Open Journal of Statistics, 2015, 5, 768-776
Published Online December 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.57076

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

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

How to cite this paper: Zhang, M.X., Qiao, J.N., Yang, H.W. and Liu, Z.X. (2015) Shrinkage Estimation of Semiparametric Model with Missing Responses for Cluster Data. *Open Journal of Statistics*, **5**, 768-776.
http://dx.doi.org/10.4236/ojs.2015.57076

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

$$Y_{ij} = Z_{ij}^{\mathrm{T}}\beta_j + X_{ij}^{\mathrm{T}}\alpha_j\left(U_{ij}\right) + \varepsilon_{ij},\ i = 1,\cdots,n_j,\ j = 1,\cdots,m, \tag{1}$$

where Y_{ij}, Z_{ij} and X_{ij} stand for the ith observation of Y, Z and X in the jth cluster. $\beta_j = \left(\beta_{1j},\cdots,\beta_{qj}\right)$ is a vector of q-dimensional unknown parametrics; $\alpha_j\left(U_{ij}\right) = \left(\alpha_{1j}\left(U_{ij}\right),\cdots,\alpha_{pj}\left(U_{ij}\right)\right)$ is a p-dimensional unknown coefficient vector. ε_{ij} is random error with mean zero and variance σ^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 δ. When $\delta = 0$, Y represents the situation of missing, and $\delta = 1$, otherwise. Suppose that responses are missing at random, δ and Y are conditionally independent, then it has

$$P\left(\delta = 1 \mid Y, X, Z, U\right) = P\left(\delta = 1 \mid X, Z, U\right).$$

Due to the practicability of the missing responses 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, $j=1,\cdots,m$. For the *i*th subjects from the *j*th cluster, let $\left\{X_{ij},Y_{ij},Z_{ij},U_{ij},\delta_{ij},i=1,\cdots,n_j,j=1,\cdots,m\right\}$ be a set of random sample from model (1), then it is easy to obtain:

$$\delta_{ij}Y_{ij}=\delta_{ij}Z_{ij}^{\mathrm{T}}\beta_j+\delta_{ij}X_{ij}^{\mathrm{T}}\alpha_j\left(U_{ij}\right)+\delta_{ij}\varepsilon_{ij}. \tag{2}$$

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

$$\delta_{ij}Y_{ij}^{*}=\delta_{ij}X_{ij}^{\mathrm{T}}\alpha_j\left(U_{ij}\right)+\delta_{ij}\varepsilon_{ij}, \tag{3}$$

where $Y_{ij}^{*}=Y_{ij}-Z_{ij}^{\mathrm{T}}\beta_j$. The coefficient vector $\alpha_j(u)=\left\{\alpha_{1j}(u),\cdots,\alpha_{pj}(u)\right\}^{\mathrm{T}}\in R^P$ is unknown but smooth function in *u* and its true value is denoted by $\alpha_{0j}(u)=\left\{\alpha_{01j}(u),\cdots,\alpha_{0pj}(u)\right\}^{\mathrm{T}}\in R^P$. Suppose that the first integer $p_0\le p$ 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 $\tilde{\alpha}_j(u)$, can be obtained by minimizing the following locally weighted least squares function:

$$Q_u\left(\alpha_j\right)=\sum_{i=1}^{n_j}\left(\delta_{ij}Y_{ij}^{*}-\delta_{ij}X_{ij}^{\mathrm{T}}\alpha_j\right)^2K_h\left(u-U_{ij}\right). \tag{4}$$

