Shrinkage Estimation of Semiparametric Model with Missing Responses for Cluster Data

Mingxing Zhang, Jiannan Qiao, Huawei Yang, Zixin Liu

Department of Mathematics and Statistics, Guizhou University of Finance and Economics, Guiyang, China
Email: zmxgraduate@163.com, qjnrunner@163.com, yang-hw2005@163.com, xinxin905@163.com

Received 23 September 2015; accepted 21 December 2015; published 24 December 2015

Copyright © 2015 by authors and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/

Abstract

This paper simultaneously investigates variable selection and imputation estimation of semiparametric partially linear varying-coefficient model in that case where there exist missing responses for cluster data. As is well known, commonly used approach to deal with missing data is complete-case data. Combined the idea of complete-case data with a discussion of shrinkage estimation is made on different cluster. In order to avoid the biased results as well as improve the estimation efficiency, this article introduces Group Least Absolute Shrinkage and Selection Operator (Group Lasso) to semiparametric model. That is to say, the method combines the approach of local polynomial smoothing and the Least Absolute Shrinkage and Selection Operator. In that case, it can conduct nonparametric estimation and variable selection in a computationally efficient manner. According to the same criterion, the parametric estimators are also obtained. Additionally, for each cluster, the nonparametric and parametric estimators are derived, and then compute the weighted average per cluster as finally estimators. Moreover, the large sample properties of estimators are also derived respectively.

Keywords

Semiparametric Partially Linear Varying-Coefficient Model, Missing Responses, Cluster Data, Group Lasso

1. Introduction

In real application, the analysis of cluster data arises in various research areas such as biomedicine and so on. Without loss of generality, the data are clustered into classes in terms of the objects which have certain similar

property. For example, focus on the same confidence interval as a cluster. Numerous parametric approaches are applied to the analysis of cluster data, and with the rapid development of computing techniques, nonparametric and semiparametric approaches have attained more and more interest. See the work of Sun et al. [1], Cai [2], Vichi [3], Yi et al. [4], Carrol [5], and He [6], among others.

Consider the semiparametric partially linear varying-coefficient model which is a useful extension of partially linear regression model and varying-coefficient model over all clusters, it satisfies
\[ Y_j = Z_j^\top \beta_j + X_j^\top \alpha_j(U_j) + \epsilon_j, \quad i = 1, \ldots, n_j, j = 1, \ldots, m, \] (1)
where \( Y_j \), \( Z_j \) and \( X_j \) stand for the \( i \)th observation of \( Y \), \( Z \) and \( X \) in the \( j \)th cluster. \( \beta_j = (\beta_{ij}, \ldots, \beta_{ij}) \) is a vector of \( q \)-dimensional unknown parametrics; \( \alpha_j(U_j) = (\alpha_{ij}(U_j), \ldots, \alpha_{ij}(U_j)) \) is a \( p \)-dimensional unknown coefficient vector. \( \epsilon_j \) is random error with mean zero and variance \( \sigma^2 \).

Obviously, when \( m = 1 \), model (1) reduces to semiparametric partially linear varying-coefficient model. A series of literature (You and Chen [7], Fan and Huang [8], Wei and Wu [9], Zhang and Lee [10]) have provided the corresponding statistic inference of such semiparametric model. In [8], Fan and Huang put forward a profile least square technique and propose generalized likelihood ratio test. In [7], You and Chen study the estimation problem when some covariates are measured with additive errors. When \( m = 1 \) and \( Z = 0 \), model (1) becomes varying-coefficient model which has been widely studied by many authors such as Fan and Zhang [11], Hastile and Tibshirani [12], Xia and Li [13], Hoover et al. [14]. When \( m = 1 \), \( p = 1 \) and \( Z = 1 \), model (1) reduces to partially linear regression model which is proposed by Engle et al. [15] when they research the influence of weather on electricity demand. See the literature of Yatchew [16], Spechman [17] and Liang et al. [18], among others.

