Robust Variable Selection for Functional Linear Regression

Guanqun Cao[1*] and Yichen Qin[2]

[1] *Auburn University, Auburn, Alabama 36849 USA; gzc0009@auburn.edu.*
[2] *University of Cincinnati, Cincinnati, Ohio 45221 USA; qinyn@ucmail.uc.edu.*
[*] *Presenting author*

Keywords. *Multiple functional predictors; Penalized estimation; Robustness; Variable selection.*

Functional linear regression has been widely used to explore the relationship between a scalar response and functional predictors. In this article, we consider the situation where multiple functional predictors are observed, but only a few of these predictors are actually useful in predicting the response. Several recent literatures investigate the variable selection in such model. For example, **?** used a penalized likelihood method that controls the sparsity of the model and the smoothness of the corresponding coefficient functions without considering any outliers. Our objective is to develop an outlying-resistant variable selection procedure to identify the important functional predictors and estimate the corresponding coefficient functions simultaneously even in presence of a significant proportion of contanimated observations.

We assume a functional linear regression, $Y_i = \alpha + \sum_{j=1}^{p} \int \beta_j(t) X_{i,j}(t) dt + \epsilon_i$. To estimate the parameter $\boldsymbol{\beta}(t) = (\beta_1(t), ..., \beta_p(t))^T$, we propose to minimize the following objective function,

$$\sum_{i=1}^{n} \phi\Big(Y_i - \alpha - \sum_{j=1}^{p} \int \beta_j(t) X_{i,j}(t) dt\Big) + \sum_{j=1}^{p} P_{\lambda,\tau}(\beta_j(t)),$$

where $\phi(r) = 1 - \exp(-r^2/c)$ is the exponential squared loss function (**?**) with tuning parameter c. Note that when $c \to \infty$, $\phi \approx r^2/c$, which corresponds to the traditional least square estimate. Due to its boundedness, ϕ can effectively limit the influence of outliers, therefore, provides robust estimate of the coefficient functions. Furthermore, we can show the $\sqrt{n}$-consistence of the proposed estimators. We further incorporate a penalty function $P_{\lambda,\tau}(\beta_j(t)) = \lambda(||\beta_j(t)||_2^2 + \tau||\beta_j''(t)||_2^2)^{1/2}$ and $||\beta_j(t)||_2^2 = \int \beta_j(t)^2 dt$, $\beta_j''(t) = \partial^2 \beta_j(t)/\partial t^2$, which controls the sparsity of the model and smoothness of coefficients. To solve the above optimization problem, we first approximate the coefficient functions $\beta_j(t)$ using the B-spline basis functions $\mathbf{b}_j(t) = (b_{j1}(t), ..., b_{jq}(t))^T$, i.e., $\beta_j(t) \approx \sum_{r=1}^{q} \gamma_{jr} b_{jr}(t)$, where γ_{jr} are the corresponding basis

Contamination Ratio		0.00	0.10	0.20	0.30
SE	Proposed Method	1.80	1.99	2.35	2.78
	Least Square	1.58	29.48	80.78	159.76
TPR	Proposed Method	0.99	0.98	0.97	0.97
	Least Square	0.80	1.00	1.00	1.00
TNR	Proposed Method	0.78	0.80	0.82	0.85
	Least Square	1.00	0.01	0.00	0.00

Table 1: Comparison of performance of the proposed method and the least squared method at difference levels of contamination

coefficients. When the functional predictors, $X_{i,j}(\cdot)$, are observed without measurement errors and at an equally spaced dense grid of points, $\{t_{j,1}, \ldots, t_{j,N_j}\}$, then $\int \beta_j(t)X_{i,j}(t)dt$ can be approximated by the Riemann sum, i.e. $\int \beta_j(t)X_{i,j}(t)dt = \mathbf{Z}_{ij}^T\boldsymbol{\gamma}_j$, where $\mathbf{Z}_{ij} = (Z_{ij,1}, \ldots, Z_{ij,q})^T$, $Z_{ij,r} = (t_{j,l} - t_{j,l-1}) \sum_l X_{i,j}(t_{j,l})b_{j,r}(t_{j,l})$ and $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jq})^T$. Hence, the functional linear model can be approximated by a typical linear regression model $Y_i \approx \alpha + \sum_{j=1}^p \mathbf{Z}_{ij}^T\boldsymbol{\gamma}_j + \epsilon_i$.

We conduct simulation to demonstrate the superior performance of the proposed method. We simulate data sets of the form $\{X_{i,1}(t), ..., X_{i,10}(t), Y_i\}$, $i = 1, ..., 1000$, where each covariate $X_{i,j}$ is observed on the set of 300 equidistant points in $(0, 300)$. In particular, the generating model is $Y_i = \alpha + \sum_{j=1}^{10} \int_0^{300} \beta_j(t)X_{i,j}(t)dt + \epsilon_i$, where $\epsilon_i \overset{i.i.d.}{\sim} (1-\delta)N(0, 0.01) + \delta N(0.01, 0.1)$ and δ is the contamination ratio. $\beta_1(t)$, $\beta_2(t)$, and $\beta_3(t)$ have Gamma-density like shape with effect sizes decreasing with increasing j, and $\beta_4(t)$ and $\beta_5(t)$ have exponential like shape with $\beta_5(t)$ being more linear. Only signals $j = 1, ..., 5$ are assumed to be relevant. We run 100 replications.

Table **??** presents the comparison of the proposed method and traditional least square method in terms of (1) square errors (SE), $\text{SE} = \sum_{j=1}^{10} \int_0^{300} (\beta_j(t) - \hat{\beta}_j(t))^2 dt$; (2) true positive rate (TPR); (3) true negative rate (TNR). As the figure shows, when there is no contamination, the proposed method outperforms the least square method slightly. As the contamination becomes more serious, the proposed method totally dominates least square method in all three categories. Therefore, we can see the clear advantage of the proposed method.

References

Gertheiss, J., Maity, A. and Staicu, A.-M. (2013). Variable Selection in Generalized Functional Linear Models, *Stat*, **2**, 86-101.

Wang, X., Jiang, Y., Huang, M. and Zhang, H. (2013). Robust Variable Selection with Exponential Squared Loss, *Journal of American Statistical Association*, **108**, 632-643.