Journal of Multivariate Analysis 192 (2022) 105100

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Subgroup analysis for high-dimensional functional regression

Xiaochen Zhang [a], Qingzhao Zhang [b,c], Shuangge Ma [d], Kuangnan Fang [b,*]

[a] Zhongtai Securities Institute for Financial Studies, Shandong University, China
[b] Department of Statistics and Data Science, School of Economics, Xiamen University, China
[c] The Wang Yanan Institute for Studies in Economics, Xiamen University, China
[d] Department of Biostatistics, Yale University, United States of America

ARTICLE INFO

Article history:
Received 9 December 2021
Received in revised form 31 August 2022
Accepted 31 August 2022
Available online 6 September 2022

AMS 2020 subject classifications:
primary 62R10
secondary 62H25

Keywords:
Functional data analysis
High-dimensional functional predictors
Subgroup analysis

ABSTRACT

Subgroup analysis for scalar data has been well studied in the literature. However, less has been done on functional data, especially on high-dimensional functional regression. In this study, we develop a high-dimensional functional regression model for simultaneous estimation and subgroup identification for a heterogeneous population. Under mild conditions, we establish the estimation and selection consistency of the proposed estimators. The proposed analysis allows the number of functional predictors and number of subgroups to increase as the sample size increases. Simulation studies demonstrate satisfactory performance of the proposed method, and it is also illustrated through a real application.

1. Introduction

Functional linear regression is a popular technique when predictors are functions and responses are scalars. It has been thoroughly studied and extensively applied. A non-exhaustive list of recent works include [1,4,14,23,34,35].

Multiple functional data arises from a collection of simultaneous recordings of several time courses for many subjects or units. There is a demand for functional variable selection in applications. Zhu et al. [39] proposed a Bayesian approach for selecting and estimating important functional predictors in a classification setting. Lian [15] conducted regression simultaneous estimation and variable selection by using the group smoothly clipped absolute deviation penalty. Fan et al. [5] proposed an additive technique for efficiently performing high-dimensional functional regression. Kong et al. [12] proposed partially functional linear models to characterize the relationship between a scalar response and both functional and scalar covariates. The methods mentioned above rely on a homogeneity assumption.

Our motivating example is a thin-film transistor liquid crystal display (TFT-LCD) dataset. The manufacturing process of TFT-LCD is comprised of hundreds of working procedures. It is challenging to conduct statistical analysis based on that large amount of raw data to get useful information for supporting operational decisions [8,25]. What we are interested in is which predictors may affect the thickness of the product. Fig. 1 shows the histogram of thickness after adjusting for the functional covariates' effects without subgroup analysis. It is easy to see that the data comes from a mixture of populations, and some unobserved latent factors may cause heterogeneity. It is not suitable to fit a standard regression model with a common intercept. This has motivated us to develop a new statistical method for simultaneous estimation and subgroup identification for a heterogeneous population.

[*] Corresponding author.
E-mail address: xmufkn@xmu.edu.cn (K. Fang).

https://doi.org/10.1016/j.jmva.2022.105100

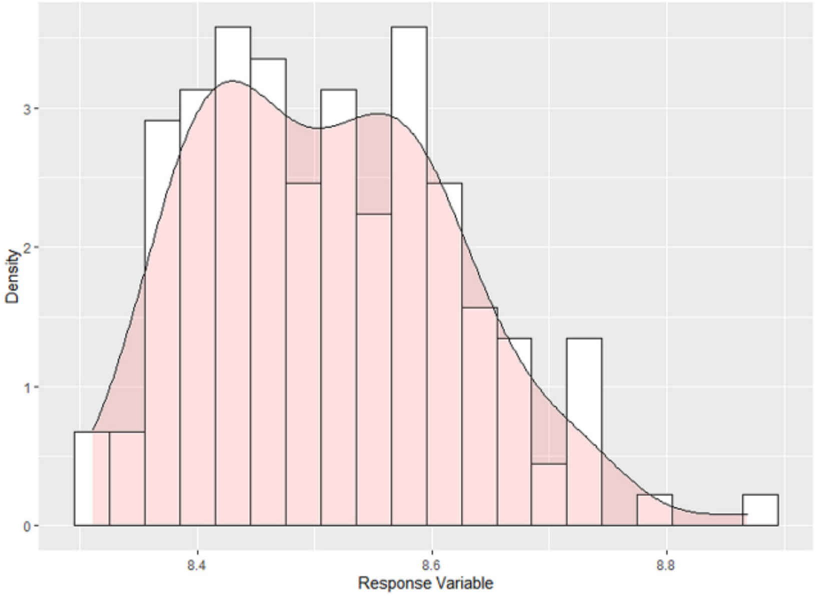

Fig. 1. Data analysis of TFT-LCD dataset: density plot of the response variable after adjusting for the covariates' effects by using the gmcp method. The dataset is collected from 149 subjects with 56 functional variables and a scalar response.

Subgroup analysis for scalar data has been well studied in the literature. Finite mixture modeling is one of the popular methods for analyzing data from a heterogeneous population [24,31]. Apart from the approaches mentioned above, Ma and Huang [20] and Zhang et al. [36] assumed that the subgroup structure is defined by group-specific intercepts. Ma et al. [21] considered a heterogeneous regression model with subject-dependent coefficients. Liu and Lin [17] proposed a heterogeneous additive partially linear model, which contains both homogeneous linear components and subject-dependent additive components. Wang et al. [30] considered a spatial automatic subgroup analysis problem for data with repeated measurements. Yan et al. [32] developed a censored linear regression model with heterogeneous treatment effects. Lu et al. [18] proposed a weighted penalized median regression approach for longitudinal data with dropouts. Hu et al. [9] considered subgroup analysis under a heterogeneous Cox model. Liu et al. [16] proposed a fused effects model for data with repeated measurements. Some other researchers [19,28,38] proposed methods to cluster subjects into subgroups based on longitudinal trajectories. Although the literature on subgroup analysis of scalar and longitudinal data is extensive, little has been done on functional data, especially under high-dimensional functional regression.

In this paper, we consider subgroup analysis for high-dimensional functional regression. This study may advance from the existing ones along with the following aspects. First, the proposed method can automatically separate observations into subgroups. A related model has been considered in Wang et al. [29], which proposed a functional partially linear regression model and identified latent subgroups of subjects by an algorithm sharing some similar spirit with K-means clustering. Our work uses the fusion penalty approach for estimation and automatically separating observations into subgroups. Besides, we can accommodate high-dimensional predictors.

Second, the proposed method can perform variable selection and identify relevant predictors along with estimating high-dimensional functional models. Under mild conditions, we establish the oracle property of the proposed estimator if the minimal sample size of subgroup satisfies $|g^0_{\min}| \gg (K + q_n s_n + q_n^2)^{1/2} n^{1/2}$, where K is the number of subgroups, q_n is the number of important predictors, s_n is the common truncation parameter, and n is the sample size. The number of functional predictors p_n, q_n, and s_n are all allowed to go to infinity. Extending theoretical results to diverging functional predictors is not trivial. It is different from Kong et al. [12], which assumed that the number of functional predictors is fixed. Besides, the existing study did not consider heterogeneity.

Third, we show that the true subgroup structure of samples can be recovered with a high probability. The number of subgroups is allowed to increase as the sample size increases. The number of subgroups K, the number of important predictors q_n, the common truncation parameter s_n, and the sample size n satisfy $K(K + q_n s_n + q_n^2)^{1/2} = o(n^{1/2})$.

The later sections are organized as follows. The model setting and methodology are described in Section 2, along with the proposed computational algorithm. In Section 3, we state technical assumptions and establish the asymptotic properties of the proposed estimator. Section 4 presents simulation studies under various scenarios to assess performance of the proposed method. We apply the proposed method to the TFT-LCD dataset in Section 5. Section 6 concludes with discussions. Additional technical results are given in Appendix.

2. Subgroup analysis for high-dimensional functional regression

2.1. Model

Suppose that the data consists of $\{x_{i1}(t), \ldots, x_{ip_n}(t), y_i\}$, $i \in \{1, \ldots, n\}$. y_i is a scalar continuous response, $\{x_{i1}(t), \ldots, x_{ip_n}(t)\}$ are p_n functional predictors, and $x_{ij}(t)$ is a real-valued, continuous, square-integrable, random curve on the compact interval T_j. p_n is allowed to go to infinity. Without loss of generality, we assume that functional predictors are centred and $T_j = T = [0, 1]$ for $j \in \{1, \ldots, p_n\}$.

In this paper, we are concerned with the following functional linear model with multiple functional observations:

$$y_i = \mu_i + \sum_{j=1}^{p_n} \int_T x_{ij}(t) b_j(t) dt + \epsilon_i, \quad i \in \{1, \ldots, n\}, \tag{1}$$

where μ_i's are unknown subject-specific intercepts, $b_j(t)$'s are square-integrable functional coefficients, and ϵ_i's are random scalar errors independent of the predictors. We assume that the first q_n functional predictors are important while the rest are not, i.e., $b_j(t) \neq 0$ for $j \in \{1, \ldots, q_n\}$ and $b_j(t) = 0$ for $j \in \{q_n + 1, \ldots, p_n\}$.

Assume $\{y_1, \ldots, y_n\}$ are from K subgroups. Here heterogeneity is modeled through μ_i's. In other words, if $\mu_i = \mu_j$, the ith observation and the jth observation are from the same subgroup. Let $\mathcal{G} = \{g_1, \ldots, g_K\}$ be the partition of $\{1, \ldots, n\}$. Then we have $\mu_i = \alpha_k$ for all $i \in g_k$, where α_k is the common value for μ_i's from subgroup g_k. We also allow the number of subgroups K to go to infinity.

2.2. Estimation

Because functional data is intrinsically infinite-dimensional, dimension reduction is critical for modeling and analysis. Functional principal component analysis (FPCA) is an important dimension reduction tool, and facilitates the conversion of inherently infinite-dimensional functional data to a finite-dimensional vector of random scores.

Let $K_j(s, t) = \mathrm{Cov}\{x_j(s), x_j(t)\}$. By Mercer's theorem, we have the spectral expansion:

$$K_j(s, t) = \sum_{k=1}^{\infty} \upsilon_{jk} \phi_{jk}(s) \phi_{jk}(t),$$

where $\upsilon_{j1} > \upsilon_{j2} > \cdots > 0$ are the eigenvalues of the linear operator associated with $K_j(s, t)$ with corresponding eigenfunctions $\{\phi_{jk}\}$. Then the Karhunen–Loève expansion of the random function $x_{ij}(t)$ in terms of the orthonormal basis $\{\phi_{jk}\}$ is $x_{ij}(t) = \sum_{k=1}^{\infty} \xi_{ijk} \phi_{jk}(t)$, where ξ_{ijk}'s are principal component scores satisfying $\mathrm{E}\left(\xi_{ijk}\right) = 0$, $\mathrm{E}\left(\xi_{ijk}^2\right) = \upsilon_{jk}$. Making use of the expansion $b_j(t) = \sum_{k=1}^{\infty} \beta_{jk} \phi_{jk}(t)$, we can rewrite Eq. (1) as:

$$y_i = \mu_i + \sum_{j=1}^{p_n} \sum_{k=1}^{\infty} \beta_{jk} \xi_{ijk} + \epsilon_i, \quad i \in \{1, \ldots, n\}.$$

In practice we do not observe the entire trajectories x_{ij}'s but have only intermittent noisy measurements. When the repeated observations are sufficiently dense for each subject, the estimates $\left\{\hat{x}_{ij} : i \in \{1, \ldots, n\}; j \in \{1, \ldots, d\}\right\}$ can be used to construct the covariance and eigenvalues/basis. Then, the empirical counterpart of $K_j(s, t)$ is $\hat{K}_j(s, t) = \sum_{k=1}^{\infty} \hat{\upsilon}_{jk} \hat{\phi}_{jk}(s) \hat{\phi}_{jk}(t)$, where $\hat{K}_j(s, t) = 1/n \sum_{i=1}^{n} \hat{x}_{ij}(s) \hat{x}_{ij}(t)$. Then, we have $\hat{\xi}_{ijk} = \int_T \hat{x}_{ij}(t) \hat{\phi}_{jk}(t) dt$.

Let s_j be the truncation parameter for the jth function, which is allowed to vary with the sample size. Denote $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{js_j})^\top$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_{p_n}^\top)^\top$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$. To further select functional predictors and identify subgroups simultaneously, we minimize the following objective function:

$$Q_{\lambda_1, \lambda_2}(\boldsymbol{\mu}, \boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \mu_i - \sum_{j=1}^{p_n} \sum_{k=1}^{s_j} \hat{\xi}_{ijk} \beta_{jk} \right)^2 + \sum_{1 \leq i < j \leq n} P\left(|\mu_i - \mu_j|, \lambda_1\right) + n \sum_{j=1}^{p_n} P\left(\|\boldsymbol{\beta}_j\|_2, c_j \lambda_2\right). \tag{2}$$

The first term in (2) is the truncated form of the standard squared loss. The second term is the sum of concave penalty functions applied to the pairwise differences of the intercepts, which is used to divide the observations into subgroups. $\lambda_1 \geq 0$ is the penalty parameter. If $\mu_i - \mu_j$ is shrunk to zero, then subjects i and j are from the same subgroup.

