Simultaneous Estimation and Variable Selection for Functional Regression Model via Rank-Based Regularization

Jieun Park* Ash Abebe† Nedret Billor†

Abstract

We propose a robust rank based variable selection method for a functional linear regression model with multiple explanatory functions and a scalar response. The procedure extends rank based group variable selection to functional variable selection and the proposed estimator is robust in the presence of outliers in predictor function space as well as response space. The performance of the proposed robust method is demonstrated with an extensive simulation study and real data examples. We prove the proposed method with a group-adaptive penalty achieves the oracle property.

Key Words: Robust, Rank-Based, Variable Selection, Functional Analysis, Oracle Property

1. Introduction

One of the important topics in statistics is the study of the relationship among variables via regression models. Linear regression analysis, in particular, is fundamental for the functional data analysis which is the analysis of infinite-dimensional variables as curves, images, and time-variant inputs. We consider a functional multiple linear regression model with p functional predictors and a continuous scalar response. When a basis expansion with d basis is used for functional predictors and functional parameters, there are $d \times p$ predictors in the multiple grouped linear regression model with p groups. Thus identifying the subset of significant predictor functions becomes a group variable selection problem rather than a single variable level selection problem.

There are various variable selection methods regarding a group structure with regularization method. Some elaborated techniques have been devised by Yuan and Lin [18] and Wang and Leng [16]. Yuan and Lin [18] used the least squared loss with the weighted group ℓ_2 penalty to suggest the group Lasso. Wang and Leng [16] obtained the consistency and the oracle property of their estimator by proposing the adaptive group Lasso. By using these least square based group variable selection methods, Gertheiss et al. [2] proposed a variable selection method for multiple functional linear regression models after converting a functional model to a grouped discrete linear model.

The aforementioned techniques are based on the least squared loss (LS) minimization. They are efficient if the true underlying distribution follows the normal distribution. However, the LS type of objective functions is vulnerable when the data contain outliers or are heavy-tailed. We develop another version of the group Lasso technique that is applied to a functional linear model with the shortcoming of LS methods removed.

We borrow a rank-based variable selection method among several trials to overcome those drawbacks and to achieve robustness in multiple linear regression models. Miakonkana et al. [11] proposed a rank-based group variable selection method with a weighted rank-based loss function same as the one in Wang and Li [17], with a group adaptive ℓ_1 norm penalty. In addition, Miakonkana et al. showed their proposed method achieves the oracle property.

*Auburn University at Montgomery, 7061 Senators Drive, Montgomery, AL 36117

†Auburn University, 221 Parker Hall, Auburn, AL 36849

Most variable selection techniques for a functional regression model were based on the regularization method with LS loss. In this point of view, Matsui and Konishi [10] used group SCAD penalty to select variables for a functional linear model with a scalar response and functional predictors. Mingotti et al. [12] proposed "Functional Lasso" for a functional response with scalar predictors by adapting Lasso method to functional linear model. Hone and Lian [5] applied the Lasso regularization method for a functional response with functional predictors to solve a linear ordinary differential equation. However, we consider the functional aspects of coefficient function $\beta(t)$ rather than applying methods for group variable selection directly. Gertheiss et al. [2] included the functional smoothness condition of coefficient functions in the penalty term while penalizing the sum of ℓ_2 norm of the coefficient functions for the generalized linear functional regression model. The LS loss function has the same drawback under the existence of outliers even though the ℓ_2 penalty selects functional variables. To overcome this, some robust loss approach has been proposed. Pannu and Billor [14] applied the least absolution deviation method to functional linear model using Gertheiss' penalty function and showed a robustness of their method. Also, we consider the smoothness property of functions to define a regularization method with a robust loss functions.

To this end, we propose a robust variable selection method for a functional linear regression model. The rank-based functional regression model is developed by modifying the work of Miakonkana et al. [11] with the penalty function in Gertheiss et al. [2]. Since the model has a weighted rank-based (RB) loss function, it has robustness in both the predictor space and the response space. The proposed model conserves the smoothness of coefficient functions while selecting significant functional variables. Also, the adaptive penalty term implies the oracle property.

In this paper, we discretize a functional linear model as a grouped linear model and introduce the proposed model in Section 2. In Section 4, we present the simulation results to examine the properties of proposed method. An application to Japanese weather data is presented in Section 5. In Appendix, we prove the oracle property of estimators from the proposed method as well.

2. Methodology

We understand a functional linear model as a grouped multiple linear regression with a finite basis expansion on a functional space. To express a functional coefficient only with the sparsity between groups, we need as many nonzero parameter inside the group as possible. Thus, we use the ℓ_2 group penalty rather than ℓ_1 group penalty. We propose a robust rank-based method for functional data after converting a functional linear model to a grouped multiple linear regression model.

2.1 Functional Linear Model with Group ℓ_2 Penalty

Consider a functional multiple linear regression model with p functional predictors and a continuous scalar response defined by Equation (1).

$$y_i = \alpha + \sum_{j=1}^{p} \int_{\mathcal{T}} X_{ij}(t)\beta_j(t)dt + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where y_i is a scalar and $X_{ij}(t)$'s on $\mathcal{T}$, the support of functional covariates, are L^2 integrable and independent with each other, $\beta_j(t)$'s are functional parameters which are also L^2 integrable, and $\varepsilon_i \overset{iid}{\sim} F$, where F is some distribution with finite Fisher information.

We use the reformulation of functional model in Equation (1) as in Gertheiss et. al. [2]. We express the functional model in Equation (1) as a discretized form over $\{t_1, \ldots, t_m\} \in \mathcal{T}$ with an appropriate basis $\{\phi(t)\}$ and an appropriate finite number of basis d. With the finite basis $\phi_{j1}, \ldots, \phi_{jd}$, the parameter function $\beta_j(t)$ can be written as a finite dimensional approximation

$$\beta_j(t) \approx \sum_{\ell=1}^{d} c_{j\ell}\phi_\ell(t). \tag{2}$$

Then we can approximate the integration in (1) as

$$\begin{aligned}
\int_{\mathcal{T}} X_{ij}(t)\beta_j(t)dt &\approx \sum_{s=1}^{m} X_{ij}(t_s)\beta_j(t_s)(t_s - t_{s-1}) && (3)\\
&\approx \sum_{\ell}\left(\sum_{s} X_{ij}(t_s)\phi_\ell(t_s)\delta_s\right)c_{j\ell} && (4)\\
&= \sum_{\ell} \Phi_{ij\ell}c_{j\ell} && (5)\\
&= \mathbf{\Phi}_{ij}^T\mathbf{c}_j && (6)
\end{aligned}$$

where $i = 1, \ldots, n, j = 1, \ldots, p, \delta_s = t_s - t_{s-1}, \mathbf{c}_j = (c_{j1}, \ldots, c_{jd})^T, \mathbf{\Phi}_{ij} = (\Phi_{ij1}, \ldots, \Phi_{ijd})^T$ and $\Phi_{ij\ell} = \sum_s X_{ij}(t_s)\phi_\ell(t_s)\delta_s$.

