Computational Statistics and Data Analysis 138 (2019) 239–259

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

# Subgroup analysis for heterogeneous additive partially linear models and its application to car sales data[☆]

Lili Liu[a], Lu Lin[a,b,*]

[a] Zhongtai Securities Institute for Financial Studies, Shandong University, Jinan, China
[b] School of Statistics, Qufu Normal University, Qufu, China

## HIGHLIGHTS

- The heterogeneous additive components are approximated by polynomial splines.
- ADMM algorithm with concave fusion penalty can automatically identify the subgroups.
- The consistency of classification and estimation is achieved.
- The algorithm is rapid and computationally stable even if the sample size is large.

ARTICLE INFO

*Article history:*
Received 16 September 2018
Received in revised form 1 April 2019
Accepted 20 April 2019
Available online 27 April 2019

*Keywords:*
Subgroup
B-spline
Heterogeneity
Clustering
Estimation consistency

## ABSTRACT

As an extension of additive partially linear model, heterogeneous additive partially linear model contains the homogeneous linear components and subject-dependent additive components, but has no group information of subject-dependent additive components. Such a model is more flexible and efficient for addressing some special issues such as precision medicine and precision marketing. A polynomial spline smoothing is used to approximate the heterogeneous additive components, and then a new clustering method is developed to automatically identify subgroups. The procedure avoids solving coefficient vector in each iterative step as in regression clustering procedures. Thus, this approach is rapid and computationally stable even if the sample size is large. Based on the clustered heterogeneous additive components, consistent estimators of the homogeneous parameters and subgroup-specific additive components are further obtained. Moreover, $\sqrt{n}$-consistency and asymptotic normality for the estimators of the parametric components are established. The simulation studies and real data analysis illustrate that the model and proposed clustering and estimation are effective in practice.

## 1. Introduction

In this paper, we study a flexible heterogeneous additive partially linear model for analyzing data from a heterogeneous population. Recently, many researchers have paid much attention to identifying heterogeneous subgroups. The precision medicine is a common application of subgroup analysis, which seeks to give precise medical treatments to heterogeneous subgroups of patients. Owing to the diversity of patients in genes, environment, age and weight and so on, the individualized treatment for different subgroups can obtain precise medical effect (see Ma and Huang, 2017). Another

[☆] This research was supported by NNSF project (11571204) of China.

[*] Corresponding author.
*E-mail address:* linlu@sdu.edu.cn (L. Lin).

https://doi.org/10.1016/j.csda.2019.04.011
0167-9473/© 2019 Elsevier B.V. All rights reserved.

wide range of real-world applications is precision marketing. Heterogeneity of marketing strategies reflects the diversity of customers' consumption behaviors and preferences. Precision marketing offers personalized customer service and is used to help enterprises increase their profits by identifying the different marketing subgroups (see You et al., 2015). Thus, correctly identifying heterogeneous subgroups to enhance the effects is a significant issue. For extending from classical linear models to a general case, we are interested in the heterogeneous additive partially linear model for subgroup analysis. The heterogeneous additive partially linear model is flexible and widely used, which can combine both parametric and nonparametric components. This model structure allows an easier interpretation of the effect of each variable and avoids the curse of dimensionality. Moreover, our proposed framework is more generic, efficient, and adaptive for incorporating linearity, nonlinearity and heterogeneity.

There has been much work on subgroup analysis in the literature. Shen and He (2015) proposed using the logistic-normal mixture model approach to deal with the heterogeneity. However, this approach requires the specified number of mixture components and an underlying distribution assumption for the model. Wang and Xia (2014) partitioned the sample space of piecewise single-index models into several regions with the knowledge of a priori classification. In addition, some penalization methods have been well developed. For example, Tibshirani et al. (2005) proposed the fused lasso method which applies the pairwise $L_1$ penalty to the regression model with ordered coefficients for clustering. The group LASSO and the fused concave penalty methods have been applied to group the effects of covariates (Guo et al., 2010; Ke et al., 2013; Shen and Huang, 2010). Our work concerns identifying subgroups of the observations, which is different from these penalty methods about studying on grouping effects of covariates.

We consider the following additive partially linear model with both homogeneous and heterogeneous components as

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + g_i(\mathbf{z}_i) + \epsilon_i, \quad i = 1, \dots, n, \tag{1.1}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)^T$ is the homogeneous coefficient vector, $g_i(\mathbf{z}_i) = g_{i0} + \sum_{j=1}^p g_{ij}(z_{ij})$ with $g_{i0} \in R$ being heterogeneous intercepts and $g_{ij} \in R$ $(i, j \geq 1)$ being unknown smooth functions, $y_i$ is the observation of $\mathbf{y} \in R$, $(\mathbf{x}_i, \mathbf{z}_i)$ is independent and identically distributed observation of $(\mathbf{x}, \mathbf{z}) \in R^q \times R^p$. The random error $\epsilon_i$ is independent of $(\mathbf{x}_i, \mathbf{z}_i)$ with $E[\epsilon_i|\mathbf{x}_i, \mathbf{z}_i] = 0$ and $\text{Var}[\epsilon_i|\mathbf{x}_i, \mathbf{z}_i] = \sigma^2$. It is assumed that $E[g_{ij}(z_{ij})] = 0$ $(i, j \geq 1)$ for identification purpose. Without loss of generality, we assume that all the covariates have mean zero. Our goal is to identify the subgroups of $g_{ij}$ for $j = 0, 1, \dots, p$, such that $g_{ij}$ is the same function in each subgroup, and then we further estimate the subgroup-specific additive functions $g_i$ and the homogeneous parameter $\boldsymbol{\beta}$.

In this paper, we use a linear combination of B-spline basis functions to approximate the nonparametric function $g_{ij}$, which is efficient and computationally simple. In the existing literature, the B-spline method has been applied to the function approximation of nonparametric and semiparametric models. For example, Sherwood and Wang (2016) approximated link functions in the single-index model by B-spline and obtained the consistent M-estimators of parametric and nonparametric components; He and Shi (1994) employed the B-spline method to estimate the nonparametric condition quantile function; B-spline approach of various models was considered by Huang et al. (2002), He et al. (2002), Elmi et al. (2011) and among others.

A challenging problem is how to identify the subgroups when the number $K$ of subgroups is unknown in advance. For linear model with unknown $K$, Ma and Huang (2017) used the concave pairwise fusion penalty approach to identify heterogeneous subgroups. When $n$ and $p$ are large, however, this method is complicated and unstable since their implementation requires iteratively storing and manipulating the entire $np$-dimensional parameters. The memory and computational cost can be extremely high. Chi and Lange (2014) studied an alternating direction method of multipliers (ADMM, Boyd et al., 2011) with the weighted $L_1$ penalty for the convex clustering. However, the $L_1$ penalty can generate biases of the estimates, and the choice of the weights can affect the quality of the clustering solution, and there is no clear rule for choosing the valid weights. Pan et al. (2013) proposed penalized regression-based clustering using the non-convex grouped truncated lasso penalty (gTLP). The non-convex gTLP performs much better than the Lasso and other $L_q$-norms, since the parameter estimation bias is largely avoided. However, it will be seen later that this algorithm is relatively complicated than ours.

In contrast to the above methods, we convert the optimization of slope coefficients to an optimization of intercept coefficients. Consequently, we can use classical clustering technique to deal with our problem. Our proposals avoid computing $np$ unknown parameters simultaneously, making the estimation and classification procedures feasible. Moreover, our methods apply to the heterogeneous additive partially linear models when each nonparametric component is approximated by a linear combination of B-spline basis functions. Then we propose an ADMM algorithm for a modified $k$-means clustering with concave penalties to automatically identify the subgroups. Afterwards, we estimate the additive components in each subgroup. Under some mild regularity conditions, we derive the convergence and the oracle property. Therefore, our results provide new insight into the heterogeneous subgroup identification for complex models and could help clinicians to tailor the treatment according to the individual characteristic of patients, thereby achieving better treatment outcomes.

The rest of the paper is organized as follows. In Section 2, we introduce the heterogeneous additive partially linear model and describe the subgroup identification procedure and the corresponding clustering algorithms in detail. Section 3 introduces the theoretical properties of the proposed procedure. Section 4 assesses the performance of our method via the Monte Carlo simulation studies. In Section 5 we discuss the application of our model to car sales data. All detailed proofs are presented in Appendix.

## 2. Methodology

### 2.1. Spline approximation

Suppose without loss of generality that each $z_j$ takes value in the closed interval $[0, 1]$. Let $0 = t_0 < t_1 < \cdots < t_{J_n} < 1$ be a partition of the interval $[0, 1]$ into subintervals $I_j = [t_j, t_{j+1}), j = 0, \ldots, J_n - 1, I_{J_n} = [t_{J_n}, 1]$, where $J_n$ increases with sample size $n$. Let $S_n$ be the space of polynomial splines of degree $l \geq 1$ consisting of functions $s$ satisfying: (i) $s$ is a polynomial of degree $l$ on each of the subintervals, and (ii) for $l \geq 2$, $s$ is $l-1$ continuously differentiable on $[0, 1]$. Denote by $\mathbf{b}(t) = (b_1(t), \ldots, b_{J_n+l}(t))^T$, a vector of normalized B-spline basis functions of degree $l$ with $J_n$ internal knots on [0,1]. The construction of B-spline basis functions can be found in Schumaker (1981) for details. Thus, for any $g_{ij}(\cdot) \in S_n$, we can write $g_{ij}(z) = \mathbf{b}(z)^T \boldsymbol{\gamma}_{ij}$, where $\boldsymbol{\gamma}_{ij} \in R^{N_n}$ with $N_n = J_n + l, j = 1, \ldots, p$. The linear combination of B-spline basis functions in $\mathbf{B}(\mathbf{z}_i) = (1, \mathbf{b}(z_{i1})^T, \ldots, \mathbf{b}(z_{ip})^T)^T$ can approximate effectively unknown nonparametric functions $g_i(\cdot)$ under the smoothness assumption (Stone, 1985). Thus, the model (1.1) can be expressed approximately by the following form

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{B}(\mathbf{z}_i)^T \boldsymbol{\gamma}_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{\gamma}_i = (\gamma_{i0}, \boldsymbol{\gamma}_{i1}^T, \ldots, \boldsymbol{\gamma}_{ip}^T)^T \in R^{N_n p+1}$ is the individual-specific spline coefficient vector. For model (2.1), we will propose a modified $k$-means clustering for subgroup identification.

Denote by $\|\mathbf{a}\| = (\sum_i |a_i|^2)^{1/2}$ the $L_2$ norm of a vector $\mathbf{a}$. We need the condition that the number of subgroups is much smaller than the sample size. Consider the following criterion:

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \mathbf{B}(\mathbf{z}_i)^T \boldsymbol{\gamma}_i)^2 + \sum_{1 \leq i < k \leq n} p(\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_k\|, \lambda), \tag{2.2}$$

where $p(t, \lambda)$ is a given penalty function with the penalty parameter $\lambda$. We then suggest using the concave penalties including the SCAD penalty (Fan and Li, 2001) and the MCP penalty (Zhang, 2010), by which we can correctly identify the number of subgroups and consistently estimate the parameters. Particularly, the SCAD penalty is defined as

$$p_\vartheta(t, \lambda) = \lambda \int_0^{|t|} \min\{1, (\vartheta - x/\lambda)_+/(\vartheta - 1)\}\, dx,$$

and the MCP penalty is expressed as

$$p_\vartheta(t, \lambda) = \lambda \int_0^{|t|} (-x/(\vartheta\lambda))_+ dx,$$

where $(x)_+ = x$ if $x > 0$ and $= 0$ otherwise, and $\vartheta$ is a parameter that controls the concavity of the penalty function.

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathbf{B}(\mathbf{z}) = (\mathbf{B}(\mathbf{z}_1), \ldots, \mathbf{B}(\mathbf{z}_n))^T$, and $\mathbf{B}_\gamma = (\mathbf{B}(\mathbf{z}_1)^T \boldsymbol{\gamma}_1, \ldots, \mathbf{B}(\mathbf{z}_n)^T \boldsymbol{\gamma}_n)^T$. Denote $\mathbf{Q}_x = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix. Write $\widetilde{\mathbf{y}} = \mathbf{Q}_x\mathbf{y} \in R^{n\times 1}$ and $\widetilde{\mathbf{B}}_\mathbf{z} = \mathbf{Q}_x\mathbf{B}(\mathbf{z}) \in R^{n\times(N_n p+1)}$. In model (2.1), for given $\boldsymbol{\gamma}$, we obtain the estimator of $\boldsymbol{\beta}$ as $\widehat{\boldsymbol{\beta}}_\gamma = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{B}_\gamma)$. When $\boldsymbol{\beta}$ is replaced by its estimator $\widehat{\boldsymbol{\beta}}_\gamma$, with some algebra, the penalized objective function in (2.2) is given by

$$L_1(\boldsymbol{\gamma}, \lambda) = \frac{1}{2}\sum_{i=1}^{n}(\widetilde{y}_i - \widetilde{\mathbf{B}}_{\mathbf{z}i}^T \boldsymbol{\gamma}_i)^2 + \sum_{1 \leq i < k \leq n} p_\vartheta(\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_k\|, \lambda). \tag{2.3}$$

The above approximate objective function will be further simplified in the following subsection.

### 2.2. Subgroup identification procedure

By the argument used above, we can see that grouping $g_i$ is equivalent to grouping $\boldsymbol{\gamma}_i$. Here we assume $K$ different subgroups $\mathcal{G}_1, \ldots, \mathcal{G}_K$ are mutually exclusive partitions of $\{1, \ldots, n\}$, satisfying $\mathcal{G} = (\mathcal{G}_1, \ldots, \mathcal{G}_K)$. Suppose $\boldsymbol{\gamma}_i = \boldsymbol{\alpha}_k$ and $g_{ij} = f_{kj}$ for all $i \in \mathcal{G}_k$, where $\boldsymbol{\alpha}_k$ and $f_{kj}$ are respectively the common value and common function in the group $\mathcal{G}_k$. However, the number $K$ of the subgroups and the subgroup $\mathcal{G}_k$ are unknown in advance. Note that $E[\widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i] = \boldsymbol{\theta}_i \in R^{N_n p+1}$, with $\boldsymbol{\theta}_i = E[\widetilde{\mathbf{B}}_\mathbf{z}\widetilde{\mathbf{B}}_\mathbf{z}^T]\boldsymbol{\gamma}_i$. Therefore, grouping $\boldsymbol{\gamma}_i$ is equivalent to grouping $\boldsymbol{\theta}_i$. The criterion (2.3) is then equivalent to the following clustering criterion:

$$\frac{1}{2}\sum_{i=1}^{n} \left\|\widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i - \boldsymbol{\theta}_i\right\|_2^2 + \sum_{i<k} p_\vartheta(\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_k\|, \lambda). \tag{2.4}$$

The equivalence between the criterions (2.3) and (2.4) is proved in Appendix.

We in the following will suggest an ADMM algorithm to identify the subgroups, and estimate the common value of $\boldsymbol{\theta}_i$'s for $i \in \mathcal{G}_k$ in objective function (2.4). Our proposed algorithm is simple and stable as it only computes $(N_n p+1)$-dimensional parameters in each iterative step. Since the penalty function is not separable in $\boldsymbol{\theta}_i$'s, a new parameter $\boldsymbol{\delta}_{ik} = \boldsymbol{\theta}_i - \boldsymbol{\theta}_k$ is introduced. Thus, the minimization problem in (2.4) becomes the constraint optimization problem as

$$L_0(\boldsymbol{\theta}, \boldsymbol{\delta}) = \frac{1}{2}\sum_{i=1}^{n} \left\|\widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i - \boldsymbol{\theta}_i\right\|_2^2 + \sum_{i<k} p_{\vartheta}(\|\boldsymbol{\delta}_{ik}\|, \lambda) \text{ subject to } \boldsymbol{\theta}_i - \boldsymbol{\theta}_k - \boldsymbol{\delta}_{ik} = 0,$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_n^T)^T$ and $\boldsymbol{\delta} = \left\{\boldsymbol{\delta}_{ik}^T, i < k\right\}^T$. The estimators of the parameters are yielded by the augmented Lagrangian

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{\upsilon}) = L_0(\boldsymbol{\theta}, \boldsymbol{\delta}) + \sum_{i<k}\langle \boldsymbol{\upsilon}_{ik}, \boldsymbol{\delta}_{ik} - \boldsymbol{\theta}_i + \boldsymbol{\theta}_k\rangle + \frac{\eta}{2}\sum_{i<k}\|\boldsymbol{\delta}_{ik} - \boldsymbol{\theta}_i + \boldsymbol{\theta}_k\|^2,$$

where $\boldsymbol{\upsilon} = \left\{\boldsymbol{\upsilon}_{ik}^T, i < k\right\}^T$ are Lagrange multipliers, $\eta$ is the penalty parameter, $\langle \mathbf{a}, \mathbf{b}\rangle = \mathbf{a}^T\mathbf{b}$ is the inner product of two same dimensional vectors $\mathbf{a}$ and $\mathbf{b}$. We use the ADMM to iteratively compute the estimators of $(\boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{\upsilon})$. For given $\boldsymbol{\delta}^m$, $\boldsymbol{\upsilon}^m$ at step $m$, we yield the following algorithm

$$\boldsymbol{\theta}^{m+1} = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{\delta}^m, \boldsymbol{\upsilon}^m), \tag{2.5}$$

$$\boldsymbol{\delta}^{m+1} = \arg\min_{\boldsymbol{\delta}} L(\boldsymbol{\theta}^{m+1}, \boldsymbol{\delta}, \boldsymbol{\upsilon}^m), \tag{2.6}$$

$$\boldsymbol{\upsilon}_{ik}^{m+1} = \boldsymbol{\upsilon}_{ik}^m + \eta(\boldsymbol{\delta}_{ik}^{m+1} - \boldsymbol{\theta}_i^{m+1} + \boldsymbol{\theta}_k^{m+1}). \tag{2.7}$$

To update $\boldsymbol{\theta}$, the minimization problem (2.5) is equivalent to minimizing

$$F(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}\left\|\widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i - \boldsymbol{\theta}_i\right\|_2^2 + \frac{\eta}{2}\sum_{i<k}\|\boldsymbol{\delta}_{ik}^m - \boldsymbol{\theta}_i + \boldsymbol{\theta}_k + \eta^{-1}\boldsymbol{\upsilon}_{ik}^m\|^2 + C,$$

where $C$ is the constant independent of $\boldsymbol{\theta}$.