However, in practice, responses may often not be available completely because of various factors. For example, some sampled units are unwilling to provide the desired information, and some investigators gather incorrect information caused by careless and so on. In that case, a commonly used technique is to introduce a new variable \( \delta \). When \( \delta = 0 \), \( Y \) represents the situation of missing, and \( \delta = 1 \), otherwise. Suppose that responses are missing at random, \( \delta \) and \( Y \) are conditionally independent, then it has
\[ P(\delta = 1|Y, X, Z, U) = P(\delta = 1|X, Z, U). \]

Due to the practicability of the missing response estimation, semiparametric partially linear varying-coefficient model with missing responses has attracted many authors’ attention, such as Chu and Cheng [19], Wei [20], Wang et al. [21] and so on.

It is worth pointing out that there is little work concerning both missing and cluster data especially in semiparametric partially linear varying-coefficient model. If ignore the difference of clusters, it leads the predictors of response values \( Y \) far away from the true values and the estimators have poor robustness. Therefore, it is necessary to take cluster data into consideration with the purpose of improving estimation efficiency. For each cluster, introduce group lasso to semiparametric partially linear varying-coefficient model respectively on the basis of complete case data. In order to automatically select variables and conduct estimation simultaneously, lasso is a popular technique which has attracted many authors’ attention such as Tibshirani [22], Zou [23] and so on. Due to the idea of lasso is to select individual derived input variable rather than the strength of groups of input variables, in this situation, it leads to select more factors as the approach of group lasso. As is shown in Yuan and Yi [24], Wang and Xia [25], Hu and Xia [26] and so on. Thus, this paper centers on the technique of group lasso in a computationally efficient manner. Further then, parametric and nonparametric components are obtained by computing the weighted average per cluster. As for the inference of estimators, the properties of asymptotic normality and consistency are also provided. And Bayesian information criterion (BIC) as tuning parameter selection criterion is used in this article.

The rest of the paper is organized as follows. The use of the applied method is given in Section 2. In Section 3, the theoretical properties are provided. Conclusions are shown in Section 4. Finally, the proofs of the main results are relegated to Appendix.

2. Semiparametric Model with the Methodology
2.1. Model with Complete-Case Data
Due to there exist missing responses, for simplicity, focus on the case where \( \delta = 1 \). That is so-called the method
of complete case data. It is assumed that there are \( m \) independent clusters, and the number of observations in the \( j \)th cluster is \( n_j, j = 1, \ldots, m \). For the \( j \)th subjects from the \( j \)th cluster, let \( \{X_{ij}, Y_{ij}, Z_{ij}, U_{ij}, \delta_{ij}, i = 1, \ldots, n_j, j = 1, \ldots, m\} \) be a set of random sample from model (1), then it is easy to obtain:

\[
\delta_j Y_{ij} = \delta_j Z_{ij}^\top \beta_j + \delta_j X_{ij}^\top \alpha_j (U_{ij}) + \delta_j \varepsilon_{ij}.
\]

(2)

In this situation, if the parametric component \( \beta_j \) is given, model (2) can be written as:

\[
\delta_j Y_{ij}^* = \delta_j X_{ij}^\top \alpha_j (U_{ij}) + \delta_j \varepsilon_{ij},
\]

(3)

where \( Y_{ij}^* = Y_{ij} - Z_{ij}^\top \beta_j \). The coefficient vector \( \alpha_j (u) = \{\alpha_{ij1}(u), \ldots, \alpha_{ijp}(u)\}^\top \in R^p \) is unknown but smooth function in \( u \) and its true value is denoted by \( \alpha_{0j}(u) = \{\alpha_{0ij1}(u), \ldots, \alpha_{0ijp}(u)\}^\top \in R^p \). Suppose that the first integer \( p_0 \leq p \) predictors are relevant and the rest are not.