The discretized version of our model is written as

$$y_i = \alpha + \sum_{j=1}^{p} \mathbf{\Phi}_{ij}^T\mathbf{c}_j + \varepsilon_i, \quad i = 1, \ldots, n \tag{7}$$

which is a grouped multiple linear regression model with p groups, d predictors in each group, and n observations. The functional linear regression model in Equation (1) becomes a discrete grouped regression model to estimate grouped parameters $\mathbf{c}_j$'s for $j = 1, \ldots, p$. Gertheiss et. al. proposed the objective function with L_2 loss and ℓ_2 group penalty.

$$\sum_{i=1}^{n}\left(y_i - \alpha - \sum_{j=1}^{p}\mathbf{\Phi}_{ij}^T\mathbf{c}_j\right)^2 + \sum_{j=1}^{p} P_{\lambda,\varphi}(\beta_j) \tag{8}$$

where

$$P_{\lambda,\varphi}(\beta_j) = \lambda(||\beta_j||_2^2 + \varphi||\beta_j''||_2^2)^{1/2}, \tag{9}$$

$||\cdot||^2$ is the functional L^2, and $\beta_j''(t) = d^2\beta_j(t)/dt^2$.

With a basis change by considering the second derivative of the coefficient functions, we can express Equation (1) as

$$y_i = \alpha + \sum_{j=1}^{p} \widetilde{\mathbf{\Phi}}_{ij}^T\widetilde{\mathbf{c}}_{\varphi,j} + \varepsilon_i \quad \text{for } i = 1, \ldots, n. \tag{10}$$

Thus, Gertheiss et al. [2] uses the following objective function to estimate $\hat{\alpha}$ and $\hat{\widetilde{\mathbf{c}}}_j$ using the group Lasso method.

$$Q_{LS}(\alpha, \widetilde{\mathbf{c}}_j) = \sum_{i=1}^{n}\left(y_i - \alpha - \sum_{j=1}^{p}\widetilde{\mathbf{\Phi}}_{ij}^T\widetilde{\mathbf{c}}_{\varphi,j}\right)^2 + \sum_{j=1}^{p}\lambda||\widetilde{\mathbf{c}}_{\varphi,j}||_2 \tag{11}$$

where adaptive penalty function is

$$P_{\lambda,\varphi}(\beta_j) = \lambda(w_j||\beta_j||_2^2 + \varphi v_j||\beta_j''||_2^2)^{1/2} \tag{12}$$

The weights w_j's and v_j's are chosen depending on data and the significancy of each coefficient function. The tuning parameter λ controls the entire penalty function and φ controls the concavity of estimated parameter functions $\beta_j(t)$'s.

2.2 Rank-Based Regression

The goal of the rank-based regression method is to estimate the coefficient vector $\boldsymbol{\beta}$ in a linear model. The rank-based method pursues also to estimate the parameter $\boldsymbol{\beta}$ under the presence of outliers. We assume that the errors are independent and identically distributed (iid) with a continuous probability density function (pdf) $f(t)$. Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ be the $n \times 1$ vector of responses, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ the $n \times p$ design matrix, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ the $n \times 1$ error vector. Then we can rewrite Equation (13) as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon. \tag{13}$$

We define a new distance measure to achieve the rank-based estimator for the coefficient vector $\boldsymbol{\beta}$ based on Jaeckel's dispersion function [6]. We follow the notations and terminology by Jaeckel [6] and Jurečková [8].

Before defining the rank-based method, we introduce the definition of a *pseudo-norm* as in Hettmansperger and McKean [4]. An operator $||\cdot||_\varphi$ is called a *pseudo-norm* if it satisfies the following four conditions.

1. $||\mathbf{u} + \mathbf{v}||_\varphi \leq ||\mathbf{u}||_\varphi + ||\mathbf{v}||_\varphi$ for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$
2. $||\alpha\mathbf{u}||_\varphi = |\alpha|||\mathbf{u}||_\varphi$ for all $\alpha \in \mathbb{R}, \mathbf{u} \in \mathbb{R}^n$
3. $||\mathbf{u}||_\varphi \geq 0$ for all $\mathbf{u} \in \mathbb{R}^n$
4. $||\mathbf{u}||_\varphi = 0$ if and only if $u_1 = \cdots = u_n$

Jaeckel's dispersion function measuring the distance between two vectors is defined by

$$D(\boldsymbol{\beta}) = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_\varphi, \tag{14}$$

where

$$||\mathbf{u}||_\varphi = \sum_{i=1}^{n} a(R(u_i))u_i, \tag{15}$$

R denotes the rank, $a(t) = \varphi(\frac{t}{n+1})$, and φ is a nondecreasing and L^2-integrable score function defined on the interval $[0, 1]$ as in Kloke and McKean [9]. Without loss of generality, we assume $\int \varphi(s)ds = 0$ and $\int \varphi^2(s)ds = 1$. Then one can check $||\cdot||_\varphi$ in Equation (15) is a pseudo-norm. A primal-dual relationship between quantile regression and rank estimation is given in Gutenbrunner and Jurecková [3].