The last term is used to identify the relevant predictors. A penalty that can produce unbiased estimation is more appealing. We use MCP in this paper, where $P(\theta, \lambda) = \lambda \int_0^{|\theta|} \{1 - x/(\lambda a_\lambda)\}_+ dx$, z_+ stands for the positive part of z, $\lambda \geq 0$ is the penalty parameter, a_λ is an additional tuning parameter. We fix $a_\lambda = 3$ as suggested in the literature [17,20,33,37]. In (2), $\| \cdot \|_2$ is the Euclidean norm, $\lambda_2 \geq 0$ is the penalty parameter, and $c_j = \sqrt{s_j}$'s are used to adjust for group sizes of parameters. With the estimators $\hat{\boldsymbol{\beta}}_j$'s, we can get $\hat{b}_j(t) = \sum_{k=1}^{s_j} \hat{\beta}_{jk} \hat{\phi}_{jk}(t)$.

2.3. Computational algorithm

To implement the proposed method, we develop an alternating direction method of multipliers (ADMM) algorithm. By introducing a new set of parameters $\eta_{ij} = \mu_i - \mu_j$ and $\boldsymbol{\eta} = \left\{\eta_{ij}, i < j\right\}^\top$, the minimization of (2) is equivalent to the constrained optimization problem:

$$U_{\lambda_1,\lambda_2}(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\eta}) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \mu_i - \sum_{j=1}^{p_n}\sum_{k=1}^{s_j}\hat{\xi}_{ijk}\beta_{jk}\right)^2 + \sum_{1\le i<j\le n} P\left(|\eta_{ij}|, \lambda_1\right) + n\sum_{j=1}^{p_n} P\left(\|\boldsymbol{\beta}_j\|_2, c_j\lambda_2\right),$$

subject to

$$\mu_i - \mu_j - \eta_{ij} = 0, \quad 1 \le i < j \le n.$$

Then the augmented Lagrangian is

$$L_\rho(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\varpi}) = U_{\lambda_1,\lambda_2}(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\eta}) + \sum_{1\le i<j\le n} \varpi_{ij}\left(\mu_i - \mu_j - \eta_{ij}\right) + \frac{\rho}{2}\sum_{1\le i<j\le n}\left(\mu_i - \mu_j - \eta_{ij}\right)^2, \tag{3}$$

where the dual parameters $\boldsymbol{\varpi} = \left\{\varpi_{ij}, i < j\right\}^\top$ are the Lagrangian multipliers, and ρ is a penalty parameter.

The ADMM algorithm minimizes the augmented Lagrangian by updating one block of parameters at a time, which consists of $\boldsymbol{\mu}, \boldsymbol{\beta}$-minimization, $\boldsymbol{\eta}$-minimization, and a dual parameter updating routine as follows.

Given $(\boldsymbol{\mu}^{(l)}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\eta}^{(l)}, \boldsymbol{\varpi}^{(l)})$ from the lth step, the update of $\boldsymbol{\beta}^{(l+1)}$ is:

$$\boldsymbol{\beta}^{(l+1)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{\sum_{j=1}^{p_n} s_j}}{\arg\min}\,\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \mu_i^{(l)} - \sum_{j=1}^{p_n}\sum_{k=1}^{s_j}\hat{\xi}_{ijk}\beta_{jk}\right)^2 + n\sum_{j=1}^{p_n} P\left(\|\boldsymbol{\beta}_j\|_2, c_j\lambda_2\right). \tag{4}$$

This is a standard group MCP problem. For more details, we refer to Huang et al. [10]. For $\boldsymbol{\mu}^{(l+1)}$,

$$\boldsymbol{\mu}^{(l+1)} = \underset{\boldsymbol{\mu}\in\mathbb{R}^n}{\arg\min} L_\rho\left(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\varpi} | \boldsymbol{\beta}^{(l+1)}, \boldsymbol{\eta}^{(l)}, \boldsymbol{\varpi}^{(l)}\right).$$

Then we have

$$\boldsymbol{\mu}^{(l+1)} = \left(\rho\boldsymbol{\Delta}^\top\boldsymbol{\Delta} + \boldsymbol{I}_n\right)^{-1}\left\{\boldsymbol{y} - \hat{\boldsymbol{\xi}}\boldsymbol{\beta}^{(l+1)} + \rho\boldsymbol{\Delta}^\top(\boldsymbol{\eta}^{(l)} - \rho^{-1}\boldsymbol{\varpi}^{(l)})\right\}, \tag{5}$$

where $\boldsymbol{\Delta} = \left\{(\boldsymbol{e}_i - \boldsymbol{e}_j), i < j\right\}^\top$, $\hat{\boldsymbol{\xi}}_i = \left(\hat{\xi}_{i11}, \ldots, \hat{\xi}_{ip_ns_n}\right)^\top$, $\hat{\boldsymbol{\xi}} = \left(\hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_n\right)^\top$, and $\boldsymbol{e}_i$ is an $n \times 1$ vector that has its ith component equal to 1 and all of its other components equal to 0.

Given $(\boldsymbol{\mu}^{(l+1)}, \boldsymbol{\beta}^{(l+1)}, \boldsymbol{\eta}^{(l)}, \boldsymbol{\varpi}^{(l)})$, the update of $\boldsymbol{\eta}$ is:

$$\eta_{ij}^{(l+1)} = \underset{\eta_{ij}\in\mathbb{R}}{\arg\min}\left\{\frac{\rho}{2}\left(\delta_{ij}^{(l+1)} - \eta_{ij}\right)^2 + P\left(|\eta_{ij}|, \lambda_1\right)\right\}$$

with $\delta_{ij}^{(l+1)} = \mu_i^{(l+1)} - \mu_j^{(l+1)} + \rho^{-1}\varpi_{ij}^{(l)}$. Then we have

$$\eta_{ij}^{(l+1)} = \begin{cases} \frac{\mathrm{ST}\left(\delta_{ij}^{(l+1)}, \lambda_1/\rho\right)}{1-1/(\gamma\rho)} & \text{if } |\delta_{ij}^{(l+1)}| \le \gamma\lambda_1 \\ \delta_{ij}^{(l+1)} & \text{if } |\delta_{ij}^{(l+1)}| > \gamma\lambda_1 \end{cases}, \tag{6}$$

where $\mathrm{ST}(t, \lambda) = \mathrm{sign}(t)\,(|t| - \lambda)_+$ is the soft thresholding operator.

Given $(\boldsymbol{\mu}^{(l+1)}, \boldsymbol{\beta}^{(l+1)}, \boldsymbol{\eta}^{(l+1)}, \boldsymbol{\varpi}^{(l)})$, the updates of dual parameters are:

$$\varpi_{ij}^{(l+1)} = \varpi_{ij}^{(l)} + \rho\left(\mu_i^{(l+1)} - \mu_j^{(l+1)} - \eta_{ij}^{(l+1)}\right), \quad 1 \le i < j \le p. \tag{7}$$

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

Step 1: Set the initial estimate $\boldsymbol{\beta}^{(0)}$ be the solution of a group Lasso problem. Let $\mu_i^{(0)} = y_i - \sum_{j=1}^{p_n}\sum_{k=1}^{s_j}\hat{\xi}_{ijk}\beta_{jk}^{(0)}$, $\eta_{ij}^{(0)} = \mu_i^{(0)} - \mu_j^{(0)}$, and $\boldsymbol{\varpi}^{(0)} = \boldsymbol{0}$.

Step 2: At iteration $l + 1$, compute $(\boldsymbol{\mu}^{(l+1)}, \boldsymbol{\beta}^{(l+1)}, \boldsymbol{\eta}^{(l+1)}, \boldsymbol{\varpi}^{(l+1)})$ by (4)–(7).

Step 3: If the stopping rule is met, terminate the algorithm. Then,

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\varpi}}) = (\boldsymbol{\mu}^{(l+1)}, \boldsymbol{\beta}^{(l+1)}, \boldsymbol{\eta}^{(l+1)}, \boldsymbol{\varpi}^{(l+1)}).$$

Otherwise, go to Step 2.

Remark 1. Following Ma and Huang [20], we stop the algorithm when $\left\|\Delta\boldsymbol{\mu}^{(l+1)} - \boldsymbol{\eta}^{(l+1)}\right\|_2 < \varepsilon$ for a small ε. Based on $\hat{\eta}$, we can get the subgroup structures. Specifically, we put y_i and y_j in the same subgroup if $\hat{\eta}_{ij} = 0$. The final intercepts of the subgroups are calculated as the mean of intercepts among each subgroup. The computer code is publicly available at https://github.com/ruiqwy/fsubgroup.

3. Theoretical properties

In this section, we provide the technical assumptions and establish the asymptotic properties of the proposed estimator. We will establish that the true subgroup structure of the samples and sparsity can be recovered. There are two types of approximations involved in the objective function (2): the estimator $\hat{\xi}_{ijk}$ of the unknown covariate ξ_{ijk} and the truncation of regression coefficients. Due to the differences between functional and scalar predictors, extending theoretical results to multiple functional regression with a diverging truncation is not trivial, especially when the number of functional predictors is permitted to diverge with the sample size. Additionally, we establish the asymptotic properties of the proposed estimator under heterogeneity, which is more challenging than under homogeneity.

Denote the minimum and maximum eigenvalues of a symmetric matrix A by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$. Let $(\boldsymbol{\mu}^0, \boldsymbol{\beta}^0)$ be the true regression parameters. Suppose that $\boldsymbol{\mu}^0$ has K distinct elements α_k^0, $1 \le k \le K$. Define $\boldsymbol{\alpha}^0 = (\alpha_1^0, \ldots, \alpha_K^0)$, and $\mathcal{G}^0 = \{g_1^0, \ldots, g_K^0\}$ with $g_k^0 = \{i : \mu_i^0 = \alpha_k^0\}$. $\boldsymbol{Z} = \{z_{ik}\}$ is a $n \times K$ matrix with $z_{ik} = 1$ if $i \in g_k$ and 0 otherwise. Then we have $\boldsymbol{\mu}^0 = \boldsymbol{Z}\boldsymbol{\alpha}^0$. Denote the cardinality of g_k^0 by $|g_k^0|$, and define $|g_{\min}^0| = \min_{1\le k\le K} |g_k^0|$, $|g_{\max}^0| = \max_{1\le k\le K} |g_k^0|$. Recall that we assume the first q_n functional predictors are important while the rest are not. Denote $\mathcal{A} = \{1, \ldots, q_n\}$ as the index set of the important variables, and $\mathcal{A}^c = \{q_n + 1, \ldots, p_n\}$. In fact, for different predictors, we may choose different truncation points to approximate the infinite sums. In this section, to simplify notation, we use a common truncation parameter s_n, which is a function of sample size n. To facilitate technical proofs, we assume the following regularity conditions.

(C1) For $j \in \{1, \ldots, p_n\}$, for any $C_0 > 0$, there exists $\varrho > 0$ such that

$$\sup_{t\in T}\left[\mathrm{E}\left\{|x_j(t)|^{C_0}\right\}\right] < \infty, \quad \sup_{s,t\in T}\left(\mathrm{E}\left[\left\{|s-t|^{-\varrho}|x_j(s) - x_j(t)|\right\}^{C_0}\right]\right) < \infty,$$

$x_j(\cdot)$ is twice continuously differentiable on T with probability one, and $\int_T \mathrm{E}\left\{x_j''(t)\right\}^4 dt < \infty$, where $x_j''(\cdot)$ denotes the second derivative of $x_j(\cdot)$. For each integer $r \ge 1$, $v_{jk}^{-r}\mathrm{E}\left(\xi_{jk}^{2r}\right)$ is bounded uniformly in k and j.

(C2) There exist positive constants C_1 and $a > 1$ such that $C_1^{-1}k^{-a} \le v_{jk} \le C_1 k^{-a}$ and $v_{jk} - v_{j,k+1} \geqslant C_1 k^{-a-1}$, where $\{v_{jk}\}$ are the eigenvalues of the covariance function for $j \in \{1, \ldots, p_n\}$ and $k \ge 1$.

(C3) $|\beta_{jk}^0| \leqslant C_2 k^{-b}$ for $k \ge 1$, $b > 2$, $j \in \{1, \ldots, p_n\}$.

(C4) The noise vector $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ has sub-Gaussian tails such that $\Pr\left(|\mathbf{a}^\top\boldsymbol{\epsilon}| > \|\mathbf{a}\|_2 x\right) \le 2\exp\left(-C_3 x^2\right)$ for any vector $\mathbf{a} \in \mathbb{R}^n$ and $x > 0$, where $0 < C_3 < \infty$.

(C5) The smoothing parameters s_n and q_n satisfy $|g_{\min}^0|^{-1} s_n^{a/2+1}\left(q_n s_n n^{1/2} + K|g_{\max}^0|^{1/2}\right) \to 0$, $s_n^{2a+2}/n \to 0$, $s_n^{2b-1}/n \to \infty$.

(C6) The tuning parameters λ_1 and λ_2 satisfy: (i) $\lambda_1 = o(1)$, $\min_{1\le k<\ell\le K}\left|\alpha_k^0 - \alpha_\ell^0\right|/\lambda_1 \to \infty$, $\left(K + q_n s_n + q_n^2\right) n |g_{\min}^0|^{-2} = o\left(\lambda_1^2\right)$; (ii) $\lambda_2 = o(1)$, $\min_{j\in\mathcal{A}}\ \|\boldsymbol{\beta}_j^0\|_2/\lambda_2 \to \infty$, $\max\left\{n s_n^a\left(K + q_n s_n + q_n^2\right)|g_{\min}^0|^{-2}, s_n^3 n^{-1}, s_n \log(p_n s_n) n^{-1}\right\} = o\left(\lambda_2^2\right)$.

