Statistics and Its Interface Volume 15 (2022) 153–154

Discussion on "Estimation of Hilbertian varying coefficient models"

Bu Zhou, Jia Guo, and Jin-Ting Zhang*

1. INTRODUCTION

Lee, Park, Hong, and Kim are to be congratulated for this interesting and excellent exposition on a general varying coefficient model with a Hilbertian response and a set of real-valued predictors. We shall refer to the authors as LPHK whenever we mention them in what follows.

Unlike the flexible generalized varying coefficient model with a Euclidean response investigated in Lee, Mammen and Park [2], the theoretical treatment of the new general varying coefficient model in this paper is abstract and complicated. LPHK have done a nice job in model identification and estimation. In particular, they proved the convergence of the backfitting algorithm that yields the estimators and derived the rates of convergence of the estimators and their asymptotic distributions. The proposed estimators are evaluated and illustrated by a simulation study and a real data application, respectively.

In this discussion, we would like to suggest an alternative way for model formulation and interesting follow-up research on significance test of component maps and analysis of variance of density functions.

2. MODEL FORMULATION

The general varying coefficient model studied by LPHK allows the predictors to enter the model linearly or nonlinearly. This gives much flexibility in model building, but it may cause some difficulties in model formulation, identification, and interpretation. Take the model formulation as an example. To describe the model, LPHK divide the d predictors into several parts in the following way:

$$(1)\qquad \underbrace{\underbrace{X_1,\ldots,X_{d_0-r}}_{\text{nonlinear}},\underbrace{X_{d_0-r+1},\ldots,X_{d_0}}_{\text{linear \& nonlinear}}}_{\text{continuous}},\ \underbrace{\underbrace{X_{d_0+1},\ldots,X_d}_{\text{linear}}}_{\text{continuous or discrete}},$$

where the predictors are ordered according to their types (without loss of generality) so that it is sufficient to use only two numbers $d_0 - r$ and d_0 to divide the whole predictor set into three parts of different types. In practice, we may first sort the predictors into the form (1) and then identify d_0 and r. Based on the "sorted" predictors, we have the following two equivalent expressions:

$$(2)\qquad \mathrm{E}(\boldsymbol{Y}|\boldsymbol{X}) = \oplus_{j=1}^{d_0}\left[\oplus_{k\in I_j} X_k \odot \boldsymbol{f}_{j,k}(X_j)\right]$$

$$(3)\qquad = \oplus_{k=d_0-r+1}^{d}\left[\oplus_{j\in \tilde{I}_k} X_k \odot \boldsymbol{f}_{j,k}(X_j)\right].$$

In the first expression (2), the first sum operation is a summation of all predictors in the nonlinear part (that can be the input of the component map $\boldsymbol{f}_{j,k}$), and the second sum operation is a summation of all predictors in the linear part and interact with the mapped value of X_j (i.e., predictors with indices in the set I_j). In the second expression (3), the first sum operation is a summation of all predictors in the linear part while the second sum operation is a summation of all predictors that can be a nonlinear part that interacts with X_k (i.e., predictors with indices in the set $\tilde{I}_k$). The equivalence of the two expressions (2) and (3) is in fact an exchange of the double summation operations, but the indices in the company sets I_j and $\tilde{I}_k$ of the second sum operations have no order as in the first sum operations, even all predictors are sorted as in (1), which further complicates the expression of the model. In what follows, we describe an alternative way for model formulation.

We suggest to formulate the condition mean $\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})$ in the following way:

$$(4)\qquad \mathrm{E}(\boldsymbol{Y}|\boldsymbol{X}) = \oplus_{j,k=1}^{d} a_{jk} X_k \odot \boldsymbol{f}_{j,k}(X_j),$$

where the indicator $a_{jk} = 1$ when the varying coefficient model contains the interaction term between the mapped value of X_j and X_k, and $a_{jk} = 0$ otherwise. Obviously, the above model formula can be simply specified by an indicator matrix $\boldsymbol{A} = (a_{ij})_{i,j=1}^{d}$ whose entries are either 1 or 0. The expression (4) includes all the d^2 pairs of predictors and nonlinear mapped values of predictors, and the entries of the indictor matrix $\boldsymbol{A}$ are used to determine if a pair is included in the model or not. The above new model formulation is very flexible in the varying coefficient model specification. For example, to exclude the terms like $X_j \odot \boldsymbol{f}_{j,j}(X_j)$ from the varying coefficient model, one just needs to set all the diagonal entries of $\boldsymbol{A}$ to be 0; to define the subset of predictors appearing in the nonlinear part, one just needs to write $\{X_j : \exists k, \text{s.t. } a_{jk} = 1\} = \{X_j : \sum_{k=1}^{d} a_{jk} > 0\}$; to define the subset of predictors appearing in the linear part, one just needs to write $\{X_k : \sum_{j=1}^{d} a_{jk} > 0\}$; to define the subset of predictors appearing in both the linear and nonlinear parts, one just needs to write $\{X_\ell :$

*Corresponding author. ORCID: 0000-0002-9677-4398.