Let φ be Wilcoxon score, that is, $\varphi\Big(\dfrac{t}{n+1}\Big) = \dfrac{t}{n+1} - \dfrac{1}{2}$. Then, Jaeckel's Wilcoxon-type dispersion function $D(\boldsymbol{\beta})$ can be written as

$$D(\boldsymbol{\beta}) = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_\varphi \tag{16}$$

$$= \frac{1}{2(n+1)} \sum_{i<j} |\varepsilon_i - \varepsilon_j| \tag{17}$$

Johnson and Peng [7] used the following objective function similar to Equation (17) for the linear regression model in Equation (13).

$$\sum_{i<j} |\varepsilon_i - \varepsilon_j| \tag{18}$$

Furthermore, to achieve robustness in the predictor space, Wang and Li [17] proposed the weighted rank-based loss function

$$\sum_{i<j} b_{ij}|\varepsilon_i - \varepsilon_j| \tag{19}$$

where

$$b_{ij} = b(\mathbf{x}_i, \mathbf{x}_j) = h(\mathbf{x}_i)h(\mathbf{x}_j), \tag{20}$$

which degrades high leverage points, where

$$h(\mathbf{x}_i) = \min\left[1, \frac{b}{(\mathbf{x}_i - \hat{\mu})^T S^{-1}(\mathbf{x}_i - \hat{\mu})}\right] \tag{21}$$

with $(\hat{\mu}, S)$ being the robust minimum volume ellipsoid estimators of the location and spread as in Wang and Li [17] and Miakonkana et al. [11].

We call this weighted Wilcoxon-type rank-based method as the *rank-based* regression method. The rank-based (RB) method estimates β by minimizing the following weighted Wilcoxon-type dispersion function as the loss function.

$$\widehat{\boldsymbol{\beta}}_{RB} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i<j} b_{ij}|\varepsilon_i - \varepsilon_j| \tag{22}$$

where b_{ij} is defined by Equation (20). We use the rank-based method as the loss function for the proposed rank-based penalized method for functional linear regression model.

2.3 Rank-Based Functional Variable Selection

We define a rank-based functional regression method with the objective function analogous to Equation (11)

$$Q_{RB}(\alpha, \widetilde{\mathbf{c}}_j) = \sum_{i<j} b_{ij}|\varepsilon_i - \varepsilon_j| + \sum_{j=1}^{p} \lambda||\widetilde{\mathbf{c}}_{\varphi,j}||_2 \tag{23}$$

where adaptive penalty function is

$$P_{\lambda,\varphi}(\beta_j) = \lambda(w_j||\beta_j||_2^2 + \varphi v_j||\beta_j''||_2^2)^{1/2} \tag{24}$$

with pairwise difference data after a proper basis expansion of functional data. We discuss the asymptotic properties of the proposed rank-based functional variable selection estimator in Appendix. We show that the ℓ_2 penalized rank-based group variable selection estimator achieves the oracle property under some regularity conditions as in Wang and Li [17] and Miakonkana et. al.[11].

3. Implementation

We borrow the idea of converting the original data (x_i, y_i) to the pairwise difference observation data $(x_{ij}, y_{ij}) = (x_i - x_j, y_i - y_j)$ for $ij = 1, \ldots, n(n-1)/2$ and $1 \le i < j \le n$ which is identical to the one used in Wang and Li [17] and Miakonkana et al. [11]. We use the R package $grplasso$ by modifying the least squared loss function to the least absolute deviation loss function. The rank based estimator can be optimized with the least absolute deviation and the group ℓ_2 penalty for the converted data. We use unpenalized estimators for λ, φ, and their group adaptive parameters as well.

4. Simulation Study

We generate sine-like functional predictors similarly as in Gertheiss [2] and generate the responses by adding errors from different kinds of distributions to the inner product between the coefficient functions $\beta(t)$'s and the functional predictors. We create the contaminated predictors which resembles the data with high leverage observations. We compare the results between LS, LAD, and RB loss with the group ℓ_2 penalties with and without optimization of smoothness.

4.1 Data Generation

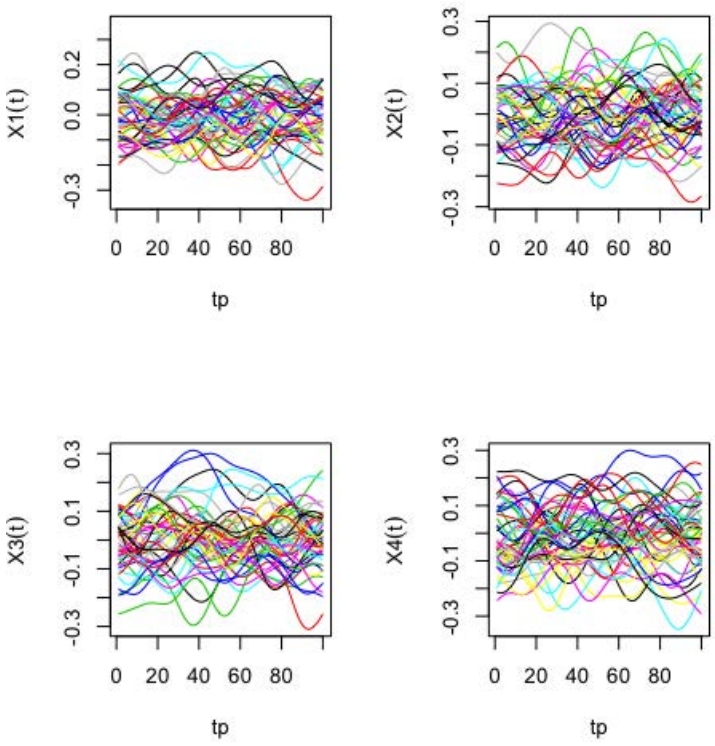

Figure 1: c_0: Predictor Functions without Contamination

Consider an example in which four functional covariates are observed at a set of 100 equidistant points in $(0, 100)$ for each sampling unit. Define for $i = 1, \ldots, n$, and $k = 1, \ldots, 4$,

$$x_{ik}(t) = \sum_{r=1}^{5} a_{ikr} \sin\Big(\frac{2\pi(5 - a_{ikr})}{150}\Big)t - m_{ikr}, k = 1, \ldots, 4 \tag{25}$$

$$y_i = \sum_{k=1}^{4} \int_0^{100} x_{i,k}(t)\beta_k(t)dt + \varepsilon_i, \quad i = 1, \ldots, n \tag{26}$$

where $a_{ikr} \sim U(0, 5)$, $m_{ikr} \sim U(0, 2\pi), i = 1, \ldots, n$, $k = 1, \ldots, 4$, $r = 1, \ldots, 5$ and $t \in [0, 100]$. Figure 1 shows the predictor functions. The true parameter functions $\beta_1(t)$