(C7) Define $\boldsymbol{U} = \begin{pmatrix} \boldsymbol{Z}^\top\boldsymbol{Z}/n & 0 \\ 0 & \mathrm{E}\left(\tilde{\boldsymbol{N}}_i\tilde{\boldsymbol{N}}_i^\top\right)\end{pmatrix}$, where $\tilde{\boldsymbol{N}}_i = \left(\xi_{i11}v_{11}^{-1/2}, \ldots, \xi_{iq_ns_n}v_{q_ns_n}^{-1/2}\right)^\top$ is a $(q_n s_n) \times 1$ vector. For some constant $0 < C_4 \le 1$ and some positive constant C_5, $C_4|g_{\min}^0|/n \le \lambda_{\min}(\boldsymbol{U}) \le C_5$.

Remark 2. Condition (C1) is a common condition in functional data analysis. We impose conditions (C2)-(C3) on the decay rates of the eigenvalues $\{v_{jk}\}$ and regression coefficients $\{\beta_{jk}^0\}$, which are similar to those adopted by [6,12,13,15,22]. The second part of condition (C2) requires that the spacings between the eigenvalues are not too small. Condition (C3) is needed only to control the tail behavior for large k. Condition (C4) assumes that the noise vector has sub-Gaussian tails. Conditions (C5)-(C7) are required for the consistency of the estimators and model selection. A main contribution is to allow the number of functional predictors to increase as the sample size increases. This distinguishes our method from those with a fixed number of functional predictors (e.g., Kong et al. [12]). The number of subgroups is also allowed to increase as the sample size increases. In the literature, it is commonly assumed that the smallest eigenvalue of $\mathrm{E}\left(\tilde{\boldsymbol{N}}_i\tilde{\boldsymbol{N}}_i^\top\right)$ is bounded by some constant C. In our model setup, $\boldsymbol{Z}^\top\boldsymbol{Z} = \mathrm{diag}\left(|g_1^0|, \ldots, |g_K^0|\right)$. By assuming $\lambda_{\min}\left\{\mathrm{E}\left(\tilde{\boldsymbol{N}}_i\tilde{\boldsymbol{N}}_i^\top\right)\right\} = C$, we have $\lambda_{\min}(\boldsymbol{U}) = \min\left(|g_{\min}^0|/n, C\right)$. Hence we assume $\lambda_{\min}(\boldsymbol{U}) \ge C_4|g_{\min}^0|/n$ for some constant $0 < C_4 \le 1$.

To facilitate theoretical analysis, we reparameterize by writing $\tilde{\beta}_{jk} = v_{jk}^{1/2}\beta_{jk}$, so that the functional principal component scores serving as predictor variables are on a common scale of variability. This reparameterization is used only for technical derivations and does not appear in the estimation procedure. Let

$$Q_n(\boldsymbol{\mu}, \tilde{\boldsymbol{\beta}}) = L_n(\boldsymbol{\mu}, \tilde{\boldsymbol{\beta}}) + \sum_{1\le i<j\le n} P\left(|\mu_i - \mu_j|, \lambda_1\right) + n\sum_{j=1}^{p_n} P\left(\|\boldsymbol{\beta}_j\|_2, \lambda_2\right),$$

where

$$L_n(\mu, \tilde{\beta}) = \frac{1}{2} \sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{p_n} \sum_{k=1}^{s_n} \left(\hat{\xi}_{ijk} v_{jk}^{-1/2} \right) \tilde{\beta}_{jk} - \mu_i \right\}^2 .$$

Theorem 1. *Under conditions (C1)-(C7) and (CA1)–(CA2) in the Appendix, if* $K \geq 2$, $|g^0_{\min}| \gg \left(K + q_n s_n + q_n^2\right)^{1/2} n^{1/2}$, *there exists a local minimizer* $\left(\hat{\mu}^\top, \check{\beta}^\top\right)^\top$ *of objective function* $Q_n(\mu, \tilde{\beta})$ *satisfying*

(i) $\Pr\left(\hat{\mu}_i = \hat{\mu}_j, \forall i, j \in g^0_k, 1 \leq k \leq \hat{K}\right) \to 1$, i.e., $\Pr\left(\hat{\mathcal{G}} = \mathcal{G}^0\right) \to 1$,

(ii) $\Pr\left\{\hat{b}_{q_n+1}(t) = \cdots = \hat{b}_{p_n}(t) = 0\right\} \to 1$,

(iii) $\left\| \left(\hat{\mu}^\top, \check{\beta}^\top\right)^\top - \left(\mu^{0\top}, \tilde{\beta}^{0\top}\right)^\top \right\|_2 = O_p\left\{\left(K + q_n s_n + q_n^2\right)^{1/2} n^{1/2} |g^0_{\min}|^{-1}\right\}$.

Remark 3. Since $|g^0_{\min}| \leq n/K$, by condition $|g^0_{\min}| \gg \left(K + q_n s_n + q_n^2\right)^{1/2} n^{1/2}$, K, q_n and s_n satisfy $K\left(K + q_n s_n + q_n^2\right)^{1/2} = o\left(n^{1/2}\right)$.

Theorem 1 shows that the true subgroup structure of the samples and sparsity can be recovered with a high probability. The estimation consistency result is expressed in terms of $\tilde{\beta}$, not the original parameter β. By Theorem 1, we can conclude that $\left\|\hat{b}_j(t) - b_j(t)\right\|^2 = O_p\left\{s_n^a\left(K + q_n s_n + q_n^2\right) n |g^0_{\min}|^{-2}\right\}$. Specifically, if the number of functional predictors p_n is fixed and the number of subgroups is one (a homogeneous data), we can get $\left\|\hat{b}_j(t) - b_j(t)\right\|^2 = O_p\left(s_n^{a+1}/n\right)$. Kong et al. [12] get the result $O_p\left\{s_n^a\left(q_n + s_n\right)/n\right\}$ under the homogeneity assumption, where q_n is the number of significant scalar covariates. If there is no scalar covariate in the model, it reduces to $O_p\left(s_n^{a+1}/n\right)$, the same as our result.

Remark 4. When the true model is homogeneous given as Eq. (1) with $\mu_1 = \cdots = \mu_n = \mu = \alpha$ and $K = 1$, the homogeneity and sparsity can be recovered with a high probability. The technical assumptions and theorem are given in Appendix.

4. Simulation analysis

In this section, we compare performance of sgmcp (the proposed estimator) with three alternatives: gmcp (standard group MCP with penalty for functional coefficients, which assumes homogeneity and has no penalty for intercepts), Oracle (under which the subgroup structure and index of significant predictors are known) and Cluster (which clusters samples based on the residuals obtained from K-means first, where the number of clusters is also chosen by BIC, and then refits the model). We consider the following function linear model:

$$y_i = \mu_i + \sum_{j=1}^{p_n} \int_0^1 x_{ij}(t) b_j(t) dt + \epsilon_i, \quad i \in \{1, \ldots, n\}.$$

The functional data is generated from the process $x_{ij}(t) = \sum_{k=1}^{3} \xi_{ijk} \phi_k(t)$, where $\xi_{ijk} \sim N(0, k^{-2})$, $\phi_1(t) = 1$, $\phi_2(t) = \sqrt{2}\cos(\pi t)$, and $\phi_3(t) = \sqrt{2}\cos(2\pi t)$. For $b_j(t)$, in terms of expansion based on $\{\phi_k\}_{k=1}^3$, we take $\beta_j = (1, 1.1, 1.2)^\top$ for $j = 1, \ldots, q_n$, and $\beta_j = (0, 0, 0)^\top$ otherwise. We set $\epsilon_i \sim N(0, 0.1)$ and $n = 100$.

To provide a good approximation to the infinite sum, we use s_j components, which explain 95% variation in $x_{ij}(t)$. For tuning parameters selection, we use a modified Bayesian information criterion (BIC) [2] defined as:

$$\mathrm{BIC}(\lambda_1, \lambda_2) = \log\left\{\sum_{i=1}^{n}\left(y_i - \hat{\mu}_i - \sum_{j=1}^{p_n}\sum_{k=1}^{s_j} \hat{\xi}_{ijk}\hat{\beta}_{jk}\right)^2 \Big/ n\right\} + C_n \frac{\log n}{n}\left(\hat{K} + q^{\#}\right),$$

where $C_n = c\log\left\{\log\left(n + \sum_{j=1}^{p_n} s_j\right)\right\}$, and $q^{\#}$ is the number of non-zero parameters. We set $c = 5$.

We comprehensively consider the following cases. In the first case, we generate data from a homogeneity model and let $\mu_i = 0$ for all $i \in \{1, \ldots, n\}$. In the second case, we generate μ_i from two different values α and $-\alpha$ with equal probabilities. We consider multiple values of α. As for the last case, the data forms three subgroups.

Performance of the estimates is measured by the following metrics. (1) Mean squared errors of μ_i: MSE $= \left(\left\|\hat{\mu} - \mu\right\|_2^2/n\right)^{1/2}$. (2) Integrated squared error of coefficient functions: ISE $= \left\{\sum_{j=1}^{p_n}\left\|\hat{b}_j(t) - b_j(t)\right\|^2\right\}^{1/2}$. (3) Sensitivity and specificity are used to evaluate feature selection. Sensitivity: the proportion of true positives being correctly identified.

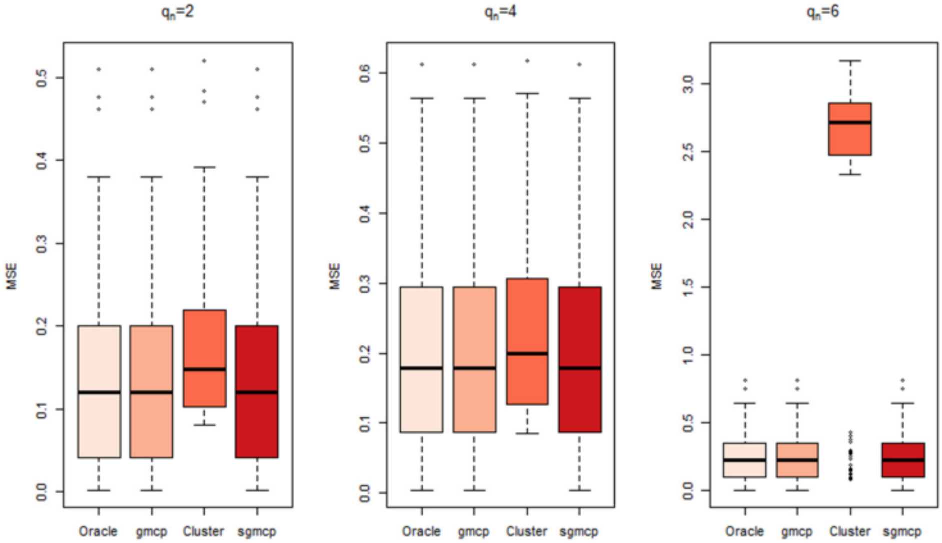

Fig. 2. MSE under Case 1 for Oracle, gmcp, Cluster, and sgmcp with $\mu_i = 0$ for all $i \in \{1, \dots, n\}$, $p_n = 100$, $q_n = 2, 4$ and 6. The data is generated from a homogeneous model and the true number of subgroup is 1. The boxplots are based on 100 Monte Carlo replicates.

Specificity: the proportion of true negatives being correctly identified. (4) The number of identified subgroups K. (5) The rand index of identified subgroups, RI, which is defined as

$$\mathrm{RI} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}}.$$

Here, TP (true positive), TN (true negative), FN (false negative) and FP (false positive) are the number of pairs of subjects in different subgroups that are assigned to different subgroups, the number of pairs from the same subgroup that are assigned to the same subgroup, the number of pairs from the same subgroup that are assigned to different subgroups, and the number of pairs from different subgroups that are assigned to the same subgroup, respectively. A higher value of RI indicates a better agreement between the identified subgroups and the true subgroup allocation. Below are the detailed settings of three cases:

Case 1: $\mu_i = 0$ for all $i \in \{1, \dots, n\}$, $p_n = 100$, $q_n = 2, 4$ and 6, meaning that the data is generated from a homogeneous model and the true number of subgroup is 1;

Case 2: $\Pr(\mu_i = \alpha) = \Pr(\mu_i = -\alpha) = 1/2$, $\alpha = 0.3, 0.5$ and 0.7, $p_n = 10, 50$ and 100, $q_n = 2$, meaning that the true number of subgroups is 2;

Case 3: $\Pr(\mu_i = 1) = \Pr(\mu_i = 0) = \Pr(\mu_i = -1) = 1/3$, $p_n = 10, 50$ and 100, $q_n = 1$ and 2, meaning that the true number of subgroups is 3.