2.2. The Kernel Least Absolute Shrinkage and Selection Operator Method

Similarity, consider the \( j \)th cluster data firstly, given any index value \( u \in [0,1] \), the estimator of \( \alpha_j (u) \), namely \( \hat{\alpha}_j (u) \), can be obtained by minimizing the following locally weighted least squares function:

\[
Q_j (\alpha_j) = \sum_{i=1}^{n_j} \left( \delta_j Y_{ij}^* - \delta_j X_{ij}^\top \alpha_j (U_{ij}) \right)^2 K_k (u - U_{ij}).
\]

(4)

According to \( \hat{\alpha}_j (u) \), define \( \hat{a}_j = \hat{\alpha}_j (U_{ij}) \) and \( \hat{B} = \{\alpha_{ij1}, \ldots, \alpha_{ijp}\}^\top \in R^{n_j \times p} \). It is clear that, \( \hat{B} \) is a nature estimator for \( \alpha_{0j} (U_{ij}) \). Furthermore, \( \hat{B} \) is also the minimizer of the following global least squares function:

\[
Q(B) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left( \delta_j Y_{ij}^* - \delta_j X_{ij}^\top \alpha_j (U_{ij}) \right)^2 K_k (U_{ij} - U_{ij})
\]

(5)

with respect to \( B = \{\alpha_{ij1}(U_{ij}), \ldots, \alpha_{ijp}(U_{ij})\}^\top \in R^{n_j \times p} \). Due to \( Q(B) \) is a quadratic function in \( B \), thus, depended on the normal equation \( \hat{\mathcal{R}} \mathcal{S} S (B) \hat{\mathcal{R}} \alpha_{ij} = 0 \) for every \( 1 \leq t \leq n_j \), its minimizer is obtained. From another aspect, for \( Q(B) \), as one can see, \( \alpha_j \) is only involved in \( Q_{ij} (\alpha_{ij}) \); see (4). Then it satisfies \( \hat{\mathcal{R}} Q (B) \hat{\mathcal{R}} \alpha_{ij} = \hat{Q}_{ij} (\alpha_{ij}) = 0 \), leading to the solution \( \hat{\alpha}_{ij} = \arg \min \alpha Q_{ij} (\alpha_{ij}) \); see (4). In that case, \( \hat{B} \) is also the minimizer of (5).

Due to it is assumed that the last \( (p - p_0) \) columns of \( B_0 \) matrix should be 0. Therefore, the goal of variable selection amounts to identifying sparse columns in matrix \( B_0 \). In order to discriminate irrelevant variable, which implies that one should identify matrix sparse solutions in \( B_0 \) in a column-wise manner. Based on the group lasso idea of Yuan and Lin [24], Wang and Xia [26], the penalized estimate is shown as follows:

\[
\hat{B}_{ij} = \{\hat{\alpha}_{ij1}(U_{ij}), \ldots, \hat{\alpha}_{ijp}(U_{ij})\}^\top = \{\hat{b}_{ij1}, \ldots, \hat{b}_{ijp}\} = \arg \min_{B \in R^{n_j \times p}} Q_{ij} (B),
\]

(6)

where \( \lambda_j = (\lambda_{ij1}, \ldots, \lambda_{ijp})^\top \in R^p \) is the tuning parameter.

\[
\hat{\alpha}_{ij} (u) = \{\hat{\alpha}_{ij1}(u), \ldots, \hat{\alpha}_{ijp}(u)\}^\top \in R^p, \hat{\beta}_{ijk} = \{\hat{\alpha}_{ij1}(U_{ij}), \ldots, \hat{\alpha}_{ijk}(U_{ij})\}^\top \in R^{n_j},
\]

(6)

\[
Q_{ij} (B) = \sum_{l=1}^{n_j} \sum_{j=1}^{m} \left( \delta_j Y_{ij}^* - \delta_j X_{ij}^\top \alpha_j (U_{ij}) \right)^2 K_k (U_{ij} - U_{ij}) + \sum_{k=1}^{p} \lambda_{ijk} \| \hat{b}_{ik} \|,
\]

(7)