According to $\tilde{\alpha}_j(u)$, define $\tilde{\alpha}_{tj}=\tilde{\alpha}_j\left(U_{tj}\right)$ and $\tilde{B}=\left(\alpha_{1j},\cdots,\alpha_{n_jj}\right)^{\mathrm{T}}\in R^{n_j\times p}$. It is clear that, $\tilde{B}$ is a nature estimator for $B_0=\left\{\alpha_{0j}\left(U_{1j}\right),\cdots,\alpha_{0j}\left(U_{n_jj}\right)\right\}^{\mathrm{T}}\in R^{n_j\times p}$. Furthermore, $\tilde{B}$ is also the minimizer of the following global least squares function:

$$Q(B)=\sum_{t=1}^{n_j}Q_{U_{tj}}\left(\alpha_{tj}\right)=\sum_{t=1}^{n_j}\sum_{i=1}^{n_j}\left\{\delta_{ij}Y_{ij}^{*}-\delta_{ij}X_{ij}^{\mathrm{T}}\alpha_{tj}\right\}^2K_h\left(U_{tj}-U_{ij}\right) \tag{5}$$

with respect to $B=\left\{\alpha_j\left(U_{1j}\right),\cdots,\alpha_j\left(U_{n_jj}\right)\right\}^{\mathrm{T}}=\left(\alpha_{1j},\cdots,\alpha_{n_jj}\right)^{\mathrm{T}}\in R^{n_j\times p}$. Due to *Q*(*B*) is a quadratic function in *B*, thus, depended on the normal equation $\partial RSS(B)/\partial\alpha_{tj}=0$ for every $1\le t\le n_j$, its minimizer is obtained. From another aspect, for *Q*(*B*), as one can see, α_{tj} is only involved in $Q_{U_{tj}}\left(\alpha_{tj}\right)$; see (4). Then it satisfies $\partial Q(B)/\partial\alpha_{tj}=\partial Q_{U_{tj}}\left(\alpha_{tj}\right)=0$, leading to the solution $\tilde{\alpha}_{tj}=\arg\min\alpha_tQ_{U_t}\left(\alpha_{tj}\right)$; see (4). In that case, $\tilde{B}$ 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:

$$\hat{B}_{\lambda_j}=\left\{\hat{\alpha}_{\lambda_j}\left(U_{1j}\right),\cdots,\hat{\alpha}_{\lambda_j}\left(U_{n_jj}\right)\right\}^{\mathrm{T}}=\left\{\hat{b}_{\lambda_j1},\cdots,\hat{b}_{\lambda_jp}\right\}=\arg\min_{B\in R^{n_j\times p}}Q_{\lambda_j}(B),$$

where $\lambda_j=\left(\lambda_{j1},\cdots,\lambda_{jp}\right)^{\mathrm{T}}\in R^P$ is the tuning parameter.

$$\hat{\alpha}_{\lambda_j}(u)=\left\{\hat{\alpha}_{\lambda_j1}(u),\cdots,\hat{\alpha}_{\lambda_jp}(u)\right\}^{\mathrm{T}}\in R^p,\hat{b}_{\lambda_jk}=\left\{\hat{\alpha}_{\lambda_jk}\left(U_{1j}\right),\cdots,\hat{\alpha}_{\lambda_jk}\left(U_{nj}\right)\right\}^{\mathrm{T}}\in R^{n_j}, \tag{6}$$

$$Q_{\lambda_j}(B)=\sum_{t=1}^{n_j}\sum_{i=1}^{n_j}\left\{\delta_{ij}Y_{ij}^{*}-\delta_{ij}X_{ij}^{\mathrm{T}}\alpha_j\left(U_{tj}\right)\right\}^2K_h\left(U_{tj}-U_{ij}\right)+\sum_{k=1}^{p}\lambda_{jk}\left\|b_{jk}\right\|, \tag{7}$$

$b_{jk}\in R^{n_j\times1}$ is the *k*th column of B, and $\|.\|$ 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 $\tilde{B}$ as the initial estimator. Let

$$\hat{B}_{\lambda_j}^m = \left(\hat{b}_{\lambda_j 1}^{(m)}, \cdots, \hat{b}_{\lambda_j p}^{(m)}\right) = \left\{\hat{\alpha}_{\lambda_j}^m\left(U_{1j}\right), \cdots, \hat{\alpha}_{\lambda_j}^m\left(U_{n_j j}\right)\right\}^{\mathrm{T}}$$