The simulation results are summarized in Tables 1–2 and Figs. 2–3 based on 100 replicates for each scenario. For all cases, specificity is 1.000(0.000), and we omit it in the table. From Figs. 2–3, we see that the performances of gmcp, sgmcp and Oracle are similar when the data is generated from a homogeneous model. The mean values of $\hat{K}$ obtained using sgmcp are 1 ($q_n = 2$), 1 ($q_n = 4$), and 1.020 ($q_n = 6$), respectively. The mean values of RI obtained using sgmcp are 1 ($q_n = 2$), 1 ($q_n = 4$), and 0.995 ($q_n = 6$), respectively. We conclude that sgmcp can recover the homogeneous model. Cluster fails to identify the homogeneous model. The mean values of $\hat{K}$ obtained using Cluster are 4.780 ($q_n = 2$), 5.130 ($q_n = 4$), and 4.860 ($q_n = 6$), respectively. The mean value of RI obtained using Cluster are 0.251 ($q_n = 2$), 0.231 ($q_n = 4$), and 0.244 ($q_n = 6$), respectively. From Table 1, we can see that for a large value of α, it is easier to identify the two subgroups with sgmcp. The proposed method leads to a more accurate subgroup structure than Cluster in terms of RI and K. Both methods get worse when p_n increases. When the value of α is large, gmcp and Cluster fail to select positive functional predictors. From Table 2, we can see that the proposed method performs better than gmcp and Cluster, and both methods get worse when p_n or q_n increases. To demonstrate performance of the proposed method in terms of subgroup identification, the results of additional simulation comparing Cases 2–3 are provided in Tables 3–4. From Table 3, we can see that as the separation parameter α increases, the mean of estimated number of subgroups becomes closer to the true value, and RI becomes closer to 1. From Table 4, we can see that as the sample size n increases, the mean of estimated number of subgroups becomes closer to the true value, and RI becomes closer to 1. We conclude that subgroup identification of the proposed method can be improved with the separation parameter α or sample size n increasing.

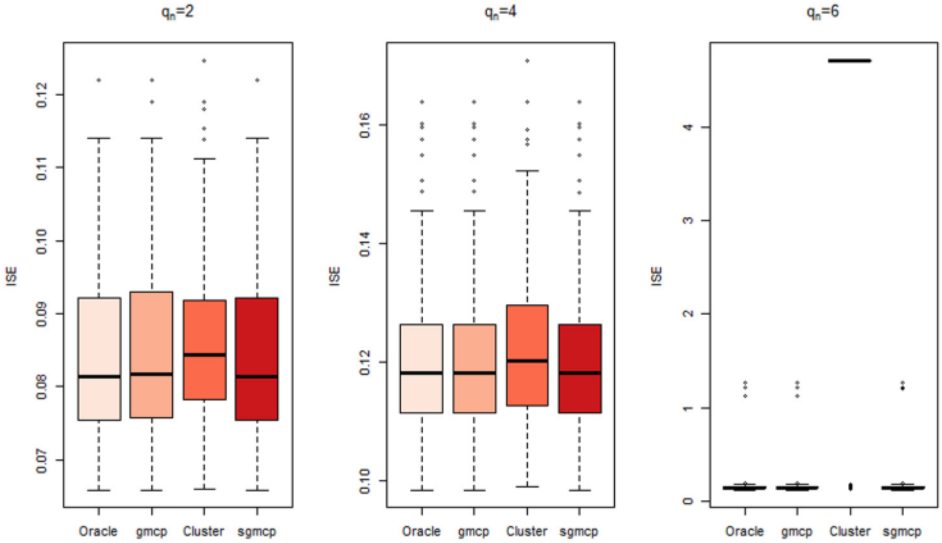

Fig. 3. ISE under Case 1 for Oracle, gmcp, Cluster, and sgmcp with $\mu_i = 0$ for all $i \in \{1, \ldots, n\}$, $p_n = 100$, $q_n = 2, 4$ and 6. The data is generated from a homogeneous model and the true number of subgroup is 1. The boxplots are based on 100 Monte Carlo replicates.

Table 1
Simulation results under Case 2 with $\Pr(\mu_i = \alpha) = \Pr(\mu_i = -\alpha) = 1/2$, $\alpha = 0.3, 0.5$ and 0.7, $p_n = 10, 50$ and 100, $q_n = 2$ based on 100 Monte Carlo replicates. The true number of subgroups is 2.

α	p_n	method	K(mean)	K(median)	RI	MSE	ISE	Sensitivity
0.3	10	Oracle	–	–	–	0.136(0.109)	0.082(0.011)	–
		gmcp	–	–	–	0.341(0.056)	0.183(0.053)	1(0)
		Cluster	4.530(1.389)	4(1.483)	0.754(0.098)	0.179(0.090)	0.142(0.047)	1(0)
		sgmcp	2.930(1.559)	2(0)	0.901(0.100)	0.181(0.098)	0.130(0.068)	1(0)
	50	Oracle	–	–	–	0.139(0.011)	0.082(0.011)	–
		gmcp	–	–	–	0.342(0.057)	0.183(0.053)	1(0)
		Cluster	4.070(1.350)	4(1.483)	0.782(0.111)	0.178(0.093)	0.136(0.046)	1(0)
		sgmcp	3.080(1.482)	2.5(0.741)	0.854(0.120)	0.204(0.100)	0.163(0.089)	1(0)
	100	Oracle	–	–	–	0.136(0.109)	0.082(0.011)	–
		gmcp	–	–	–	0.341(0.056)	0.183(0.053)	1(0)
		Cluster	3.750(1.321)	4(1.483)	0.806(0.118)	0.173(0.093)	0.130(0.046)	1(0)
		sgmcp	3.120(1.458)	2.5(0.741)	0.840(0.117)	0.209(0.098)	0.172(0.087)	1(0)
0.5	10	Oracle	–	–	–	0.136(0.109)	0.082(0.011)	–
		gmcp	–	–	–	0.526(0.039)	0.282(0.089)	1(0)
		Cluster	2.730(1.325)	2(0)	0.928(0.117)	0.157(0.102)	0.111(0.057)	1(0)
		sgmcp	2.270(0.649)	2(0)	0.966(0.073)	0.172(0.117)	0.118(0.080)	1(0)
	50	Oracle	–	–	–	0.136(0.109)	0.082(0.011)	–
		gmcp	–	–	–	0.526(0.039)	0.478(0.670)	0.920(0.273)
		Cluster	2.710(1.233)	2(0)	0.907(0.156)	0.258(0.377)	0.316(0.714)	0.920(0.273)
		sgmcp	2.590(1.006)	2(0)	0.930(0.092)	0.207(0.124)	0.169(0.136)	1(0)
	100	Oracle	–	–	–	0.136(0.109)	0.082(0.011)	–
		gmcp	–	–	–	0.526(0.039)	1.188(1.182)	0.630(0.485)
		Cluster	3.270(1.588)	2(0)	0.799(0.232)	0.663(0.686)	1.068(1.273)	0.630(0.485)
		sgmcp	2.620(0.951)	2(0)	0.913(0.097)	0.220(0.126)	0.193(0.144)	1(0)
0.7	10	Oracle	–	–	–	0.136(0.109)	0.082(0.011)	–
		gmcp	–	–	–	0.717(0.030)	2.514(0.659)	0.090(0.288)
		Cluster	5.050(1.366)	5(1.483)	0.562(0.136)	1.514(0.444)	2.484(0.753)	0.090(0.288)
		sgmcp	2.090(0.494)	2(0)	0.986(0.052)	0.158(0.120)	0.110(0.107)	1(0)
	50	Oracle	–	–	–	0.143(0.115)	0.083(0.011)	–
		gmcp	–	–	–	0.719(0.033)	2.720(0.000)	0(0)
		Cluster	5.040(0.953)	5(1.483)	0.521(0.013)	1.657(0.118)	2.720(0.000)	0(0)
		sgmcp	2.700(1.474)	2(0)	0.944(0.100)	0.202(0.140)	0.176(0.179)	1(0)
	100	Oracle	–	–	–	0.139(0.110)	0.082(0.011)	–
		gmcp	–	–	–	0.717(0.030)	2.720(0.000)	0(0)
		Cluster	4.800(0.910)	5(1.483)	0.521(0.013)	1.652(0.121)	2.720(0.000)	0(0)
		sgmcp	2.790(1.328)	2(0)	0.928(0.108)	0.229(0.146)	0.217(0.212)	1(0)

Table 2
Simulation results under Case 3 with $\Pr(\mu_i = 1) = \Pr(\mu_i = 0) = \Pr(\mu_i = -1) = 1/3$, $p_n = 10, 50$ and 100, $q_n = 1$ and 2 based on 100 Monte Carlo replicates. The true number of subgroups is 3.

q_n	p_n	method	K(mean)	K(median)	RI	MSE	ISE	Sensitivity
1	10	Oracle	–	–	–	0.110(0.076)	0.057(0.010)	–
		gmcp	–	–	–	0.823(0.030)	0.842(0.756)	0.680(0.469)
		Cluster	4.100(1.403)	3(0)	0.859(0.165)	0.468(0.460)	0.700(0.852)	0.680(0.469)
		sgmcp	3.640(1.124)	3(0)	0.942(0.075)	0.206(0.117)	0.176(0.170)	1(0)
	50	Oracle	–	–	–	0.110(0.076)	0.057(0.010)	–
		gmcp	–	–	–	0.823(0.030)	1.104(0.799)	0.520(0.502)
		Cluster	4.220(1.353)	4(1.483)	0.805(0.175)	0.634(0.495)	0.987(0.908)	0.520(0.502)
		sgmcp	3.710(1.266)	3(0)	0.906(0.094)	0.255(0.143)	0.232(0.194)	1(0)
	100	Oracle	–	–	–	0.110(0.076)	0.057(0.010)	–
		gmcp	–	–	–	0.823(0.030)	1.231(0.791)	0.440(0.499)
		Cluster	4.150(1.242)	4(1.483)	0.775(0.176)	0.717(0.489)	1.133(0.900)	0.440(0.499)
		sgmcp	3.730(1.221)	3(0)	0.886(0.110)	0.279(0.154)	0.267(0.206)	1(0)
2	10	Oracle	–	–	–	0.137(0.108)	0.082(0.012)	–
		gmcp	–	–	–	0.831(0.035)	2.698(0.217)	0.010(0.100)
		Cluster	5.350(1.009)	5(1.483)	0.616(0.038)	1.646(0.190)	2.695(0.248)	0.010(0.100)
		sgmcp	3.830(1.484)	3(0.741)	0.867(0.098)	0.327(0.136)	0.359(0.231)	1(0)
	50	Oracle	–	–	–	0.218(0.104)	0.412(0.190)	–
		gmcp	–	–	–	0.831(0.035)	2.698(0.217)	0.010(0.100)
		Cluster	5.070(1.018)	5(1.483)	0.613(0.039)	1.640(0.190)	2.695(0.248)	0.010(0.100)
		sgmcp	3.750(1.395)	3(1.483)	0.798(0.117)	0.410(0.149)	0.468(0.243)	1(0)
	100	Oracle	–	–	–	0.218(0.104)	0.412(0.190)	–
		gmcp	–	–	–	0.831(0.035)	2.698(0.217)	0.010(0.100)
		Cluster	4.790(1.047)	5(1.483)	0.610(0.040)	1.634(0.189)	2.695(0.248)	0.010(0.100)
		sgmcp	4.010(1.560)	4(1.483)	0.780(0.127)	0.427(0.157)	0.486(0.244)	1(0)

Table 3
Additional simulation results for the proposed method under two subgroups and three subgroups with $p_n = 10$ based on 100 Monte Carlo replicates. As the separation parameter α increases, the mean of estimated number of subgroups becomes closer to the true value, and RI becomes closer to 1.

Setting	α	K(mean)	K(median)	RI
Two subgroups with	0.3	2.930(1.559)	2(0)	0.901(0.100)
$\Pr(\mu_i = \alpha) = \Pr(\mu_i = -\alpha) = 1/2$	0.5	2.270(0.649)	2(0)	0.966(0.073)
	0.7	2.090(0.494)	2(0)	0.986(0.052)
	1	2.040(0.400)	2(0)	0.992(0.040)
	1.5	2.040(0.400)	2(0)	0.992(0.039)
Three subgroups with $q_n = 1$	1	3.640(1.124)	3(0)	0.942(0.075)
and $\Pr(\mu_i = \alpha) = \Pr(\mu_i = 0) = \Pr(\mu_i = -\alpha) = 1/3$	1.5	3.460(1.176)	3(0)	0.960(0.073)
	3	3.210(0.701)	3(0)	0.971(0.068)

Table 4
Additional simulation results for the proposed method under Cases 2–3 with $p_n = 10$ based on 100 Monte Carlo replicates. As the sample size n increases, the mean of estimated number of subgroups becomes closer to the true value, and RI becomes closer to 1.

Setting	n	K(mean)	K(median)	RI
Case 2 with $\alpha = 0.5$	100	2.270(0.649)	2(0)	0.966(0.073)
	200	2.210(0.742)	2(0)	0.990(0.045)
	300	2.160(0.677)	2(0)	0.994(0.028)
Case 3 with $q_n = 1$	100	3.640(1.124)	3(0)	0.942(0.075)
	200	3.260(0.613)	3(0)	0.986(0.026)
	300	3.140(0.493)	3(0)	0.994(0.013)

5. Application

The TFT-LCDs extend over various applications such as office-automation, electric home appliances, transportations, and more [27]. In this section, we apply the proposed method to the TFT-LCD dataset. Since TFT-LCDs have excellent features such as a low profile, lightweight, low operating-voltage, low power-consumption, full color capabilities, large area, and higher resolution, they now play a leading role in various flat-panel electronic display devices [27]. Since the technological environment has become increasingly competitive due to globalization's rapid speed, fast and accurate estimation is essential to a successful delivery of devices in a timely manner [3]. Data analysis can help the semiconductor industry make better use of product information and improve product quality.