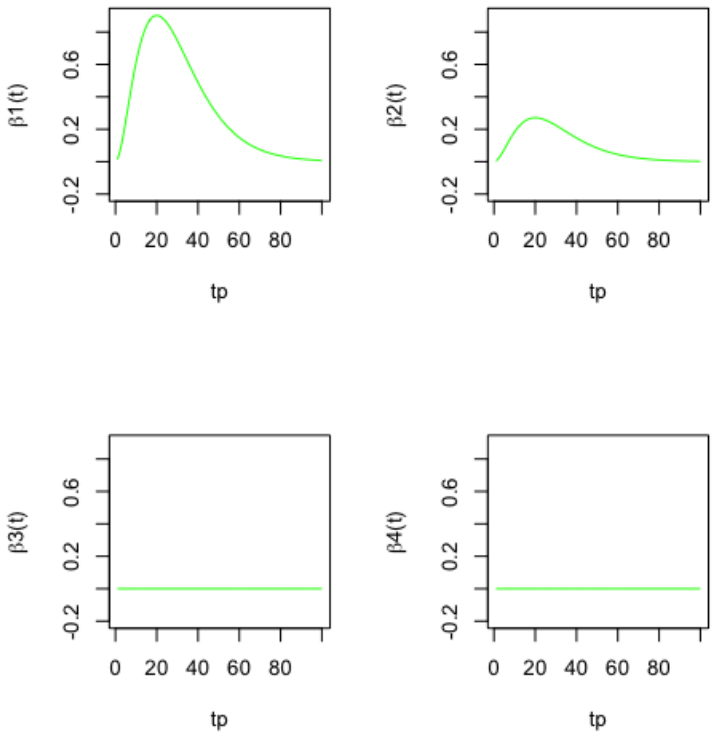

Figure 2: True Parameter $\beta(t)$ Curves

and $\beta_2(t)$ are γ distribution density curves with different stretches and $\beta_3(t) = \beta_4(t) = 0$ as shown in Figure 2.

To see the effect of the weights b_{ij}, we generate the contaminated data in the predictor space. We use the contamination criteria in Fraiman and Muniz [1]. We use three types of 15% contamination for each predictor function with the contamination size constant $M = 5$ with asymmetric contamination (c_1) in Figure 3 compared to no x contamination (c_0) in Figure 1. They are generated by the following definition.

- No Contamination(c_0):

$$z_{i,k}^{no}(t) = x_{i,k}(t)$$

- Symmetric Contamination(c_1):

$$z_{i,k}^{a}(t) = x_{i,k}(t) + c\sigma M$$

where $c \sim Bernoulli(0.15)$, $M = 5$, and σ is a random variable independent of c which is 1 or -1 with probability 0.5.

Thus, we consider asymmetric contamination in the x direction and three kinds of y direction errors, the standard normal (en), t_3 (et3) and the mixed normal errors (em). We compare the results between LS and RB loss functions with different combinations of contaminations. Also, we check the oracle property of the proposed method with adaptivity by simulating for two different sample sizes $n = 100$ and $n = 150$. For each case, we assess the average number of significant predictors (model size or degree of freedom) and the mean model error or the average of the root mean squared error of β ($RMSE(\beta)$) over the 100 runs.

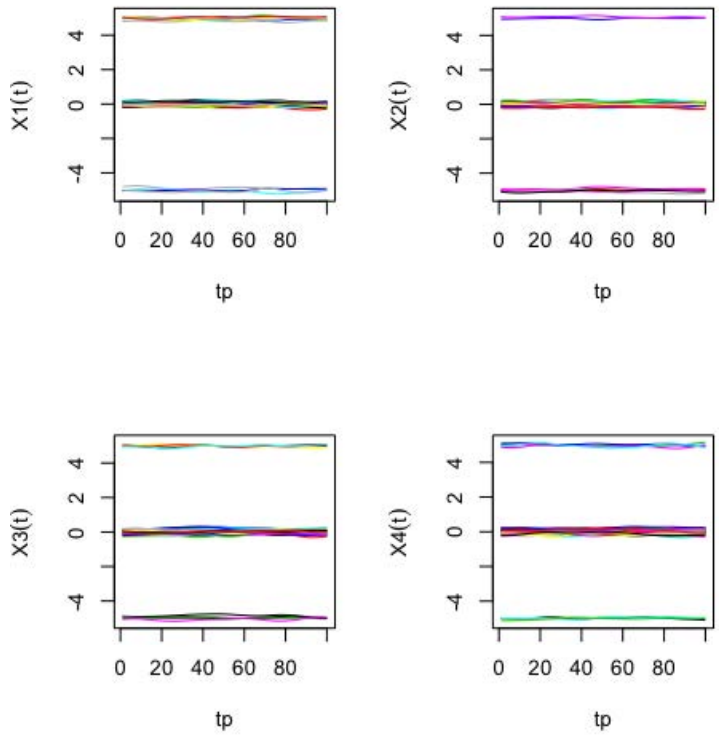

Figure 3: c_1: Predictor Functions with 15% Asymmetric Contamination

4.2 Robustness and Oracle Property

		LS						RB					
		x1	x2	x3	x4	Model.Size	Model.Error	x1	x2	x3	x4	Model.Size	Model.Error
c_0	en	1	1	0.34	0.41	2.75	0.0645	1	1	0.33	0.37	2.7	0.067
	em	1	0.93	0.48	0.33	2.74	0.109	1	1	0.37	0.39	2.76	0.072
	t_3	1	1	0.44	0.40	2.84	0.093	1	1	0.37	0.46	2.83	0.079
	Oracle	1	1	0	0	2	0	1	1	0	0	2	0

Table 1: Comparison under y Outliers Based on $RMSE(\beta)$

Table 1 shows the comparison between LS and RB methods under the presence of the outliers in the response space. LS performs better than RB method under the standard normal error since LS has the smaller model error. With the presence of outliers in the response space, RB estimates better than LS. RB and LS have 0.072 and 0.109 as the model errors, respectively. LS fails to detect the second variable in 7%, however, RB detects the second one as significant variable in 100%. The rank-based method also has a smaller model error for t_3 error with a smaller model size. This simulation result says the rank-based method performs better under the presence of response outliers. Figure 4 shows the performance difference between LS and RB under Huber mixed normal errors by choosing λ which minimizes $RMSE(\beta)$. RB method gives better precision with narrower estimated clouds than LS. The estimation using the cross-validation in Table 2 gives similar results with smaller model errors by the rank-based method.