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

$$\sum_{t=1}^{n_j}\sum_{i=1}^{n_j}\left\{\delta_{ij}Y_{ij}^* - \delta_{ij}X_{ij}^{\mathrm{T}}\alpha_j\left(U_{tj}\right)\right\}^2 K_h\left(U_{tj}-U_{ij}\right) + \sum_{k=1}^{p}\lambda_{jk}\frac{\left\|b_{jk}\right\|^2}{\left\|\hat{b}_{\lambda_{jk}}^{(m)}\right\|}$$

$$= \sum_{t=1}^{n_j}\left(\sum_{i=1}^{n_j}\left\{\delta_{ij}Y_{ij}^* - \delta_{ij}X_{ij}^{\mathrm{T}}\alpha_j\left(U_t\right)\right\}^2 K_h\left(U_{tj}-U_{ij}\right) + \sum_{k=1}^{p}\lambda_{jk}\frac{\alpha_{jk}^2\left(U_{tj}\right)}{\left\|\hat{b}_{\lambda_{jk}}^{(m)}\right\|}\right),$$

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

$$\hat{\alpha}_{\lambda_j}^{(m+1)}\left(U_{tj}\right) = \left(\sum_{t=1}^{n_j}\delta_{ij}X_{ij}X_{ij}^{\mathrm{T}}K_h\left(U_{tj}-U_{ij}\right) + D^{(m)}\right)^{-1} \times \left(\sum_{i=1}^{n_j}\delta_{ij}X_{ij}Y_{ij}^*K_h\left(U_{tj}-U_{ij}\right)\right) \tag{8}$$

where $D^{(m)}$ is a $p\times p$ diagonal matrix with its *k*th diagonal component given by $\lambda_{jk}\Big/\left\|\hat{b}_{\lambda_{jk}}^m\right\|$, $k=1,\cdots,p$.

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

$$\hat{\alpha}_c^{glasso} = \left(\hat{\alpha}_1(U), \cdots, \hat{\alpha}_{P_0}(U)\right) = \frac{1}{m}\frac{1}{n_j}\left(\sum_{s=1}^{m}\sum_{i=1}^{n_j}\alpha_{is1}(U), \cdots, \sum_{s=1}^{m}\sum_{i=1}^{n_j}\alpha_{isp}(U)\right),$$

where $\alpha_{isp}(U)$ means $\alpha_P(U)$ in *s*th cluster of *i*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 β are also derived. As is shown:

$$Q_\lambda(\beta) = \sum_{i=1}^{n_j}\left\{\delta_{ij}Y_{ij} - \delta_{ij}X_{ij}^{\mathrm{T}}\hat{\alpha}_c^{glasso} - \sum_{k=1}^{q}\delta_{ij}Z_{ij}^{\mathrm{T}}\beta_k\right\}^2 K_h\left(u-U_{ij}\right) + \sum_{k=1}^{q}\lambda_k\left\|\beta_k\right\|^2, \tag{9}$$

where β_j 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 β is given by

$$\hat{\beta}_c^{glasso} = \left(\hat{\beta}_1, \cdots, \hat{\beta}_{q_0}\right) = \frac{1}{m}\frac{1}{n_j}\left(\sum_{s=1}^{m}\sum_{i=1}^{n_j}\beta_{is1}, \cdots, \sum_{s=1}^{m}\sum_{i=1}^{n_j}\beta_{isq}\right). \tag{10}$$

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 Ω. Its density function $f(.)$ is Lipschitz continuous and bounded away from 0 on its support.

Assumption 2. For each $U\in\Omega$, $E\left(ZZ^{\mathrm{T}}\,|\,U\right)$ is non-singular. $E\left(XX^{\mathrm{T}}\,|\,U\right)$, $E\left(ZZ^{\mathrm{T}}\,|\,U\right)$ and $E\left(XZ^{\mathrm{T}}\,|\,U\right)$

are all Lipschitz continuous. And they have bounded second order derivatives on [0, 1].

