Journal of the Korean Statistical Society ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Robust adaptive model selection and estimation for partial linear varying coefficient models in rank regression

Xiaofei Sun [a], Kangning Wang [a,b,*], Lu Lin [b]

[a] School of Statistics, Shandong Technology and Business University, Yantai, China
[b] Institute of Financial Studies, Shandong University, Jinan, China

ARTICLE INFO

Article history:
Received 7 May 2017
Accepted 7 September 2017
Available online xxxx

AMS 2000 subject classifications:
62G05
62E20
62J02

Keywords:
Partial linear varying coefficient models
Rank regression
Selection consistency
Oracle property
Robustness and efficiency

ABSTRACT

Partial linear varying coefficient models are often used in real data analysis for a good balance between flexibility and parsimony. In this paper, we propose a robust adaptive model selection method based on the rank regression, which can do simultaneous coefficient estimation and three types of selections, i.e., varying and constant effects selection, relevant variable selection. The new method has superiority in robustness and efficiency by inheriting the advantage of the rank regression approach. Furthermore, consistency in the three types of selections and oracle property in estimation are established as well. Simulation studies also confirm our method.

1. Introduction

Partial linear varying coefficient models (PLVCM) (Ahmad, Leelahanon, & Li, 2005; Fan & Huang, 2005; Kai, Li, & Zou, 2011) are often considered for its good balance between flexibility and parsimony. There are a large number of literatures on the estimation and variable selection for PLVCM. For estimation, we refer Ahmad et al. (2005), Fan and Huang (2005), Kai et al. (2011), Wang, Zhu, and Zhou (2009), Zhang, Zhao, and Liu (2013) and Zhou and Liang (2009). For variable selection, examples include but are not limited to Li and Liang (2008), Wang, Li, and Huang (2008), Wang and Lin (2016), Wang and Xia (2009), Zhang et al. (2013), Zhao and Xue (2009) and Zhao, Zhang, Liu, and Lv (2014).

The most important assumption in the aforementioned methods is to assume that the subset of variables having constant or varying effect on the response is known in advance, or say, the true model structure is determined. This assumption underlies the construction of the estimators and investigation of their theoretical properties in the existing methods. However, in the application, it is unreasonable to artificially determine which subset of variables has constant or varying effect on the response.

To solve the above problem, Hu and Xia (2012), Leng (2009), Noh and Keilegom (2012) and Xia, Zhang, and Tong (2004) proposed some methods to identify the partial linear structure. Furthermore, Tang, Wang, Zhu, and Song (2012) proposed unified methods, which can select the relevant variables and partial linear structure simultaneously.

However, the aforementioned methods are mainly built upon mean regression or likelihood based methods, which can be adversely influenced by outliers or heavy-tail distributions. Although, Tang et al. (2012) gave a quantile regression method

* Corresponding author at: School of Statistics, Shandong Technology and Business University, Yantai, China.
E-mail address: wkn1986@126.com (K. Wang).

https://doi.org/10.1016/j.jkss.2017.09.003
1226-3192/

which is robust, it has limitations in terms of efficiency. Furthermore, the method in Tang et al. (2012) needed an iterative two-step procedure that is very inconvenient in the application. Hence, it would be highly desirable to develop an efficient and robust adaptive method that can simultaneously conduct model identification and estimation in one step.

Recently, Wang, Kai, and Li (2009) proposed a novel procedure for the varying coefficient model based on rank regression and demonstrated that the new method is highly efficient across a wide class of error distributions and possesses comparable efficiency in the worst case scenario compared with mean regression. Similar conclusions on rank regression have been further confirmed in Feng, Zou, Wang, Wei, and Chen (2015), Leng (2010), Sun and Lin (2014), and the references therein.

Therefore, motivated by the above discussion, we propose a robust adaptive model selection procedure in the rank regression setting, which can do simultaneous coefficient estimation and three types of selections, i.e., varying and constant effects selection, relevant variable selection. Specifically, we first embed the PLVCM into a varying coefficient model and use the spline method to approximate unknown functions. Then, a two-fold SCAD (Fan & Li, 2001) penalty is employed to discriminate the nonzero components as well as linear components from the nonlinear ones by penalizing both the coefficient functions and their first derivatives. The new adaptive selection procedure has superiority in robustness and efficiency by inheriting the advantage of the rank regression approach. Furthermore, consistency in the three types of selections and oracle property in estimation are established as well. Although, Feng et al. (2015) also proposed a penalized rank regression procedure, their method is only for selecting the relevant variables, which is completely different from our method.