\( b_{ik} \in R^{n_j \times 1} \) is the kth column of B, and \( \| \cdot \| \) means the usual Euclidean norm.
2.3. Local Quadratic Approximation

It is well known that, there exist many computational algorithms for the lasso-type problems such as local quadratic approximation, the least angle regression and many others. For simplicity, this article describes here an easy implementation based on the idea of the local quadratic approximation. Specifically, the implementation is based on an iterative algorithm with $\hat{B}$ as the initial estimator. Let

$$
\hat{B}^m_{ij} = (\hat{b}^{(m)}_{ij}, \ldots, \hat{b}^{(m)}_{ij}) = \left\{ \hat{a}^{(m)}_{ij}(U_{ij}), \ldots, \hat{a}^{(m)}_{ij}(U_{n_{ij}}) \right\}^T
$$

be the Klasso estimate obtained in the $m$th iteration $j$ cluster. Then, the loss function in (6) can be locally approximated by

$$
\sum_{j=1}^{n_j} \sum_{t=1}^{n_t} \left\{ \delta_y Y_j \delta_x Y_j \alpha_j(U_y) \right\}^2 K_h(U_y - U_y) + \sum_{k=1}^{n_k} \| \hat{b}_{ik} - \hat{b}^{(m)}_{ik} \|,
$$

whose minimizer is given by $\hat{B}^{(m+1)}_{ij}$ with the $th$ row given by

$$
\hat{\alpha}_{ij}^{(m+1)}(U_y) = \left( \sum_{t=1}^{n_t} \delta_x Y_j \delta_x Y_j K_h(U_y - U_y) + D^{(m)} \right)^{-1} \times \left( \sum_{t=1}^{n_t} \delta_x X_j \delta_x X_j K_h(U_y - U_y) \right),
$$

where $D^{(m)}$ is a $p \times p$ diagonal matrix with its $k$th diagonal component given by $\lambda_k / \| \hat{b}^{(m)}_{ik} \|$, $k = 1, \ldots, p$.

Furthermore, for each cluster and each group, by using weighted mean idea to gain the finally estimator of coefficient vector $\alpha(U)$. That is, the finally estimator of $\alpha(U)$ can be given by

$$
\hat{\alpha}_{v}^{\text{classo}}(U) = \left( \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} \hat{\alpha}_{ij}^{(m)}(U) \right) / \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} \hat{\alpha}_{ij}^{(m)}(U),
$$

where $\alpha_{iv}(U)$ means $\alpha_p(U)$ in $i$th cluster of $v$th subject.

2.4. Estimation of Parametric Component

In terms of the above estimator of nonparametric component and according to the same criterion, the lasso estimation of parametric components $\beta$ are also derived. As is shown:

$$
Q_2(\beta) = \sum_{j=1}^{n_j} \left\{ \delta_y Y_j \delta_y X_j \hat{\alpha}_{v}^{\text{classo}} - \sum_{k=1}^{n_k} \delta_x Z_j \hat{\beta}_{v} \right\}^2 K_h(u - U_y) + \sum_{k=1}^{n_k} \| \hat{\beta}_{v} \|,
$$

where $\beta_{ij}$ is a coefficient vector of size $q$. Under its assumption, there are $q_0$ predictors relevant and the rest are not. Similarity, following the idea of local quadratic approximation and weighted mean the finally estimator of $\beta$ is given by

$$
\hat{\beta}_{v}^{\text{classo}} = \left( \hat{\beta}_1, \ldots, \hat{\beta}_{q_0} \right) = \frac{1}{m \sum_{j=1}^{n_j}} \left( \sum_{j=1}^{n_j} \beta_{v} \right),
$$

3. Theoretical Properties

3.1. Technical Conditions

The following assumptions are needed to prove the theorems for the proposed estimation methods.

Assumption 1. The random variable $U$ has a bounded support $\Omega$. Its density function $f(\cdot)$ is Lipschitz continuous and bounded away from 0 on its support.

