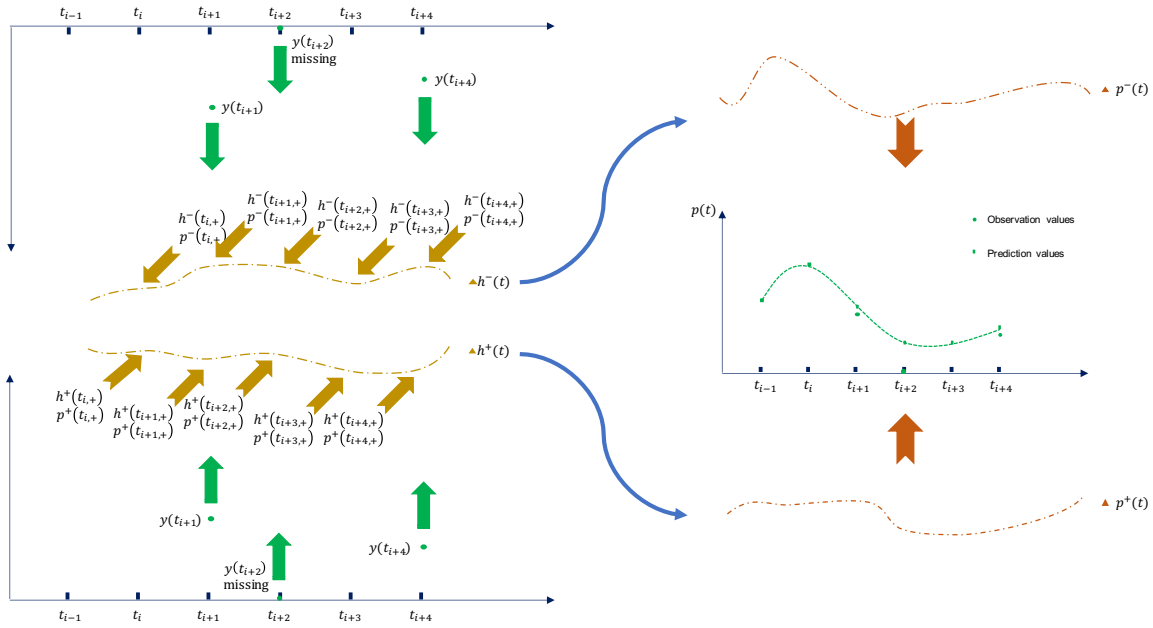

Figure 1: A concise diagram of deep learning model

There are two procedures to use deep learning method to interpolate the data.

1. Feed all data into the model, and learn the time trend using the deep learning model.
2. Use the tuned-parameter model to interpolate the data.

I used Pytorch[2] to implement the deep learning model. And the interpolation results of deep learning model is shown in Fig. 2.

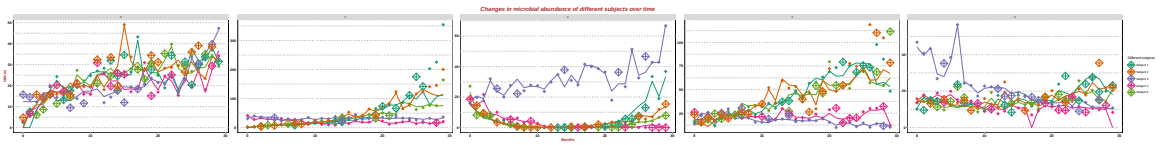

Figure 2: Results of deep learning interpolation.
The information shown in solid points are used to train the model. However, the rest information was not used for the training process.

3.2 B spline interpolation

B-spline is widely used in nonparametric models in biomedical data. B-splines can be defined by construction by means of the Cox-de Boor recursion formula. Given a knot sequence $\ldots, t_0, t_1, t_2, \ldots$, then the Bsplines of order 1 are defined by Eq. 3.1.

$$B_{i,1}(x) := \begin{cases} 1 & \text{if} \quad t_i \leq x < t_{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

These satisfy $\sum_i B_{i,1}(x) = 1$ for all x because for any x exactly one of the $B_{i,1}(x) = 1$, and all the others are zero. The higher order B-splines are defined by recursion, shown in Eq. 3.2 and Eq. 3.3.

$$\omega_{i,k}(x) := \begin{cases} \frac{x - t_i}{t_{i+k} - t_i}, & t_{i+k} \neq t_i \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

$$B_{i,k+1}(x) := \omega_{i,k}(x) B_{i,k}(x) + [1 - \omega_{i+1,k}(x)] \, B_{i+1,k}(x) \tag{3.3}$$

I use gam[3] to do the B-spline interpolation. And the interpolation results of B-spline model is shown in Fig. 3.

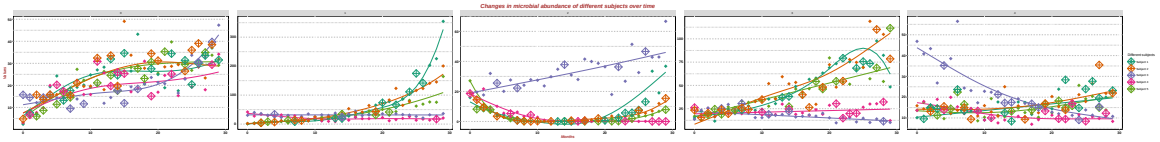

Figure 3: Results of B spline interpolation.
The information shown in solid points are fed into the spline model. However, the rest information was not used for this process.

3.3 Comparsion of the different interpolation methods

We can find out from Fig. 2 and Fig. 3 that the B spline method seems too smooth that it can not represent the fluctuation of microbiome time series. And we use mean difference to compare the interpolation results of different methods, which is shown in Eq. 3.4, and the results are shown in Table 2.

$$\text{Mean Difference} = \frac{1}{n} \sum_{i=1}^{n} |p_i - o_i| \tag{3.4}$$

Table 2: Mean Difference of different interpolation methods

	Deep Learning	B Spline
Mean Difference	3.205	4.039

4 Conclusion

From the previous figures and tables, we can see that the deep learning methods perform better in the interpolation task. The main reason for it is that the deep learning method will learn the time trend of all the subjects, and use the learned information to interpolate the results. However, the spline will only use the information of a single subject. However, the deep learning method will cost far more time than spline method. Therefore, we need to decide whether speed or precision should be the priority before choosing the methods.

References

[1] Ashley R Coenen, Sarah K Hu, Elaine Luo, Daniel Muratore, and Joshua S Weitz. A primer for microbiome time-series analysis. *Frontiers in genetics*, 11, 2020.

[2] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[3] Trevor Hastie. *gam: Generalized Additive Models*, 2020. R package version 1.20.