As mentioned earlier, the manufacturing process of TFT-LCD is comprised of hundreds of working procedures. It can be roughly divided into the following four processed: thin-film transistor (TFT), color filter (CF), cell, and module. TFT

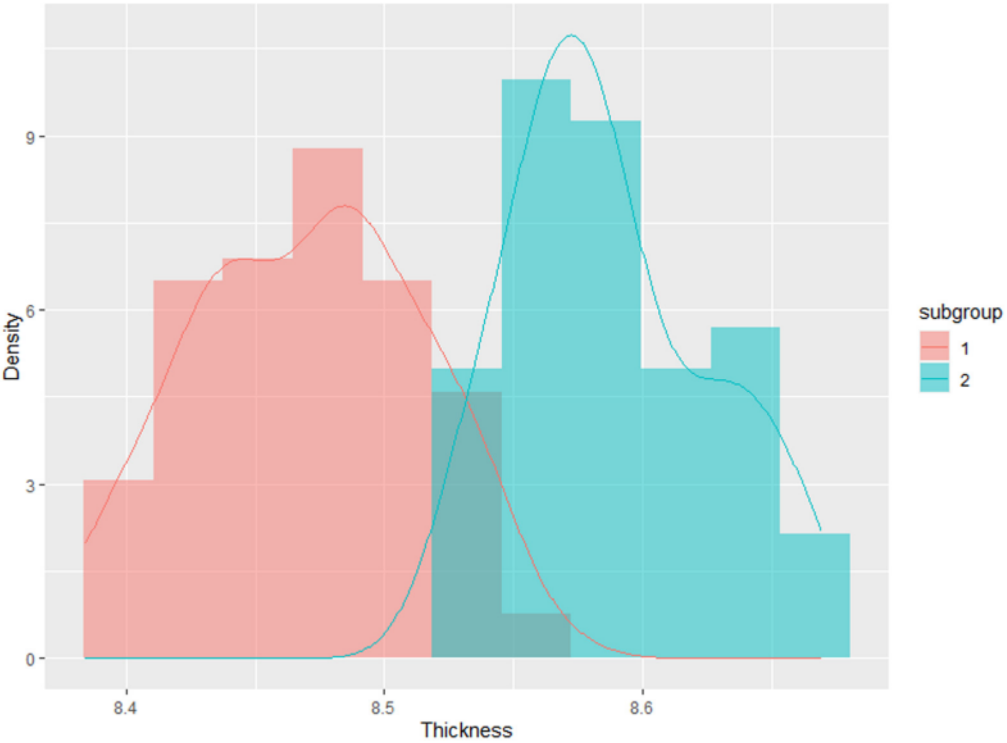

Fig. 4. Density plot of thickness after adjusting for the covariates' effects in each of the two identified subgroups by using the sgmcp method. The distribution is more homogeneous within each subgroup in Fig. 1. The dataset is collected from 149 subjects with 56 functional variables and a scalar response.

and CF are fabrication processes. The cell process is to assemble TFT and CF into LCD panels. The module process then assembles LCD panels with other necessary parts to complete final TFT-LCD products [26].

The data we analyze is collected in the TFT process. In this process, circuit is connected to glass substrate to form a TFT board. Glass substrate is coated with an organic film, and what we are concerned about is its thickness. This dataset is collected from 149 subjects. 56 variables are collected, such as temperature, gas, liquid flow, and power with manufacturing time. The values of these variables are recorded over time, and so these variables can be regarded as functional. Without loss of generality, we transform the data for proprietary information protection. To eliminate the effect of the large numerical difference between those variables, we first conduct global rescaling. Following Happ and Greven [7], we use the rescaled elements $w_j^{1/2}x_j(t)$ to build models with $w_j = \left[\int_T \widehat{\mathrm{Var}}\left\{x_j(t_j)\right\} dt_j\right]^{-1}$.

The histogram of thickness after adjusting for the effects of the covariates by using the gmcp method is shown in Fig. 1. It still shows multiple modes among the samples. Therefore, in addition to identifying the affected processes, we also need to group the samples.

With the proposed method, the samples are divided into two subgroups. The subgroup sizes are 97 and 52, respectively. The estimated values of the intercepts are 8.471 and 8.587, respectively. Though the difference may seem small, it is comparable to $\sum_{j=1}^{56}\int_T x_{ij}(t)b_j(t)dt$, the standard deviation of which is 0.132. We present the histogram after adjusting for the effects of the covariates in Fig. 4. We see that the distribution is more homogeneous within each subgroup in Fig. 1. With the alternatives, Cluster divides the samples into four subgroups. As for variable selection, gmcp selects two predictors: S4_ACT_CALIFE01_L and S4_ACT_DC01VOL; Cluster selects one predictor: S4_ACT_CALIFE01_L; and sgmcp selects four predictors: S4_ACT_DC01PWR, S4_ACT_DC01VOL, S4_ACT_TMPOS and ST_ACT_CALIFE01_L. Following Zhang et al. [36], we fit logistic regression of the subgroups (1 for subgroup 1 and 0 for subgroup 2) against the four variables selected by sgmcp, and report the coefficient estimation in Fig. 5. We see that S4_ACT_DC01PWR and S4_ACT_DC01VOL may affect subgrouping. For a new subject, we can predict the response according to this logistic regression.

6. Conclusions

This paper has introduced the methodology and an effective estimation algorithm for subgroup analysis for high-dimensional functional regression. The proposed method can automatically divide observations into subgroups and simultaneously perform variable selection to identify relevant predictors. The objective function includes three parts.

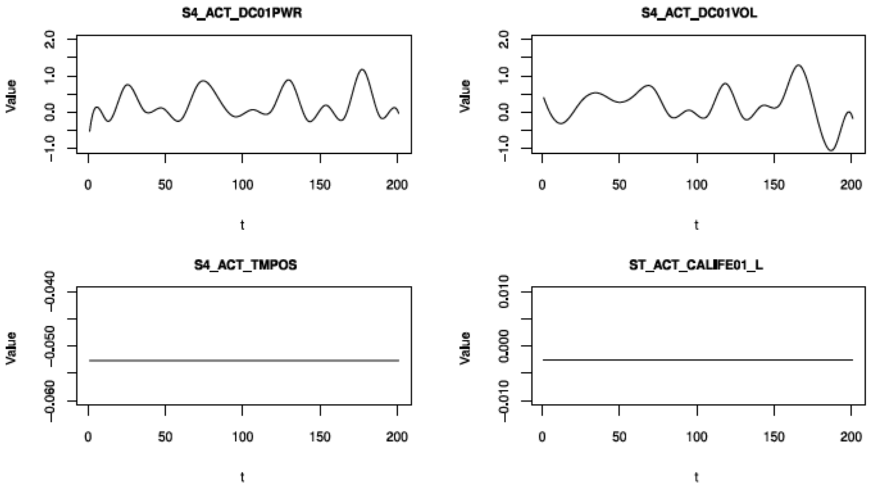

Fig. 5. Plots of the estimated coefficients in the logistic regression of the subgroups (1 for subgroup 1 and 0 for subgroup 2) against the four variables selected by sgmcp (S4_ACT_DC01PWR, S4_ACT_DC01VOL, S4_ACT_TMPOS and ST_ACT_CALIFE01_L).

The first term is the standard squared loss. The second term is the sum of concave penalty functions on the pairwise differences of the intercepts, whose main function is subgrouping. And the last term is used to identify relevant predictors. The ADMM technique has been used to realize the proposed method, and we choose the tuning parameters by BIC. We allow the number of functional predictors and number of subgroups to go to infinity when establishing the oracle property of the proposed estimator. The simulation studies have illustrated that the proposed method is effective in practice.

In this paper, we have assumed that heterogeneity can be modeled through subject-specific intercepts. Extension to the scenario where heterogeneity is modeled through functional coefficients is also worth pursuing. Though Ma et al. [21] have considered a heterogeneous regression model by assuming that the coefficients of treatment variables are subject-dependent, the corresponding algorithm and proof of oracle property will be more complex for functional predictors. A critical statistical challenge for functional coefficients arises from identifying subregions correctly for a heterogeneous population without any subgroup information. Another line of research is robust estimation, for example, against outliers in response and heteroscedasticity in regression by taking advantage of Huber loss.

CRediT authorship contribution statement

Xiaochen Zhang: Methodology, Software, Formal analysis, Writing – original draft. **Qingzhao Zhang:** Conceptualization, Methodology. **Shuangge Ma:** Conceptualization, Writing – review & editing. **Kuangnan Fang:** Supervision, Resource, Conceptualization, Writing – review & editing.

Acknowledgments

We thank the Editor, Associate Editor, and two reviewers for insightful comments and suggestions. This study was supported by the National Natural Science Foundation of China (72071169, 11971404), Humanity and Social Science Youth Foundation of Ministry of Education of China (20YJC910004), National Social Science Foundation of China (21&ZD146), National Science Foundation (1916251), and China Postdoctoral Science Foundation (2021M701999).

Appendix

Regularity assumptions are described in Appendix A. In Appendix B, we state some useful Lemmas and the proof of Theorem 1. Technical assumptions and Theorem 2 for the homogeneous model are given in Appendix C.

Appendix A. Regularity assumptions

Write $\|\alpha\|$ and $\int_T \alpha\beta$ (or $\langle\alpha,\beta\rangle$) for $\left\{\int_T \alpha^2(t)dt\right\}^{1/2}$ and $\int_T \alpha(t)\beta(t)dt$, where $\alpha(\cdot)$ and $\beta(\cdot)$ are square-integrable functions on T. The following condition (CA1) concerns the design on which x_{ij} is observed and the local linear smoother $\hat{x}_{ij}$. Condition (CA2) allows the smooth estimate $\hat{x}_{ij}$ serve as well as the true x_{ij} in asymptotic analysis.

(CA1) For $j \in \{1, \ldots, p_n\}$, $\left\{t_{ijl}, l \in \{1, \ldots, m_{ij}\}\right\}$ are deterministic and ordered increasingly for $i \in \{1, \ldots, n\}$. There exist densities g_{ij} uniformly smooth over i, satisfying $\int_T g_{ij}(t)dt = 1$ and $0 < c_1 < \inf_i \left\{\inf_{t\in T} g_{ij}(t)\right\} < \sup_i \left\{\sup_{t\in T} g_{ij}(t)\right\} < c_2 < \infty$ that generate t_{ijl} according to $t_{ijl} = G_{ij}^{-1}\left\{l/(m_{ij}+1)\right\}$, where G_{ij}^{-1} is the inverse of $G_{ij}(t) = \int_{-\infty}^{t} g_{ij}(s)ds$. For each $j \in \{1, \ldots, d\}$, there exists a common sequence of bandwidth h_j such that $0 < c_1 < \inf_i h_{ij}/h_j < \sup_i h_{ij}/h_j < c_2 < \infty$, where h_{ij} is the bandwidth for the smoothed trajectories $\hat{x}_{ij}$. The kernel density function is smooth and compactly supported.

(CA2) Let $T = [a_0, b_0]$, $t_{ij0} = a_0$, $t_{ij,m_{ij}+1} = b_0$, $\Delta_{ij} = \sup\left\{t_{ij,l+1} - t_{ij,l}, l \in \{0, \ldots, m_{ij}\}\right\}$ and $m_j = m_j(n) = \inf_{i\in\{1,\ldots,n\}} m_{ij}$. For $j \in \{1, \ldots, p_n\}$, $\sup_i \Delta_{ij} = O\left(m_j^{-1}\right)$, $h_j \sim m_j^{-1/5}$, $m_j n^{-5/4} \to \infty$, where we denote $0 < \lim a_n/b_n < \infty$ by $a_n \sim b_n$.

Appendix B. Auxiliary lemmas and Proof of Theorem 1

Lemma 1. *Define the following notations, for $k, k_1, k_2 \in \{1, \ldots, s_n\}$ and $j, j_1, j_2 \in \{1, \ldots, p_n\}$,*

$$\theta_{jk}^{(1)} = \sum_{i=1}^{n}\left(\hat{\xi}_{ijk} - \xi_{ijk}\right)^2 v_{jk}^{-1}, \quad \theta_{k_1k_2}^{j_1j_2(2)} = \sum_{i=1}^{n}\left(\hat{\xi}_{ij_1k_1}\hat{\xi}_{ij_2k_2} - \xi_{ij_1k_1}\xi_{ij_2k_2}\right)\left(v_{j_1k_1}v_{j_2k_2}\right)^{-1/2},$$

$$\theta_{k_1k_2}^{j_1j_2(3)} = \sum_{i=1}^{n}\left\{\xi_{ij_1k_1}\xi_{ij_2k_2} - \mathrm{E}\left(\xi_{j_1k_1}\xi_{j_2k_2}\right)\right\}\left(v_{j_1k_1}v_{j_2k_2}\right)^{-1/2}, \quad \theta_{k_1k_2}^{j_1j_2(4)} = \theta_{k_1k_2}^{j_1j_2(2)} + \theta_{k_1k_2}^{j_1j_2(3)}.$$

Under conditions (C1), (C2), (C5) and (CA1)-(CA2), we have

$$\theta_{jk}^{(1)} = O_p\left(k^{a+2}\right), \quad \theta_{k_1k_2}^{j_1j_2(2)} = O_p\left(k_1^{a/2+1}n^{1/2} + k_2^{a/2+1}n^{1/2}\right),$$

$$\theta_{k_1k_2}^{j_1j_2(3)} = O_p\left(n^{1/2}\right), \quad \theta_{k_1k_2}^{j_1j_2(4)} = O_p\left(k_1^{a/2+1}n^{1/2} + k_2^{a/2+1}n^{1/2}\right),$$

where the $O_p(\cdot)$ and $o_p(\cdot)$ terms are uniform for $k, k_1, k_2 \in \{1, \ldots, s_n\}$ and $j, j_1, j_2 \in \{1, \ldots, p_n\}$.