Table 3 and 4 show the results with outliers in both the predictor space and the response space. RB performs better and more robust than RB with outliers in the predictor space with

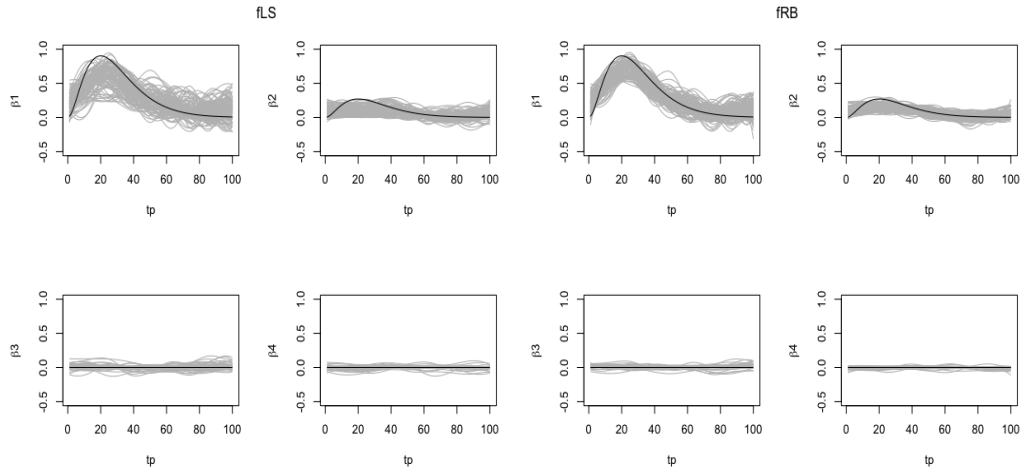

Figure 4: Estimated $\beta(t)$ under Huber Mixed Normal Errors by $RMSE(\beta)$

		LS						RB					
		x1	x2	x3	x4	Model.Size	Model.Error	x1	x2	x3	x4	Model.Size	Model.Error
c_0	en	1	1	0.63	0.68	3.31	0.0658	1	1	0.54	0.62	2.674	0.0648
	em	1	0.95	0.62	0.52	3.19	0.1102	1	1	0.59	0.51	3.1	0.0765
	t_3	1	0.98	0.61	0.59	3.18	0.0959	1	1	0.53	0.58	3.11	0.0818

Table 2: Comparison under y Outliers Based on CV

n	y	ls.df	ls.m(β)	ls.sd(β)	rb.df	rb.m(β)	rb.sd(β)
100	en	3.98	2.059	0.087	3.92	2.23	0.225
	em	3.86	2.395	0.382	3.95	2.383	0.368
150	en	4	2.004	0.09	3.71	2.139	0.268
	em	3.95	2.131	0.221	3.66	2.109	0.227

Table 3: c_1 Adapt0 by CV

a smaller average of $RMSE(\beta)$ and closer to the true model size in Table 3.

n	y	ls.df	ls.m(β)	ls.sd(β)	rb.df	rb.m(β)	rb.sd(β)
100	en	2.28	2.105	0.034	2.59	2.255	0.17
	em	2.11	2.284	0.261	2.44	2.342	0.227
150	en	2.27	2.095	0.032	2.2	2.23	0.143
	em	2.21	2.153	0.1	2.18	2.203	0.104

Table 4: c_1 Adapt2 by CV

We can check the oracle property of the proposed method with Adapt2 in Table 4. The result for RB with $n = 150$ has the model size closer to the true model size and a smaller average of $RMSE(\beta)$ than the one with $n = 100$.

5. Real Data Application: Weather Data

We apply the proposed rank-based method to analyze weather data in Matsui and Konishi [10] available in Chronological Scientific Tables 2005. The weather data includes monthly observed average temperatures (TEMP), average atmospheric pressure (PRESSURE), time of daylight (DAYLIGHT), average humidity (HUMIDITY), and annual total precipitation at 79 stations from 1971 to 2000 in Japan. We assume the annual total precipitation is a response variable depending four predictor functions, TEMP, PRESSURE, DAYLIGHT, and HUMIDITY in Figure 5 since these four predictors are trajectories over time. Sawant

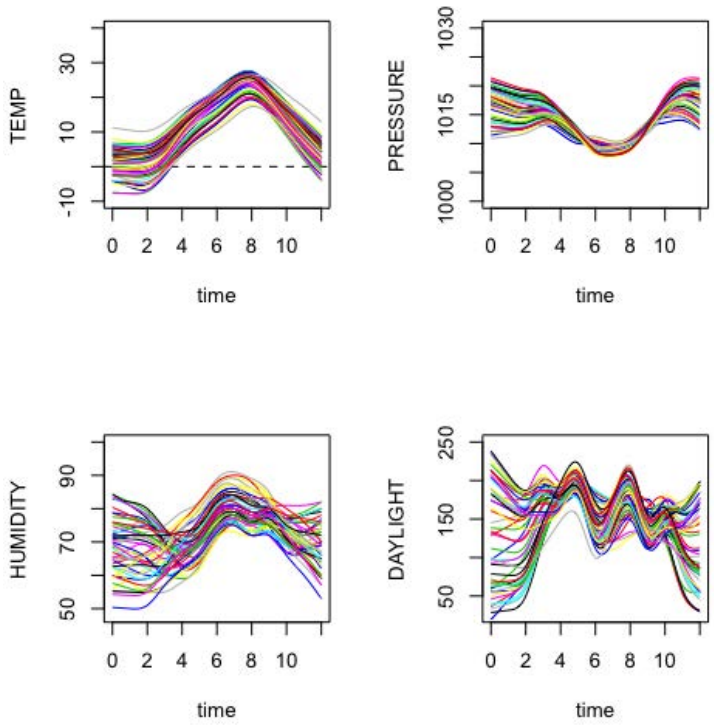

Figure 5: The Predictors of Weather Data

[15] shows the curves of TEMP and PRESSURE for the 78th and 79th observations and

