Computational Statistics and Data Analysis 103 (2016) 384–400

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Robust shrinkage estimation and selection for functional multiple linear model through LAD loss

Lele Huang[a,d], Junlong Zhao[b,*], Huiwen Wang[a,e], Siyang Wang[c]

[a] *School of Economics and Management, Beihang University, China*
[b] *School of Statistics, Beijing Normal University, China*
[c] *School of Statistics and Mathematics, Central University of Finance and Economics, China*
[d] *Credit Card Center, China Everbright Bank, China*
[e] *Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations, China*

ARTICLE INFO

Article history:
Received 3 October 2015
Received in revised form 6 April 2016
Accepted 27 May 2016
Available online 8 June 2016

Keywords:
Functional linear model
Variables selection
Robustness
Multiple functional predictors

ABSTRACT

In functional data analysis (FDA), variable selection in regression model is an important issue when there are multiple functional predictors. Most of the existing methods are based on least square loss and consequently sensitive to outliers in error. Robust variable selection procedure is desirable. When functional predictors are considered, both non-data-driven basis (e.g. B-spline) and data-driven basis (e.g. functional principal component (FPC)) are commonly used. The data-driven basis is flexible and adaptive, but it raise some difficulties, since the basis must be estimated from data.

Since least absolute deviation (LAD) loss has been proven robust to the outliers in error, we propose in this paper a robust variable selection with data-driven basis FPC and LAD loss function. The asymptotic results are established for both fixed and diverging p. Our results include the existing results as special cases. Simulation results and a real data example confirm the effectiveness of the proposed method.

1. Introduction

The past few years has witnessed the rapid development of functional data analysis (FDA) (Ramsay and Silverman, 1997, 2002; Ferraty and Vieu, 2006; Horváth and Kokoszka, 2012; Fan et al., 2015, etc.). Functional linear model, which has been widely used in the functional data analysis (Ramsay and Silverman, 1997, 2002; Hall and Horowitz, 2007; Cai and Hall, 2006), is defined as follows.

$$Y = \alpha + \int_F \beta(t)\mathbf{X}(t)dt + \epsilon, \tag{1}$$

where $\mathbf{X}(t)$ is a functional predictor, α and ϵ are scalars with ϵ being independent of $\mathbf{X}(t)$ and F is an interval. The slope function $\beta(t)$ is to be estimated. Many results have been published on model (1), where only one functional predictor is presented (Hall and Hooker, 2014; Comte et al., 2012; García-Portugués et al., 2014; Shang and Cheng, 2015).

In practice, it is often the case that we need to consider the relationship between response Y and multiple functional predictors $\mathbf{X}_j(t)$, $1 \leq j \leq p$, with more sophisticated relationship beyond linear relationship. Therefore, similar to the

[*] Corresponding author.
E-mail address: zjlczh@126.com (J. Zhao).

http://dx.doi.org/10.1016/j.csda.2016.05.017

generalization of linear model to additive model, model (1) can been extended naturally to the functional additive model (James and Silverman, 2005; Müller and Yao, 2008; Fan et al., 2015). Particularly, when the link functions are identity, functional additive model becomes the following functional linear additive model

$$Y = \sum_{j=1}^{p} \int_F \beta_{0j}(t)\mathbf{X}_j(t)dt + \epsilon, \tag{2}$$

where $E(\mathbf{X}_j(t)) = 0$ and $E\mathbf{X}_j^2(t) < \infty$ for all $t \in F, 1 \le j \le p$, ϵ is a scalar of median 0 and independent of $\mathbf{X}_j(t)$'s. The sample paths of $\{\mathbf{X}_j(t) : t \in F\}$ belong to space $L^2(F)$, the set of all square integrable functions defined on F. The functional coefficients $\beta_{0j}(t), 1 \le j \le p$ are to be estimated.

For functional (linear) additive model with $p > 1$, variable selection is an important but challenging issue. Zhu et al. (2010) proposed a Bayesian method for the linear additive model in classification problem and Fan et al. (2015) proposed a method called functional additive regression (FAR) which can select variables when p is much larger than n.

Robustness is an important issue in variable selection. For linear model, it has been shown that the efficiency of variable selection methods, such as LASSO (Tibshiranit, 1996, etc.), can be significantly reduced by outliers (Fan et al., 2014; Zhao et al., 2013). And many robust variable selection methods, based on LAD-loss or quantile regression, etc., have been developed (e.g. Wang et al., 2007; Belloni et al., 2011; Li and Zhu, 2008; Fan et al., 2014, etc.). However, in the framework of FDA, robustness is less considered in variable selection for regression models. Most of the existing works, based on least square loss function, are sensitive to the outliers in error or heavy-tailed error (e.g. Lian, 2013; Fan et al., 2015; Chen et al., 2011, etc.). In this paper, we attempt to develop variable selection method that is insensitive to outliers in error or heavy tailed error ϵ.

However, different from linear model, functional data is nonparametric in nature. A commonly used strategy in FDA is to expand the functional data with respect to certain basis, similar to nonparametric or semiparametric statistics. In the literature, both non-data-driven and data-driven approaches are commonly used. The former one includes B-spline or other given orthogonal basis, (see Li and Hsing, 2007, Cardot et al., 2005, Lian and Li, 2014), while the latter one includes functional principal component (FPC) basis, which is a popular tool in FDA, see Cai and Hall (2006); Hall and Horowitz (2007) for details.