For given $\boldsymbol{\delta}^m$, $\boldsymbol{\upsilon}^m$, we set the derivative $\partial F(\boldsymbol{\theta})/\partial\boldsymbol{\theta}_i = 0$ to derive the update of $\boldsymbol{\theta}_i$. The update $\boldsymbol{\theta}_i^{m+1}$ is computed analytically as

$$\boldsymbol{\theta}_i^{m+1} = \frac{1}{1+n\eta}\mathbf{A}_i + \frac{\eta}{1+n\eta}\sum_{i=1}^{n}\widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i, \tag{2.8}$$

where

$$\mathbf{A}_i = \widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i + \sum_{i<k}(\eta\boldsymbol{\delta}_{ik}^m + \boldsymbol{\upsilon}_{ik}^m) - \sum_{k<i}(\eta\boldsymbol{\delta}_{ki}^m + \boldsymbol{\upsilon}_{ki}^m). \tag{2.9}$$

To update $\boldsymbol{\delta}$, we need to minimize the function

$$\frac{\eta}{2}\|\boldsymbol{\delta}_{ik} - \boldsymbol{\pi}_{ik}^m\|^2 + \sum_{i<k} p_{\vartheta}(\|\boldsymbol{\delta}_{ik}\|, \lambda).$$

Thus for the MCP penalty with $\vartheta > 1/\eta$, the update $\boldsymbol{\delta}_{ik}^{m+1}$ is

$$\boldsymbol{\delta}_{ik}^{m+1} = \begin{cases} \dfrac{S(\boldsymbol{\pi}_{ik}^m, \lambda/\eta)}{1 - 1/(\vartheta\eta)} & \|\boldsymbol{\pi}_{ik}^m\| \le \lambda\vartheta, \\ \boldsymbol{\pi}_{ik}^m & \|\boldsymbol{\pi}_{ik}^m\| > \lambda\vartheta. \end{cases} \tag{2.10}$$

For the SCAD penalty with $\vartheta > 1/\eta + 1$, the solution is

$$\boldsymbol{\delta}_{ik}^{m+1} = \begin{cases} S(\boldsymbol{\pi}_{ik}^m, \lambda/\eta) & \|\boldsymbol{\pi}_{ik}^m\| \le \lambda + \lambda/\eta, \\ \dfrac{S(\boldsymbol{\pi}_{ik}^m, \vartheta\lambda/((\vartheta-1)\eta))}{1 - 1/((\vartheta-1)\eta)} & \lambda + \lambda/\eta < \|\boldsymbol{\pi}_{ik}^m\| \le \lambda\vartheta, \\ \boldsymbol{\pi}_{ik}^m & \|\boldsymbol{\pi}_{ik}^m\| > \lambda\vartheta, \end{cases} \tag{2.11}$$

where $\boldsymbol{\pi}_{ik}^m = \boldsymbol{\theta}_i^m - \boldsymbol{\theta}_k^m - \eta^{-1}\boldsymbol{\upsilon}_{ik}^m$, and $S(\mathbf{x}, t) = (1 - t/\|\mathbf{x}\|)_+\mathbf{x}$ is groupwise soft thresholding operator. Finally, the Lagrange multiplier $\boldsymbol{\upsilon}_{ik}$ is updated by (2.7). It is worth noting that the observations $i$ and $k$ are classified into the same group if $\widehat{\boldsymbol{\delta}}_{ik} = 0$. After the subgroup $\mathcal{G}_k$ is identified, we obtain the estimated number of subgroups $\widehat{K}$, and the estimated subgroups $\widehat{\mathcal{G}}_1, \ldots, \widehat{\mathcal{G}}_{\widehat{K}}$.

Thus, the estimators of $\boldsymbol{\beta}$ and $(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{\widehat{K}})$ are

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}_1, \ldots, \widehat{\boldsymbol{\alpha}}_{\widehat{K}}) = \arg\min_{\boldsymbol{\beta}, \boldsymbol{\alpha}_k} \frac{1}{2}\sum_{k=1}^{\widehat{K}}\sum_{i\in\widehat{\mathcal{G}}_k}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{B}(\mathbf{z}_i)^T\boldsymbol{\alpha}_k\right)^2. \tag{2.12}$$

The matrix $(\widetilde{\mathbf{B}}_{\mathbf{z}}\widetilde{\mathbf{B}}_{\mathbf{z}}^T)$ is nonsingular usually (see Li, 2000), then we can also estimate $\boldsymbol{\gamma}_i$ after getting $\boldsymbol{\theta}_i$. According to $\widehat{\boldsymbol{\theta}}_i = E[\widetilde{\mathbf{B}}_{\mathbf{z}}\widetilde{\mathbf{B}}_{\mathbf{z}}^T]\widehat{\boldsymbol{\gamma}}_i$, the estimator of $\boldsymbol{\gamma}_i$ is $\widehat{\boldsymbol{\gamma}}_i = (1/n\sum_{i=1}^n \widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{\mathbf{B}}_{\mathbf{z}i}^T)^{-1}\widehat{\boldsymbol{\theta}}_i$. Then we can plug $\widehat{\boldsymbol{\gamma}}_i$ back into $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y}-\mathbf{B}_{\widehat{\gamma}})$ to get the estimator of $\boldsymbol{\beta}$, where $\mathbf{B}_{\widehat{\gamma}} = (\mathbf{B}(\mathbf{z}_1)^T\widehat{\boldsymbol{\gamma}}_1, \ldots, \mathbf{B}(\mathbf{z}_n)^T\widehat{\boldsymbol{\gamma}}_n)^T$. The estimator of $\boldsymbol{\alpha}_k$ for the $k$th group is $\widehat{\boldsymbol{\alpha}}_k = |\widehat{\mathcal{G}}_k|^{-1}\sum_{i\in\widehat{\mathcal{G}}_k}\widehat{\boldsymbol{\gamma}}_i$, where $|\widehat{\mathcal{G}}_k|$ is the cardinality of $\widehat{\mathcal{G}}_k$.

Then by the identifiability condition $E[g_{ij}(z_{ij})] = 0$, the centered spline estimator of nonparametric function $g_{ij}$ for group $\widehat{\mathcal{G}}_k$ is

$$\widehat{f}_{kj}(z_{ij}) = \mathbf{B}(z_{ij})^T\widehat{\boldsymbol{\alpha}}_{kj} - \frac{1}{|\widehat{\mathcal{G}}_k|}\sum_{i\in\widehat{\mathcal{G}}_k}\mathbf{B}(z_{ij})^T\widehat{\boldsymbol{\alpha}}_{kj}, \tag{2.13}$$

for $k = 1\cdots, \widehat{K}, j = 1\cdots, p$. The estimator of heterogeneous intercept $g_{i0}$ for group $\widehat{\mathcal{G}}_k$ is $\widehat{f}_{k0} = \widehat{\alpha}_{k0} + |\widehat{\mathcal{G}}_k|^{-1}\sum_{i\in\widehat{\mathcal{G}}_k}\sum_{j=1}^p \mathbf{B}(z_{ij})^T\widehat{\boldsymbol{\alpha}}_{kj}$. Finally, the estimator of $g_i(\mathbf{z}_i)$ for $i \in \widehat{\mathcal{G}}_k$ is $\widehat{f}_k(\mathbf{z}_i) = \widehat{f}_{k0} + \sum_{j=1}^p \widehat{f}_{kj}(z_{ij})$.

### 2.3. Algorithm

Based on the above analysis, the algorithm consists of the following steps:

**Algorithm 1** ADMM for the clustering

```
Require: Initialize $\boldsymbol{\delta}^0$ and $\boldsymbol{v}^0$
for $m = 1, 2, \cdots$ do
    for $i = 1, \cdots, n$ do
        Compute $\mathbf{A}_i$ using (2.9)
        Compute $\boldsymbol{\theta}_i^{m+1}$ using (2.8)
    end for
    for $i, k = 1, \cdots, n$ and $i < k$ do
        Compute $\boldsymbol{\delta}_{ik}^{m+1}$ using (2.10) or (2.11)
        Compute $\boldsymbol{v}_{ik}^{m+1}$ using (2.7)
    end for
    if convergence criterion is met then
        Stop and denote the last iteration by $\widehat{\boldsymbol{\theta}}$
    else
        m=m+1
    end if
end for
After identifying $\mathcal{G}_k$, estimate $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}_k)$ using (2.12) with $k = 1, \cdots, \widehat{K}$
```

## 3. Theoretical properties

In this section, we establish the theoretical properties for the proposed procedure. First, we briefly review some definitions and results that are related to function analysis. Let $m$ be a positive integer and $\nu \in (0, 1]$, $\kappa = m + \nu > 1.5$. Define $\mathcal{H}$ as the collection of functions $g(\cdot)$ on [0,1] whose $m$th derivative $g^{(m)}(\cdot)$ satisfies the Hölder condition of order $\nu$: $|g^{(m)}(z') - g^{(m)}(z)| \le c|z' - z|^\nu$, $0 \le z, z' \le 1$, where $c$ is some positive constant. By the result of de Boor (2001, p.149), for any $g_i(\cdot) \in \mathcal{H}$, there exists individual-specific spline coefficient vector $\boldsymbol{\gamma}_i = (\gamma_{i0}, \boldsymbol{\gamma}_{i1}^T, \ldots, \boldsymbol{\gamma}_{ip}^T)^T \in R^{N_n p+1}$, such that $\|\mathbf{B}(\mathbf{z}_i)^T\boldsymbol{\gamma}_i - g_i(\mathbf{z}_i)\|_\infty = O(J_n^{-\kappa})$. $\|\phi\|_\infty = \sup_m |\phi(m)|$ is the supremum norm of a function $\phi$ on [0,1]. Let $|\mathcal{G}_{min}| = \min_{1\le k\le K}|\mathcal{G}_k|$ and $|\mathcal{G}_{max}| = \max_{1\le k\le K}|\mathcal{G}_k|$ be the minimum and maximum group sizes respectively. For any vector $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_p)^T$, denote the supremum norm by $\|\boldsymbol{\zeta}\|_\infty = \max_{1\le j\le p}|\zeta_j|$, and for any matrix $\mathbf{A} = (A_{ij})_{i=1,j=1}^{s,t}$, denote $\|\mathbf{A}\|_\infty = \max_{1\le i\le s}\sum_{j=1}^t |A_{ij}|$. For any measurable function $\phi$ on $[0, 1]^p$, define the norm $\|\phi\|_n^2 = \frac{1}{n}\sum_{i=1}^n \phi^2(\boldsymbol{z}_i)$. To avoid confusion, let $\boldsymbol{\beta}^0$ be the true parameter value and $f_k^0$ be the true additive function for group $\mathcal{G}_k$, $k = 1, \ldots, K$. Additionally, let $\boldsymbol{\gamma}_i^0$ be the true spline coefficient. Write $\boldsymbol{\theta}_i^0 = E[\widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i] = E[\widetilde{\mathbf{B}}_{\mathbf{z}}\widetilde{\mathbf{B}}_{\mathbf{z}}^T]\boldsymbol{\gamma}_i^0$.

Now we consider the heterogeneous additive partially linear regression with the oracle information that the true group structure $(\mathcal{G}_1, \ldots, \mathcal{G}_K)$ is known, the oracle estimation of $\boldsymbol{\theta}_i$, $i = 1, \ldots, n$ will be

$$(\widehat{\boldsymbol{\theta}}_1^{or}, \ldots, \widehat{\boldsymbol{\theta}}_n^{or}) = \underset{\boldsymbol{\theta}_i, i=1,\ldots,n}{\arg\min}\ \frac{1}{2}\sum_{k=1}^K\sum_{i\in\mathcal{G}_k}\left\|\widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i - \boldsymbol{\theta}_i\right\|_2^2,$$

and correspondingly, the common values of $\widehat{\boldsymbol{\theta}}_i^{or}$ for each subgroup are given by

$$(\widehat{\boldsymbol{\xi}}_1^{or}, \ldots, \widehat{\boldsymbol{\xi}}_K^{or}) = \mathop{\arg\min}_{\boldsymbol{\xi}_k, k=1,\ldots,K} \frac{1}{2} \sum_{k=1}^{K} \sum_{i \in \mathcal{G}_k} \left\| \widetilde{\mathbf{B}}_{\mathbf{z}i} \widetilde{y}_i - \boldsymbol{\xi}_k \right\|_2^2 .$$

The oracle estimators $\widehat{\boldsymbol{\beta}}^{or}$ and $\widehat{\boldsymbol{\alpha}}_k^{or}$, $k = 1, \ldots, K$ are obtained by minimizing (2.12) with the known true group structure. Hence, the oracle estimator $\widehat{f}_{kj}^{or}(z_{ij})$ is given by (2.13) with $\widehat{\boldsymbol{\alpha}}_{kj}^{or}$ instead of $\widehat{\boldsymbol{\alpha}}_{kj}$, $j = 1, \ldots, p$. Accordingly, we obtain the oracle estimator of the nonparametric function as $\widehat{f}_k^{or}(\mathbf{z}_i) = \widehat{f}_{k0}^{or} + \sum_{j=1}^{p} \widehat{f}_{kj}^{or}(z_{ij})$ for $i \in \mathcal{G}_k$ and $k = 1, \ldots, K$.

The following conditions are introduced for analyzing the asymptotic behavior of $\widehat{\boldsymbol{\beta}}$ and $\widehat{g}_i$.

(C1) Each nonparametric function $g_{ij} \in \mathcal{H}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. The number of internal knots $J_n$ satisfies $|\mathcal{G}_{max}|^{1/(2\kappa)} \leq J_n \leq |\mathcal{G}_{min}|^{1/3}$

(C2) The random vector $\mathbf{x}$ satisfies $c_1\|\boldsymbol{\omega}\|^2 \leq \boldsymbol{\omega}^T E(\mathbf{x}\mathbf{x}^T|\mathbf{z})\boldsymbol{\omega} \leq c_2\|\boldsymbol{\omega}\|^2$ for any vector $\boldsymbol{\omega} \in R^p$, where $c_1$ and $c_2$ are some positive constants.

(C3) The distribution of $\mathbf{z}$ is absolutely continuous and its density $f$ is bounded away from zero and infinity on $[0, 1]^p$. There exist positive constants $M_1$ and $M_2$ such that $\sup_{il} |\widetilde{\mathbf{B}}_{\mathbf{z}_{il}}| \leq M_1$, for $1 \leq i \leq n$, $1 \leq l \leq N_n p$ and $\sup_{ij} |\mathbf{X}_{ij}| \leq M_2$, for $1 \leq i \leq n$, $1 \leq j \leq p$.

(C4) The projection function $\Gamma_k(\mathbf{z}) = E(\mathbf{x}|\mathbf{z} = z)$ has the additive form $\Gamma_k(\mathbf{z}) = \Gamma_{k1}(z_1) + \cdots + \Gamma_{kp}(z_p)$, where $\Gamma_{kj} \in \mathcal{H}$, $E[\Gamma_{kj}(z_j)] = 0$ and $E[\Gamma_{kj}(z_j)]^2 < \infty$, for $k = 1, \ldots, K, j = 1, \ldots, p$.

(C5) The function $\rho_\gamma(t) = \lambda^{-1}p_\lambda(t, \lambda)$ is a symmetric, non-decreasing and concave in $t$ for $t \in [0, \infty)$. It is constant for $t \geq a\lambda$ with some constant $a > 0$, and $\rho(0) = 0$. $\rho'(t)$ exists and is continuous except for a finite number values of $t$ and $\rho(0+) = 1$.

(C6) The noise vector $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ has sub-Gaussian tails such that $P(|\boldsymbol{a}^T\boldsymbol{\epsilon}| > \|\boldsymbol{a}\|x) \leq 2\exp(-c_3x^2)$ for any vector $\boldsymbol{a} \in R^n$ and $x > 0$, where $c_3$ is a positive constant.