Assumption 2. For each $U \in \Omega$, $E(ZZ^T \mid U)$ is non-singular. $E(XX^T \mid U)$, $E(ZZ^T \mid U)$ and $E(XZ^T \mid U)$
are all Lipschitz continuous. And they have bounded second order derivatives on \([0, 1]\).

Assumption 3. There is an \(s > 2\) such that \(E\|X\|^s < \infty\) and \(E\|\hat{X}\|^s < \infty\) and for some \(s < 2 - s^{-1}\) such that \(n^{2-1/s} \to \infty\).

Assumption 4. \(\{\alpha_j(\cdot), j = 1, \ldots, p\}\) have continuous second derivatives in \(U \in \Omega\).

Assumption 5. The function \(K(\cdot)\) is a symmetric density function with compact support.

Lemma 1. Suppose that the Assumptions of \((A1)-(A5)\) hold, \(h \propto n_j^{-\delta}\), \(n_j^{1/10} a_{nj} \to 0\), then it satisfies

\[
n_j^{-1} \sum_{i=1}^{n_j} \hat{\alpha}_{ij}(U_{ij}) - \alpha_i(U_{ij}) = 0 \text{ in } (n_j^{-4/5}).
\]

Lemma 2. If \((A1)-(A5)\), \(h \propto n_j^{-\delta}\), \(n_j^{1/10} a_{nj} \to 0\), and \(n_j^{1/10} b_{nj} \to \infty\), then \(P\{\|\hat{b}_{nj}\| = 0\} \to 1\) for any \(p_0 < k \leq p\).

The proof of Lemma 1 and Lemma 2 can be shown in Wang and Xia [25].

### 3.2. Basic Theorems

Suppose that the Assumptions \((A1)-(A5)\) hold. For \(j\)th cluster, let \(X_{ajj} = \{X_{ij1}, \ldots, X_{ijp}\}^T \in R^{p}\),

\[
X_{ajj} = (X_{a1j}, \ldots, X_{ajj})^T \in R^{p\times r} \text{ and } \hat{\alpha}_{ajj}(u) = (\hat{\alpha}_{ajj1}(u), \ldots, \hat{\alpha}_{ajjp}(u)) \in R^{p\times r},
\]

\[
\hat{\alpha}_{ajj}(u) = (\hat{\alpha}_{ajj1}(u), \ldots, \hat{\alpha}_{ajjp}(u)) \in R^{p\times r}.
\]

Denote \(a_{nj} = \max\{\lambda_{kj} 1 \leq k \leq p_0\}, \ b_{nj} = \min\{\lambda_{kj} p_0 + 1 \leq k \leq p\}.

Theorem 1. Assume \((A1)-(A5)\), \(h \propto n_j^{-\delta}\), \(n_j^{1/10} a_{nj} \to 0\), and \(n_j^{1/10} b_{nj} \to \infty\), then we have

\[
P\{\sup_{u \in [0,1]} \|\hat{\alpha}_{ajj}(u) - \hat{\alpha}_{a_{nj}}(u)\| = 0\} \to 1 \text{ for any } p_0 < k \leq p.
\]

With the purpose of considering the oracle property, define the oracle estimators as follows:

\[
\hat{\alpha}_{a_{nj}}(u) = \left(1 \frac{1}{n_j} \sum_{i=1}^{n_j} \delta_{ajj} X_{a_{nj}} Y_{a_{nj}}^{k} K_h \left(U_{ij} - u\right)\right) \times \left(1 \frac{1}{n_j} \sum_{i=1}^{n_j} \delta_{ajj} X_{ajj} Y_{a_{nj}}^{k} K_h \left(U_{ij} - u\right)\right).
\]

Theorem 2. Suppose that the assumptions are satisfied, if \(h \propto n_j^{-\delta}\), \(n_j^{1/10} a_{nj} \to 0\), and \(n_j^{1/10} b_{nj} \to \infty\), then it is easy to see that

\[
\sup_{u \in [0,1]} \|\hat{\alpha}_{a_{nj}}(u) - \hat{\alpha}_{a_{nj}}(u)\| = 0 \text{ in } (n_j^{-2/5}).
\]