The rest of this paper is organized as follows. In Section 2, we introduce the new method and investigate its theoretical properties and related implementation issues. Numerical studies are reported in Section 3. All the technical proofs are provided in Appendix.

2. Robust adaptive model selection in rank regression

2.1. Two-fold penalization rank regression

Suppose that the observed full data set is

$$\boldsymbol{D}_n = \{D_i = (Y_i, \boldsymbol{X}_i, U_i), i = 1, \ldots, n\}, \tag{2.1}$$

where Y_i is the response of the ith observation, $\boldsymbol{X}_i = (X_i^{(1)}, \ldots, X_i^{(p)})^T \in \mathbb{R}^p$ is the covariate vector, and assume index variable $U_i \in [0, 1]$ without loss of generality. Then PLVCM with underlying true partial linear structure and irrelevant variables must have the following form:

$$Y_i = \sum_{k\in\mathcal{A}_V} X_i^{(k)}\alpha_k(U_i) + \sum_{k\in\mathcal{A}_C} X_i^{(k)}\beta_k + \sum_{k\in\mathcal{A}_Z} X_i^{(k)}0(U_i) + \epsilon_i, \tag{2.2}$$

where unknown sets $\mathcal{A}_V$, $\mathcal{A}_C$ and $\mathcal{A}_Z$ are the index sets for varying effects, nonzero constant effects and zero effects, respectively, they are mutually exclusive and satisfy $\mathcal{A}_V \bigcup \mathcal{A}_C \bigcup \mathcal{A}_Z = \{1, \ldots, p\}$.

Thus, given data set $\boldsymbol{D}_n$, our main aim is to identify the index sets $\mathcal{A}_V$, $\mathcal{A}_C$, $\mathcal{A}_Z$, and estimate the nonzero coefficients $\alpha_k(u),\ k \in \mathcal{A}_V$ and $\beta_k,\ k \in \mathcal{A}_C$ efficiently and robustly.

As the partial linear structure is unknown in advance, the PLVCM (2.2) can be embedded into the following varying coefficient model:

$$Y_i = X_i^{(1)}\alpha_1(U_i) + \cdots + X_i^{(p)}\alpha_p(U_i) + \epsilon_i. \tag{2.3}$$

Thus, if $\alpha_k(u) \equiv 0$ for $u \in [0, 1]$, $X^{(k)}$ is an irrelevant variable, otherwise, if derivative $\alpha_k'(u) \equiv 0$ for $u \in [0, 1]$, then $X^{(k)}$ has constant effect, otherwise, $\alpha_k(u)$ is a varying function. Therefore, problem becomes that of determining which $\alpha_k(\cdot)$s are zero functions and which $\alpha_k'(\cdot)$s are zero functions.

Then, we can use the polynomial splines to approximate the $\alpha_k(\cdot)$s. Let $0 = \tau_0 < \tau_1 < \cdots < \tau_{K_n} < \tau_{K_n+1} = 1$ be a partition of $[0, 1]$ into $K_n + 1$ subintervals $I_{nj} = [\tau_j, \tau_{j+1}), j = 0, \ldots, K_n - 1$, and $I_{nK_n} = [\tau_{K_n}, \tau_{K_n+1}]$, where $K_n = n^{\vartheta}$ with $0 < \vartheta < 0.5$ is a positive integer such that $\max_{1\le j\le K_n+1}|\tau_j - \tau_{j-1}| = O(n^{-\vartheta})$. Let $\mathcal{F}_n$ be the space of polynomial splines of degree $D \geq 1$ consisting of functions f satisfying: (i) the restriction of f to I_{nj} is a polynomial of degree D for $0 \leq j \leq K_n$; (ii) f is $(D-1)$-times continuously differentiable on $[0, 1]$ (Schumaker, 1981). There exist B-spline basis functions $\boldsymbol{B}(\cdot) = (B_{1,D}(\cdot), \ldots, B_{d_n,D}(\cdot))^T$ for $\mathcal{F}_n$, where $d_n = K_n + D + 1$ (Schumaker, 1981). Then $\alpha_k(u)$ can be approximated as

$$\alpha_k(u) \approx \sum_{j=1}^{d_n} B_{j,D}(u)\gamma_{k,j} = \boldsymbol{B}(u)^T\boldsymbol{\gamma}_k, \tag{2.4}$$

where $\boldsymbol{\gamma}_k = (\gamma_{k,1}, \ldots, \gamma_{k,d_n})^T$.

Please cite this article in press as: Sun, X., et al., Robust adaptive model selection and estimation for partial linear varying coefficient models in rank regression. Journal of the Korean Statistical Society (2017), https://doi.org/10.1016/j.jkss.2017.09.003.