The conditions above are analogous in spirit to those in Wang et al. (2011), which are commonly assumed in additive partially linear models with the additive functions approximated by B-spline basis functions. The first assumption in Conditions (C1) is a smoothness assumption, which is typical for a B-spline approximation to functions satisfying the Hölder's condition, (see Stone, 1986). The second assumption in Condition (C1) provides the rate of growth of the dimension of the spline spaces relative to the subgroup size. Moreover, it implies the sparsity of the index set $\mathcal{G}$. Condition (C2) implies that the eigenvalues of $E(\mathbf{x}\mathbf{x}^T|\mathbf{z})$ are bounded away from 0 and infinity. Condition (C3) is a boundedness condition on the covariates, which is common assumptions in asymptotic analysis of nonparametric problems, (see Liu et al., 2011; Stone, 1985). Condition (C4) is required for the convergence rate of the nonparametric parts. Conditions (C5) and (C6) are commonly used in penalized method for high dimensional settings. The concave penalties such as MCP and SCAD satisfy Condition (C5) (see Ma and Huang, 2017).

**Theorem 1.** *Assume the conditions C1–C6 hold, then the oracle estimator $\widehat{\boldsymbol{\beta}}^{or}$ has the following asymptotic properties:*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0) \to N(0, \Phi^{-1}\Sigma\Phi^{-1}),$$

$$m_k^{-1} \sum_{i \in \mathcal{G}_k} (\widehat{f}_k^{or}(\mathbf{z}_i) - f_k^0(\mathbf{z}_i))^2 = O_p\left\{J_n/m_k\right\}, \ k = 1, \ldots, K,$$

*where $\Phi = E[\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}^T]$, $\Sigma = E[\epsilon^2\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}^T]$, $\widetilde{\mathbf{x}} = \mathbf{x} - E(\mathbf{x}|\mathbf{z} = \mathbf{z})$ and $m_k = |\mathcal{G}_k|$ is the cardinality of $\mathcal{G}_k$.*

The theorem shows that both the oracle estimators of the homogeneous parameter $\boldsymbol{\beta}$ and nonparametric component $f_k$ are consistent. It is worth noting that the oracle estimator $\widehat{\boldsymbol{\beta}}^{or}$ has the classical $\sqrt{n}$-convergence rate and is normally distributed asymptotically, but the convergence rate of the oracle estimator $\widehat{f}_k^{or}$ is slower than $\sqrt{n}$, it is because the estimator $\widehat{f}_k^{or}$ is obtained only by the data with indices in $\mathcal{G}_k$.

**Theorem 2.** *Assume the conditions C1–C6 hold, we have that with probability at least $1 - 2KN_npn^{-c_1}$, the following holds:*

$$\|((\widehat{\boldsymbol{\xi}}_1^{or} - \boldsymbol{\xi}_1^0)^T, \ldots, (\widehat{\boldsymbol{\xi}}_K^{or} - \boldsymbol{\xi}_K^0)^T)^T\|_\infty \leq \psi_n,$$

*where $\psi_n = c|\mathcal{G}_{min}|^{-1}\sqrt{n\log n}$, and $c$, $c_1$ are some positive constants, $\boldsymbol{\xi}_k^0$ is the true common value of $\boldsymbol{\theta}_i^0$'s from group $\mathcal{G}_k$. Consequently, the oracle estimator $\widehat{\boldsymbol{\theta}}_i^{or}$ satisfies*

$$\|((\widehat{\boldsymbol{\theta}}_1^{or} - \boldsymbol{\theta}_1^0)^T, \ldots, (\widehat{\boldsymbol{\theta}}_n^{or} - \boldsymbol{\theta}_n^0)^T)^T\|_\infty \leq \psi_n,$$

*and $\sup_i \|\widehat{\boldsymbol{\theta}}_i^{or} - \boldsymbol{\theta}_i^0\| \leq \phi_n$, where $\phi_n = \sqrt{N_np}\psi_n$.*

Since $|\mathcal{G}_{min}| \leq n/K$ and Condition (C1) $J_n \leq |\mathcal{G}_{min}|^{1/3}$, by letting $|\mathcal{G}_{min}| = \kappa n/K$ and $J_n = \delta|\mathcal{G}_{min}|^{1/3}$ for some constant $0 < \kappa, \delta \leq 1$, the bound is $\psi_n = c\kappa^{-1}K\sqrt{\log n/n}$, and $\phi_n = c(K/\kappa)^{5/6}\sqrt{p\delta}\sqrt{\log n/(n^{2/3})}$. Moreover, if $K$ and $p$ are fixed quantities and $0 < c < \infty$, then $\psi_n = C_1\sqrt{\log n/n}$ and $\phi_n = C_2\sqrt{\log n/(n^{2/3})}$ for some constant $0 < C_1, C_2 < \infty$. It is seen from the above result that the bound $\psi_n \to 0$ and $\phi_n \to 0$ as $n \to \infty$, meaning that the oracle estimator $\widehat{\boldsymbol{\theta}}_i^{or}$

is consistent with probability tending to one. In addition, let $a_n = \min_{i\in\mathcal{G}_s,k\in\mathcal{G}_t,s\neq t}\|\boldsymbol{\theta}_i^0-\boldsymbol{\theta}_k^0\| = \min_{s\neq t}\|\boldsymbol{\xi}_s^0-\boldsymbol{\xi}_t^0\|$ be the minimal difference between any two subgroups. We have the following theorem.

**Theorem 3.** *Under the conditions C1–C6, if $a_n > C\lambda$ and $\lambda \gg \phi_n$ for some positive constant $C$, then the estimators $\widehat{\boldsymbol{\theta}}_1,\ldots,\widehat{\boldsymbol{\theta}}_n$ obtained by minimizing the objective function (2.4) satisfy*

$$P(\widehat{\boldsymbol{\theta}}_1=\widehat{\boldsymbol{\theta}}_1^{or},\ldots,\widehat{\boldsymbol{\theta}}_n=\widehat{\boldsymbol{\theta}}_n^{or})\to 1 \ \text{ as } n\to\infty.$$

Let $\widehat{\boldsymbol{\xi}}_k$ be the estimated common value such that $\widehat{\boldsymbol{\theta}}_i=\widehat{\boldsymbol{\xi}}_k$ for all $i\in\mathcal{G}_k$. Theorem 3 indicates the oracle property, in other words, the oracle least squares estimator with knowing the true group structure is the local minimizer of the objective function (2.4) with probability close to one. Thus, the true groups can be recovered with the estimated common value for the group $\mathcal{G}_k$ given as $\widehat{\boldsymbol{\xi}}_k=\widehat{\boldsymbol{\theta}}_i^{or}$ for $i\in\mathcal{G}_k$, that is, $P(\widehat{\boldsymbol{\xi}}_1=\widehat{\boldsymbol{\xi}}_1^{or},\ldots,\widehat{\boldsymbol{\xi}}_K=\widehat{\boldsymbol{\xi}}_K^{or})\to 1$ and $P(\widehat{K}=K)\to 1$. This property together with Theorem 2 implies that our estimators $(\widehat{\boldsymbol{\theta}}_1^T,\ldots,\widehat{\boldsymbol{\theta}}_n^T)^T$ have the consistency of classification that can correctly identify the true subgroups of a heterogeneous population. Based on the above results and the relationship between $\widehat{\boldsymbol{\theta}}_1,\ldots,\widehat{\boldsymbol{\theta}}_n$ and $\widehat{\boldsymbol{\gamma}}_1,\ldots,\widehat{\boldsymbol{\gamma}}_n$, we have the following corollary regarding the asymptotic property of estimators $\widehat{\boldsymbol{\beta}}$ and $(\widehat{f}_1,\ldots,\widehat{f}_K)$.

**Corollary 1.** *Assume the conditions C1–C6 hold, we have that as $n\to\infty$, the following asymptotic results hold:*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\to N(0,\Phi^{-1}\Sigma\Phi^{-1}),$$

$$m_k^{-1}\sum_{i\in\mathcal{G}_k}(\widehat{f}_k(\mathbf{z}_i)-f_k^0(\mathbf{z}_i))^2 = O_p\{J_n/m_k\},\ k=1,\ldots,K,$$

$$\|m_1^{-1}\sum_{i\in\mathcal{G}_1}(\widehat{f}_1(\mathbf{z}_i)-f_1^0(\mathbf{z}_i))^2,\ldots,m_K^{-1}\sum_{i\in\mathcal{G}_K}(\widehat{f}_K(\mathbf{z}_i)-f_K^0(\mathbf{z}_i))^2\|_\infty \le O_p\{J_n/|\mathcal{G}_{min}|\}$$

*with $\Phi$ and $\Sigma$ given in Theorem 1.*

This theorem gives the asymptotic normality and the convergence rate $\sqrt{n}$ for estimating $\boldsymbol{\beta}$ and a slower convergence rate than $\sqrt{n}$ for estimating $f_k$.

## 4. Simulation

In this section, our goal is to examine the finite sample behavior of the newly proposed method by simulation study, and compare it with the methods that ignore the heterogeneity. But we do not compare ours with published methods because there are no other estimation and identification methods for heterogeneous additive partially linear model in the existing literature, to the best of our knowledge. We evaluate the performance of the estimators $\widehat{\boldsymbol{\beta}}$, $\widehat{g}(\mathbf{z})=(\widehat{g}_1(z_1),\ldots,\widehat{g}_n(z_n))^T$ and $\widehat{K}$ by the sample mean, median and standard deviation (sd) and the square roots of MSE. Here the square roots of MSE for $\widehat{\boldsymbol{\beta}}$ and $\widehat{g}(\mathbf{z})$ are defined as $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|/\sqrt{q}$ and $\|\widehat{g}(\mathbf{z})-g(\mathbf{z})\|/\sqrt{n}$ for each replication respectively (see Ma and Huang, 2017). Moreover, the percentage of $\widehat{K}$ equaling to the true number of subgroups based on 100 replications is employed to illustrate the clustering accuracy.

Cubic B-spline is used to estimate all the nonparametric functions $g_{ij}$, and the modified Bayesian Information Criterion (BIC) (Wang et al., 2007) is employed to select the tuning parameter $\lambda$, i.e., it is chosen by minimizing

$$\mathrm{BIC}(\lambda)=\log\left[\sum_{i=1}^{n}\left\|\widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i-\widehat{\boldsymbol{\theta}}_i(\lambda)\right\|_2^2/(nN_np)\right]+C_n\widehat{K}(\lambda)\frac{\log n}{n},$$

where $C_n$ is a positive number depending on $n$. Following Ma and Huang (2017), it is chosen as $C_n=c\log(\log(n))$, where $c$ is a positive constant. In the simulation procedure, the fixed $\vartheta=3$ and $\eta=1$ are used, two different sample sizes $n=100$ and 200 are considered, and all the simulation results are obtained via 100 replications.

**Example 1.** We begin with the following heterogeneous additive partial linear model designed similarly to that of Carroll et al. (1997):

$$y_i=\mathbf{x}_i^T\boldsymbol{\beta}+g_i(z_i)+\epsilon_i,\quad i=1,\ldots,n, \tag{4.1}$$

where the covariates $\mathbf{x}_i=(x_{i1},x_{i2})^T$ are sampled from the 2-dimensional normal distribution with mean 0, variance 1 and correlation coefficient $\rho=0.3$, $z_i$ are independently from $N(0,1)$ truncated to the interval $[0,1]$, the error terms $\epsilon_i$ are independent and identically distributed as $N(0,0.5^2)$. We set the homogeneous parameter $\boldsymbol{\beta}=(2,2)^T$ and randomly assign the heterogeneous additive functions $g_i$ to two subgroups with equal probabilities: $P(i\in\mathcal{G}_1)=P(i\in\mathcal{G}_2)=0.5$, and $g_i=g_1^0$ for $i\in\mathcal{G}_1$, $g_i=g_2^0$ for $i\in\mathcal{G}_2$. The following two different cases about $g_1^0$ and $g_2^0$ are considered:

(1) $g_1^0$ and $g_2^0$ have similar functional forms: $g_1^0(z_i)=\alpha\sin(\frac{\pi(\sqrt{3}z_i-A)}{C-A})$ and $g_2^0(z_i)=-\alpha\sin(\frac{\pi(\sqrt{3}z_i-A)}{C-A})$, where $A=\sqrt{3}/2-1.645/\sqrt{2}$ and $C=\sqrt{3}/2+1.645/\sqrt{2}$, and $\alpha$ is set to two different values $\alpha=1,2$.

(2) $g_1^0$ and $g_2^0$ have different functional forms: $g_1^0(z_i)=2\sin(\frac{\pi(\sqrt{3}z_i-A)}{C-A})$ and $g_2^0(z_i)=-2\cos(\sqrt{3}z_i)+2\cos(\sqrt{3}z_i)^2$, where the values of A and C are the same as those in case (1).

**Table 1**
Simulation results for Example 1.

| | | n = 100 | | | | n = 200 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Median | sd | per | Mean | Median | sd | per |
| Case (1) | SACD | 1.95 | 2.00 | 0.2611 | 0.93 | 1.99 | 2.00 | 0.1000 | 0.99 |
| $\alpha = 1$ | MCP | 1.97 | 2.00 | 0.2642 | 0.93 | 2.02 | 2.00 | 0.1407 | 0.98 |
| Case (1) | SCAD | 1.98 | 2.00 | 0.2453 | 0.94 | 2.02 | 2.00 | 0.1407 | 0.98 |
| $\alpha = 2$ | MCP | 1.99 | 2.00 | 0.2245 | 0.95 | 2.02 | 2.00 | 0.1407 | 0.98 |
| Case (2) | SCAD | 1.91 | 2.00 | 0.2876 | 0.91 | 2.01 | 2.00 | 0.1737 | 0.97 |
| | MCP | 1.91 | 2.00 | 0.2876 | 0.91 | 2.01 | 2.00 | 0.1737 | 0.97 |

**Table 2**
Simulation results for Example 1.

| | | $\widehat{\boldsymbol{\beta}}$ | | | $\widehat{g}(\mathbf{z})$ | | |
|---|---|---|---|---|---|---|---|
| $n$ | Method | Case (1)<br>$\alpha = 1$ | Case (1)<br>$\alpha = 2$ | Case (2) | Case (1)<br>$\alpha = 1$ | Case (1)<br>$\alpha = 2$ | Case (2) |
| 100 | SCAD | 0.0250<br>(0.0135) | 0.0276<br>(0.0194) | 0.0264<br>(0.0139) | 0.0648<br>(0.0176) | 0.0817<br>(0.0512) | 0.0851<br>(0.0383) |
| | MCP | 0.0252<br>(0.0136) | 0.0280<br>(0.0208) | 0.0268<br>(0.0143) | 0.0655<br>(0.0184) | 0.0836<br>(0.0550) | 0.0845<br>(0.0375) |
| | ORACLE | 0.0246<br>(0.0131) | 0.0241<br>(0.0136) | 0.0244<br>(0.0138) | 0.0641<br>(0.0162) | 0.0664<br>(0.0170) | 0.0658<br>(0.0191) |
| 200 | SCAD | 0.0170<br>(0.0091) | 0.0175<br>(0.0092) | 0.0165<br>(0.0092) | 0.0460<br>(0.0130) | 0.0494<br>(0.0176) | 0.0632<br>(0.0295) |
| | MCP | 0.0169<br>(0.0091) | 0.0175<br>(0.0092) | 0.0163<br>(0.0091) | 0.0461<br>(0.0131) | 0.0494<br>(0.0176) | 0.0617<br>(0.0287) |
| | ORACLE | 0.0170<br>(0.0092) | 0.0173<br>(0.0091) | 0.0151<br>(0.0085) | 0.0453<br>(0.0122) | 0.0482<br>(0.0118) | 0.0456<br>(0.0122) |

The simulation results are reported in Table 1. From Table 1 we can see that the median of $\widehat{K}$ over the 100 replications is 2, the true number of subgroups for all cases, and the mean values are very close to 2 for both MCP and SCAD methods. Moreover, the standard deviation becomes smaller and the mean gets closer to the true value 2, and the percentage of correctly selecting the number of subgroups becomes larger as the sample size $n$ increases.

The estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{g}(\mathbf{z}) = (\widehat{g}_1(z_1), \ldots, \widehat{g}_n(z_n))^T$ by the MCP and SCAD are calculated based on the replications with the estimated number of groups equal to two. For the oracle estimators, they are calculated based on the 100 replications. Table 2 reports the sample mean and the standard deviation (in parentheses) of the square root of the mean squared errors (MSE) for $\widehat{\boldsymbol{\beta}}$ and $\widehat{g}(\mathbf{z})$ for the SCAD, MCP and the oracle estimators. Table 2 shows that the SCAD and MCP methods have very small MSE and sd for all the cases, which are close to those of the oracle estimators. This supports the oracle property established in Theorem 3. Moreover, as the sample size $n$ increases, the MSE values of the estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{g}(\mathbf{z})$ tend to zero, implying the consistency of the estimators. This result supports our consistency result in Corollary 1. Fig. 1 presents the boxplots of the estimators $\widehat{\boldsymbol{\beta}}$ by SCAD (white) and MCP (gray) with $n = 200$ for the three cases, respectively. From the three plots, we observe that the median of $\widehat{\boldsymbol{\beta}}$ is the true value 2 for SCAD and MCP methods, and the biases are very small in all replications. Moreover, we have $\lambda = 0.102$, $a_n = \min_{i\in\mathcal{G}_s,k\in\mathcal{G}_t,s\neq t} \|\boldsymbol{\theta}_i^0 - \boldsymbol{\theta}_k^0\| = 0.408$, $|\mathcal{G}_{min}|^{-1}\sqrt{n\log n} = 0.429$, and $\sqrt{N_n p}|\mathcal{G}_{min}|^{-1}\sqrt{n\log n} = 0.858$ with $n = 100$ and $\alpha = 1$ in case (1) based on the 100 replications. The conditions in Theorem 3 $a_n > C\lambda$ and $\lambda \gg \phi_n$ can be satisfied by some positive constant $c$. Moreover, the other simulation results can also satisfy the condition.