Assumption 3. There is an $s>2$ such that $E\|X\|^{2s}<\infty$ and $E\|Z\|^{2s}<\infty$ and for some $\varepsilon<2-s^{-1}$ such that $n^{2\varepsilon-1}h\to\infty$.

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

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

Lemma 1. *Suppose that the Assumptions of* (A1)-(A5) *hold,* $h\infty n_j^{-1/5}$, *and* $n_j^{11/10}a_{n_j}\to 0$, *then it satisfies*

$$n_j^{-1}\sum_{t=1}^{n_j}\left\|\hat{\alpha}_{\lambda_j}(U_{tj})-\alpha_0(U_{tj})\right\|^2=0_p\left(n_j^{-4/5}\right).$$

Lemma 2. *If* (A1)-(A5), $h\infty n_j^{-1/5}$, $n_j^{11/10}a_{n_j}\to 0$, *and* $n_j^{11/10}b_{n_j}\to\infty$, *then* $P\left(\left\|\hat{b}_{\lambda_j k}\right\|=0\right)\to 1$ *for any* $p_0<k\le p$.

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 $X_{iaj}=\left(X_{i1j},\cdots,X_{ip_0j}\right)^{\mathrm{T}}\in R^{p_0}$,

$X_{ibj}=\left(X_{i(p_0+1)j},\cdots,X_{ipj}\right)^{\mathrm{T}}\in R^{p-p_0}$ and $\hat{\alpha}_{a\lambda_j}(u)=\left(\hat{\alpha}_{\lambda_j 1}(u),\cdots,\hat{\alpha}_{\lambda_j p_0}(u)\right)\in R^{p_0}$,

$\hat{\alpha}_{b\lambda_j}(u)=\left(\hat{\alpha}_{\lambda_j p_0+1}(u),\cdots,\hat{\alpha}_{\lambda_j p}(u)\right)\in R^{p-p_0}$. Denote $a_{n_j}=\max\left\{\lambda_{kj}1\le k\le p_0\right\}$, $b_{n_j}=\min\left\{\lambda_{kj}p_0+1\le k\le p\right\}$.

Theorem 1. *Assume* (A1)-(A5), $h\infty n_j^{-1/5}$, $n_j^{11/10}a_{n_j}\to 0$, *and* $n_j^{11/10}b_{n_j}\to\infty$, *then we have*

$P\left(\sup_{u\in[0,1]}\left\|\hat{\alpha}_{b,\lambda_j}(u)\right\|=0\right)\to 1$ *for any* $p_0<k\le p$.

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

$$\hat{\alpha}_{ora}(u)=\left\{\frac{1}{n_j}\sum_{i=1}^{n_j}\delta_{iaj}X_{iaj}X_{iaj}^{\mathrm{T}}K_h(U_{ij}-u)\right\}^{-1}\times\left\{\frac{1}{n_j}\sum_{i=1}^{n_j}\delta_{iaj}X_{iaj}Y_{ij}^{*}K_h(U_{ij}-u)\right\}.$$

Theorem 2. *Suppose that the assumptions are satisfied, if* $h\infty n_j^{-1/5}$, $n_j^{11/10}a_{n_j}\to 0$, *and* $n_j^{11/10}b_{n_j}\to\infty$, *then it is easy to see that*

$$\sup_{u\in[0,1]}\left\|\hat{\alpha}_{a,\lambda_j}(u)-\hat{\alpha}_{ora}(u)\right\|=0_p\left(n^{-2/5}\right).$$

3.3. Tuning Parameter Selection

In the case where $n_j^{11/10}a_{n_j}\to 0$ and $n_j^{11/10}b_{n_j}\to\infty$, the optimal convergence rata 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:

$$\lambda_{jk}=\frac{\lambda_0}{n_j^{-1/2}\left\|\tilde{\alpha}_{jk}\right\|},\tag{11}$$