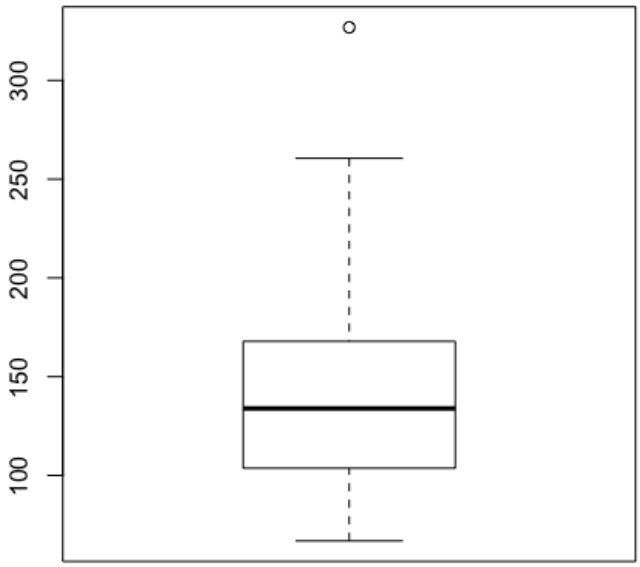

Figure 6: Boxplot of the Response, Annual Average Precipitation

the curves for the 1st, 2nd, and 3rd observations for the HUMIDITY are outliers. We see an outlier in the response on the box plot in Figure 6. The weather data set has outliers in both the predictor space and the response space. We approach to find the relation between predictor functions and the continuous discrete response using the multiple functional linear model. First, we find the λ and φ which minimize the objective functions with LS and RB loss functions by 10 fold cross-validation. We estimate the coefficient parameter functions for predictor functions using the optimal λ and φ. In Adapt0, LS detects all predictors as significant and RB chooses three predictors except PRESSUE in Table 5.

In Figure 7, TEMP and HUMIDITY have a positive effect to the response value since their estimated coefficient functions are positive over the range. In TEMP, LS estimate gives increasing weight over time, but RB estimate has a constant weight over time. Under RB method, Daylight is negative from January to August and positive after August. The estimates of PRESSURE are close or identical to zero compared to other estimated

		TEMP	PRESSURE	HUMIDITY	DAYLIGHT
Adapt0	LS	✓	✓	✓	✓
	RB	✓	--	✓	✓
Adapt1	LS	--	✓	--	--
	RB	✓	--	✓	✓
Adapt2	LS	--	✓	--	--
	RB	✓	--	✓	✓

Table 5: Relevant Predictors for Weather Data

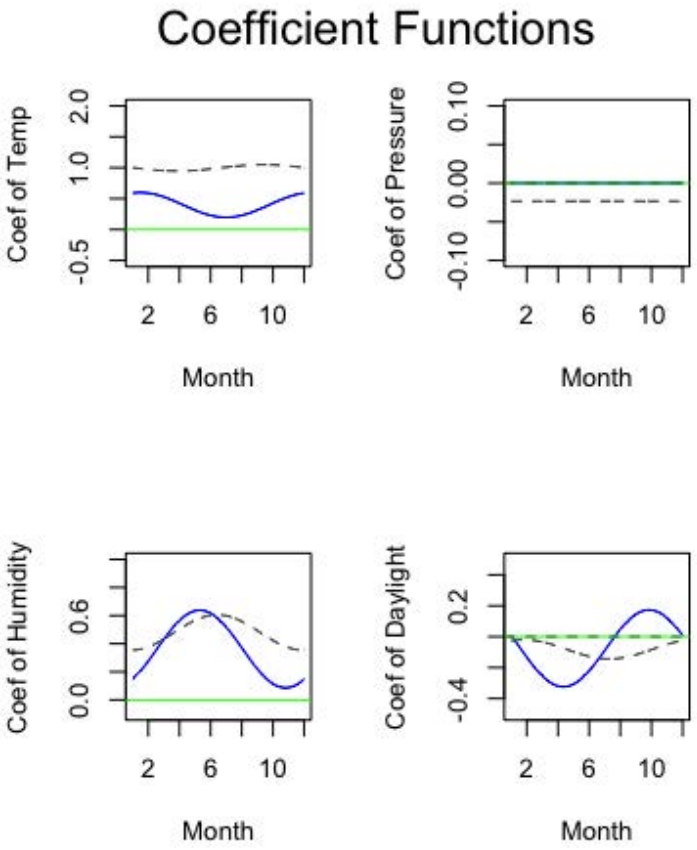

Figure 7: Estimated Weather Coefficients with Adapt0

coefficient functions in both methods. Adapt1 and Adapt2 choose only one coefficient for PRESSURE for LS. However, RB chooses the same three predictors as significant. We can check the oracle property of RB method with the adaptivity in the objective function. The mean values of prediction error over 10-folds with Adapt1 is 160.91 compared to 214.52 with Adapt0 (without adaptivity).

6. Conclusion

We established the rank-based method for functional linear model with a weighted Wilcoxon objective function penalized by ℓ_2 group penalty. By using the group ℓ_2 penalty, we can obtain only between-group sparsity to express a functional coefficient precisely by taking as many nonzero coefficients as possible for all basis functions. The resulting estimator has the oracle property with robustness in both the predictor and the response space. It selects variables and estimate the parameter functions simultaneously. However, it is challenging to find the optimal tuning parameter with CV and other criteria, SIC, BIC, GACV, or GCV depend on the combination of the number basis for function, the sample size, and errors in the response. One extension of the proposed method is to establish a proper relation between the number of basis, the sample size, and errors to find the optimal tuning parameters for rank-based loss function with SIC, BIC, GACV, or GCV.

References

[1] Ricardo Fraiman and Graciela Muniz, *Trimmed means for functional data*, Test **10** (2001), no. 2, 419–440.

[2] Jan Gertheiss, Arnab Maity, and Ana-Maria Staicu, *Variable selection in generalized functional linear models*, Stat **2** (2013), no. 1, 86–101.

[3] Christoph Gutenbrunner and J Jurecková, *Regression rank scores and regression quantiles*, The Annals of Statistics (1992), 305–330.

[4] Thomas P Hettmansperger and Joseph W McKean, *Robust nonparametric statistical methods*, Arnold, 1998.

[5] Zhaoping Hong and Heng Lian, *Inference of genetic networks from time course expression data using functional regression with lasso penalty*, Communications in Statistics-Theory and Methods **40** (2011), no. 10, 1768–1779.

[6] Louis A Jaeckel, *Estimating regression coefficients by minimizing the dispersion of the residuals*, The Annals of Mathematical Statistics (1972), 1449–1458.

[7] Brent A Johnson and Limin Peng, *Rank-based variable selection*, Journal of Nonparametric Statistics **20** (2008), no. 3, 241–252.

[8] Jana Jureckova, *Nonparametric estimate of regression coefficients*, The Annals of Mathematical Statistics (1971), 1328–1338.

[9] John Kloke and Joseph W McKean, *Nonparametric statistical methods using r*, Chapman and Hall/CRC, 2014.

[10] Hidetoshi Matsui and Sadanori Konishi, *Variable selection for functional regression models via the l1 regularization*, Computational Statistics & Data Analysis **55** (2011), no. 12, 3304–3310.