If we do not consider the heterogeneity in additive components, the parameters estimated by the nonlinear least squares without cluster analysis can be misleading. To illustrate this point, in Fig. 2, we plot the values of the true additive component $g_i^0(z_i)$ (black dotted lines), the estimated additive component $\widehat{g}_i(z_i)$ (red dotted lines) by our methods and the estimated additive component $\widehat{g}_i^{nls}(z_i)$ (gray dotted lines) by the nonlinear least squares against values of $z_i$. The first one is plotted by using the 98 replications which have two estimated subgroups by the SCAD method for $\alpha = 2$ and $n = 200$, and the second one is plotted by using the 97 replications which have two estimated subgroups by the SCAD method for case (2) and $n = 200$. We can see that the fitted function curves by our methods are close to the true curves in the two plots. However, the fitted function curves by the nonlinear least squares are far away from the true ones.

**Example 2.** In this experiment, we consider the heterogeneous models with three dimensional variables as

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \sum_{j=1}^{3} g_{ij}(z_{ij}) + \epsilon_i, \quad i = 1, \ldots, n, \tag{4.2}$$

where $\mathbf{x}_i$, $\epsilon_i$ and $\boldsymbol{\beta}$ are generated in the same way as in Example 1. Let $\mathbf{z}_i = (z_{i1}, z_{i2}, z_{i3})^T$ in which $z_{i1}, z_{i2}, z_{i3}$ are generated independently from $N(0, 1)$ truncated to the interval [0,1], $i = 1, \ldots, n$. We randomly divide $g_i$ into two

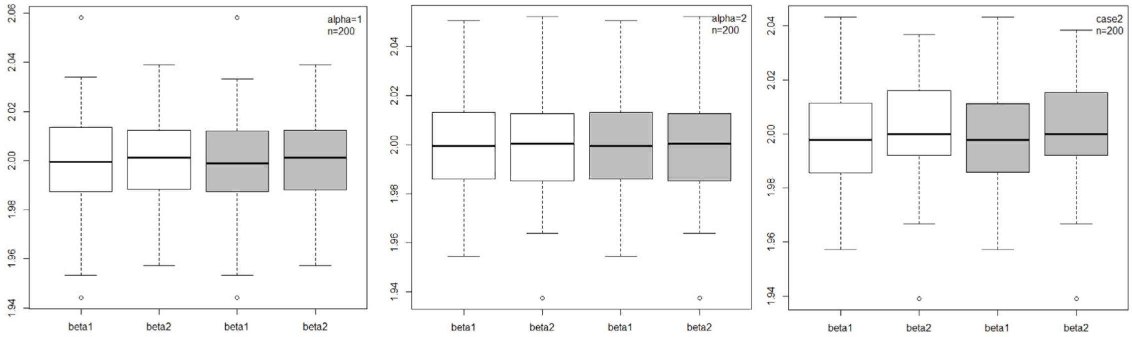
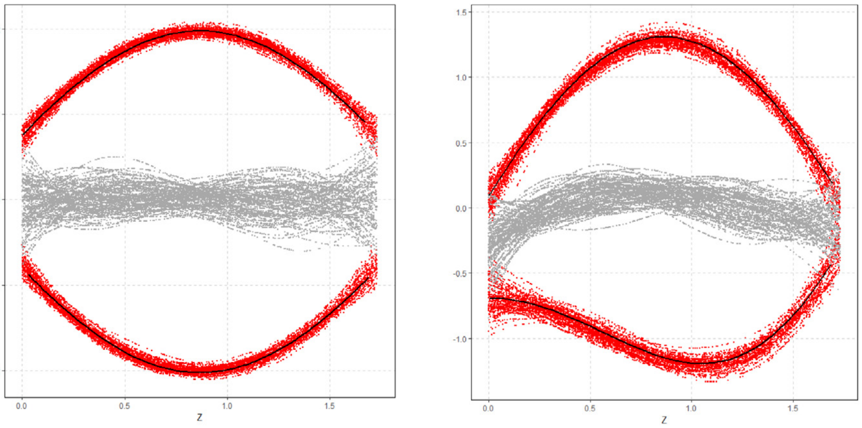


**Fig. 1.** The boxplots of the estimators $\widehat{\boldsymbol{\beta}}$ by SCAD (white) and MCP (gray) with $n = 200$ for three cases in Experiment 1.



(a) $\alpha = 2$

(b) case 2

**Fig. 2.** Plots of $g_i(z_i)$ (black dotted lines), $\widehat{g}_i(z_i)$ (red dotted lines) and $\widehat{g}_i^{nls}(z_i)$ (gray dotted lines) against values of $z_i$ for $n = 200$ in Experiment 1 . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 3**
Simulation results for Example 2.

| | | n = 100 | | | | n = 200 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Median | sd | per | Mean | Median | sd | per |
| $\alpha = 1$ | SCAD | 1.97 | 2.00 | 0.2642 | 0.93 | 2.01 | 2.00 | 0.1000 | 0.99 |
| | MCP | 2.01 | 2.00 | 0.2657 | 0.93 | 2.01 | 2.00 | 0.1000 | 0.99 |
| $\alpha = 2$ | SCAD | 2.02 | 2.00 | 0.2000 | 0.96 | 2.02 | 2.00 | 0.1407 | 0.98 |
| | MCP | 2.03 | 2.00 | 0.2227 | 0.95 | 2.02 | 2.00 | 0.1407 | 0.98 |

subgroups with equal probabilities, i.e. $P(i \in \mathcal{G}_1) = P(i \in \mathcal{G}_2) = 0.5$, and $g_i(\mathbf{z}_i) = g_{11}(z_{i1}) + g_{12}(z_{i2}) + g_{13}(z_{i3})$ for $i \in \mathcal{G}_1$, $g_i(\mathbf{z}_i) = g_{21}(z_{i1}) + g_{22}(z_{i2}) + g_{23}(z_{i3})$ for $i \in \mathcal{G}_2$. For multiple variables, here we only consider that heterogeneous additive components have the similar function form for different subgroups. We set $g_{11}(z_{i1}) = \alpha_{11}\sin(\pi z_{i1})/(2-\sin(\pi z_{i1}))$, $g_{12}(z_{i2}) = \alpha_{12}\sin(\pi z_{i2})^2$, $g_{13}(z_{i3}) = \alpha_{13}\sin(\pi z_{i3})$ for $i \in \mathcal{G}_1$, and $g_{21}(z_{i1}) = \alpha_{21}\sin(\pi z_{i1})/(2-\sin(\pi z_{i1}))$, $g_{22}(z_{i2}) = \alpha_{22}\sin(\pi z_{i2})^2$, $g_{23}(z_{i3}) = \alpha_{23}\sin(\pi z_{i3})$ for $i \in \mathcal{G}_2$. We let $\alpha_{1j} = \alpha$ and $\alpha_{2j} = -\alpha$ for $j = 1, 2, 3$. and $\alpha$ is set to be 1, 2 for different signal-noise ratios.

Table 3 reports the mean, median and standard deviation (sd) of $\widehat{K}$ and the percentage of $\widehat{K}$ equaling to the true number of subgroups by MCP and SCAD. It is seen that all the average numbers of $\widehat{K}$ are close to the true value 2 for different values of $\alpha$, and the median of $\widehat{K}$ is 2. This implies that our method can correctly identify the subgroups. Moreover, the percentage of correctly selecting the number of subgroups increases as the sample size $n$ increases.

Denote $\widehat{g}(\mathbf{z}) = (\sum_{j=1}^{3}\widehat{g}_{1j}(z_{1j}), \ldots, \sum_{j=1}^{3}\widehat{g}_{nj}(z_{nj}))^T$, and the estimators of three nonparametric components are $\widehat{g}_j(\mathbf{z}_j) = (\widehat{g}_{1j}(z_{1j}), \ldots, \widehat{g}_{nj}(z_{nj}))^T$ with $j = 1, 2, 3$, respectively. The true nonparametric component $g_1(\mathbf{z}_1)$ consists of $g_{11}(z_{i1})$ for $i \in \mathcal{G}_1$ and $g_{21}(z_{i1})$ for $i \in \mathcal{G}_1$, $g_2(\mathbf{z}_2)$ and $g_3(\mathbf{z}_3)$ are similar. Table 4 reports the sample mean and standard deviation (in parentheses) of the square roots of MSE for $\widehat{\boldsymbol{\beta}}$, $\widehat{g}_1(\mathbf{z}_1)$, $\widehat{g}_2(\mathbf{z}_2)$, $\widehat{g}_3(\mathbf{z}_3)$ and $\widehat{g}(\mathbf{z})$. By Table 4, we can compare the performance

**Table 4**
Simulation results for Example 2.

| | | $\alpha = 1$ | | | $\alpha = 2$ | | |
|---|---|---|---|---|---|---|---|
| $n$ | Method | SCAD | MCP | ORACLE | SCAD | MCP | ORACLE |
| 100 | $\boldsymbol{\beta}$ | 0.0363 | 0.0363 | 0.0306 | 0.0395 | 0.0405 | 0.0257 |
| | | (0.0249) | (0.0249) | (0.0162) | (0.0919) | (0.0934) | (0.0136) |
| | $g_1(\mathbf{z}_1)$ | 0.1037 | 0.1037 | 0.0842 | 0.1245 | 0.1249 | (0.0915) |
| | | (0.0888) | (0.0888) | (0.0341) | (0.1480) | (0.1566) | (0.0630) |
| | $g_2(\mathbf{z}_2)$ | 0.1435 | 0.1435 | 0.1242 | 0.1665 | 0.1693 | 0.1369 |
| | | (0.1343) | (0.1343) | (0.0648) | (0.1639) | (0.1795) | (0.0992) |
| | $g_3(\mathbf{z}_3)$ | 0.1161 | 0.1162 | 0.1053 | 0.1636 | 0.1650 | 0.1317 |
| | | (0.0540) | (0.0540) | (0.0300) | (0.1390) | (0.1434) | (0.0397) |
| | $g(\mathbf{z})$ | 0.1612 | 0.1613 | 0.1428 | 0.2066 | 0.2073 | 0.1461 |
| | | (0.0571) | (0.0570) | (0.0186) | (0.3256) | (0.3278) | (0.0176) |
| 200 | $\boldsymbol{\beta}$ | 0.02025 | 0.02027 | 0.02003 | 0.01769 | 0.01770 | 0.01690 |
| | | (0.00985) | (0.00984) | (0.00987) | (0.01084) | (0.01083) | (0.00918) |
| | $g_1(\mathbf{z}_1)$ | 0.05345 | 0.05340 | 0.05165 | 0.07406 | 0.07405 | 0.07072 |
| | | (0.01325) | (0.01312) | (0.01249) | (0.01696) | (0.01696) | (0.01272) |
| | $g_2(\mathbf{z}_2)$ | 0.07396 | 0.07395 | 0.07395 | 0.08395 | 0.08396 | 0.08157 |
| | | (0.01533) | (0.01531) | (0.01506) | (0.01684) | (0.01683) | (0.01565) |
| | $g_3(\mathbf{z}_3)$ | 0.07258 | 0.07246 | 0.07143 | 0.10510 | 0.10507 | 0.10282 |
| | | (0.01211) | (0.01191) | (0.01150) | (0.01600) | (0.01599) | (0.01426) |
| | $g(\mathbf{z})$ | 0.10224 | 0.10218 | 0.10039 | 0.11536 | 0.11535 | 0.10948 |
| | | (0.01416) | (0.01402) | (0.01220) | (0.02037) | (0.02039) | (0.01098) |

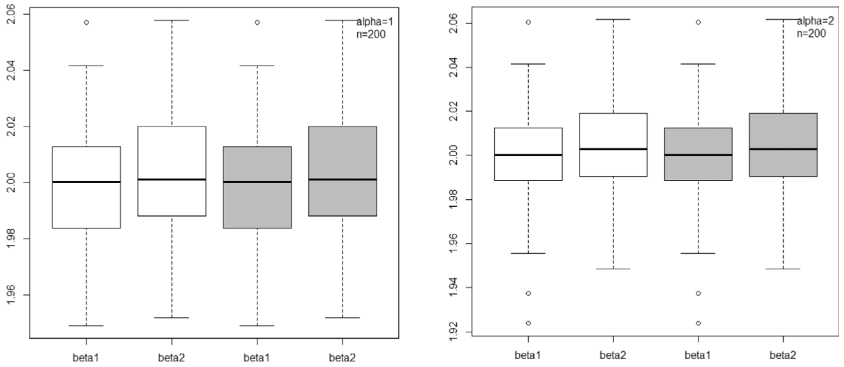


**Fig. 3.** The boxplots of $\widehat{\boldsymbol{\beta}}$ by using SCAD (white) and MCP (gray) with $n = 200$ for $\alpha = 1$ and $\alpha = 2$, respectively in Example 2.

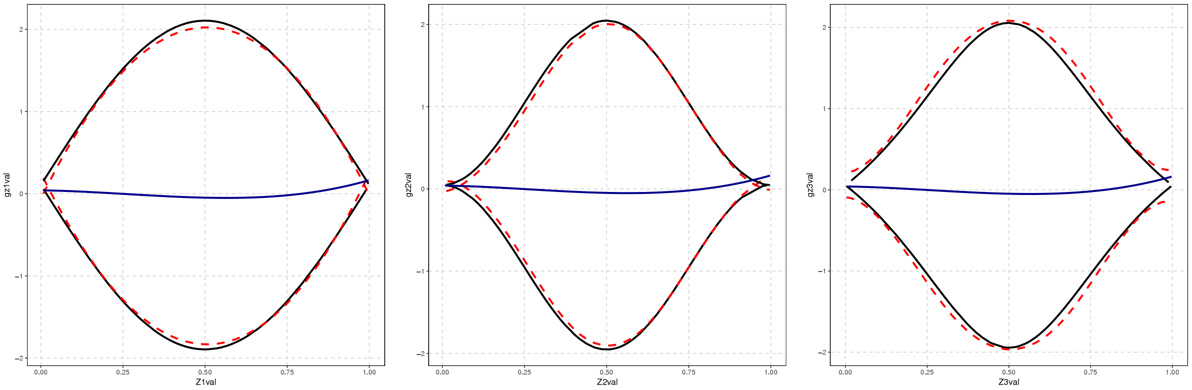
of the estimators by our method and the oracle estimators. The mean and sd are similar for MCP and SCAD, and they are also close to the corresponding values of the oracle estimators. Furthermore, our method is robust to different choices of $\alpha$. It comes as no surprise that all the performance measures generally improve with increased sample sizes. We observe that the MSE values are small and decrease as $n$ increases for both MCP and SCAD, suggesting the consistency of estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{g}(\mathbf{z})$.

Fig. 3 shows the boxplots of parameter estimator $\widehat{\boldsymbol{\beta}}$ by SCAD (white) and MCP (gray) for $n = 200$ respectively. We observe that variation in the estimators by SCAD and MCP is similar. For both MCP and SCAD methods, the median value of $\widehat{\boldsymbol{\beta}}$ is very close to the true parameter $\boldsymbol{\beta} = (2, 2)^T$, and $\widehat{\boldsymbol{\beta}}$ has much smaller bias 0.02 in almost all replications for $\alpha = 1$ and $\alpha = 2$.

Fig. 4 is drawn to examine the behavior of three nonparametric function estimates $\widehat{g}_1(z)$, $\widehat{g}_2(z)$, $\widehat{g}_3(z)$. As mentioned above, the true function $g_1(z)$ consists of $g_{11}(z)$ and $g_{21}(z)$, $g_2(z)$ and $g_3(z)$ are similar. Fig. 4 displays the average estimated curves of three nonparametric functions by our methods (red dashed lines), the corresponding true function curves (black solid lines) and the corresponding average estimated curves by the nonlinear OLS (blue dotted lines) over the 98 replications which have two estimated groups by SCAD for $\alpha = 2$ and $n = 200$. We can see that the difference between the true functions and the estimated functions $\widehat{g}_1(z)$, $\widehat{g}_2(z)$, $\widehat{g}_3(z)$ by our method is barely visible, which implies that there is little bias. However, the estimated function curves by the OLS are far away from the true curves.