where $\tilde{\alpha}_{jk}$ is the kth column of the unpenalized estimate $\tilde{B}$ in jth cluster. Since α_{jk} is an estimator with $\lambda_{jk}=0$, the results of **Theorem 1** and **Theorem 2** can be applied. Thus, as long as $\lambda_0 n_j^{11/10}\to 0$ but $\lambda_0 n_j^{3/2}\to\infty$, one can conclude that $n_j^{11/10}a_{n_j}\to 0$ and $n_j^{11/10}b_{n_j}\to\infty$. 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, λ_0 is defined as follows:

$$BIC_{\lambda_j}=\log\left(RSS_{\lambda_j}\right)+df_{\lambda_j}\times\frac{\log(n_j h)}{n_j h},\tag{12}$$

where df_{λ_j} is the number of varying coefficients identified by B_{λ_j}. RSS_{λ_j} is

$$RSS_{\lambda_j} = n_j^{-2} \sum_{t=1}^{n_j} \sum_{i=1}^{n_j} \left\{ \delta_{ij} Y_{ij}^* - \delta_{ij} X_{ij}^{\mathrm{T}} \hat{\alpha}_{\lambda_j} \left(U_{tj} \right) \right\}^2 K_h \left(U_{tj} - U_{ij} \right). \tag{13}$$

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 $R = \left\{ k_1, \cdots, k_{p^*} \right\}$ as an arbitrary model with a total of $0 \le p^* \le p$ nonzero coefficients (*i.e.* $X_{ijk_1}, \cdots, X_{ijk_{p^*}}$). Then, $R_T = \{1, \cdots, p_0\}$ means the true model and $R_{\lambda_j} = \left\{ k : \left\| \hat{\alpha}_{\lambda_j k} \right\| > 0 \right\}$ denotes the model identified by the proposed estimate $\hat{B}_{\lambda_j}$. Consequently, $R_{\hat{\lambda}_j}$ represents the model identified by $\hat{B}_{\hat{\lambda}}$.

Theorem 3. *Selection Consistency. Suppose that Assumptions* (*A*1)-(*A*5) *hold, the tuning parameter* $\hat{\lambda}_j$ *selected by the BIC criterion can indeed identify the true model consistency, i.e.* $P\left(R_{\hat{\lambda}_j} = R_T \right) \to 1$ *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

[1] Sun, Y., Li, J.L. and Zhang, W.Y. (2012) Estimation and Model Selection in a Class of Semiparametric Models for Cluster Data. *Annals of the Institute of Statistical Mathematics*, **64**, 835-856. http://dx.doi.org/10.1007/s10463-011-0342-9

[2] Cai, J.W. (2005) Semiparametric Models for Clustered Recurrent Event Data. *Life Data Analysis*, **11**, 405-425. http://dx.doi.org/10.1007/s10985-005-2970-y

[3] Vichi, M. (2008) Fitting Semiparametric Clustering Models to Dissimilarity Data. *Advances in Data Analysis and Classification*, **2**, 121-161. http://dx.doi.org/10.1007/s11634-008-0025-4

[4] Yi, G.Y., He, W.Q. and Liang, H. (2011) Semiparametric Marginal and Association Regression Methods for Clustered Binary Data. *Annals of the Institute of Statistical Mathematics*, **63**, 511-533. http://dx.doi.org/10.1007/s10463-009-0239-z

[5] Carroll, R., Maity, A., Mammen, E. and Yu, K. (2009) Efficient Semiparametric Marginal Estimation for the Partially Linear Additive Model for Longitudinal/Clustered Data. *Statistics in Biosciences*, **1**, 10-31. http://dx.doi.org/10.1007/s12561-009-9000-7

[6] He, S., Wang, F. and Sun, L.Q. (2013) A Semiparametric Additive Rates Model for Clustered Recurrent Event Data, Acta Mathematicae Applicatae Sinica. *English. Series*, **29**, 55-62. http://dx.doi.org/10.1007/s10255-011-0093-7

[7] You, J.H. and Chen, G.M. (2006) Estimation of a Semiparametric Varying-Coefficient Partially Linear Errors-in-Variables Model. *Journal of Multivariate Analysis*, **97**, 324-341. http://dx.doi.org/10.1016/j.jmva.2005.03.002

[8] Fan, J.Q. and Huang, T. (2005) Profile Likelihood Inferences on Semiparametric Varying-Coefficient Partially Linear

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

[9] Wei, C.H. and Wu, X.Z. (2008) Profile Lagrange Multiplier Test for Partially Linear Varying-Coefficient Regression Models. *Journal of Systems Science & Mathematical Sciences*, **28**, 416-424.