$(\sum_{j=1}^{d} a_{j\ell})(\sum_{k=1}^{d} a_{\ell k}) > 0\}$; and the subset of predictors appearing only in the nonlinear or linear part can be expressed as $\{X_\ell : (\sum_{j=1}^{d} a_{j\ell}) = 0 \text{ and } (\sum_{k=1}^{d} a_{\ell k}) > 0\}$ or $\{X_\ell : (\sum_{j=1}^{d} a_{j\ell}) > 0 \text{ and } (\sum_{k=1}^{d} a_{\ell k}) = 0\}$, respectively. Admittedly, it needs some further investigation to check whether the above new model formulation can simplify the notations or better facilitate the analysis of the varying coefficient model.

3. TESTING THE SIGNIFICANCE OF COMPONENT MAPS

As mentioned by LPHK in their concluding remark, an important problem is to identify significant component maps in the varying coefficient model. To this end, one may consider the hypothesis testing problem $H_0 : \boldsymbol{f}_{j,k} = \mathbf{0}$ for a given j and all $k \in I_j$. Equivalently, we can write $H_0 : \boldsymbol{f}_j = \mathbf{0}$, where $\boldsymbol{f}_j = (\boldsymbol{f}_{j,k} : k \in I_j)$ as defined by LPHK. Following Zhang, Guo and Zhou [4], Zhang et al. [5], and Smaga and Zhang [3] among others, a natural test statistic can be constructed using the squared L^2-norm of the backfitting estimator $\hat{\boldsymbol{f}}_j$, that is $T_n = \sum_{k\in I_j} \|\hat{\boldsymbol{f}}_{j,k}\|_{\mathbb{H}}^2$ where $\|\hat{\boldsymbol{f}}_{j,k}\|_{\mathbb{H}}^2$ denotes the squared L^2-norm of $\hat{\boldsymbol{f}}_{j,k}$ as defined by LPHK. To conduct the test, one may approximate the null distribution of T_n analytically or nonparametrically. For example, based on the theoretic results established by LPHK in Theorem 3, one may show that the test statistic T_n and a chi-square mixture have the same normal or non-normal limiting distribution. This shows that one may approximate the null distribution of T_n using the Welch–Satterthwaite chi-square approximation as done in Zhang, Guo and Zhou [4] and Zhang et al. [5] among others. Alternatively, one may approximate the null distribution of T_n using a parametric bootstrap approach or a nonparametric bootstrap approach. Further studies in this direction are interesting and warranted.

4. ANALYSIS OF VARIANCE OF DENSITY FUNCTIONS

To demonstrate the performance of the proposed backfitting estimators of the varying coefficient model, LPHK present a simulation study and a real data example where density functions are analyzed. Delicado [1] and Smaga and Zhang [3] considered the analysis of variance (ANOVA) problem of density functions, which tests if k underlying mean density functions are the same based on k given samples of the random density functions. This ANOVA problem can be treated as a special case of the general varying coefficient model considered by LPHK, where the only predictors are the k dummy variables indicating which group the random density functions belong to. Under this special varying coefficient model, the ANOVA problem can be tested via applying the above L^2-norm-based test to check if the coefficients of the k dummy variables are the same. A question arises naturally. In terms of size control and power, which approach is preferred? Further studies in this direction are also interested and warranted.

ACKNOWLEDGEMENTS

Bu Zhou was supported by the National Natural Science Foundation of China under Grant No. 11901520 and the Zhejiang Provincial Natural Science Foundation of China under Grant No. LY21A010007. Jia Guo was supported by the National Natural Science Foundation of China under Grants No. 11901522. Jin-Ting Zhang was financially supported by the National University of Singapore Academic Research grant R-155-000-187-114.

Received 26 March 2021

REFERENCES

[1] Delicado, P. (2007). Functional k-sample problem when data are density functions. *Computational Statistics* **22** 391–410. MR2336343

[2] Lee, Y. K., Mammen, E. and Park, B. U. (2012). Flexible generalized varying coefficient regression models. *The Annals of Statistics* **40** 1906–1933. MR3015048

[3] Smaga, L. and Zhang, J.-T. (2020). Linear hypothesis testing for weighted functional data with applications. *Scandinavian Journal of Statistics* **47** 493–515. MR4157148

[4] Zhang, J.-T., Guo, J. and Zhou, B. (2017). Linear hypothesis testing in high-dimensional one-way MANOVA. *Journal of Multivariate Analysis* **155** 200–216. MR3607891

[5] Zhang, J.-T., Guo, J., Zhou, B. and Cheng, M.-Y. (2020). A simple two-sample test in high dimensions based on L^2-norm. *Journal of American Statistical Association* **115** 1011–1027. MR4107696

Bu Zhou
School of Statistics and Mathematics
Zhejiang Gongshang University
China
E-mail address: bu.zhou@u.nus.edu

Jia Guo
School of Management
Zhejiang University of Technology
China
E-mail address: jia.guo@u.nus.edu

Jin-Ting Zhang
Department of Statistics and Applied Probability
National University of Singapore
Singapore
E-mail address: stazjt@nus.edu.sg