Lemma 1 quantifies the asymptotic orders of several important types of expressions that will be encountered in the proofs of our lemmas and main theorems. The asymptotic properties of $\hat{\xi}_{ijk}$ are well studied by Kong et al. [12], and we omit the detailed proof here.

Recall that we reparameterize by writing $\tilde{\beta}_{jk} = v_{jk}^{1/2}\beta_{jk}$, $\hat{\boldsymbol{\beta}}$ denotes the estimate of $\boldsymbol{\beta}$, and $\check{\boldsymbol{\beta}}$ denotes the estimate of $\tilde{\boldsymbol{\beta}}$. Define $\tilde{\boldsymbol{W}}_1$ as the $n \times (K + q_n s_n)$ matrix with the ith row

$$\left(z_{i1}, \ldots, z_{iK}, \xi_{i11}v_{11}^{-1/2}, \ldots, \xi_{i1s_n}v_{1s_n}^{-1/2}, \ldots, \xi_{iq_n1}v_{q_n1}^{-1/2}, \ldots, \xi_{iq_ns_n}v_{q_ns_n}^{-1/2}\right).$$

Moreover, define $\check{\boldsymbol{W}}_1$ as the $n \times (K + q_n s_n)$ matrix with the ith row

$$\left(z_{i1}, \ldots, z_{iK}, \hat{\xi}_{i11}v_{11}^{-1/2}, \ldots, \hat{\xi}_{i1s_n}v_{1s_n}^{-1/2}, \ldots, \hat{\xi}_{iq_n1}v_{q_n1}^{-1/2}, \ldots, \hat{\xi}_{iq_ns_n}v_{q_ns_n}^{-1/2}\right).$$

The next lemma characterizes the eigenvalues of $\check{\boldsymbol{W}}_1$. The essential difference between Lemma 3 in Kong et al. [12] and the following lemma is that we allow the number of functional predictors q_n to grow with the sample size n, while Kong et al. [12] assumed it fixed.

Lemma 2. *Under conditions (C1), (C2), (C5), (C7) and (CA1)-(CA2),*

$$\left|\lambda_{\min}\left(\check{\boldsymbol{W}}_1^\top\check{\boldsymbol{W}}_1/|g_{\min}^0|\right) - \lambda_{\min}\left(\boldsymbol{U}n/|g_{\min}^0|\right)\right| = o_p(1), \quad \left|\lambda_{\max}\left(\check{\boldsymbol{W}}_1^\top\check{\boldsymbol{W}}_1/n\right) - \lambda_{\max}(\boldsymbol{U})\right| = o_p(1).$$

Proof. Let $\|A\|_1$ denote the L_1 norm of matrix A. We have:

$$\begin{aligned}|\lambda_{\min}\left(\check{\boldsymbol{W}}_1^\top\check{\boldsymbol{W}}_1/|g_{\min}^0|\right) - \lambda_{\min}\left(\boldsymbol{U}n/|g_{\min}^0|\right)\Big| &\leq \left\|\check{\boldsymbol{W}}_1^\top\check{\boldsymbol{W}}_1/|g_{\min}^0| - \boldsymbol{U}n/|g_{\min}^0|\right\|_1 \\ \leq |g_{\min}^0|^{-1}O_p\Bigg[q_n\sum_{k_1=1}^{s_n}\left|\theta_{k_1s_n}^{j_1j_2(4)}\right| &+ \sum_{l=1}^{K}\sum_{i=1}^{n}\left\{\left(\hat{\xi}_{ijs_n} - \xi_{ijs_n}\right)v_{js_n}^{-1/2}z_{il}\right\} + \sum_{l=1}^{K}\sum_{i=1}^{n}\xi_{ijk}v_{jk}^{-1/2}z_{il} \\ &+ \sum_{j=1}^{q_n}\sum_{k=1}^{s_n}\sum_{i=1}^{n}\left\{\left(\hat{\xi}_{ijk} - \xi_{ijk}\right)v_{jk}^{-1/2}z_{il}\right\} + \sum_{j=1}^{q_n}\sum_{k=1}^{s_n}\sum_{i=1}^{n}\xi_{ijk}v_{jk}^{-1/2}z_{il}\Bigg].\end{aligned} \tag{B.1}$$

Recall that $z_{il} = 1$ if $i \in g_l$ and 0 otherwise. Then we have

$$\left|\sum_{i=1}^{n}\left\{\left(\hat{\xi}_{ijk}-\xi_{ijk}\right)v_{jk}^{-1/2}z_{il}\right\}\right| \le |g_l^0|^{1/2}\left\{\sum_{i=1}^{n}\left(\hat{\xi}_{ijk}-\xi_{ijk}\right)^2 v_{jk}^{-1}\right\}^{1/2} = O_p\left(|g_l^0|^{1/2}k^{a/2+1}\right). \tag{B.2}$$

Since $\mathrm{E}\left[\sum_{i\in g_l^0}\left\{\xi_{ijk}-\mathrm{E}\left(\xi_{ijk}\right)\right\}v_{jk}^{-1/2}\right]^2 \le |g_l^0|\mathrm{E}\left\{\xi_{ijk}-\mathrm{E}\left(\xi_{ijk}\right)\right\}^2 v_{jk}^{-1} = |g_l^0|$, we have

$$\sum_{i\in g_l^0}\left\{\xi_{ijk}-\mathrm{E}\left(\xi_{ijk}\right)\right\}v_{jk}^{-1/2} = O_p\left(|g_l^0|^{1/2}\right).$$

Thus

$$\left|\sum_{i=1}^{n}\xi_{ijk}v_{jk}^{-1/2}z_{il}\right| = \left|\sum_{i=1}^{n}\left\{\xi_{ijk}z_{il}-\mathrm{E}\left(\xi_{ijk}z_{il}\right)\right\}v_{jk}^{-1/2}\right| = \left|\sum_{i\in g_l^0}\left\{\xi_{ijk}-\mathrm{E}\left(\xi_{ijk}\right)\right\}v_{jk}^{-1/2}\right| = O_p\left(|g_l^0|^{1/2}\right). \tag{B.3}$$

By Lemma 1, (B.1)–(B.3) and condition (C5), we have:

$$\begin{aligned}
&\left|\lambda_{\min}\left(\check{\boldsymbol{W}}_1^\top\check{\boldsymbol{W}}_1/|g_{\min}^0|\right)-\lambda_{\min}\left(\boldsymbol{U}n/|g_{\min}^0|\right)\right| \\
\le&|g_{\min}^0|^{-1}O_p\left[q_n\sum_{k_1=1}^{s_n}\left|\theta_{k_1s_n}^{j_1j_2(4)}\right| + \sum_{l=1}^{K}\sum_{i=1}^{n}\left\{\left(\hat{\xi}_{ijs_n}-\xi_{ijs_n}\right)v_{js_n}^{-1/2}z_{il}\right\} + \sum_{l=1}^{K}\sum_{i=1}^{n}\xi_{ijk}v_{jk}^{-1/2}z_{il}\right. \\
&\left.+\sum_{j=1}^{q_n}\sum_{k=1}^{s_n}\sum_{i=1}^{n}\left\{\left(\hat{\xi}_{ijk}-\xi_{ijk}\right)v_{jk}^{-1/2}z_{il}\right\} + \sum_{j=1}^{q_n}\sum_{k=1}^{s_n}\sum_{i=1}^{n}\xi_{ijk}v_{jk}^{-1/2}z_{il}\right] \\
=&|g_{\min}^0|^{-1}O_p\left\{q_n\sum_{k_1=1}^{s_n}\left(k_1^{a/2+1}n^{1/2}+s_n^{a/2+1}n^{1/2}\right) + \sum_{l=1}^{K}|g_l^0|^{1/2}s_n^{a/2+1} + \sum_{l=1}^{K}|g_l^0| + q_n|g_{\max}^0|^{1/2}\sum_{k=1}^{s_n}k^{a/2+1}\right. \\
&\left.+q_ns_n|g_{\max}^0|^{1/2}\right\} \\
=&|g_{\min}^0|^{-1}O_p\left(q_ns_n^{a/2+2}n^{1/2}+Ks_n^{a/2+1}|g_{\max}^0|^{1/2}+K|g_{\max}^0|^{1/2}+q_n|g_{\max}^0|^{1/2}{s_n}^{a/2+2}+q_ns_n|g_{\max}^0|^{1/2}\right) \\
=&O_p\left\{|g_{\min}^0|^{-1}s_n^{a/2+1}\left(q_ns_nn^{1/2}+K|g_{\max}^0|^{1/2}\right)\right\} = o_p(1).
\end{aligned}$$

Similarly, we can get $\left|\lambda_{\max}\left(\check{\boldsymbol{W}}_1^\top\check{\boldsymbol{W}}_1/n\right)-\lambda_{\max}(\boldsymbol{U})\right| = o_p(1)$. This completes the proof. □

Define $\tilde{\eta}_1^0 = \left(\alpha^0, \tilde{\beta}_1^0, \ldots, \tilde{\beta}_{q_n}^0\right)$. Lemma 3 concerns the asymptotic order of $\Gamma_1 = P_{\check{W}_1}\left(\boldsymbol{y}-\check{\boldsymbol{W}}_1\tilde{\eta}_1^0\right)$, where $P_{\check{W}_1} = \check{\boldsymbol{W}}_1\left(\check{\boldsymbol{W}}_1^\top\check{\boldsymbol{W}}_1\right)^{-1}\check{\boldsymbol{W}}_1^\top$.

Lemma 3. *Under conditions (C1), (C2), (C3), (C5) and (CA1)-(CA2),* $\|\Gamma_1\|_2^2 = O_p\left(K+q_ns_n+q_n^2\right)$.

Proof. Note that

$$\begin{aligned}
\|\Gamma_1\|_2^2 &= \left\|P_{\check{W}_1}(\boldsymbol{y}-\check{\boldsymbol{W}}_1\tilde{\eta}_1^0)\right\|_2^2 = \left\|P_{\check{W}_1}\left\{\epsilon+\kappa+\left(\tilde{\boldsymbol{W}}_1-\check{\boldsymbol{W}}_1\right)\tilde{\eta}_1^0\right\}\right\|_2^2 \\
&\le O\left\{\left\|P_{\check{W}_1}\epsilon\right\|_2^2+\left\|P_{\check{W}_1}\kappa\right\|_2^2+\left\|P_{\check{W}_1}\left(\tilde{\boldsymbol{W}}_1-\check{\boldsymbol{W}}_1\right)\tilde{\eta}_1^0\right\|_2^2\right\},
\end{aligned} \tag{B.4}$$

where $\kappa = (\kappa_1, \ldots, \kappa_n)^\top$ and $\kappa_i = \sum_{j=1}^{q_n}\sum_{k=s_n+1}^{\infty}\xi_{ijk}\beta_{jk}^0$.

Since $\left\|P_{\check{W}_1}\epsilon\right\|_2^2 = \epsilon^\top P_{\check{W}_1}\epsilon$,

$$\mathrm{E}\left(\epsilon^\top P_{\check{W}_1}\epsilon\right) = \mathrm{E}\left\{\mathrm{E}\left(\epsilon^\top P_{\check{W}_1}\epsilon \mid \check{\boldsymbol{W}}_1\right)\right\} = \mathrm{E}\left[\mathrm{tr}\left\{P_{\check{W}_1}\mathrm{E}\left(\epsilon^\top\epsilon\right)\right\}\right] = \sigma^2\,\mathrm{tr}\left(P_{\check{W}_1}\right) = \sigma^2\left(q_ns_n+K\right) = O\left(q_ns_n+K\right),$$

and $\mathrm{Var}\left(\epsilon^\top P_{\check{W}_1}\epsilon\right) = \mathrm{E}\left\{\mathrm{Var}\left(\epsilon^\top P_{\check{W}_1}\epsilon|\check{\boldsymbol{W}}_1\right)\right\} + \mathrm{Var}\left\{\mathrm{E}\left(\epsilon^\top P_{\check{W}_1}\epsilon \mid \check{\boldsymbol{W}}_1\right)\right\} = O_p(q_n s_n + K)$, we get

$$\left\|P_{\check{W}_1}\epsilon\right\|_2^2 = O_p(q_n s_n + K). \tag{B.5}$$

As $\mathrm{Var}\left(\sum_{k=s_n+1}^{\infty}\xi_{ijk}\beta_{jk}^0\right) = \sum_{k=s_n+1}^{\infty} v_{jk}\left(\beta_{jk}^0\right)^2 = O\left(\sum_{k=s_n+1}^{\infty}k^{-1}k^{-2b}\right) = O\left(s_n^{-2b}\right)$, we have $\mathrm{Var}\left(\kappa_i\right) = O\left(q_n s_n^{-2b}\right)$, $\mathrm{E}\left(\kappa_i^2\right) = O\left(q_n s_n^{-2b}\right)$. Thus, $\|\kappa\|_2^2 = O_p\left(nq_n s_n^{-2b}\right)$. Then, we have:

$$\left\|P_{\check{W}_1}\kappa\right\|_2^2 \le \|\kappa\|_2^2 = O_p\left(nq_n s_n^{-2b}\right). \tag{B.6}$$