[10] Zhang, W., Lee, S.Y. and Song, X. (2002) Local Polynomial Fitting in Semivarying Coefficient Models. *Journal of Multivariate Analysis*, **82**, 166-188. http://dx.doi.org/10.1006/jmva.2001.2012

[11] Fan, J.Q. and Zhang, W.Y. (1999) Statistical Estimation in Varying-Coefficient Models. *Annals of Statistics*, **27**, 1491-1581. http://dx.doi.org/10.1214/aos/1017939139

[12] Hastile, T.J. and Tibshirani, R.J. (1993) Varying-Coefficient Models (With Discussion). *Journal of the Royal Statistical Society*: *Series B*, **55**, 757-796.

[13] Xia, Y.C. and Li, W.K. (1999) On the Estimation and Testing of Functional-Coefficient Linear Models. *Statistica Sinica*, **9**, 737-757.

[14] Hoover, D.R., Rice, J.A., Wu, C.O. and Yang, L.P. (1998) Nonparametric Smoothing Estimates of Time-Varying Coefficient Models with Longitudinal Data. *Biometrika*, **85**, 809-822. http://dx.doi.org/10.1093/biomet/85.4.809

[15] Engle, R.F., Granger, W.J., Rice, J. and Weiss, A. (1996) Semiparametric Estimates of the Relation between Weather and Electricity Techniques. *Journal of the American Statistical Association*, **80**, 310-319. http://dx.doi.org/10.1080/01621459.1986.10478274

[16] Yatchew, A. (1997) An Elementary Estimator of the Partial Linear Model. *Economics Letters*, **57**, 135-143. http://dx.doi.org/10.1016/S0165-1765(97)00218-8

[17] Speckman, P. (1988) Kernel Smoothing in Partial Linear Models. *Journal of the Royal Statistical Society*: *Series B*, **50**, 413-416.

[18] Liang, H. (2006) Estimation in Partially Linear Models and Numerical Comparisons. *Computational Statistics & Data Analysis*, **50**, 675-687. http://dx.doi.org/10.1016/j.csda.2004.10.007

[19] Chu, C. and Cheng, P. (1995) Nonparametric Regression Estimation with Missing Data. *Journal of Statistical Planning and Inference*, **48**, 85-99. http://dx.doi.org/10.1016/0378-3758(94)00151-K

[20] Wei, C.H. (2010) Estimation in Partially Linear Varying-Coefficient Errors-in-Variables Models with Missing Responses. *Acta Mathematica Scientia*, **30**, 1042-1054.

[21] Wang, Q., Linton, O. and Hardle, W. (2007) Semiparametric Regression Analysis with Missing Response at Random. *Journal of Multivariate Analysis*, **98**, 334-345. http://dx.doi.org/10.1016/j.jmva.2006.10.003

[22] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*: *Series B*, **58**, 267-288.

[23] Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429. http://dx.doi.org/10.1198/016214506000000735

[24] Yuan, M. and Lin, Y. (2006) Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society*: *Series B*, **68**, 49-67. http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x

[25] Wang, H.S. and Xia, Y.C. (2009) Shrinkage Estimation of the Varying Coefficient Model. *Journal of the American Statistical Association*, **104**, 747-757. http://dx.doi.org/10.1198/jasa.2009.0138

[26] Hu, T. and Xia, Y.C. (2010) Adaptive Semivarying Coefficient Model Selection. *Statistica Sinica*, **22**, 575-599.