**Example 3.** In this experiment, we consider the homogeneous model as

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + g_i(z_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{4.3}$$



(c) $g_1(z)$ (d) $g_2(z)$ (e) $g_3(z)$

**Fig. 4.** Curve estimates for the additive functions $g_1(z)$, $g_2(z)$ and $g_3(z)$ in Example 2.

**Table 5**
Simulation results for Example 3.

| | | Mean | Median | sd | per |
|---|---|---|---|---|---|
| $n = 100$ | SACD | 1.01 | 1.00 | 0.100 | 0.99 |
| | MCP | 1.02 | 1.00 | 0.141 | 0.98 |
| $n = 200$ | SCAD | 1.00 | 1.00 | 0.000 | 1.00 |
| | MCP | 1.01 | 1.00 | 0.100 | 0.99 |

**Table 6**
Simulation results for Example 3.

| | n = 100 | | n = 200 | |
|---|---|---|---|---|
| Method | $\boldsymbol{\beta}$ | $g(\mathbf{z})$ | $\boldsymbol{\beta}$ | $g(\mathbf{z})$ |
| SCAD | 0.0245 | 0.0887 | 0.0184 | 0.0639 |
| | (0.0135) | (0.0148) | (0.0097) | (0.0111) |
| MCP | 0.0244 | 0.0889 | 0.0183 | 0.0638 |
| | (0.0136) | (0.0147) | (0.0097) | (0.0110) |
| ORACLE | 0.0246 | 0.0889 | 0.0184 | 0.0639 |
| | (0.0134) | (0.0148) | (0.0097) | (0.0111) |

where $\mathbf{x}_i$, $z_i$, $\epsilon_i$ and $\boldsymbol{\beta}$ are simulated in the same way as in Example 1, the nonparametric function is chosen as the following exponential function

$$g_i(z) = 10\left\{\exp(-3.25z) - 4\exp(-6.5z) + 3\exp(-9.75z)\right\},$$

which is more complicated than those used in the previous samples. We use our proposed approach to fit the model with possible heterogeneity in the additive components. The simulation results are presented in Tables 5 and 6.

Table 5 presents the sample mean, median and standard deviation (sd) of $\widehat{K}$ and the percentage of $\widehat{K}$ equaling to the true number of subgroups by MCP and SCAD with $n = 100$ and $n = 200$. We have the similar results as the above. The sample median of $\widehat{K}$ is the true value 1 for both methods. Moreover, the selection results are more accurate as $n$ increases.

Table 6 presents the sample mean and standard deviation (in parentheses) of the square roots of MSE for $\widehat{\boldsymbol{\beta}}$ and $\widehat{g}(\mathbf{z})$. The MSE and sd are similar for both MCP and SCAD, and they are very close to the corresponding values of the oracle estimators. Furthermore, the MSE values of $\widehat{\boldsymbol{\beta}}$ and $\widehat{g}_i$ decrease as $n$ increases for both MCP and SCAD, implying the consistency of the parameter estimator and the nonparametric estimator. These results indicate that our proposed method works well for the homogeneous model. On the other hand, the estimation consistency of $\widehat{\boldsymbol{\beta}}$ is further reflected by the boxplots in Fig. 5. Similar to the findings in the above, it can be seen that the median and mean of estimator $\widehat{\boldsymbol{\beta}}$ are very close to the true value of the parameter in Fig. 5.

Fig. 6 shows the average estimated curves of the additive function $\widehat{g}_i$ by our methods (red dashed lines) over the 99 replications which have one estimated group by SCAD for $n = 200$ and the true function (black solid lines). It can be seen that the estimated function $\widehat{g}_i$ is very close to the true one, demonstrating the estimation accuracy.

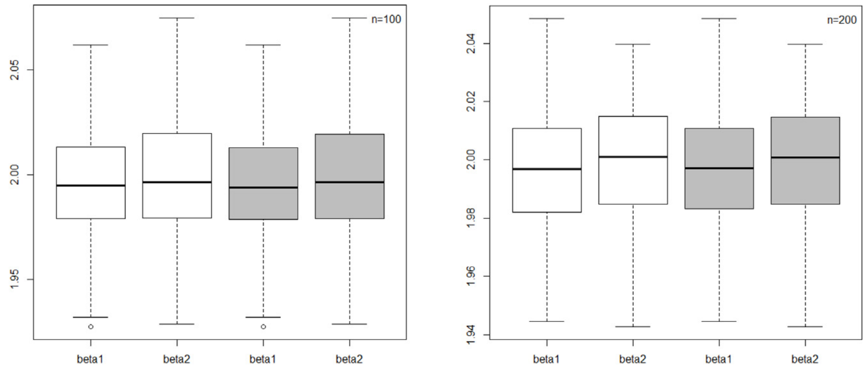


**Fig. 5.** The boxplots of parameter estimator $\widehat{\boldsymbol{\beta}}$ by SCAD (white) and MCP (gray) with $n = 100$ and $n = 200$ in Example 3.

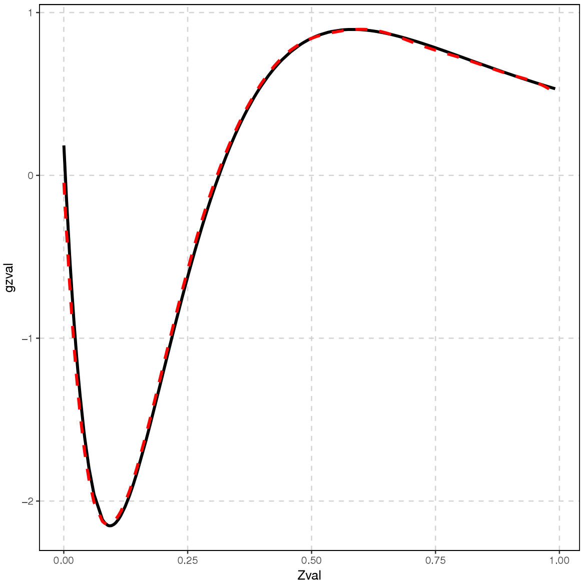


**Fig. 6.** Curve estimates for the additive function $g_i$. The black solid curve is the true function; The red dashed curve is the average estimated curve by our method with $n = 200$ in Example 3 . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

## 5. Application

In this section, we apply our method to the problem of car sales. The car sales data of USA is used to judge whether potential consumers will buy a Japanese car or an American car according to their several characteristics. The dataset can be accessed at http://www-stat.wharton.upenn.edu/~waterman/fsw/baur/assigtxt.htm. The dataset contains 263 consumption records, and each record consists of a binary response $y_i$ ($0 =$ the Japanese car, $1 =$ the American car) and the covariate vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{i5})^T$, $i = 1, \ldots, 263$, where $x_{i1}$ is the consumer's age from 18 to 60, $x_{i2}$ is the consumer's gender ($0 =$ female, $1 =$ male), $x_{i3}$ is the consumer's marital status ($0 =$ single, $1 =$ married), $x_{i4}$ is the favorable size of a consumer ($0 =$ small, $1 =$ medium, $2 =$ large), and $x_{i5}$ is the type of car ($0 =$ work, $1 =$ sporty, $2 =$ family). All of the predictors are centered and standardized, and the response variable is log-transformed as $\widetilde{y}_i = \log((a + y_i)/(b - y_i))$ with $a = 0.01$ and $b = 1.01$.

To see possible heterogeneity in the regression model, we first fit the dataset by a homogeneous linear regression model as $\widetilde{y}_i = \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i$, and obtain the mean square error $\text{MSE} = \sum_{i=1}^n(\widetilde{y}_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}^{ols})^2/n = 15.76$, where $\widehat{\boldsymbol{\beta}}^{ols}$ is the estimated value of $\boldsymbol{\beta}$ from OLS. Such a relatively large value of MSE indicates that homogeneous linear fitting may be unreasonable. Before improving the regression modeling, we check if there is the heterogeneity among possible regressions. By the classification analysis, we see that about 78% people who prefer the small cars choose the Japanese cars, while only 39% people who prefer the large size cars choose the Japanese cars. On the other hand, about 66% single people choose the Japanese cars, while only 51% married people choose the Japanese cars. These results are consistent with the empirical analysis, as we all know, Japanese cars are usually smaller than the American ones. There are 212 individuals with the age $x_{i1} \leqslant 35$, and among these people, 62 percent people buy the Japanese cars. Among 51 individuals with the age $x_{i1} > 35$, only 33 percent people buy the Japanese cars. However, for the predictors $x_2$ and $x_5$, the percentage of buying the Japanese cars or American cars has a very small difference. Through the above analysis, we judge that the model should be of heterogeneity, specially for covariates $x_1$, $x_3$ and $x_4$.

Thus, the following heterogeneous additive partially linear model is used to fit the dataset:

$$\widetilde{y}_i = \beta_0 + x_{i2}\beta_1 + x_{i5}\beta_2 + g_{i1}(x_{i1}) + g_{i2}(x_{i3}) + g_{i3}(x_{i4}) + \epsilon_i,\ \ i = 1, \ldots, 263. \tag{5.1}$$

We use cubic splines to estimate the first additive component and fit categorical explanatory variables $x_{i3}$ and $x_{i4}$ in a heterogeneous linear way, and finally employ our proposal together with the SCAD to identify the subgroups. To determine the number of knots in the B-splines approximation of the nonparametric component, we try the numbers of knots from 2 to 9 for $g_{i1}(x_{i1})$ and choose the number that gives the smallest mean squared error. As a result, two knots are used for $g_{i1}(x_{i1})$.

According to the model above, the 263 data are classified into 9 subgroups denoted by $\mathcal{G}_j$, $j = 1, \ldots, 9$, which respectively contain 17, 15, 38, 26, 19, 43, 18, 36, and 51 data. Consequently, the resulting empirical model is $\widehat{E}(\widetilde{y}_i|x) =$

$$\begin{cases} -0.58 - 0.020x_{i2} + 0.022x_{i5} + \mathbf{B}^T(x_{i1})\widehat{\boldsymbol{\gamma}}_1 - 3.323x_{i3} - 0.002x_{i4} & \text{if } i \in \mathcal{G}_1, \\ -0.58 - 0.020x_{i2} + 0.022x_{i5} + \mathbf{B}^T(x_{i1})\widehat{\boldsymbol{\gamma}}_2 - 3.096x_{i3} + 0.803x_{i4} & \text{if } i \in \mathcal{G}_2, \\ -0.58 - 0.020x_{i2} + 0.022x_{i5} + \mathbf{B}^T(x_{i1})\widehat{\boldsymbol{\gamma}}_3 + 3.945x_{i3} + 1.543x_{i4} & \text{if } i \in \mathcal{G}_3, \\ -0.58 - 0.020x_{i2} + 0.022x_{i5} + \mathbf{B}^T(x_{i1})\widehat{\boldsymbol{\gamma}}_4 + 1.715x_{i3} + 1.801x_{i4} & \text{if } i \in \mathcal{G}_4, \\ -0.58 - 0.020x_{i2} + 0.022x_{i5} + \mathbf{B}^T(x_{i1})\widehat{\boldsymbol{\gamma}}_5 + 2.584x_{i3} - 2.297x_{i4} & \text{if } i \in \mathcal{G}_5, \\ -0.58 - 0.020x_{i2} + 0.022x_{i5} + \mathbf{B}^T(x_{i1})\widehat{\boldsymbol{\gamma}}_6 + 0.897x_{i3} + 2.730x_{i4} & \text{if } i \in \mathcal{G}_6, \\ -0.58 - 0.020x_{i2} + 0.022x_{i5} + \mathbf{B}^T(x_{i1})\widehat{\boldsymbol{\gamma}}_7 + 2.164x_{i3} - 0.931x_{i4} & \text{if } i \in \mathcal{G}_7, \\ -0.58 - 0.020x_{i2} + 0.022x_{i5} + \mathbf{B}^T(x_{i1})\widehat{\boldsymbol{\gamma}}_8 - 2.878x_{i3} - 2.381x_{i4} & \text{if } i \in \mathcal{G}_8, \\ -0.58 - 0.020x_{i2} + 0.022x_{i5} + \mathbf{B}^T(x_{i1})\widehat{\boldsymbol{\gamma}}_9 - 3.656x_{i3} + 0.002x_{i4}, & \text{if } i \in \mathcal{G}_9, \end{cases}$$

where

$$\begin{aligned} &\widehat{\boldsymbol{\gamma}}_1 = (0.715, 0.706, 0.802, 0.549)^T, \widehat{\boldsymbol{\gamma}}_2 = (0.642, 0.725, 0.544, 0.711)^T, \\ &\widehat{\boldsymbol{\gamma}}_3 = (1.706, 1.241, 2.507, -0.213)^T, \widehat{\boldsymbol{\gamma}}_4 = (0.531, 0.724, 0.469, 0.614)^T, \\ &\widehat{\boldsymbol{\gamma}}_5 = (1.213, 0.592, 1.790, -0.016)^T, \widehat{\boldsymbol{\gamma}}_6 = (-0.145, -0.076, -0.192, -0.181)^T, \\ &\widehat{\boldsymbol{\gamma}}_7 = (-0.623, -0.835, -0.936, 1.099)^T, \widehat{\boldsymbol{\gamma}}_8 = (-0.839, -1.054, -0.488, -1.583)^T, \\ &\widehat{\boldsymbol{\gamma}}_9 = (-1.337, -1.291, -1.426, -1.103)^T, \end{aligned}$$

and the B-splines are constructed using the function “create.bspline.basis” in R, and all the components of $\mathbf{B}(x_{i1})$ are positive.

From the empirical model, we have the following findings:

(1) An interesting result is that the MSE of the empirical model is significantly reduced to a very small value 0.0008. Thus, taking into account the subgroup structure leads to a notable improvement of the model fitting.

(2) We find from the empirical model that the 17, 15, 38, 26 and 19 data correspond to the people who buy American cars, and among these people, the 17 customers are single and like the small and large cars, the 15 customers are single and like the medium car, the 38 customers are married and like the medium car, the 26 customers are married and like the large car, the 19 customers are married and like the small car. From the parameter estimation of empirical model, we can see that single status and small size are negatively related to buying an American car, and married status and large, medium size are positively related to buying an American car. Moreover, almost all components of the estimators $\widehat{\boldsymbol{\gamma}}_1, \widehat{\boldsymbol{\gamma}}_2, \widehat{\boldsymbol{\gamma}}_3, \widehat{\boldsymbol{\gamma}}_4, \widehat{\boldsymbol{\gamma}}_5$ are positive. All of the above results imply that the people who either prefer the large and medium size cars or are married or are older tend to buy the American cars.

(3) On the other hand, the 43, 18, 36, 51 data correspond to the people who buy Japanese cars, among these people, the 43 customers are single and like the small car, the 18 customers are single and like the medium car, the 36 customers are married and like the medium car, the 51 customers are married and like the small and large cars. Similar to the analysis above, it is seen that single status and small size are positively related to buying a Japanese car, and married status and medium size are negatively related to buying a Japanese car. Moreover, all the components of the estimators $\widehat{\boldsymbol{\gamma}}_6, \widehat{\boldsymbol{\gamma}}_7, \widehat{\boldsymbol{\gamma}}_8, \widehat{\boldsymbol{\gamma}}_9$ are negative. These observations imply that the people who either prefer the small size car or are single or are younger are inclined to buy the Japanese cars.

(4) The estimators of homogeneous parameter estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are −0.020 and 0.022, respectively, both are very small, implying consumer's gender and the type of car have little effect on buying Japanese cars or American cars. Therefore, the features of the gender and the type of car are not subject-specific.

The results of findings (2) and (3) are summarized into Table 7. “+” means that the covariates are positively related to $\widetilde{y}$, while “−” means that the covariates are negatively related to $\widetilde{y}$.

**Table 7**
Results of findings (2) and (3) for application.

| $\widetilde{y}$ | Group | People | $x_{i3}$ | $x_{i4}$ | coef of $x_{i3}$ | coef of $x_{i4}$ | coef of $x_{i1}$ |
|---|---|---|---|---|---|---|---|
| American cars | $\mathcal{G}_1$ | 17 | Single | Small and large | $-$ | $-$ | $+$ |
| | $\mathcal{G}_2$ | 15 | Single | Medium | $-$ | $+$ | $+$ |
| | $\mathcal{G}_3$ | 38 | Married | Medium | $+$ | $+$ | $+$ |
| | $\mathcal{G}_4$ | 26 | Married | Large | $+$ | $+$ | $+$ |
| | $\mathcal{G}_5$ | 19 | Married | small | $+$ | $-$ | $+$ |
| Japanese cars | $\mathcal{G}_6$ | 43 | Single | Small | $+$ | $+$ | $-$ |
| | $\mathcal{G}_7$ | 18 | Single | Medium | $+$ | $-$ | $-$ |
| | $\mathcal{G}_8$ | 36 | Married | Medium | $-$ | $-$ | $-$ |
| | $\mathcal{G}_9$ | 51 | Married | small and large | $-$ | $+$ | $-$ |

The results of the customer classification by our empirical model are the same as those by the direct empirical classification discussed before. Furthermore, our empirical model presents the quantifiable outcomes and the quantifiable relationship between the response $y$ and the characteristics $x_j$. Thus, our proposal can successfully classify the customers, and give the correct decision and prediction. On this basis, enterprises can make precise marketing strategies for different customer categories, which target the high potential customers in several ways, including hobby of car size, age and marital status. In addition, the case study shows that the heterogeneous additive partially linear framework is efficient and can help enterprises in planning their precision marketing.