For $P_{\check{W}_1}\left(\tilde{\boldsymbol{W}}_1 - \check{\boldsymbol{W}}_1\right)\tilde{\eta}_1^0$, we have:

$$\begin{aligned}
\left\|P_{\check{W}_1}\ \left(\tilde{\boldsymbol{W}}_1 - \check{\boldsymbol{W}}_1\right)\tilde{\eta}_1^0\right\|_2^2 &\le \left\|\left(\tilde{\boldsymbol{W}}_1 - \check{\boldsymbol{W}}_1\right)\tilde{\eta}_1^0\right\|_2^2 = O\left[\sum_{i=1}^{n}\left\{\sum_{j=1}^{q_n}\sum_{k=1}^{s_n}\left(\hat{\xi}_{ijk} - \xi_{ijk}\right)\beta_{jk}^0\right\}^2\right]\\
&\le O\left[2q_n\sum_{i=1}^{n}\sum_{j=1}^{q_n}\left\{\sum_{k=1}^{s_n}\left(\tilde{\xi}_{ijk} - \xi_{ijk}\right)\beta_{jk}^0\right\}^2 + 2q_n\sum_{i=1}^{n}\sum_{j=1}^{q_n}\left\{\sum_{k=1}^{s_n}\left(\hat{\xi}_{ijk} - \tilde{\xi}_{ijk}\right)\beta_{jk}^0\right\}^2\right]\\
&\le O\left\{q_n\sum_{j=1}^{q_n}\sum_{i=1}^{n}\|x_{ij}\|^2 O_p\left(\sum_{k=1}^{s_n}k^{-b}kn^{-1/2}\right)^2\right\} + O_p\left[q_n\sum_{j=1}^{q_n}\sum_{i=1}^{n}\left\{\|\hat{x}_{ij} - x_{ij}\|^2\left(\sum_{k=1}^{s_n}k^{-b}\right)^2\right\}\right] = O_p\left(q_n^2\right).
\end{aligned} \tag{B.7}$$

By conditions (C4) and (B.4)–(B.7), we have:

$$\|\Gamma_1\|_2^2 \le O_p\left(q_n s_n + K + nq_n s_n^{-2b} + q_n^2\right) = O_p\left(K + q_n s_n + q_n^2\right).$$

This completes the proof. □

Proof of Theorem 1. When the true subgroup memberships of samples are known, that is, $\mathcal{G}^0 = \{g_1^0, \ldots, g_K^0\}$ and $\boldsymbol{Z}$ are known, the oracle estimators for $\boldsymbol{\mu}$ and $\tilde{\boldsymbol{\beta}}$ are:

$$\left(\hat{\boldsymbol{\mu}}^o, \check{\boldsymbol{\beta}}^o\right) = \underset{\boldsymbol{\mu}\in\mathcal{M}_G,\ \tilde{\boldsymbol{\beta}}\in\mathbb{R}^{p_n s_n}}{\arg\min}\ L_n(\boldsymbol{\mu}, \tilde{\boldsymbol{\beta}}) + n\sum_{j=1}^{p_n}P\left(\|\boldsymbol{\beta}_j\|_2, \lambda_2\right),$$

where

$$L_n(\boldsymbol{\mu}, \tilde{\boldsymbol{\beta}}) = \frac{1}{2}\sum_{i=1}^{n}\left\{y_i - \sum_{j=1}^{p_n}\sum_{k=1}^{s_n}\left(\hat{\xi}_{ijk}v_{jk}^{-1/2}\right)\tilde{\beta}_{jk} - \mu_i\right\}^2,$$

and $\mathcal{M}_G$ is the subspace of $\mathbb{R}^n$ defined as

$$\mathcal{M}_G = \left\{\boldsymbol{\mu}\in\mathbb{R}^n : \mu_i = \mu_j,\ \text{for any } i, j\in g_k^0, 1\le k\le K\right\}.$$

Correspondingly, the oracle estimators for the common intercepts $\boldsymbol{\alpha}$ and $\tilde{\boldsymbol{\beta}}$ are

$$\left(\hat{\boldsymbol{\alpha}}^o, \check{\boldsymbol{\beta}}^o\right) = \underset{\boldsymbol{\alpha}\in\mathbb{R}^K,\ \tilde{\boldsymbol{\beta}}\in\mathbb{R}^{p_n s_n}}{\arg\min}\ L_n(\boldsymbol{Z\alpha}, \tilde{\boldsymbol{\beta}}) + n\sum_{j=1}^{p_n}P\left(\|\boldsymbol{\beta}_j\|_2, \lambda_2\right),$$

with $\hat{\boldsymbol{\mu}}^o = \boldsymbol{Z}\hat{\boldsymbol{\alpha}}^o$. For the simplicity of notation, we denote $\tilde{\eta} = \left(\boldsymbol{\alpha}^\top, \tilde{\boldsymbol{\beta}}^\top\right)^\top$ and $Q_n^o\left(\tilde{\eta}\right) = L_n(\boldsymbol{Z\alpha}, \tilde{\boldsymbol{\beta}}) + n\sum_{j=1}^{p_n}P\left(\|\boldsymbol{\beta}_j\|_2, \lambda_2\right)$.

The proof includes two steps. In Step 1, we establish properties of the oracle estimators $\check{\eta}^o = \left(\hat{\boldsymbol{\alpha}}^{o\top}, \check{\boldsymbol{\beta}}^{o\top}\right)^\top$. The oracle estimators are theoretical constructions useful for stating properties of the proposed estimators. In Step 2, we show that $\check{\xi} = \left(\hat{\boldsymbol{\mu}}^{o\top}, \check{\boldsymbol{\beta}}^{o\top}\right)^\top$ with $\hat{\boldsymbol{\mu}}^o = \boldsymbol{Z}\hat{\boldsymbol{\alpha}}^o$ is a local minimizer of the proposed penalized objective function $Q_n(\boldsymbol{\mu}, \tilde{\boldsymbol{\beta}})$ with probability approaching one.

Step 1. We first constrain $L_n(\boldsymbol{Z\alpha}, \tilde{\boldsymbol{\beta}})$ on the subspace where the true zero parameters are set as 0, that is $\left\{\tilde{\boldsymbol{\eta}} = \left(\boldsymbol{\alpha}^\top, \tilde{\boldsymbol{\beta}}^\top\right)^\top \in \mathbb{R}^{K+p_n s_n} : \tilde{\boldsymbol{\beta}}_{\mathcal{A}^c} = 0\right\}$. Define $\tilde{\boldsymbol{\eta}}_1 = \left(\boldsymbol{\alpha}^\top, \tilde{\boldsymbol{\beta}}_{\mathcal{A}}^\top\right)^\top$ and

$$\bar{L}_n(\tilde{\boldsymbol{\eta}}_1) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{q_n}\sum_{k=1}^{s_n}\hat{\xi}_{ijk}v_{jk}^{-1/2}\tilde{\beta}_{jk} - \boldsymbol{z}_i^\top\boldsymbol{\alpha}\right)^2.$$

We now show that there exists a local minimizer $\check{\boldsymbol{\eta}}_1^{oo}$ of $\bar{L}_n(\tilde{\boldsymbol{\eta}}_1)$ such that $\left\|\check{\boldsymbol{\eta}}_1^{oo} - \tilde{\boldsymbol{\eta}}_1^0\right\|_2 = O_p(\alpha_n)$, where $\tilde{\boldsymbol{\eta}}_1^0 = \left(\boldsymbol{\alpha}^{0\top}, \tilde{\boldsymbol{\beta}}_{\mathcal{A}}^{0\top}\right)^\top$ and $\alpha_n = \left(K + q_n s_n + q_n^2\right)^{1/2} n^{1/2}|g_{\min}^0|^{-1}$. If we can prove that for any $\varepsilon > 0$, there exists a large constant C_6 such that

$$\Pr\left\{\inf_{\|\boldsymbol{u}\|_2 = C_6}\bar{L}_n(\tilde{\boldsymbol{\eta}}_1^0 + \alpha_n\boldsymbol{u}) > \bar{L}_n(\tilde{\boldsymbol{\eta}}_1^0)\right\} > 1 - \epsilon, \tag{B.8}$$

then $\bar{L}_n(\tilde{\boldsymbol{\eta}}_1)$ has a local minimizer $\check{\boldsymbol{\eta}}_1^{oo}$ that satisfies $\left\|\check{\boldsymbol{\eta}}_1^{oo} - \tilde{\boldsymbol{\eta}}_1^0\right\|_2 = O_p(\alpha_n)$.

We have

$$\begin{aligned}\bar{L}_n(\tilde{\boldsymbol{\eta}}_1^0 + \alpha_n\boldsymbol{u}) - \bar{L}_n(\tilde{\boldsymbol{\eta}}_1^0) &= \frac{1}{2}\left\|Y - \check{\boldsymbol{W}}_1^\top\left(\tilde{\boldsymbol{\eta}}_1^0 + \alpha_n\boldsymbol{u}\right)\right\|_2^2 - \frac{1}{2}\left\|Y - \check{\boldsymbol{W}}_1^\top\tilde{\boldsymbol{\eta}}_1^0\right\|_2^2 \\ &= \frac{1}{2}\alpha_n^2\boldsymbol{u}^\top\left(\check{\boldsymbol{W}}_1^\top\check{\boldsymbol{W}}_1\right)\boldsymbol{u} - \Gamma_1^\top\check{\boldsymbol{W}}_1\alpha_n\boldsymbol{u} \geqslant \frac{1}{2}\alpha_n^2\boldsymbol{u}^\top\left(\check{\boldsymbol{W}}_1^\top\check{\boldsymbol{W}}_1\right)\boldsymbol{u} - \|\Gamma_1\|_2\,\lambda_{\max}^{1/2}\left(\check{\boldsymbol{W}}_1^\top\check{\boldsymbol{W}}_1\right)\alpha_n\|\boldsymbol{u}\|_2.\end{aligned} \tag{B.9}$$

Let $r_n = \left(K + q_n s_n + q_n^2\right)^{1/2}$. According to Lemma 2, we have

$$\frac{1}{2}\alpha_n^2\boldsymbol{u}^\top\left(\check{\boldsymbol{W}}_1^\top\check{\boldsymbol{W}}_1\right)\boldsymbol{u} \geqslant \frac{1}{2}\lambda_{\min}\left(\check{\boldsymbol{W}}_1^\top\check{\boldsymbol{W}}_1\right)\alpha_n^2\|\boldsymbol{u}\|_2^2 \geqslant C_4|g_{\min}^0|\alpha_n^2\|\boldsymbol{u}\|_2^2 = C_4 n|g_{\min}^0|^{-1}r_n^2\|\boldsymbol{u}\|_2^2. \tag{B.10}$$

Besides, we have

$$\left|-\|\Gamma_1\|_2\,\lambda_{\max}^{1/2}\left(\check{\boldsymbol{W}}_1^\top\check{\boldsymbol{W}}_1\right)\alpha_n\|\boldsymbol{u}\|_2\right| \leq C_7 r_n n^{1/2}\alpha_n\|\boldsymbol{u}\|_2 = C_7 n|g_{\min}^0|^{-1}r_n^2\|\boldsymbol{u}\|_2, \tag{B.11}$$

where C_7 is a positive constant. Allowing $C_6 = \|\boldsymbol{u}\|_2$ to be large enough, (B.11) is dominated by (B.10), which is positive. This proves (B.8).

Denote $\check{\boldsymbol{\eta}}^{oo} = \left(\check{\boldsymbol{\eta}}_1^{oo\top}, \check{\boldsymbol{\beta}}_{\mathcal{A}^c}^{oo\top}\right)^\top \in \mathbb{R}^{K+p_n s_n}$ with $\check{\boldsymbol{\beta}}_{\mathcal{A}^c}^{oo} = 0$. Next, we show that $\check{\boldsymbol{\eta}}^{oo}$ is a local minimizer of $Q_n^o(\tilde{\boldsymbol{\eta}})$ over the whole space $\mathbb{R}^{K+p_n s_n}$. By the Karush–Kuhn–Tucker conditions, it suffices to show that $\check{\boldsymbol{\eta}}^{oo}$ satisfies the following conditions:

$$S_l^*\left(\check{\boldsymbol{\eta}}^{oo}\right) = 0, \quad l \in \{1, \ldots, K\}, \tag{B.12}$$

$$\boldsymbol{S}_j\left(\check{\boldsymbol{\eta}}^{oo}\right) + nP'\left(\|\hat{\boldsymbol{\beta}}_j^{oo}\|_2, \lambda_2\right)\frac{\hat{\boldsymbol{\beta}}_j^{oo}}{\|\hat{\boldsymbol{\beta}}_j^{oo}\|_2} = 0, \quad j \in \mathcal{A}, \tag{B.13}$$

$$\left\|\boldsymbol{S}_j(\check{\boldsymbol{\eta}}^{oo})\right\|_2 \leq \lambda_2 n, \quad j \in \mathcal{A}^c, \tag{B.14}$$

where

$$S_l^*(\tilde{\boldsymbol{\eta}}) = \partial\left\{\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{l=1}^{K}z_{il}\alpha_l - \sum_{j=1}^{p_n}\sum_{k=1}^{s_n}\hat{\xi}_{ijk}\beta_{jk}\right)^2\right\}\Big/\partial\alpha_l,$$

$$\boldsymbol{S}_j(\tilde{\boldsymbol{\eta}}) = \partial\left\{\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{l=1}^{K}z_{il}\alpha_l - \sum_{j=1}^{p_n}\sum_{k=1}^{s_n}\hat{\xi}_{ijk}\beta_{jk}\right)^2\right\}\Big/\partial\boldsymbol{\beta}_j,$$

$\hat{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}$, $\check{\boldsymbol{\beta}}$ is the estimate of $\tilde{\boldsymbol{\beta}}$, and $\tilde{\beta}_{jk} = v_{jk}^{1/2}\beta_{jk}$.