The FAR method of Fan et al. (2015) can do variable selection for the case p being much larger than n in functional linear additive model (2). However, their method, based on a set of non-data-driven bases, denoted as $\{b_l(t), l = 1, 2 \ldots, \}$, needs restrictive assumptions. Let $X_{ij}(t), 1 \le i \le n$ be the *i.i.d.* observations of $\mathbf{X}_j(t)$ for $1 \le j \le p$. A key assumption for FAR is that all elements in $\{X_{ij}(t), \beta_j(t), 1 \le i \le n, 1 \le j \le p\}$, are uniformly well approximated by first $q_n = o(n)$ basis $\{b_1(t), \ldots, b_{q_n}(t)\}$ (see (A) in Condition 1 of Fan et al. (2015)), such that the approximation error is simultaneously small and neglectable. Note that the same set of bases is used for all $1 \le j \le p$. This assumption may be mild when p is fixed, but it is quite strong for large p, especially when p is diverging. In addition, the selection of such basis $\{b_l(t), \ l = 1, 2, \ldots, \}$ in practice is not easy.

On the other hand, data-driven basis seems more flexible and adaptive, which enables us to approximate $\{X_{ij}(t), \beta_j(t), 1 \le i \le n\}$ with different basis $\{\hat{b}_{jl}(t), l = 1, 2, \ldots, \}$ for different j, where $\hat{b}_{jl}(t)$'s are estimated from data. However, additional challenges raise for data-driven basis. Since the bases $\{\hat{b}_{jl}(t), l = 1, 2 \cdots\}, 1 \le j \le p$, are estimated from data, we need them to be uniformly close to its population version to guarantee the consistency. This will be challenging when p is diverging. FPC is a popular data-driven basis in functional data analysis, since one can know the loss in variance information for truncated number of FPCs (Hall and Horowitz, 2007).

FPC basis may have some advantages in variable selection compared with non-data-driven basis. For any given orthonormal basis $\{m_{jk}(t), k = 1, 2, \ldots, \}$, where $1 \le j \le p$, we have $\mathbf{X}_j(t) = \sum_{k=1}^{\infty} \eta_{jk} m_{jk}(t)$ and $\beta_{0j}(t) = \sum_{k=1}^{\infty} b_{jk} m_{jk}(t)$, where η_{jk}'s and b_{jk}'s are the coefficients. Consequently, the functional linear additive model becomes linear model $Y = \sum_{j=1}^{p} \sum_{k=1}^{\infty} \eta_{jk} b_{jk} + \epsilon$, where b_{jk}'s are unknown parameters to be estimated. To estimate the coefficients, we truncate the summands, keeping only the first L_j terms for each $1 \le j \le p$, which results in the model $Y = \sum_{j=1}^{p} \sum_{k=1}^{L_j} \eta_{jk} b_{jk} + \tilde{\epsilon}$, where $\tilde{\epsilon} = \sum_{j=1}^{p} \sum_{k>L_j} \eta_{jk} b_{jk} + \epsilon$ being the error. Then after truncation, we have the approximation error $(\sum_{k>L_j} b_{jk}^2)^{1/2}$ with respect to nonzero $\beta_{0j}(t)$. Moreover, truncation increases the noise level of the error, since ϵ is independent of $\mathbf{X}_j(t)$'s. Large L_j's will reduce the approximation error and the noise level of $\tilde{\epsilon}$, which is helpful in improving the efficiency of the estimate. However, large L_j's also have side effect in view of variable selection. Since we aim to select $\mathbf{X}_j(t), 1 \le j \le p$, the coefficients $\{b_{jk}, k = 1, 2, \ldots, L_j\}$ should be treated as a group in the penalty term. It is known from the literature of group lasso for linear model that convergence rate of the estimators $\{\hat{b}_{jk}, 1 \le k \le L_j, 1 \le j \le p\}$ towards $\{b_{jk}, 1 \le k \le L_j, 1 \le j \le p\}$ is affected significantly by the largest group size $\max_{1 \le j \le p} L_j$, e.g. the convergence rate in ℓ_2 norm being of the order $\sqrt{\max_{1 \le j \le p} L_j/n} + \sqrt{\log p/n}$ (e.g. Negahban et al., 2012). Therefore, it is important to keep good balance on group size, the noise level of $\tilde{\epsilon}$ and the approximation error.

FPC basis may have some merits in this setting. For simplicity of illustration, we take $\{m_{jk}(t), l = 1, 2 \cdots, 1 \le j \le p\}$ to be the population version of the FPC. If both $\beta_{0j}(t)$'s and total variances can be approximated well by the first few FPCs, then noise level, group size and approximation error will be well controlled simultaneously. And it can be expected that FPC basis will have good performance in variable selection. In the case where the first few FPCs (or equivalently L_j's are small) still approximate total variance well and its approximation to nonzero β_{0j} is not very bad (or equivalently vector $(b_{jk}, 1 \le k \le L_j)$ is not very close to zero), we see that noise level is still controlled under some mild conditions of $\|\beta_{0j}\|$. Moreover, for any