### 3.3. Tuning Parameter Selection

In the case where \(n_j^{1/10} a_{nj} \to 0\) and \(n_j^{1/10} b_{nj} \to \infty\), the optimal convergence rate can be obtained and the true model can be consistently identified. Due to there exists a great challenge to select \(p\) shrinkage parameters, thus as shown in Zou [23], Wang and Xia [25], simplify the tuning parameters as follows:

\[
\lambda_{kj} = \frac{\lambda_{0}}{n_j^{-1/2} \|\hat{\alpha}_{ajj}\|},
\]

where \(\hat{\alpha}_{ajj}\) is the \(k\)th column of the unpenalized estimate \(\hat{B}\) in \(j\)th cluster. Since \(\hat{\alpha}_{ajj}\) is an estimator with \(\lambda_{kj} = 0\), the results of Theorem 1 and Theorem 2 can be applied. Thus, as long as \(\lambda_{0} n_j^{1/10} \to 0\) but \(\lambda_{0} n_j^{1/10} \to \infty\), one can conclude that \(n_j^{1/10} a_{nj} \to 0\) and \(n_j^{1/10} b_{nj} \to \infty\). Furthermore, the original \(p\)-dimensional problem about \(\lambda \in R^p\) becomes a univariate problem regarding \(\lambda_{0} \in R\). According to BIC-type criterion, \(\lambda_{0}\) is defined as follows:

\[
\text{BIC}_{\lambda_{kj}} = \log (\text{RSS}_{\lambda_{kj}}) + df_{\lambda_{kj}} \times \frac{\log(n_{j} h)}{n_{j} h},
\]
where \( df_{\lambda_j} \) is the number of varying coefficients identified by \( B_{\lambda_j} \). \( RSS_{\lambda_j} \) is

\[
RSS_{\lambda_j} = n_j^{-2} \sum_{j=1}^{n_j} \sum_{i=1}^{n_j} \left( \delta_{ij} Y_{ij} - \hat{\delta}_{ij} X_{ij}^T \hat{\beta}_{\lambda_j} (U_{ij}) \right)^2 K_k \left( U_{ij} - U_{\cdot j} \right).
\]

(13)

Obviously, the effective sample size \( n_j h \) is used instead of the original sample size \( n_j \). Further then, the tuning parameter can be given by

\[
\hat{\lambda}_j = \arg \min BIC_{\lambda_j}.
\]

Note that \( R = \{k_1, \cdots, k_p\} \) as an arbitrary model with a total of \( 0 \leq p^* \leq p \) nonzero coefficients (i.e. \( X_{ik1}, \cdots, X_{ikp} \)). Then, \( R_T = \{1, \cdots, p_0\} \) means the true model and \( R_{\lambda_j} = \{k : \| \hat{\beta}_{(i,k)} \| > 0 \} \) denotes the model identified by the proposed estimate \( \hat{B}_{\lambda_j} \). Consequently, \( R_{\lambda_j} \) represents the model identified by \( \hat{B}_{\lambda_j} \).

**Theorem 3.** Selection Consistency. Suppose that Assumptions (A1)-(A5) hold, the tuning parameter \( \hat{\lambda}_j \) selected by the BIC criterion can indeed identify the true model consistency, i.e. \( P \left( R_{\lambda_j} = R_T \right) \to 1 \) as \( n_j \to \infty \).

4. Conclusion

In this paper, it mainly discusses the shrinkage estimation of semiparametric partially linear varying-coefficient model under the circumstance that there exist missing responses for cluster data. Combined the idea of complete-case data, this paper introduces group lasso into semiparametric model with different cluster respectively. The new method simultaneously conducts variable selection and model estimation. Meanwhile, the technique not only reduces biased results but also improves the estimation efficiency. Finally, combined the idea of weighted mean, the nonparametric and parametric estimators are derived. The BIC criterion as tuning parameter selection is well applied in this artice. Furthermore, the properties of asymptotic normality and consistency are also derived theoretically.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (61472093). This support is greatly appreciated.