[27] Hunter, D.R. and Li, R. (2005) Variable Selection Using MM Algorithms. *Annals of Statistics*, **33**, 1617-1642. http://dx.doi.org/10.1214/009053605000000200

Appendix

Proof of Theorem 1

Proof. Based on **Lemma 2** and as shown in Hunter and Li [27], one can know that $\left\|\hat{b}_{\lambda_j k}^m\right\| \to \left\|\hat{b}_{\lambda_j k}\right\|$ for each $1 \le k \le p$. Then, as long as $m \to \infty$, one can see that when $k \le p_0$ then $\left\|\hat{b}_{\lambda_j k}^m\right\|$ converge to a positive number, otherwise, $\left\|\hat{b}_{\lambda_j k}^m\right\|$ converge to 0. Denote D_{aa}^m as the upper $p_0 \times p_0$ diagonal submatrix of $n_j^{-1}D^{(m)}$ and D_{bb}^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_{aa}^m must converge to some finite number while D_{bb}^m diverge to infinity in the case where $m \to \infty$.

For simplify, we follow (8) and $\hat{\alpha}_{\lambda_j}(u)$ can be rewritten as $\hat{\alpha}_{\lambda_j}(u) = \left\{\omega(u) + n_j^{-1}D^m\right\}^{-1}\zeta(u)$, where $\omega(u)$ is a 2×2 block matrix given by $\left\{\omega_{aa}(u), \omega_{ab}(u); \omega_{ba}(u), \omega_{bb}(u)\right\}$ and

$\zeta(u) = \left\{\zeta_a^{\mathrm{T}}(u), \zeta_b^{\mathrm{T}}(u)\right\}^{\mathrm{T}} \in R^p$, with $\omega_{aa}(u) = n_j^{-1}\sum_{i=1}^{n_j} \delta_{iaj}X_{iaj}X_{iaj}^{\mathrm{T}}K_h(U_{ij}-u)$,

$\omega_{bb}(u) = n_j^{-1}\sum_{i=1}^{n_j}\delta_{ibj}X_{ibj}X_{ibj}^{\mathrm{T}}K_h(U_{ij}-u)$, $\omega_{ab}(u) = n_j^{-1}\sum_{i=1}^{n_j}\delta_{iaj}X_{iaj}X_{ibj}^{\mathrm{T}}K_h(U_{ij}-u) = \omega_{ba}^{(m)}(u)$,

$\zeta_a^{\mathrm{T}}(u) = n_j^{-1}\sum\delta_{iaj}X_{iaj}Y_{ij}^*K_h(u-U_{ij})$, and $\zeta_b^{\mathrm{T}}(u) = n_j^{-1}\sum\delta_{ibj}X_{ibj}Y_{ij}^*K_h(u-U_{ij})$. If $\left\{\omega(u)+D^{(m)}\right\}^{-1}$ is given by $\left\{\Xi_{aa}^{(m)}, \Xi_{ab}^{(m)}, \Xi_{ba}^{(m)}, \Xi_{bb}^{(m)}\right\}$ one obtains

$$\Xi_{aa}^{(m)}(u) = \left(\omega_{aa}(u) + D_{aa}^{(m)} - \omega_{ab}(u)\left\{\omega_{bb}(u)+D_{bb}^{(m)}\right\}^{-1}\omega_{ba}\right)^{-1},$$

$$\Xi_{ab}^{(m)}(u) = -\left\{\omega_{aa}(u)+D_{aa}^{(m)}\right\}^{-1}\omega_{ab}(u)\times\left(\omega_{aa}(u)+D_{aa}^{(m)}-\omega_{ab}(u)\left\{\omega_{bb}(u)+D_{bb}^{(m)}\right\}^{-1}\omega_{ba}\right)^{-1},$$

$$\Xi_{ba}^{(m)}(u) = -\left\{\omega_{bb}(u)+D_{bb}^{(m)}\right\}^{-1}\omega_{ba}(u)\times\left(\omega_{bb}(u)+D_{bb}^{(m)}-\omega_{ba}(u)\left\{\omega_{aa}(u)+D_{aa}^{(m)}\right\}^{-1}\omega_{ab}\right)^{-1},$$

$$\Xi_{bb}^{(m)}(u) = \left(\omega_{bb}(u)+D_{bb}^{(m)}-\omega_{ba}(u)\left\{\omega_{aa}(u)+D_{aa}^{(m)}\right\}^{-1}\omega_{ab}\right)^{-1}.$$