## Appendix. Proofs of main results

### A.1. Proof of the equality in (2.3)

In the following, we prove $\widetilde{y}_i - \widetilde{\mathbf{B}}_{\mathbf{z}i}^T\boldsymbol{\gamma}_i$ is equivalent to $y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{B}(\mathbf{z}_i)^T\boldsymbol{\gamma}_i$. By the symbols in Section 2, we have

$$\mathbf{B}(\mathbf{z}) = (\mathbf{B}(\mathbf{z}_1), \ldots, \mathbf{B}(\mathbf{z}_n))^T = \begin{pmatrix} \mathbf{B}(\mathbf{z}_1)^T \\ \cdots \\ \mathbf{B}(\mathbf{z}_n)^T \end{pmatrix} \in R^{n\times(N_n p+1)},$$

$$\mathbf{B}_\gamma = (\mathbf{B}(\mathbf{z}_1)^T\boldsymbol{\gamma}_1, \ldots, \mathbf{B}(\mathbf{z}_n)^T\boldsymbol{\gamma}_n)^T = \begin{pmatrix} \mathbf{B}(\mathbf{z}_1)^T\boldsymbol{\gamma}_1 \\ \cdots \\ \mathbf{B}(\mathbf{z}_n)^T\boldsymbol{\gamma}_n \end{pmatrix} \in R^{n\times 1}.$$

Since the estimator of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{B}_\gamma)$. Thus $y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{B}(\mathbf{z}_i)^T\boldsymbol{\gamma}_i$ can be expressed by matrix with $n$ samples: $\mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{B}_\gamma) - \mathbf{B}_\gamma$. Denote $\mathbf{Q}_x = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ and $\mathbf{I}_n$ is the $n \times n$ identity matrix. Let $\widetilde{\mathbf{y}} = \mathbf{Q}_x\mathbf{y} \in R^{n\times 1}$ and $\widetilde{\mathbf{B}}_\mathbf{z} = \mathbf{Q}_x\mathbf{B}(\mathbf{z}) \in R^{n\times(N_n p+1)}$. Thus, for the $i_{th}$ sample, we obtain $\widetilde{y}_i - \widetilde{\mathbf{B}}_{\mathbf{z}i}^T\boldsymbol{\gamma}_i$ is equivalent to $y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{B}(\mathbf{z}_i)^T\boldsymbol{\gamma}_i$ based on $\boldsymbol{\beta}$ consistently estimated.

### A.2. Proof of the equivalence between the criterions (2.3) and (2.4)

Let

$$L_1(\boldsymbol{\gamma}, \lambda) = \frac{1}{2}\sum_{i=1}^{n}(\widetilde{y}_i - \widetilde{\mathbf{B}}_{\mathbf{z}i}^T\boldsymbol{\gamma}_i)^2 + \sum_{1\le i<k\le n} p_\vartheta(\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_k\|, \lambda). \tag{2.3}$$

$$L_2(\boldsymbol{\theta}, \lambda) = \frac{1}{2}\sum_{i=1}^{n}\left\|\widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i - \boldsymbol{\theta}_i\right\|_2^2 + \sum_{i<k} p_\vartheta(\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_k\|, \lambda). \tag{2.4}$$

Assume the coefficients $\boldsymbol{\gamma}_i$ and $\boldsymbol{\gamma}_j$ are in the same group $\mathcal{G}_k$ $k = 1, \ldots, K$ after minimizing (2.3). By the relationship between $\boldsymbol{\theta}_i$ and $\boldsymbol{\gamma}_i$, the minimizers of (2.4) are $\boldsymbol{\theta}_i = E[\widetilde{\mathbf{B}}_\mathbf{z}\widetilde{\mathbf{B}}_\mathbf{z}^T]\boldsymbol{\gamma}_i$ and $\boldsymbol{\theta}_j = E[\widetilde{\mathbf{B}}_\mathbf{z}\widetilde{\mathbf{B}}_\mathbf{z}^T]\boldsymbol{\gamma}_j$. Thus, we have

$$0 \le \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\| = \|E(\widetilde{\mathbf{B}}_\mathbf{z}\widetilde{\mathbf{B}}_\mathbf{z}^T)(\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j)\| \le \|E(\widetilde{\mathbf{B}}_\mathbf{z}\widetilde{\mathbf{B}}_\mathbf{z}^T)\|\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\| = 0,$$

implying $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\| = 0$. Thus, $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ are also in the same group and vice versa for $\boldsymbol{\gamma}_i$ and $\boldsymbol{\gamma}_j$. Assume the coefficients $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ are in the same group $\mathcal{G}_k$ $k = 1, \ldots, K$ after minimizing (2.4). Then

$$0 \le \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\| = \|E[(\widetilde{\mathbf{B}}_\mathbf{z}\widetilde{\mathbf{B}}_\mathbf{z}^T)^{-1}](\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)\| \le \|E[(\widetilde{\mathbf{B}}_\mathbf{z}\widetilde{\mathbf{B}}_\mathbf{z}^T)^{-1}]\|\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\| = 0.$$

So we have $\|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\| = 0$, meaning that $\boldsymbol{\gamma}_i$ and $\boldsymbol{\gamma}_j$ are also in the same group. Thus, the optimization problem (2.3) is equivalent to (2.4) in terms of identifying subgroups.

### *A.3. Proof of Theorem 1*

Recall that for any functions $f_i \in \mathcal{H}$, we can have $\boldsymbol{\gamma}_i$ and an additive spline function $\widetilde{f}_i = \mathbf{B}(\mathbf{z}_i)^T \widetilde{\boldsymbol{\gamma}}_i$ such that $\|\widetilde{f}_i - f_i^0\|_\infty = O(h^\kappa)$, where $h = \frac{1}{J_n+1}$ is the distance between neighboring knots (see Liu et al., 2011). Write

$$\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \widetilde{f}_i)^2, \qquad \widehat{L}(\boldsymbol{\zeta}) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{B}(\mathbf{z}_i)^T\boldsymbol{\alpha}_k\right)^2,$$

where $\boldsymbol{\zeta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$, $K$ and $\mathcal{G}_k$ are known in advance. The oracle estimator $\widehat{\boldsymbol{\beta}}^{or}$ and $\widehat{\boldsymbol{\alpha}}^{or} = (\widehat{\boldsymbol{\alpha}}_1^{orT}, \ldots, \widehat{\boldsymbol{\alpha}}_K^{orT})^T$ are obtained by minimizing $\widehat{L}(\boldsymbol{\zeta})$. For notational simplicity, we write $M_i^0 = \mathbf{x}_i^T\boldsymbol{\beta}^0 + f_k^0(\mathbf{z}_i)$, and $\widetilde{M}_i = \mathbf{x}_i^T\widetilde{\boldsymbol{\beta}} + \widetilde{f}_k(\mathbf{z}_i) = \mathbf{x}_i^T\widetilde{\boldsymbol{\beta}} + \mathbf{B}(\mathbf{z}_i)^T\widetilde{\boldsymbol{\alpha}}_k$ for $i \in \mathcal{G}_k$.

**Lemma 1.** *Under the conditions (C1)–(C6), $\sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \to N(\mathbf{0}, \Psi^{-1}\Sigma_1\Psi^{-1})$, where $\Psi = E(\mathbf{x}\mathbf{x}^T)$, and $\Sigma_1 = E(\epsilon^2\mathbf{x}\mathbf{x}^T)$.*

**Lemma 2.** *Under the conditions (C1)–(C6), there exist some positive constants $c_k$ and $C$ such that $\|A_{nk}^{-1}\|_2 \le c_k$ and $\|V_n^{-1}\|_2 \le C$, a.s., where $k = 1, \ldots, K$, $A_{nk} = \frac{1}{m_k}\sum_{i\in\mathcal{G}_k}\mathbf{B}(\mathbf{z}_i)\mathbf{B}(\mathbf{z}_i)^T$ and $V_n = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$.*

**Proof.** The proof of Lemmas 1 and 2 is referred to Lemma A.1 and Lemma A.2 of Liu et al. (2011).

By Taylor expansion theorem, we have

$$\widehat{\boldsymbol{\alpha}}_k - \widetilde{\boldsymbol{\alpha}}_k = -\left(\left.\frac{\partial^2\widehat{L}(\boldsymbol{\zeta})}{\partial\boldsymbol{\alpha}_k\partial\boldsymbol{\alpha}_k^T}\right|_{\boldsymbol{\alpha}_k=\bar{\boldsymbol{\alpha}}_k}\right)^{-1}\left.\frac{\partial\widehat{L}(\boldsymbol{\zeta})}{\partial\boldsymbol{\alpha}_k}\right|_{\boldsymbol{\alpha}_k=\widetilde{\boldsymbol{\alpha}}_k},$$

where $\bar{\boldsymbol{\alpha}}_k$ is between $\widehat{\boldsymbol{\alpha}}_k$ and $\widetilde{\boldsymbol{\alpha}}_k$. Take the derivative of $\widehat{L}(\boldsymbol{\zeta})$ with respect to $\boldsymbol{\alpha}_k$, $k = 1, \ldots, K$, then we yield

$$\begin{aligned}\left.\frac{\partial\widehat{L}(\boldsymbol{\zeta})}{\partial\boldsymbol{\alpha}_k}\right|_{\boldsymbol{\alpha}_k=\widetilde{\boldsymbol{\alpha}}_k} &= -\sum_{i\in\mathcal{G}_k}\left(y_i - \widetilde{M}_i\right)\mathbf{B}(\mathbf{z}_i)\\ &= \sum_{i\in\mathcal{G}_k}(M_i^0 - y_i)\mathbf{B}(\mathbf{z}_i) + \sum_{i\in\mathcal{G}_k}(\widetilde{f}_k(\mathbf{z}_i) - f_k^0(\mathbf{z}_i))\mathbf{B}(\mathbf{z}_i) + \sum_{i\in\mathcal{G}_k}\mathbf{x}_i^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\mathbf{B}(\mathbf{z}_i).\end{aligned}$$

Note that $\left\|-\frac{1}{m_k}\sum_{i\in\mathcal{G}_k}(y_i - M_i^0)\mathbf{B}(\mathbf{z}_i)\right\| = \left\{\sum_{j=1}^p\sum_{t=-\rho+1}^{J_n}(\frac{1}{m_k}\sum_{i\in\mathcal{G}_k}\epsilon_i B_{j,t}(z_{ij}))^2\right\}^{\frac{1}{2}}$,

$$E\left[\sum_{j=1}^{p}\sum_{t=-\rho+1}^{J_n}(\frac{1}{m_k}\sum_{i\in\mathcal{G}_k}\epsilon_i B_{j,t}(z_{ij}))^2\right] \le C\frac{J_n}{m_k}.$$

Thus we obtain $\left\|\frac{1}{m_k}\sum_{i\in\mathcal{G}_k}(M_i^0 - y_i)\mathbf{B}(\mathbf{z}_i)\right\| = O_p(\sqrt{J_n/m_k})$. By Condition 4 and Lemma 1, using the similar argument, we have

$$\left\|\frac{1}{m_k}\sum_{i\in\mathcal{G}_k}(\widetilde{f}_k(\mathbf{z}_i) - f_k^0(\mathbf{z}_i))\mathbf{B}(\mathbf{z}_i)\right\| = O_p(\sqrt{J_n}h^\kappa),$$

$$\left\|\frac{1}{m_k}\sum_{i\in\mathcal{G}_k}(\mathbf{x}_i^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0))\mathbf{B}(\mathbf{z}_i)\right\| = O_p(\sqrt{J_n/m_k}).$$

we consequently have that

$$\left\|\frac{1}{m_k}\left.\frac{\partial\widehat{L}(\boldsymbol{\zeta})}{\partial\boldsymbol{\alpha}}\right|_{\boldsymbol{\alpha}_k=\widetilde{\boldsymbol{\alpha}}_k}\right\| = O_p(J_n^{1/2}(h^\kappa + m_k^{-1/2})).$$

According to Lemma 2, $\|A_{nk}^{-1}\|_2 = O_p(1)$, so

$$\|\widehat{\boldsymbol{\alpha}}_k^{or} - \widetilde{\boldsymbol{\alpha}}_k\| \le \|A_{nk}^{-1}\|_2\left\|\frac{1}{m_k}\left.\frac{\partial\widehat{L}(\boldsymbol{\zeta})}{\partial\boldsymbol{\alpha}}\right|_{\boldsymbol{\alpha}_k=\widetilde{\boldsymbol{\alpha}}_k}\right\| = O_p(J_n^{1/2}(h^\kappa + m_k^{-1/2})).$$

$$\left.\frac{\partial\widehat{L}(\boldsymbol{\zeta})}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} = \sum_{i=1}^{n}(M_i^0 - y_i)\mathbf{x}_i + \sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k}(\widetilde{f}_k(\mathbf{z}_i) - f_k^0(\mathbf{z}_i))\mathbf{x}_i + \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$$

Similarly, we obtain

$$\left\| \frac{1}{n} \sum_{i=1}^{n} ((M_i^0 - y_i)\mathbf{x}_i) \right\| = O_p(\sqrt{J_n/n}),$$

$$\left\| \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{G}_k} (\widetilde{f}_k(\mathbf{z}_i) - f_k^0(\mathbf{z}_i))\mathbf{x}_i \right\| = O_p(\sqrt{J_n h^{\kappa}}),$$

$$\left\| \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i \mathbf{x}_i^T (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)) \right\| = O_p(\sqrt{J_n/n}).$$

By Lemma 2, we yield

$$\|\widehat{\boldsymbol{\beta}}^{or} - \widetilde{\boldsymbol{\beta}}\| \le \|V_n^{-1}\|_2 \left\| \frac{1}{n} \frac{\partial \widehat{L}(\boldsymbol{\zeta})}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}} \right\| = O_p(J_n^{1/2}(h^{\kappa} + n^{-1/2})).$$

As a result, we finally obtain that for $k = 1, \dots, K$,

$$\begin{aligned} \|(\widehat{f}_k^{or} - f_k^0)\|_{m_k}^2 &\le \|(\widehat{f}_k^{or} - \widetilde{f}_k)\|_{m_k}^2 + \|(\widetilde{f}_k - f_k^0)\|_{m_k}^2 \\ &\le c\|(\widehat{\boldsymbol{\alpha}}_k^{or} - \widetilde{\boldsymbol{\alpha}}_k)\|_2^2 + \|(\widetilde{f}_k - f_k^0)\|_{m_k}^2 \\ &\le O_p\left\{J_n(h^{\kappa} + m_k^{-1/2})^2\right\} + O_p(h^{2\kappa}) \\ &= O_p\{J_n/m_k\}. \end{aligned}$$

where $\|\phi\|_{m_k}^2 = m_k^{-1} \sum_{i \in \mathcal{G}_k} \phi^2(\mathbf{z}_i)$ for any measurable functions $\phi$. Thus, we have

$$\|m_1^{-1} \sum_{i \in \mathcal{G}_1} (\widehat{f}_1^{or}(\mathbf{z}_i) - f_1^0(\mathbf{z}_i))^2, \dots, m_K^{-1} \sum_{i \in \mathcal{G}_K} (\widehat{f}_K^{or}(\mathbf{z}_i) - f_K^0(\mathbf{z}_i))^2\|_{\infty} \le O_p\{J_n/|\mathcal{G}_{min}|\}.$$

Thus, $\|(\widehat{f}_1^{or} - f_1^0), \dots, (\widehat{f}_K^{or} - f_K^0)\|_{\infty} \le O_p\{J_n/|\mathcal{G}_{min}|\}$. According to Lemma 1 and the above result,

$$\begin{aligned} \|\boldsymbol{\beta}^{or} - \boldsymbol{\beta}^0\| &\le \|\boldsymbol{\beta}^{or} - \widetilde{\boldsymbol{\beta}}\| + \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| \\ &\le O_p\left\{J_n^{1/2}(h^{\kappa} + n^{-1/2})\right\} + O_p(n^{-1/2}) \\ &= O_p\left\{J_n^{1/2}(h^{\kappa} + n^{-1/2})\right\}. \end{aligned}$$

**Lemma 3.** *Under (C1)–(C6), we have*

$$\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{G}_k} (\widehat{f}_k^{or}(\mathbf{z}_i) - f_k^0(\mathbf{z}_i))\widetilde{\mathbf{x}}_i = o_p(1/\sqrt{n}), \tag{A.1}$$

$$\frac{1}{n} \sum_{i=1}^{n} \widetilde{\mathbf{x}}_i \Gamma(\mathbf{z}_i)^T (\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0) = o_p(1/\sqrt{n}). \tag{A.2}$$

**Proof.** Since the proofs of (A.1) and (A.2) are similar, here we focus on (A.1) to save space. Denote $\psi(\mathbf{z}, f_k) = f_k(\mathbf{z})\widetilde{\mathbf{x}}$, using the fact that

$$E\left[\psi(z, \widehat{f}_k^{or}) - \psi(z, f_k^0)\right]^2 = E\left[(\widehat{f}_k^{or} - f_k^0)(\mathbf{z}_i)\widetilde{\mathbf{x}}_i\right]^2 \le O(\|\widehat{f}_k^{or} - f_k^0\|^2).$$

According to Lemma A.2 of Huang (1999), we have

$$\mathcal{A}_1(\delta_k) = \left\{\psi(\cdot, \widehat{f}_k^{or}) - \psi(\cdot, f_k^0) : \|\widehat{f}_k^{or} - f_k^0\| \le \delta_k\right\} \text{ is } c\left\{(J_n + \rho)\log(\delta_k/\epsilon) + \log(\delta_k^{-1})\right\}.$$

and the corresponding entropy integral $J_{\square}(\delta_k, \mathcal{A}_1(\delta_k), \|\cdot\|_2) \le cJ_n^{1/2}\|\widehat{f}_k^{or} - f_k^0\|_2 = O_p(J_n/\sqrt{m_k})$ In addition, by Lemma 7 of Stone (1986), $\|\widehat{f}_k^{or} - f_k^0\|_{\infty} \le cJ_n^{1/2}\|\widehat{f}_k^{or} - f_k^0\|_2 = O_p(J_n/\sqrt{m_k})$ and by Lemma 3.4.2 of van der Vaart and Wellner (1996), we have that , for $\pi_k = \sqrt{m_k/J_n}$

$$
\begin{aligned}
&E\left|\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k}(\widehat{f}_k^{or}(\mathbf{z}_i) - f_k^0(\mathbf{z}_i))\widetilde{\mathbf{x}}_i - E[(\widehat{f}_k^{or}(\mathbf{z}_i) - f_k^0(\mathbf{z}_i))\widetilde{\mathbf{x}}_i]\right| \\
&\le \frac{1}{\sqrt{n}}Kc_1\pi_k^{-1}((J_n+\rho)^{1/2} + \log^{1/2}(\pi_k))\left[1 + \frac{c_2\pi_k^{-1}((J_n+\rho)^{1/2} + \log^{1/2}(\pi_k))}{\pi_k^{-2}\sqrt{m_k}}c_3\right] \\
&\le O(1)n^{-1/2}\pi_k^{-1}((J_n+\rho)^{1/2} + \log^{1/2}(\pi_k)).
\end{aligned}
$$

By Condition (C1) that $J_n \le |\mathcal{G}_{min}|^{1/3}$, lead to $\pi_k^{-1}J_n^{1/2} = \frac{J_n}{\sqrt{m_k}} \le \frac{|\mathcal{G}_{min}|^{1/3}}{\sqrt{m_k}} \to 0$, as $n \to \infty$. Thus, $O(\pi_k^{-1}J_n^{1/2}) = o(1)$, then we have

$$
E\left|\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k}(\widehat{f}_k^{or}(\mathbf{z}_i) - f_k^0(\mathbf{z}_i))\widetilde{\mathbf{x}}_i - E[(\widehat{f}_k^{or}(\mathbf{z}_i) - f_k^0(\mathbf{z}_i))\widetilde{\mathbf{x}}_i]\right| = o(1/\sqrt{n}).
$$

This together with $E[(\widehat{f}_k^{or}(\mathbf{z}_i) - f_k^0(\mathbf{z}_i))\widetilde{\mathbf{x}}_i] = 0$, leads to (A.1).

