Statistics and Its Interface Volume 15 (2022) 159–160

Discussion on "Estimation of Hilbertian varying coefficient models"

Jianhua Z. Huang*, Jiangli Wang, and Huiya Zhou

We would like to congratulate Dr. Lee, Dr. Park, Dr. Hong, and Dr. Kim for developing a rather comprehensive theory and methodology for a general type of varying coefficient regression model applicable when the response variable takes values in a Hilbert space. It adds a powerful new tool in a statistician's toolbox. The practical relevance of the new theory and methodology is nicely illustrated in a real data application of predicting electricity consumption. While we enjoyed very much reading the theoretical part of the paper, we shall focus our discussion on the real data example presented in Section 4.3 of the paper.

The response variable is a functional variable, the monthly average of the daily consumption trajectory. The three predictor variables are temperature, cloudiness, and a binary variable indicating weekday or weekend. A natural question is whether the varying coefficient model (4.5) is an adequate model. Are there other variables that are useful for prediction? For example, the season indicator (e.g. winter, spring, summer, or fall) may be a useful predictor. Only additive effects of the temperature and the cloudiness enter the model. Is there an interaction effect? Temperature varies during the day and thus, if more detailed temperature data (e.g., hourly) is available and used as a functional predictor, can the prediction error be reduced? There are many tools in classical statistics for model checking, such as residual plots, leverage plots. It is of interest to know what tools are available when the response variable is Hilbertian.

It may be more straightforward to apply two separate additive models, one to the weekend data, one to the weekday data. The varying coefficient model (4.5) was applied to the weekend and weekday data combined. What is the advantage of this unified model? Does it have better prediction performance than using two separate additive models? Compared to using two separate additive models, are there additional underlying assumptions for using the varying coefficient model?

Some data pre-processing steps were preformed before fitting the model to data and comparing the prediction performance. Specifically, the hourly electricity consumption was first normalized, then smoothed by employing the local linear smoothing to obtain a function on $[0, 24]$, and finally re-normalized so that the function has unit integral. The processed data were used as the response variable for model fitting and as the target for prediction. What is the purpose of normalization? It may be practically more interesting to predict the actual hourly consumption instead of the consumption pattern (i.e., the shape of the daily consumption trajectory), for which normalization would distort the prediction goal. On the other hand, suppose there is a good reason to predict the monthly average of daily consumption pattern. A more natural normalization approach would be to first obtain the daily consumption pattern through normalization, and then take the average over the whole month; of course, the average needs to be defined appropriately. The approach taken in the paper is to first obtain the monthly average of the (non-normalized) consumption trajectory and then normalize. In what situation would one use the prediction of such a quantity? The data smoothing step may introduce bias and also make the induced variable hard to interpret. How does one interpret the functional data (e.g., $Y_i(t)$ in Section 4.3) resulting from the pre-processing process? Can the prediction of $Y_i(t)$ be converted to a prediction of a quantity that is easily interpretable, such as hourly electricity consumption?

One advantage of the general framework is to allow non-traditional definition of vector addition and scalar multiplication in the Hilbert space, which is convenient for modeling probability densities as a response variable (e.g., Section 4.1). However, the coefficient functions or their additive components in such a varying coefficient model may not be easy to interpret. For instance, how do we interpret the coefficient functions or their additive components in model (4.5)? The binary variable weekday/weekend is coded as $\pm 1/2$. If it is coded as a 0/1-valued dummy variable, how would the interpretation change? Because of the non-traditional definition of vector addition and scalar multiplication, different coding schemes may yield dramatic different shapes of the coefficient functions.

Suppose, for simplicity, we only consider the effect of temperature and weekend/weekday, model (4.5) can be simplified to

$$\mathbb{E}(Y|X_1, Z) = f_{1,3}(X_1) \oplus Z \odot f_{1,4}(X_1).$$

Consider two coding schemes mentioned above, and introduce a new notation to denote the corresponding coding

*Corresponding author.

variable, i.e., $Z^{(1)} = \pm 1/2$, $Z^{(2)} = 0/1$, and weekdays correspond to $Z^{(1)} = 1/2$, $Z^{(2)} = 1$. We can write the corresponding models as

$$\text{(Model 1:)} \quad \mathbb{E}(Y|X_1, Z^{(1)}) = f_{1,3}^{(1)}(X_1) \oplus Z^{(1)} \odot f_{1,4}^{(1)}(X_1),$$

$$\text{(Model 2:)} \quad \mathbb{E}(Y|X_1, Z^{(2)}) = f_{1,3}^{(2)}(X_1) \oplus Z^{(2)} \odot f_{1,4}^{(2)}(X_1).$$

The coefficient functions in these two models have a simple relationship:

$$f_{1,3}^{(2)} = f_{1,3}^{(1)} \ominus \frac{1}{2} \odot f_{1,4}^{(1)},$$

$$f_{1,4}^{(2)} = f_{1,4}^{(1)}.$$

If the traditional definition of vector addition and scalar multiplication is used, such relationship is easily interpretable. In particular, $f_{1,3}^{(2)}$ is $f_{1,3}^{(1)}$ shifted by one half of the values of the function $f_{1,4}^{(1)}$. Such a simple explanation does not hold when the non-traditional definition of vector addition and scalar multiplication is used.

To conclude, we are grateful to the Editors-in-Chief, the guest editors of this special issue, and the authors for providing us with this opportunity to discuss an inspiring work. Our discussion has been mainly on practical issues of applying the new methodology. We look forward to hearing some insights from the authors on these practical issues.

Received 18 April 2021

Jianhua Z. Huang
Texas A&M University
College Station, TX 77843
USA
E-mail address: jianhua@stat.tamu.edu

Jiangli Wang
University of Science and Technology of China
Hefei, Anhui 230027
China
E-mail address: wjl91@mail.ustc.edu.cn

Huiya Zhou
Renmin University of China
Beijing 100872
China
E-mail address: freedom00y@ruc.edu.cn