[11] Guy-vanie M Miakonkana, Brice M Nguelifack, and Asheber Abebe, *Rank-based group variable selection*, Journal of Nonparametric Statistics **28** (2016), no. 3, 550–562.

[12] Nicola Mingotti, Lillo Rodríguez, Rosa Elvira, and Juan Romo Urroz, *Lasso variable selection in functional regression*, Statistics and Econometrics Series 13, Working paper (2013), 13–14.

[13] Joshua D Naranjo and TP Hettmansperger, *Bounded influence rank regression*, Journal of the Royal Statistical Society: Series B (Methodological) **56** (1994), no. 1, 209–220.

[14] Jasdeep Pannu and Nedret Billor, *Robust group-lasso for functional regression model*, Communications in Statistics-Simulation and Computation **46** (2017), no. 5, 3356–3374.

[15] Pallavi Sawant, Nedret Billor, and Hyejin Shin, *Functional outlier detection with robust functional principal component analysis*, Computational Statistics **27** (2012), no. 1, 83–102.

[16] Hansheng Wang and Chenlei Leng, *A note on adaptive group lasso*, Computational statistics & data analysis **52** (2008), no. 12, 5277–5286.

[17] Lan Wang and Runze Li, *Weighted wilcoxon-type smoothly clipped absolute deviation method*, Biometrics **65** (2009), no. 2, 564–571.

[18] Ming Yuan and Yi Lin, *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68** (2006), no. 1, 49–67.

A. Oracle Property in RB Loss with Adaptive Group ℓ_2 Penalty

A.1 Oracle Property on Discrete Multiple Linear Model

We consider the estimation consistency, the variable selection consistency and the oracle property for the rank-based group variable selection with ℓ_2 penalty.

We show that the group ℓ_2 penalized rank-based variable selection estimator has the oracle property under some regularity conditions. In this section, we follow the definition and notation as Miakonkana et al.[11] and Wang and Li [17]. We assume that only the first $k_0 \leq K$ groups are significant, that is, $||\boldsymbol{\beta}_k||_2 \neq 0$ for $k \leq k_0$ and $||\boldsymbol{\beta}_k||_2 = 0$ for $k > k_0$. Denote $\boldsymbol{\beta}_0$ the true parameter, $\boldsymbol{\beta}_a$ the vector containing all relevant groups and $\boldsymbol{\beta}_b$ the vector of all irrelevant groups. Let $\hat{\boldsymbol{\beta}}_a$ and $\hat{\boldsymbol{\beta}}_b$ be their corresponding penalized rank-based estimator.

The following regularity conditions will be assumed.

C1. The errors ϵ_i are iid with a density function f that is absolute continuous and has a finite fisher informations. That is,

$$I(f) = \int_{-\infty}^{\infty} \Big[\frac{f'(e)}{f(e)}\Big]^2 f(e)de < \infty$$

C2. The matrices $\mathbf{X}$ and $\mathbf{WX}$ satisfy the Huber's condition.

C3. $n^{-1}\mathbf{X}'\mathbf{WX} \xrightarrow{\text{P}} \mathbf{C}$, and $n^{-1}\mathbf{X}'\mathbf{X} \xrightarrow{\text{P}} \mathbf{\Sigma}$ are positive definite matrices.

given by

$$\mathbf{C} = \frac{1}{2}\int\int(\mathbf{x}_2 - \mathbf{x}_1)(\mathbf{x}_2 - \mathbf{x}_1)'b(\mathbf{x}_1, \mathbf{x}_2)dM(\mathbf{x}_2)dM(\mathbf{x}_1)$$
$$\mathbf{V} = \int\{\int(\mathbf{x}_2 - \mathbf{x}_1)b(\mathbf{x}_1, \mathbf{x}_2)dM(\mathbf{x}_2)\}\{\int(\mathbf{x}_2 - \mathbf{x}_1)b(\mathbf{x}_1, \mathbf{x}_2)dM(\mathbf{x}_2)\}'dM(\mathbf{x}_1)$$
$$\mathbf{\Sigma} = \frac{1}{2}\int\int(\mathbf{x}_2 - \mathbf{x}_1)(\mathbf{x}_2 - \mathbf{x}_1)'dM(\mathbf{x}_2)dM(\mathbf{x}_1)$$

and $M(\mathbf{x})$ denotes the CDF of $\mathbf{x}$, $\mathbf{X}$ is a matrix whose rows are $\mathbf{x}_i$, and the entries ω_{ij} of the matrix $\mathbf{W}$ are defined like in Naranjo and Hettmansperger (1994)[13], defined by

$$\omega_{ij} = \begin{cases} n^{-1}b_{ij} & \text{if} \quad i \neq j \\ n^{-1}\sum_{k\neq i} b_{ij} & \text{if} \quad i = j \end{cases} \tag{27}$$

We derive conditions for model selection and estimation consistency when when the sample size n increases.

Following the notation in Wang and Leng (2008)[16] define

$$a_n = \max\{\lambda_{kj} : 1 \leq j \leq k; k \leq k_0\} \text{ and } b_n = \min\{\lambda_{kj} : 1 \leq j \leq k; k > k_0\},$$

and $H(\mathbf{x}, y)$ be the joint distribution between the covariate $\mathbf{x}$ and the response variable y.

Theorem 1. *Let* $(y_1, \mathbf{x}_1), \cdots, (y_n, \mathbf{x}_n)$ *be independent and identically distributed from* $H(\mathbf{x}, y)$*. Assume the regularity conditions* C1–C3*.*

a. If $\sqrt{n}a_n \xrightarrow{P} 0$ *then* $||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0||_2 = O_p(n^{-1/2})$

b. If $\sqrt{n}a_n \xrightarrow{P} 0$ *and* $\sqrt{n}b_n \xrightarrow{P} \infty$ *then* $\hat{\boldsymbol{\beta}}_b \xrightarrow{P} 0$

c. Under local shrinking contamination, $H_n^*(\mathbf{x}, y)$*,* $\sqrt{n}(\hat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a) \xrightarrow{D} N(\eta, \tau^2 C_{11}^{-1} V_{11} C_{11}^{-1})$

To prove Theorem 1, we define the following expressions defined in Wang and Li [17] with the group ℓ_2 penalty.

$$
\begin{aligned}
Q_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i<j} b_{ij}|\epsilon_i - \epsilon_j| + n\sum_{k=1}^{K}\Big(\sum_{j=1}^{p_k}(\lambda_{kj}\theta_{kj})^2\Big)^{1/2} \\
D_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i<j} b_{ij}|\epsilon_i - \epsilon_j| \\
S_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i<j} b_{ij}(\mathbf{x}_i - \mathbf{x}_j)sgn((y_i - y_j) - (\mathbf{x}_i - \mathbf{x}_j)'\boldsymbol{\theta}) \\
A_n(\boldsymbol{\theta}) &= (2\sqrt{3\tau})^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\mathbf{X}'\mathbf{W}\mathbf{X}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - (\boldsymbol{\theta} - \boldsymbol{\theta}_0)'S_n(\boldsymbol{\theta}_0) + D_n(\boldsymbol{\theta}_0)
\end{aligned}
$$

Every above expression is identical to the one in Wang and Li [17] except the group ℓ_2 penalty. We can borrow the result of the following lemma.