By the condition (C4) and the above results, there exists an empirically centered additive function $\widetilde{\Gamma}_{kj}$ such that $\|\widetilde{\Gamma}_{kj} - \Gamma_{kj}\|_\infty = O_p(h^\kappa)$. Define a function class $\mathcal{M}_n = \left\{M(\mathbf{x},\mathbf{z}) = f_k(\mathbf{z}) + \mathbf{x}^T\boldsymbol{\beta}, f_k \in \mathcal{G}, k = 1,\ldots,K\right\}$. For simplicity of notation, denote $\widehat{M}_i = \widehat{M}(\mathbf{x}_i,\mathbf{z}_i) = \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}^{or} + \widehat{f}_k^{or}(\mathbf{z}_i) = \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}^{or} + \mathbf{B}(\mathbf{z}_i)^T\widehat{\boldsymbol{\alpha}}_k^{or}$, and $\widehat{M}_{\boldsymbol{\nu}i} = \widehat{M}_i + \boldsymbol{\nu}^T(\mathbf{x}_i - \widetilde{\Gamma}_k(\mathbf{z}_i))$, and $\widetilde{\mathbf{x}}_i = \mathbf{x}_i - \Gamma_k(\mathbf{z}_i)$, where $i \in \mathcal{G}_k$, and $k = 1,\ldots,K$. When $\boldsymbol{\nu} = 0$, the function $\widehat{M}_{\boldsymbol{\nu}i}$ minimizes $L_n(M) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k}(y_i - M(\mathbf{x}_i,\mathbf{z}_i))^2$, for all $M(\mathbf{x},\mathbf{z}) \in \mathcal{M}_n$. Accordingly, we have

$$
\begin{aligned}
\mathbf{0} = \left.\frac{\partial\widehat{L}(\widehat{M}_{\boldsymbol{\nu}})}{\partial\boldsymbol{\nu}}\right|_{\boldsymbol{\nu}=\mathbf{0}} &= -\sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k}(y_i - \widehat{M}_i)(\mathbf{x}_i - \Gamma_k(\mathbf{z}_i)) \\
&= -\sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k}(y_i - \widehat{M}_i)\widetilde{\mathbf{x}}_i + O_p(h^\kappa) \\
&= -\sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k}\epsilon_i\widetilde{\mathbf{x}}_i + \sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k}(\widehat{M}_i - M_i^0)\widetilde{\mathbf{x}}_i + O_p(h^\kappa).
\end{aligned}
$$

Note that the second term $\sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k}(\widehat{M}_i - M_i^0)\widetilde{\mathbf{x}}_i$ is rewritten as

$$
(\sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^T)(\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0) + (\sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k}\widetilde{\mathbf{x}}_i\Gamma_k(\mathbf{z}_i)^T)(\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0) + \sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k}(\widehat{f}_k^{or}(\mathbf{z}_i) - f_k^0(\mathbf{z}_i))\widetilde{\mathbf{x}}_i.
$$

This in conjunction with Lemma 3, yields

$$
\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k}(\widehat{M}_i - M_i^0)\widetilde{\mathbf{x}}_i = \left\{E(\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^T) + o_p(1)\right\}(\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0) + o_p(\frac{1}{\sqrt{n}}).
$$

By Condition 4, one has

$$
\mathbf{0} = -\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k}\epsilon_i\widetilde{\mathbf{x}}_i + \left\{E(\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^T) + o_p(1)\right\}(\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0) + o_p(\frac{1}{\sqrt{n}}).
$$

We consequently have that

$$
\sqrt{n}(\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0) \to N(0, \Phi^{-1}\Sigma\Phi^{-1}), \quad \Phi = E[\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}^T], \quad \Sigma = E[\epsilon^2\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}^T].
$$

### *A.4. Proof of Theorem 2*

Let $\|\cdot\|$ be the Euclidean norm. In addition, given $\mathbf{z} = (z_1,\ldots,z_p)^T$, define $\mathcal{G}$ as the collection of functions $g(\mathbf{z}) = \sum_{j=1}^{p}g_j(z_j)$, $g_j \in \mathcal{H}$, and $E[g_j(z_j)] = 0$. Assume $\mathbf{B}_j(\mathbf{z}_{ij}), j = 1,\ldots,p$ is the centered version spline basis (see Liu et al., 2011), which is convenient for asymptotic analysis.

$$\begin{aligned}
&\|((\widehat{\boldsymbol{\xi}}_1^{or}-\boldsymbol{\xi}_1^0)^T,\ldots,(\widehat{\boldsymbol{\xi}}_K^{or}-\boldsymbol{\xi}_K^0)^T)^T\|_\infty\\
&=\left\|\left(\frac{1}{m_1}\sum_{i\in\mathcal{G}_1}(\widetilde{\mathbf{B}}_{\mathbf{z}_i}\widetilde{y}_i-E[\widetilde{\mathbf{B}}_{\mathbf{z}_i}\widetilde{y}_i])^T,\ldots,\frac{1}{m_K}\sum_{i\in\mathcal{G}_K}(\widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i-E[\widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i])^T\right)^T\right\|_\infty\\
&\le\frac{1}{|\mathcal{G}_{min}|}\left\|\left(\sum_{i\in\mathcal{G}_1}(\widetilde{\mathbf{B}}_{\mathbf{z}_i}\widetilde{y}_i-E[\widetilde{\mathbf{B}}_{\mathbf{z}_i}\widetilde{y}_i])^T,\ldots,\sum_{i\in\mathcal{G}_K}(\widetilde{\mathbf{B}}_{\mathbf{z}_i}\widetilde{y}_i-E[\widetilde{\mathbf{B}}_{\mathbf{z}_i}\widetilde{y}_i])^T\right)^T\right\|_\infty.
\end{aligned}$$

Denote the $l$-element of $\widetilde{\mathbf{B}}_{\mathbf{z}i}$ by $\widetilde{B}_{zil}$, $l=1,\ldots,N_np$.

$$\begin{aligned}
&P\left(\left\|\left(\sum_{i\in\mathcal{G}_1}(\widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i-E[\widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i])^T,\ldots,\sum_{i\in\mathcal{G}_K}(\widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i-E[\widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i])^T\right)^T\right\|_\infty>c\sqrt{n\log n}\right)\\
&\le\sum_{k=1}^{K}\sum_{l=1}^{Np}P\left(\left|\sum_{i\in\mathcal{G}_k}(\widetilde{B}_{zil}\widetilde{y}_i-E[\widetilde{B}_{zil}\widetilde{y}_i])\right|>c\sqrt{n\log n}\right)\\
&\le\sum_{k=1}^{K}\sum_{l=1}^{Np}P\left(\left|\sum_{i\in\mathcal{G}_k}(\widetilde{B}_{zil}\widetilde{y}_i-E[\widetilde{B}_{zil}\widetilde{y}_i])\right|>c\sqrt{m_k\log n}\right)\\
&\le\sum_{k=1}^{K}\sum_{l=1}^{Np}P\left(\left|\frac{1}{m_k}\sum_{i\in\mathcal{G}_k}\widetilde{B}_{zil}\widetilde{y}_i-E[\widetilde{B}_{zil}\widetilde{y}_i]\right|>c\sqrt{\log n}/\sqrt{m_k}\right).
\end{aligned}$$

By Condition (C3) that $\sup_{il}|\widetilde{\mathbf{B}}_{\mathbf{z}_{il}}|\le M_1$ for $1\le i\le n$ and $1\le l\le N_np$, then we have $|\widetilde{\mathbf{B}}_{\mathbf{z}_{il}}\widetilde{y}_i|\le M_1$. By Bernstein's Inequality (Lemma 2.2.9, van der Vaart and Wellner, 1996) in conjunction with the union bound of probability, for any $c>0$, there exist some positive constants $c_1>0$, $c_2>0$ such that

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{l=1}^{Np}P\left(\left|\frac{1}{m_k}\sum_{i\in\mathcal{G}_k}\widetilde{B}_{zil}\widetilde{y}_i-E[\widetilde{B}_{zil}\widetilde{y}_i]\right|>c\sqrt{\log n}/\sqrt{m_k}\right)\\
&\le 2KN_np\exp(-c_1cm_k\log n/m_k)\\
&=2KN_npn^{-c_2}.
\end{aligned}$$

Thus, we have

$$\|((\widehat{\boldsymbol{\xi}}_1^{or}-\boldsymbol{\xi}_1^0)^T,\ldots,(\widehat{\boldsymbol{\xi}}_K^{or}-\boldsymbol{\xi}_K^0)^T)^T\|_\infty\le c|\mathcal{G}_{min}|^{-1}\sqrt{n\log n}.$$

### *A.5. Proof of Theorem 3*

The proof of Theorem 3 is referred to Theorem 4.2 of Ma and Huang (2016). Define

$$\begin{aligned}
&L_n(\boldsymbol{\theta})=\frac{1}{2}\sum_{i=1}^{n}\left(\widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i-\boldsymbol{\theta}_i\right)^2,\quad P_n(\boldsymbol{\theta})=\sum_{i<k}\rho(\|\boldsymbol{\theta}_i-\boldsymbol{\theta}_k\|,\lambda),\\
&L_n^{\mathcal{G}}(\boldsymbol{\xi})=\frac{1}{2}\sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k}\left(\widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i-\boldsymbol{\xi}_i\right)^2,\quad P_n^{\mathcal{G}}(\boldsymbol{\xi})=\sum_{i<k}m_im_k\rho(\|\boldsymbol{\xi}_i-\boldsymbol{\xi}_k\|,\lambda),
\end{aligned}$$

and denote

$$Q_n(\boldsymbol{\theta})=L_n(\boldsymbol{\theta})+P_n(\boldsymbol{\theta}),\quad Q_n^{\mathcal{G}}(\boldsymbol{\theta})=L_n^{\mathcal{G}}(\boldsymbol{\theta})+P_n^{\mathcal{G}}(\boldsymbol{\theta}).$$

Denote $\mathcal{M}_{\mathcal{G}}=\left\{\boldsymbol{\theta}:\boldsymbol{\theta}_i=\boldsymbol{\theta}_j,\ \text{for any } i,j\in\mathcal{G}_k, 1\le k\le K\right\}$. Let $T:\mathcal{M}_{\mathcal{G}}\to R^K$ be the mapping such that $T(\boldsymbol{\theta})$ is the $K$ vector whose $k$th coordinate equals the common value of $\theta_i$ for $i\in\mathcal{G}_k$. and Let $T^*:R^n\to R^K$ be the mapping such that $T^*(\boldsymbol{\theta})=\left\{|\mathcal{G}_k|^{-1}\sum_{i\in\mathcal{G}_k}\boldsymbol{\theta}_i\right\}_{k=1}^K$. We can see $T(\boldsymbol{\theta})=T^*(\boldsymbol{\theta})$ when $\boldsymbol{\theta}\in\mathcal{M}_{\mathcal{G}}$. By calculation, for $\boldsymbol{\theta}\in\mathcal{M}_{\mathcal{G}}$, we have $P_n(\boldsymbol{\theta})=P_n^{\mathcal{G}}(T(\boldsymbol{\theta}))$ and for $\boldsymbol{\delta}\in R^K$, $P_n(T^{-1}(\boldsymbol{\delta}))=P_n^{\mathcal{G}}(\boldsymbol{\delta})$. Thus, we obtain

$$Q_n(\boldsymbol{\theta})=Q_n^{\mathcal{G}}(T(\boldsymbol{\theta})),\ Q_n^{\mathcal{G}}(\boldsymbol{\delta})=Q_n(T^{-1}(\boldsymbol{\delta})).\tag{A.3}$$

By Theorem 2, there is an event $E_1$ such that $\|((\widehat{\boldsymbol{\xi}}_1^{or}-\boldsymbol{\xi}_1^0)^T, \ldots, (\widehat{\boldsymbol{\xi}}_K^{or}-\boldsymbol{\xi}_K^0)^T)^T\|_\infty \le \psi_n$, and $\|((\widehat{\boldsymbol{\theta}}_1^{or}-\boldsymbol{\theta}_1^0)^T, \ldots, (\widehat{\boldsymbol{\theta}}_n^{or}-\boldsymbol{\theta}_n^0)^T)^T\|_\infty \le \psi_n$ where $\psi_n = c|\mathcal{G}_{min}|^{-1}\sqrt{n\log n}$, and $c$ is some positive constant. Moreover, $\sup_i \|\widehat{\boldsymbol{\theta}}_i^{or} - \boldsymbol{\theta}_i^0\| \le \sqrt{N_n p}\psi_n$. Write $\sqrt{N_n p}\psi_n = \phi_n$. Consider the neighborhood of $\boldsymbol{\theta}^0$:

$$\Theta = \left\{\boldsymbol{\theta} : \sup_i \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^0\| \le \phi_n\right\}.$$

Thus, $((\widehat{\boldsymbol{\theta}}_1^{or})^T, \ldots, (\widehat{\boldsymbol{\theta}}_n^{or})^T)^T \in \Theta$ on the event $E_1$. For any $\boldsymbol{\theta}$, denote $\boldsymbol{\theta}^* = T^{-1}(T^*(\boldsymbol{\theta}))$. Then $((\widehat{\boldsymbol{\theta}}_1^{or})^T, \ldots, (\widehat{\boldsymbol{\theta}}_n^{or})^T)^T$ is a strictly local minimizer of the objective function (2.4) with probability approaching 1 through the following two steps:

(1) On the event $E_1$, $Q_n(\boldsymbol{\theta}^*) > Q_n(\widehat{\boldsymbol{\theta}}^{or})$ for any $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\theta}^* \neq \widehat{\boldsymbol{\theta}}^*$.

(2) There is an event $E_2$ such that on $E_1 \cap E_2$, there is a neighborhood of $\widehat{\boldsymbol{\theta}}^{or}$, denoted by $\Theta_n$, such that $Q_n(\boldsymbol{\theta}) \ge Q_n(\boldsymbol{\theta}^*)$ for any $\boldsymbol{\theta} \in \Theta_n \cap \Theta$ for sufficiently large $n$.

Thus, by the result in (1) and (2), we have $Q_n(\boldsymbol{\theta}) > Q_n(\widehat{\boldsymbol{\theta}}^{or})$ for any $\boldsymbol{\theta} \in \Theta_n \cap \Theta$ and $\boldsymbol{\theta} \neq \widehat{\boldsymbol{\theta}}^*$, so $((\widehat{\boldsymbol{\theta}}_1^{or})^T, \ldots, (\widehat{\boldsymbol{\theta}}_n^{or})^T)^T$ is a strictly local minimizer of $Q_n(\boldsymbol{\theta})$ given (2.4) on $E_1 \cap E_2$.