Due to each diagonal component of D_{aa}^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 $\Xi_{ba}^{(m)}(u)$ and $\Xi_{bb}^{(m)}(u)$ converge to 0 uniformly on [0, 1] as $m \to \infty$. It is easy to see that

$$\hat{\alpha}_{\lambda_j,b}^{m+1} = \Xi_{ba}^{(m)}(u)\zeta_a^{\mathrm{T}}(u) + \Xi_{bb}^{(m)}(u)\zeta_b^{\mathrm{T}}(u),$$

where $\zeta_a^{\mathrm{T}}(u)$, and $\zeta_b^{\mathrm{T}}(u)$ are uniformly bounded. Obviously, $\hat{\alpha}_{\lambda_j,k}^m \to 0$ as $m \to \infty$ when $p_0 < k \le p$. Therefore, $\sup\left\|\hat{\alpha}_{\lambda_j k}(u)\right\| = 0$ for every $p_0 < k \le p$. It completes the proof of Theorem 1.

Proof of Theorem 2

Proof. As is well known, $\hat{\alpha}_{a\lambda_j}(u)$ is the solution of the following equation

$$\frac{1}{n_j}\sum_{i=1}^{n_j}\delta_{iaj}X_{iaj}\left(\delta_{ij}Y_{ij}^* - \delta_{iaj}X_{iaj}^{\mathrm{T}}\hat{\alpha}_{a\lambda_j}\right)K_h(U_{ij}-u) + n_j^{-1}\sum_{k=1}^{p_0}\lambda_{jk}\frac{\hat{b}_{\lambda_j k}}{\left\|\hat{b}_{\lambda_j k}\right\|} = 0.$$

That is to say, $\hat{\alpha}_{a\lambda_j}$ satisfies

$$\hat{\alpha}_{a\lambda_j}(u)=\left\{\hat{\Theta}(u)\right\}^{-1}\times\left\{\frac{1}{n_j}\sum_{i=1}^{n_j}\delta_{iaj}X_{iaj}Y_{ij}^{*}K_h\left(U_{ij}-u\right)+\frac{1}{n_j}\sum_{k=1}^{p_0}\lambda_{jk}\frac{\hat{b}_{\lambda_j k}}{\left\|\hat{b}_{\lambda_j k}\right\|}\right\}=0,$$

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

$$\begin{aligned}&\max\left\|\hat{\alpha}_{a\lambda_j}(u)-\hat{\alpha}_{ora}(u)\right\|\\&=\max\left\|\left\{\hat{\Theta}(u)\right\}^{-1}\left(\frac{1}{n_j}\sum_{k=1}^{p_0}\lambda_{jk}\frac{\hat{b}_{\lambda_j k}}{\left\|\hat{b}_{\lambda_j k}\right\|}\right)\right\|\le\hat{\lambda}_{\max}\left\|\left(\frac{1}{n_j}\sum_{k=1}^{p_0}\lambda_{jk}\frac{\hat{b}_{\lambda_j k}}{\left\|\hat{b}_{\lambda_j k}\right\|}\right)\right\|\\&\le\frac{\hat{\lambda}_{\max}}{n_j}\sum_{k=1}^{p_0}\lambda_{jk}\le\frac{\hat{\lambda}_{\max}p_0a_{n_j}}{n_j}=0_p\left(n_j^{-21/10}\right),\end{aligned}$$

where $\hat{\lambda}_{\max}=\sup_u\lambda_{\max}\left\{f(u)\Theta(u)\right\}$ with $\lambda_{\max}(A)$ represents the maximal eigenvalue of an arbitrary positive definite matrix A. Notice that $n_j^{-21/10}=0\left(n_j^{-2/5}\right)$, as a result it completes the proof of **Theorem 2**. □