References


Models. *Bernoulli*, 11, 1031-1057. [dx.doi.org/10.3150/bj/1137421639]


Appendix

Proof of Theorem 1

Proof. Based on Lemma 2 and as shown in Hunter and Li [27], one can know that \( \| \hat{\beta}_{\alpha,j} \| \rightarrow \| \hat{\beta}_{\alpha,j} \| \) for each \( 1 \leq k \leq p \). Then, as long as \( m \to \infty \), one can see that when \( k \leq p_0 \) then \( \| \hat{\beta}_{\alpha,j} \| \) converge to a positive number, otherwise, \( \| \hat{\beta}_{\alpha,j} \| \) converge to 0. Denote \( D_{aj}^m \) as the upper \( p_0 \times p_0 \) diagonal submatrix of \( n_j^{-1}D^{(m)} \) and \( D_{bj}^m \) as the lower \((p-p_0)\times(p-p_0)\) diagonal submatrix of \( n_j^{-1}D^{(m)} \). From the definition of \( D^{(m)} \), it is remarkable that each diagonal component of \( D_{aj}^m \) must converge to some finite number while \( D_{bj}^m \) diverge to infinity in the case where \( m \to \infty \).

For simplify, we follow (8) and \( \hat{\alpha}_{ij}(u) \) can be rewritten as \( \hat{\alpha}_{ij}(u) = \{ \omega(u) + n_j^{-1}D^{(m)} \}^{-1} \zeta(u) \), where \( \omega(u) \) is a \( 2 \times 2 \) block matrix given by \( \{ \omega_{aa}(u), \omega_{ab}(u); \omega_{ba}(u), \omega_{bb}(u) \} \) and 
\[
\zeta(u) = (\zeta_a^T(u), \zeta_b^T(u))^T \in \mathbb{R}^p, \quad \text{with} \quad \omega_{aa}(u) = n_j^{-1} \sum_{i=1}^{n_j} \delta_{ij}Y_iX_i^T K_b(U_y - u), \\
\omega_{ab}(u) = n_j^{-1} \sum_{i=1}^{n_j} \delta_{ij}X_iY_i^T K_b(U_y - u), \\
\omega_{ba}(u) = n_j^{-1} \sum_{i=1}^{n_j} \delta_{ij}X_i^T Y_i K_b(U_y - u) = \omega_{ba}(u), \\
\zeta_a^T(u) = n_j^{-1} \sum_{i=1}^{n_j} \delta_{ij}X_iY_i^T K_b(u - U_y), \quad \text{and} \quad \zeta_b^T(u) = n_j^{-1} \sum_{i=1}^{n_j} \delta_{ij}X_i^T Y_i K_b(u - U_y). \]
If \( \{ \omega(u) + D^{(m)} \}^{-1} \) is given by \( \{ \Xi^{(m)}_{aa}, \Xi^{(m)}_{ab}, \Xi^{(m)}_{ba}, \Xi^{(m)}_{bb} \} \) one obtains 
\[
\Xi^{(m)}_{aa}(u) = \left( \omega_{aa}(u) + \omega_{bb}(u) \right) \left( \omega_{ba}(u) + \omega_{bb}(u) \right)^{-1} \omega_{ba}(u), \\
\Xi^{(m)}_{ab}(u) = -\left( \omega_{aa}(u) + \omega_{bb}(u) \right)^{-1} \omega_{ab}(u) \left( \omega_{ba}(u) + \omega_{bb}(u) \right)^{-1} \omega_{ba}(u), \\
\Xi^{(m)}_{ba}(u) = -\left( \omega_{ab}(u) + \omega_{bb}(u) \right)^{-1} \omega_{ba}(u) \left( \omega_{ba}(u) + \omega_{bb}(u) \right)^{-1} \omega_{ba}(u), \\
\Xi^{(m)}_{bb}(u) = \left( \omega_{ab}(u) + \omega_{bb}(u) \right) \left( \omega_{ba}(u) + \omega_{bb}(u) \right)^{-1} \omega_{ba}(u). 
\]