Lemma 1. *Under assumptions C1–C3,*

i. for all $\epsilon > 0$ *and* $c > 0$*,*

$$
[\sup_{\sqrt{n}||\boldsymbol{\theta}-\boldsymbol{\theta}_0||\leq c} |D_n(\boldsymbol{\theta}) - A_n(\boldsymbol{\theta})| \geq \epsilon] \xrightarrow{P} 0
$$

under either H *or* H_n^**,*

ii. $n^{-1/2}S_n(\boldsymbol{\theta}_0) \xrightarrow{D} N(0, \mathbf{V}/3)$ *under* H*,*

iii. $n^{-1/2}S_n(\boldsymbol{\theta}_0) \xrightarrow{D} N(\eta, \mathbf{V}/3)$ *under* H_n^**.*

We follow the same logic to Miakonkana et al. for the proof of Theorem 1 with the group adaptive ℓ_2 penalty instead of the group and element-wise adaptive ℓ_1 penalty.

Proof. To prove part (a), it is sufficient to show that $\forall \epsilon > 0$, there exists a large constant C such that

$$
P\Big(\inf_{||u||=C} Q_n(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{u}) > Q_n(\boldsymbol{\theta}_0)\Big) \geq 1 - \epsilon
$$

where $\mathbf{u}$ is a vector of dimension p. Since $Q_n(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$, this implies that with probability at least $1 - \epsilon$ the penalized estimator lies in the ball $\{\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{u} : ||\mathbf{u}|| \leq C\}$. Let $G_n(\mathbf{u}) = Q_n(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{u}) - Q_n(\boldsymbol{\theta}_0)$. Denote by u_{kj} the component of $\mathbf{u}$ corresponding to θ_{kj}.

By Lemma 1,

$$
\begin{aligned}
G_n(\mathbf{u}) &= (2\sqrt{3})^{-1}\mathbf{u}'[n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X}]\mathbf{u} - \mathbf{u}'n^{-1/2}S_n(\boldsymbol{\theta}_0) \\
&\quad + n\sum_{k=1}^{K}\Big[\Big(\sum_{j=1}^{p_k}(\lambda_{kj}(\theta_{kj}+n^{-1/2}u_{kj}))^2\Big)^{1/2} - \Big(\sum_{j=1}^{p_k}(\lambda_{kj}\theta_{kj})^2\Big)^{1/2}\Big] + o_p(1) \\
&\geq (2\sqrt{3})^{-1}\mathbf{u}'[n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X}]\mathbf{u} - \mathbf{u}'n^{-1/2}S_n(\boldsymbol{\theta}_0) - \sqrt{n}\sum_{k=1}^{k_0}\Big(\sum_{j=1}^{p_k}(\lambda_{kj}u_{kj})^2\Big)^{1/2} + o_p(1) \\
&= (2\sqrt{3})^{-1}\mathbf{u}'[n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X}]\mathbf{u} - \mathbf{u}'O_p(1) - \sqrt{n}\sum_{k=1}^{k_0}\Big(\sum_{j=1}^{p_k}(\lambda_{kj}u_{kj})^2\Big)^{1/2} + o_p(1) \\
&\geq (2\sqrt{3})^{-1}\mathbf{u}'[n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X}]\mathbf{u} - \mathbf{u}'O_p(1) - k_0\sqrt{n}a_n(||\mathbf{u}||_2) + o_p(1).
\end{aligned}
$$

Note that $n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X} \xrightarrow{\text{P}} \mathbf{C}$, a positive definite matrix, and $\sqrt{n}a_n \xrightarrow{\text{P}} 0$. Therefire, for n sufficiently large, the first term on the right hand side of the ineuqality above dominates. $G_n(\mathbf{u})$ can be made positive when the size of ball C is chosen to be sufficiently large. We now prove part (b). Suppose that $\widehat{\boldsymbol{\theta}}_b \neq 0$, $\forall n \in \mathbb{N}$. Let k be such that $k_0 < k < K$ and $\widehat{\theta}_{kj} \neq 0$ for some j such that $1 \leq j \leq p_k$. Since $Q_n(\boldsymbol{\theta})$ is differentiable at any point, except the origin, $\widehat{\theta}_{kj}$ must be solution of the equation

$$
0 = n^{-3/2}\sum_{i<j} b_{ij}(\mathbf{x}_{ik} - \mathbf{x}_{jk})sgn(y_i - y_j) - (\mathbf{x}_i - \mathbf{x}_j)'\boldsymbol{\theta}) + \sqrt{n}\lambda_{kj}sgn(\theta_{kj}).
$$

Now, by the consistency of $\widehat{\boldsymbol{\theta}_n}$ and part (ii.) of lemma 1, the first term of the right hand side of the equation above is $O_p(1)$. In addition, $\sqrt{n}b_n \xrightarrow{P} \infty$ implies that $\sqrt{n}\lambda_{kj} \xrightarrow{P} \infty$. So the equation does not hold for large values of n, as we assume that $\widehat{\theta}_{kj} \neq 0$. Therefore, $\widehat{\boldsymbol{\theta}}_b \xrightarrow{P} 0$.

The proof of part (c) is identical to the proof of Theorem 2 given in the Web Appendix of Wang and Li (2009)[17], and will therefore be omitted here.

□

A.2 Oracle Property on Functional Linear Model

We convert the functional linear model in Equation (1) to the discretized model in Equation (23) considering the functional group adaptive penalty. Similarly, we can see the oracle property of the rank-based estimates with the adaptive group ℓ_2 penalty for functional linear model.