In the following we prove the result (1). We first show $P_n^{\mathcal{G}}(T^*(\boldsymbol{\theta})) = C_n$, for any $\boldsymbol{\theta} \in \Theta$ where $C_n$ is a constant. Let $T^*(\boldsymbol{\theta}) = \boldsymbol{\xi} = (\boldsymbol{\xi}_1^T, \ldots, \boldsymbol{\xi}_K^T)^T$ and $\boldsymbol{\xi}^0 = (\boldsymbol{\xi}_1^{0T}, \ldots, \boldsymbol{\xi}_K^{0T})^T$. It suffices to show that $\|\boldsymbol{\xi}_k - \boldsymbol{\xi}_{k'}\| > a\lambda$ for all $k$ and $k'$. Then by Condition (C5) which the concave penalties such as MCP and SCAD satisfy, we have $\rho(\|\boldsymbol{\xi}_k - \boldsymbol{\xi}_{k'}\|)$ is a constant, and as a result $P_n^{\mathcal{G}}(T^*(\boldsymbol{\theta}))$ is a constant. Since $\|\boldsymbol{\xi}_k - \boldsymbol{\xi}_{k'}\| \ge \|\boldsymbol{\xi}_k^0 - \boldsymbol{\xi}_{k'}^0\| - 2\sup_k \|\boldsymbol{\xi}_k - \boldsymbol{\xi}_k^0\|$, and

$$\begin{aligned}\sup_k \|\boldsymbol{\xi}_k - \boldsymbol{\xi}_k^0\|^2 &= \sup_k \|\sum_{i\in\mathcal{G}_k} \boldsymbol{\theta}_i/|\mathcal{G}_k| - \boldsymbol{\xi}_k^0\|^2 = \sup_k \|\sum_{i\in\mathcal{G}_k}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^0)/|\mathcal{G}_k|\|^2\\ &\le \sup_k |\mathcal{G}_k|^{-1}\sum_{i\in\mathcal{G}_k} \|(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^0)\|^2 \le \sup_i \|(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^0)\|^2 \le \phi_n^2,\end{aligned}$$

then by the assumption $a_n > a\lambda \gg \phi_n$ for all $k$ and $k'$, $\|\boldsymbol{\xi}_k - \boldsymbol{\xi}_{k'}\| \ge a_n - 2\phi_n > a\lambda$. so we have $P_n^{\mathcal{G}}(T^*(\boldsymbol{\theta})) = C_n$, and $Q_n^{\mathcal{G}}(T^*(\theta)) = L_n^{\mathcal{G}}(T^*(\theta)) + C$ for all $\boldsymbol{\theta} \in \Theta$. Since $\widehat{\boldsymbol{\xi}}^{or}$ is the unique global minimizer of $L_n^{\mathcal{G}}(\boldsymbol{\xi})$, then $L_n^{\mathcal{G}}(T^*(\theta)) > L_n^{\mathcal{G}}(\widehat{\boldsymbol{\xi}}^{or})$ for all $T^*(\boldsymbol{\theta}) \neq \widehat{\boldsymbol{\xi}}^{or}$ and thus $Q_n^{\mathcal{G}}(T^*(\theta)) > Q_n^{\mathcal{G}}(\widehat{\boldsymbol{\xi}}^{or})$ for all $T^*(\boldsymbol{\theta}) \neq \widehat{\boldsymbol{\xi}}^{or}$. By (A.3), we have $Q_n^{\mathcal{G}}(\widehat{\boldsymbol{\xi}}^{or}) = Q_n(\widehat{\boldsymbol{\theta}}^{or})$ and $Q_n^{\mathcal{G}}(T^*(\boldsymbol{\theta})) = Q_n(T^{-1}(T^*(\boldsymbol{\theta}))) = Q_n(\boldsymbol{\theta}^*)$. Thus, $Q_n(\boldsymbol{\theta}^*) > Q_n(\widehat{\boldsymbol{\theta}}^{or})$ for all $\boldsymbol{\theta}^* \neq \widehat{\boldsymbol{\theta}}^{or}$, and the result (1) is proved.

Next we prove the result in (2). For a positive sequence $t_n$, let $\Theta_n = \left\{\boldsymbol{\theta}_i : \sup_i \|\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_i^{or}\| \le t_n\right\}$. For $\boldsymbol{\theta} \in \Theta_n \cap \Theta$, by Taylor's expansion, we have

$$Q_n(\boldsymbol{\theta}) - Q_n(\boldsymbol{\theta}^*) = \Gamma_1 + \Gamma_2,$$

where

$$\begin{aligned}\Gamma_1 &= -\sum_{i=1}^n \left(\widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{y}_i - \boldsymbol{\theta}_i^m\right)(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*),\\ \Gamma_2 &= \sum_{i=1}^n \frac{\partial P_n(\boldsymbol{\theta}^m)}{\partial \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*).\end{aligned}$$

in which $\boldsymbol{\theta}_i^m = \varsigma\boldsymbol{\theta}_i + (1-\varsigma)\boldsymbol{\theta}_i^*$ for some $\varsigma \in (0, 1)$. Moreover,

$$\begin{aligned}\Gamma_2 &= \lambda\sum_{j>i} \rho'(\|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m\|)\|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m\|^{-1}(\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m)^T(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*)\\ &+ \lambda\sum_{j<i} \rho'(\|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m\|)\|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m\|^{-1}(\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m)^T(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*)\\ &= \lambda\sum_{j>i} \rho'(\|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m\|)\|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m\|^{-1}(\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m)^T(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*)\\ &+ \lambda\sum_{i<j} \rho'(\|\boldsymbol{\theta}_j^m - \boldsymbol{\theta}_i^m\|)\|\boldsymbol{\theta}_j^m - \boldsymbol{\theta}_i^m\|^{-1}(\boldsymbol{\theta}_j^m - \boldsymbol{\theta}_i^m)^T(\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*)\\ &= \lambda\sum_{j>i} \rho'(\|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m\|)\|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m\|^{-1}(\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m)^T[(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*) - (\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*)],\end{aligned}$$

When $i, j \in \mathcal{G}_k$, $\boldsymbol{\theta}_i^* = \boldsymbol{\theta}_j^*$, and $\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m$ has the same sign as $\boldsymbol{\theta}_i - \boldsymbol{\theta}_j$. Thus,

$$\begin{aligned}\Gamma_2 &= \lambda\sum_{k=1}^K \sum_{i,j\in\mathcal{G}_k, i<j} \rho'(\|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m\|)\|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m\|^{-1}(\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m)^T(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)\\ &+ \lambda\sum_{k<k'} \sum_{i\in\mathcal{G}_k, j\in\mathcal{G}_{k'}} \rho'(\|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m\|)\|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m\|^{-1}(\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m)^T[(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*) - (\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*)],\end{aligned}$$

As shown in the result (1), $\sup_i \|\boldsymbol{\theta}_i^* - \boldsymbol{\theta}_i^0\|^2 = \sup_k \|\boldsymbol{\xi}_k - \boldsymbol{\xi}_k^0\|^2 \le \phi_n^2$, since $\boldsymbol{\theta}_i^m = \varsigma\boldsymbol{\theta}_i + (1-\varsigma)\boldsymbol{\theta}_i^*$,

$$\sup_i \|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_i^0\| \le \varsigma\phi_n + (1-\varsigma)\phi_n = \phi_n. \tag{A.4}$$

then for $k \ne k'$, $i \in \mathcal{G}_k, j \in \mathcal{G}_{k'}$,

$$\|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m\| \ge \min_{i\in\mathcal{G}_k, j\in\mathcal{G}_{k'}} \|\boldsymbol{\theta}_i^0 - \boldsymbol{\theta}_j^0\| - 2\max_i \|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_i^0\| \ge a_n - 2\phi_n > a\lambda,$$

thus $\rho'(\|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m\|) = 0$. Thus,

$$\Gamma_2 = \lambda \sum_{k=1}^{K} \sum_{i,j\in\mathcal{G}_k, i<j} \rho'(\|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m\|)\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|.$$

By the same reasoning as (A.3), we obtain $\sup_i \|\boldsymbol{\theta}_i^* - \widehat{\boldsymbol{\theta}}_i^{or}\| \le \sup_i \|\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_i^{or}\|$. so

$$\begin{aligned}\sup_i \|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m\| &\le 2\sup_i \|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_i^*\| \le 2\sup_i \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*\| \\ &\le 2(\sup_i \|\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_i^{or}\| + \sup_i \|\boldsymbol{\theta}_i^* - \widehat{\boldsymbol{\theta}}_i^{or}\|) \le 4\sup_i \|\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_i^{or}\| \le 4t_n.\end{aligned}$$

Thus, $\rho'(\|\boldsymbol{\theta}_i^m - \boldsymbol{\theta}_j^m\|) \ge \rho'(4t_n)$ by concavity of $\rho(\cdot)$. As a result,

$$\Gamma_2 \ge \lambda \sum_{k=1}^{K} \sum_{i,j\in\mathcal{G}_k, i<j} \rho'(4t_n)\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|. \tag{A.5}$$

$$\begin{aligned}\Gamma_1 = -\mathbf{w}^T(\boldsymbol{\theta} - \boldsymbol{\theta}^*) &= -\sum_{k=1}^{K} \sum_{i,j\in\mathcal{G}_k} \frac{\mathbf{w}_i^T(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)}{|\mathcal{G}_k|} \\ &= -\sum_{k=1}^{K} \sum_{i,j\in\mathcal{G}_k} \frac{\mathbf{w}_i^T(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)}{2|\mathcal{G}_k|} - \sum_{k=1}^{K} \sum_{i,j\in\mathcal{G}_k} \frac{\mathbf{w}_i^T(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)}{2|\mathcal{G}_k|} \\ &= -\sum_{k=1}^{K} \sum_{i,j\in\mathcal{G}_k} \frac{(\mathbf{w}_j - \mathbf{w}_i)^T(\boldsymbol{\theta}_j - \boldsymbol{\theta}_i)}{2|\mathcal{G}_k|} \\ &= -\sum_{k=1}^{K} \sum_{i,j\in\mathcal{G}_k, i<j} \frac{(\mathbf{w}_j - \mathbf{w}_i)^T(\boldsymbol{\theta}_j - \boldsymbol{\theta}_i)}{|\mathcal{G}_k|}.\end{aligned}$$

where $\mathbf{w} = (\mathbf{w}_1^T, \dots, \mathbf{w}_n^T)^T = \left[(\widetilde{\mathbf{B}}_{\mathbf{z}1}\widetilde{y}_1 - \boldsymbol{\theta}_1^m)^T, \dots, (\widetilde{\mathbf{B}}_{\mathbf{z}n}\widetilde{y}_n - \boldsymbol{\theta}_n^m)^T\right]^T$. Write $\widetilde{\boldsymbol{\epsilon}} = \mathbf{Q}_x\boldsymbol{\epsilon} \in R^{n\times 1}$, since $\mathbf{w}_i = \widetilde{\mathbf{B}}_{\mathbf{z}i}(\widetilde{\boldsymbol{\epsilon}}_i + \widetilde{\mathbf{B}}_{\mathbf{z}i}^T(\boldsymbol{\gamma}_i^0 - \boldsymbol{\gamma}_i^m)) = \widetilde{\mathbf{B}}_{\mathbf{z}i}\widetilde{\boldsymbol{\epsilon}}_i + \boldsymbol{\theta}_i^0 - \boldsymbol{\theta}_i^m$, and then $\sup_i \|\mathbf{w}_i\| \le \sup_i \left\{\|\widetilde{\mathbf{B}}_{\mathbf{z}i}\|\|\mathbf{Q}_{xi}\|\|\boldsymbol{\epsilon}\|_\infty + \|\boldsymbol{\theta}_i^0 - \boldsymbol{\theta}_i^m\|\right\}$, By Condition (C6),

$$P(\|\boldsymbol{\epsilon}\|_\infty > \sqrt{2c_1^{-1}}\sqrt{\log n}) \le \sum_{i=1}^{n} P(|\epsilon_i| > \sqrt{2c_1^{-1}}\sqrt{\log n}) \le 2n^{-1}.$$

Thus there is an event $E_2$ such that $\sup_i \|\mathbf{w}_i\| \le \sqrt{N_n p}M_1\sqrt{n}M_2 2c_1^{-1}\sqrt{\log n} + \phi_n$. Then

$$\begin{aligned}&\frac{(\mathbf{w}_j - \mathbf{w}_i)^T(\boldsymbol{\theta}_j - \boldsymbol{\theta}_i)}{|\mathcal{G}_k|} \\ \le& |\mathcal{G}_{min}|^{-1}\|\mathbf{w}_j - \mathbf{w}_i\|\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_i\| \le 2|\mathcal{G}_{min}|^{-1}\sup_i \|\mathbf{w}_i\|\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_i\| \\ \le& 2|\mathcal{G}_{min}|^{-1}(\sqrt{N_n p}M_1\sqrt{n}M_2\sqrt{2c_1^{-1}}\sqrt{\log n} + \phi_n)\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_i\|.\end{aligned}$$

Since $\phi_n = c\sqrt{N_n p}|\mathcal{G}_{min}|^{-1}\sqrt{n\log n}$ and $p = o(n)$, $\lambda \gg \phi_n$. Let $t_n = o(1)$, then $\rho'(4t_n) \to 1$. Therefore, we have

$$\begin{aligned}&Q_n(\boldsymbol{\theta}) - Q_n(\boldsymbol{\theta}^*) = \Gamma_1 + \Gamma_2 \\ \ge& \sum_{k=1}^{K} \sum_{i,j\in\mathcal{G}_k, i<j} [\lambda\rho'(4t_n) - 2|\mathcal{G}_{min}|^{-1}(\sqrt{N_n p}M_1\sqrt{n}M_2\sqrt{2c_1^{-1}}\sqrt{\log n} + \phi_n)]\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\| \ge 0,\end{aligned}$$

for sufficiently large $n$, so that the result (2) is proved.

## References

de Boor, C., 2001. A Practical Guide To Splines, revised ed. Springer, New York.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. 3 (1122).

Carroll, R.J., Fan, J., Gijbels, I., Wand, M.P., 1997. Generalized partially linear single-index models. J. Amer. Statist. Assoc. 92, 477–489.

Chi, E.C., Lange, K., 2014. Splitting methods for convex clustering. J. Comput. Graph. Statist. 24, 994–1013.

Elmi, A., Ratcliffe, S.J., Parrey, S., Guo, W.S., 2011. A B-spline based semiparametric nonlinear mixed effects model. J. Comput. Graph. Statist. 20, 492–509.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360.

Guo, F.J., Levina, E., Michailidis, G., Zhu, J., 2010. Pairwise variable selection for high-dimensional model-based clustering. Biometrics 66, 793–804.

He, X., Shi, P., 1994. Convergence rate of b-spline estimators of nonparametric conditional quantile functions. J. Nonparametr. Stat. 3, 299–308.

He, X., Zhu, Z., Fung, W.K., 2002. Estimationin a semiparametric model for longitudinal data with unspecidied dependence structure. Biometrika 89, 579–590.

Huang, J., 1999. Efficient estimation of the partly linear additive cox model. Ann. Statist. 27, 1536–1563.

Huang, J., Wu, C., Zhou, L., 2002. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. Biometrika 89, 111–128.

Ke, T., Fan, J., Wu, Y., 2013. Homogeneity in regression. J. Amer. Statist. Assoc. 110, 175–194.

Li, Q., 2000. Efficient estimation of additive partially linear models. Internat. Econom. Rev. 41 (4), 1073–1092.

Liu, X., Wang, L., Liang, H., 2011. Estimation and variable selection for semiparametric additive partial linear models. Statist. Sin. 21, 1225–1248.

Ma, S., Huang, J., 2016. Estimating subgroup-specific treatment effects via concave fusion. (Accepted Manuscript).

Ma, S., Huang, J., 2017. A concave pairwise fusion approach to subgroup analysis. J. Amer. Statist. Assoc. 112, 410–423.

Pan, Wei, Xiaotong, Shen, Binghui, Liu, 2013. Cluster analysis: unsupervised learning via supervised learning with a non-convex penalty. J. Mach. Learn. Res. 14 (1), 1865–1889.

Schumaker, L.L., 1981. Spline Functions: Basic Theory. Wiley, New York.

Shen, J., He, X., 2015. Inference for subgroup analysis with a structured logistic-normal mixture model. J. Amer. Statist. Assoc. 110, 303–312.

Shen, X., Huang, H.C., 2010. Grouping pursuit through a regularization solution surface. J. Amer. Statist. Assoc. 105, 727–739.

Sherwood, B., Wang, L., 2016. Partially linear additive quantile regression in ultra-high dimension. Ann. Statist. 44 (1), 288–317.

Stone, C.J., 1985. Additive regression and other nonparametric models. Ann. Statist. 13, 689–705.

Stone, C.J., 1986. The dimensionality reduction principle for generalized additive models. Ann. Statist. 14, 590–606.

Tibshirani, S., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. J. R. Statist. Soc. B 67, 91–108.

van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes. Springer, New York.

Wang, H., Li, R., Tsai, C.L., 2007. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika 94, 553–568.

Wang, L., Liu, X., Liang, H., Carroll, R.J., 2011. Estimation and variable selection for generalized additive partial linear models. Ann. Statist. 39, 1827–1851.

Wang, T., Xia, Y., 2014. A piecewise single-index model for dimension reduction. Technometrics 56, 312–324.

You, Z., Si, Y.W., Zhang, D., Zeng, X.X., Leung, S.C.H., Li, T., 2015. A decision-making framework for precision marketing. Expert Syst. Appl. 42, 3357–3367.

Zhang, C., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. 38, 894–942.