Due to each diagonal component of \( D_{aj}^m \) must converge to some finite number while \( D_{bb}^m \) diverge to infinity in the case where \( m \to \infty \), thus each component of \( \Xi^{(m)}_{aa}(u) \) and \( \Xi^{(m)}_{bb}(u) \) converge to 0 uniformly on \([0, 1]\) as \( m \to \infty \). It is easy to see that 
\[
\hat{\alpha}_{aj,k}^{n+1} = \Xi^{(m)}_{aa}(u) \zeta_a^T(u) + \Xi^{(m)}_{bb}(u) \zeta_b^T(u), 
\]
where \( \zeta_a^T(u) \) and \( \zeta_b^T(u) \) are uniformly bounded. Obviously, \( \hat{\alpha}_{aj,k}^{n+1} \to 0 \) as \( m \to \infty \) when \( p_0 < k \leq p \). Therefore, \( \sup \| \hat{\alpha}_{aj,k}(u) \| = 0 \) for every \( p_0 < k \leq p \). It completes the proof of Theorem 1.

Proof of Theorem 2

Proof. As is well known, \( \hat{\alpha}_{aj,k}(u) \) is the solution of the following equation 
\[
\frac{1}{n_j} \sum_{i=1}^{n_j} \delta_{ij}X_i \left( \delta_{ij}Y_i - \delta_{ij}X_i \hat{\alpha}_{aj,k} \right) K_b(U_y - u) + n_j^{-1} \sum_{k=1}^{p_0} \hat{\beta}_{aj,k} \| \hat{\beta}_{aj,k} \| = 0. 
\]
That is to say, \( \hat{\alpha}_{aj,k} \) satisfies
\[ \hat{a}_{\alpha j}(u) = \left[ \hat{\Theta}(u) \right]^{-1} \times \left\{ \frac{1}{n_j} \sum_{i=1}^{n_j} \hat{y}_i X_{aj}^T Y_{aj} K_h \left( U_{aj} - u \right) + \frac{1}{n_j} \sum_{k=1}^{p_j} \lambda_{jk} \frac{\hat{b}_{j,k}}{\|\hat{b}_{j,k}\|} \right\} = 0, \]

where \( \hat{\Theta}(u) = n_j^{-1} \sum_{i=1}^{n_j} K_h \left( u - U_{aj} \right) \delta_y X_{aj} X_{aj}^T \). By Lemma 2 and combined with the oracle estimator \( \hat{a}_{\alpha j}(u) \), it satisfies

\[
\max \left\| \hat{a}_{\alpha j}(u) - \hat{a}_{\alpha j}(u) \right\| = \max \left\{ \hat{\Theta}(u)^{-1} \left[ \frac{1}{n_j} \sum_{k=1}^{p_j} \lambda_{jk} \frac{\hat{b}_{j,k}}{\|\hat{b}_{j,k}\|} \right] \right\} \leq \hat{\lambda}_{\max} \left[ \frac{1}{n_j} \sum_{k=1}^{p_j} \lambda_{jk} \frac{\hat{b}_{j,k}}{\|\hat{b}_{j,k}\|} \right] \leq \frac{\hat{\lambda}_{\max} \alpha_{aj}}{n_j} n_j^{-21/10},
\]

where \( \hat{\lambda}_{\max} = \sup \lambda_{\max} \{ f(u) \Theta(u) \} \) with \( \lambda_{\max}(A) \) represents the maximal eigenvalue of an arbitrary positive definite matrix A. Notice that \( n_j^{-21/10} = O\left(n_j^{-23/5}\right) \), as a result it completes the proof of Theorem 2. \( \square \)