Obviously, we have $S_l^*\left(\check{\boldsymbol{\eta}}^{oo}\right) = 0$ for $l \in \{1, \ldots, K\}$, and (B.12) holds. If $\min_{j\in\mathcal{A}}\|\hat{\boldsymbol{\beta}}_j^{oo}\|_2 \geq a_\lambda\lambda_2$, we have $P'\left(\|\hat{\boldsymbol{\beta}}_j^{oo}\|_2, \lambda_2\right) = 0$, and certainly (B.13) holds. Note that $\|\hat{\boldsymbol{\beta}}_j^{oo}\|_2 \geq \|\boldsymbol{\beta}_j^0\|_2 - \|\hat{\boldsymbol{\beta}}_j^{oo} - \boldsymbol{\beta}_j^0\|_2$. We have $\min_{j\in\mathcal{A}}\|\boldsymbol{\beta}_j^0\|_2/\lambda_2 \to \infty$ under condition (C6). Since $\|\check{\boldsymbol{\beta}}_j^{oo} - \tilde{\boldsymbol{\beta}}_j^0\|_2 = O_p(\alpha_n)$ and $\tilde{\beta}_{jk}^0 = v_{jk}^{1/2}\beta_{jk}^0$, we have $\|\hat{\boldsymbol{\beta}}_j^{oo} - \boldsymbol{\beta}_j^0\|_2 = O_p\left(s_n^{a/2}\alpha_n\right) = o_p\left(\lambda_2\right)$. Thus $\min_{j\in\mathcal{A}}\|\hat{\boldsymbol{\beta}}_j^{oo}\|_2/\lambda_2 \to \infty$ in probability. Then we get $\Pr\left(\|\hat{\boldsymbol{\beta}}_j^{oo}\|_2 \geq a_\lambda\lambda_2 \text{ for } j \in \{1, \ldots, q_n\}\right) \to 1$. Then (B.13) follows.

Now we prove (B.14). It suffices to show that

$$\Pr\left\{\max_{j\in\mathcal{A}^c}\left\|\boldsymbol{S}_j(\breve{\eta}^{oo})\right\|_2 > \lambda_2 n\right\} \to 0.$$

Denote the kth element of $\boldsymbol{S}_j(\breve{\eta}^{oo})$ as $S_{jk}(\breve{\eta}^{oo})$. We have

$$\begin{aligned} S_{jk}(\breve{\eta}^{oo}) &= -\sum_{i=1}^{n}\hat{\xi}_{ijk}\left(y_i-\sum_{l=1}^{K}z_{il}\hat{\alpha}_l-\sum_{j=1}^{p_n}\sum_{k=1}^{s_n}\hat{\xi}_{ijk}v_{jk}^{-1/2}\breve{\beta}_{jk}^{oo}\right) = -\sum_{i=1}^{n}\left(\xi_{ijk}+\hat{\xi}_{ijk}-\xi_{ijk}\right) \\ &\quad\times\left(y_i-\sum_{l=1}^{K}z_{il}\hat{\alpha}_l-\sum_{j=1}^{p_n}\sum_{k=1}^{s_n}\hat{\xi}_{ijk}v_{jk}^{-1/2}\breve{\beta}_{jk}^{oo}\right) \\ &= -\sum_{i=1}^{n}\left(\xi_{ijk}+\hat{\xi}_{ijk}-\xi_{ijk}\right)\left\{\kappa_i+\epsilon_i+\sum_{l=1}^{K}z_{il}\left(\alpha_l^0-\hat{\alpha}_l\right)\right. \\ &\quad\left.+\sum_{j=1}^{p_n}\sum_{k=1}^{s_n}\xi_{ijk}v_{jk}^{-1/2}\left(\tilde{\beta}_{jk}^{0}-\breve{\beta}_{jk}^{oo}\right)+\sum_{j=1}^{p_n}\sum_{k=1}^{s_n}\left(\xi_{ijk}-\hat{\xi}_{ijk}\right)v_{jk}^{-1/2}\breve{\beta}_{jk}^{oo}\right\}. \end{aligned}$$

Let $\boldsymbol{\xi}_{\cdot jk}$ be the $n\times 1$ vector with the ith element ξ_{ijk}, $\hat{\boldsymbol{\xi}}_{\cdot jk}$ be the $n\times 1$ vector with the ith element $\hat{\xi}_{ijk}$, and $\Gamma_2 = \boldsymbol{\kappa}+\left(\tilde{\boldsymbol{W}}_1-\breve{\boldsymbol{W}}_1\right)^{\top}\tilde{\eta}_1^0+\breve{\boldsymbol{W}}_1\left(\tilde{\eta}_1^0-\breve{\eta}_1^{oo}\right)$. Then we have

$$S_{jk}(\breve{\eta}^{oo}) = -\left(\boldsymbol{\xi}_{\cdot jk}+\hat{\boldsymbol{\xi}}_{\cdot jk}-\boldsymbol{\xi}_{\cdot jk}\right)^{\top}(\Gamma_2+\boldsymbol{\epsilon}) = -\boldsymbol{\xi}_{\cdot jk}^{\top}\Gamma_2-\boldsymbol{\xi}_{\cdot jk}^{\top}\boldsymbol{\epsilon}-\left(\hat{\boldsymbol{\xi}}_{\cdot jk}-\boldsymbol{\xi}_{\cdot jk}\right)^{\top}\Gamma_2-\left(\hat{\boldsymbol{\xi}}_{\cdot jk}-\boldsymbol{\xi}_{\cdot jk}\right)^{\top}\boldsymbol{\epsilon}.$$

Hence, it follows that

$$\begin{aligned} \Pr\left\{\max_{j\in\mathcal{A}^c}\left\|S_j(\breve{\eta}^{oo})\right\|_2 > \lambda_2 n\right\} &\le \Pr\left\{\max_{j\in\mathcal{A}^c}\max_{k\in\{1,\ldots,s_n\}}\left|S_{jk}(\breve{\eta}^{oo})\right| > \lambda_2 n s_n^{-1/2}\right\} \\ &\le \Pr\left(\max_{j\in\mathcal{A}^c}\max_{k\in\{1,\ldots,s_n\}}\|\boldsymbol{\xi}_{\cdot jk}\|_2\|\Gamma_2\|_2 > \lambda_2 n s_n^{-1/2}/4\right) \\ &\quad+\Pr\left(\max_{j\in\mathcal{A}^c}\max_{k\in\{1,\ldots,s_n\}}|\boldsymbol{\xi}_{\cdot jk}^{\top}\boldsymbol{\epsilon}| > \lambda_2 n s_n^{-1/2}/4\right) \\ &\quad+\Pr\left(\max_{j\in\mathcal{A}^c}\max_{k\in\{1,\ldots,s_n\}}\|\hat{\boldsymbol{\xi}}_{\cdot jk}-\boldsymbol{\xi}_{\cdot jk}\|_2\|\Gamma_2\|_2 > \lambda_2 n s_n^{-1/2}/4\right) \\ &\quad+\Pr\left(\max_{j\in\mathcal{A}^c}\max_{k\in\{1,\ldots,s_n\}}\|\hat{\boldsymbol{\xi}}_{\cdot jk}-\boldsymbol{\xi}_{\cdot jk}\|_2\|\boldsymbol{\epsilon}\|_2 > \lambda_2 n s_n^{-1/2}/4\right) \\ &:= \mathrm{P}_1+\mathrm{P}_2+\mathrm{P}_3+\mathrm{P}_4. \end{aligned} \tag{B.15}$$

From the proof of Lemma 3, we have $\|\Gamma_2\|_2^2 = O_p\left(r_n^2\right) = o_p\left(n s_n^{-1}\lambda_2^2/16\right)$ under condition (C6). As $\sum_{i=1}^{n}\xi_{ijk}^2 = \sum_{i=1}^{n}\left(\int x_{ij}\phi_{jk}\right)^2 \le \sum_{i=1}^{n}\left(\|x_{ij}\|^2\|\phi_{jk}\|^2\right) = O_p(n)$, we have $\max_{j\in\mathcal{A}^c}\max_{k\in\{1,\ldots,s_n\}}\sum_{i=1}^{n}\xi_{ijk}^2 = O_p(n)$. Then we have $\max_{j\in\mathcal{A}^c}\max_{k\in\{1,\ldots,s_n\}}\|\boldsymbol{\xi}_{\cdot jk}\|_2\ \|\Gamma_2\|_2 = o_p\left(\lambda_2 n s_n^{-1/2}/4\right)$. Thus

$$\mathrm{P}_1 = o(1). \tag{B.16}$$

If we combine conditions (C1), (C2) and (C4), we can get

$$\begin{aligned} \Pr\left(\left|\sum_{i=1}^{n}\xi_{ijk}\epsilon_i\right| > \lambda_2 n s_n^{-1/2}/4\right) &= \Pr\left(\left|\sum_{i=1}^{n}\xi_{ijk}v_{jk}^{-1/2}v_{jk}^{1/2}\epsilon_i\right| > \lambda_2 n s_n^{-1/2}/4\right) \le \Pr\left(C_8\left|\sum_{i=1}^{n}v_{jk}^{1/2}\epsilon_i\right| > \lambda_2 n s_n^{-1/2}/4\right) \\ &\le \Pr\left(C_9 k^{-a/2}\left|\sum_{i=1}^{n}\epsilon_i\right| > \lambda_2 n s_n^{-1/2}/4\right) \le 2\exp\left(-\frac{C_3\lambda_2^2 n k^a}{16C_9^2 s_n}\right), \end{aligned}$$

where C_8 and C_9 are positive constants. Then,

$$\begin{aligned} \mathrm{P}_2 &= \Pr\left(\max_{j\in\mathcal{A}^c}\max_{k\in\{1,\ldots,s_n\}}\left|\boldsymbol{\xi}_{\cdot jk}^{\top}\boldsymbol{\epsilon}\right| > \lambda_2 n s_n^{-1/2}/4\right) \le 2p_n s_n\exp\left(-\frac{C_3\lambda_2^2 n}{16C_9^2 s_n}\right) \\ &= 2\exp\left[\log\left(p_n s_n\right)\left\{1-\frac{C_3\lambda_2^2 n}{16C_9^2 s_n\log\left(p_n s_n\right)}\right\}\right] \to 0 \end{aligned} \tag{B.17}$$

under condition (C6). Moreover, we have

$$\max_{j\in\mathcal{A}^c}\max_{k\in\{1,\ldots,s_n\}}\sum_{i=1}^{n}\left(\hat{\xi}_{ijk}-\xi_{ijk}\right)^2=O_p\left(s_n^2\right).$$

Thus $\max_{j\in\mathcal{A}^c}\max_{k\in\{1,\ldots,s_n\}}\|\hat{\boldsymbol{\xi}}_{\cdot jk}-\boldsymbol{\xi}_{\cdot jk}\|_2\|\varGamma_2\|_2=o_p\left(\lambda_2 n s_n^{-1/2}/4\right)$. Then

$$\mathrm{P}_3=o(1). \tag{B.18}$$

Noting that $\|\boldsymbol{\epsilon}\|_2=\left(\sum_{i=1}^{n}\epsilon_i^2\right)^{1/2}=O_p\left(n^{1/2}\right)$ and $\max_{j\in\mathcal{A}^c}\max_{k\in\{1,\ldots,s_n\}}\|\hat{\boldsymbol{\xi}}_{\cdot jk}-\boldsymbol{\xi}_{\cdot jk}\|_2=O_p(s_n)$, we have $\max_{j\in\mathcal{A}^c}\max_{k\in\{1,\ldots,s_n\}}\|\hat{\boldsymbol{\xi}}_{\cdot jk}-\boldsymbol{\xi}_{\cdot jk}\|_2\|\boldsymbol{\epsilon}\|_2=O_p\left(s_n n^{1/2}\right)=o_p\left(\lambda_2 n s^{-1/2}/4\right)$ by condition (C6). Then

$$\mathrm{P}_4=\Pr\left(\max_{j\in\mathcal{A}^c}\max_{k\in\{1,\ldots,s_n\}}\|\hat{\boldsymbol{\xi}}_{\cdot jk}-\boldsymbol{\xi}_{\cdot jk}\|_2\|\boldsymbol{\epsilon}\|_2>\lambda_2 n s_n^{-1/2}/4\right)\to 0. \tag{B.19}$$

By (B.15)–(B.19), (B.14) follows. Hence $\check{\eta}^{oo\top}=(\check{\eta}_1^{oo\top},\mathbf{0}^\top)^\top$ is a local minimizer of $Q_n^o(\tilde{\eta})$ over the whole space $R^{K+p_n s_n}$. It is easy to deduce that $\Pr\left\{\hat{b}_{q_n+1}^{oo}(t)=\cdots=\hat{b}_{p_n}^{oo}(t)=0\right\}\to 1$, where $\hat{b}_j^{oo}(t)=\sum_{k=1}^{s_n}\hat{\beta}_{jk}^{oo}\hat{\phi}_{jk}(t)$ is the corresponding functional coefficient.

Step 2. Recall that the oracle estimators is denoted by $\check{\eta}^o=\left(\hat{\alpha}^{o\top},\check{\beta}^{o\top}\right)^\top$, and

$$Q_n(\boldsymbol{\mu},\tilde{\boldsymbol{\beta}})=L_n(\boldsymbol{\mu},\tilde{\boldsymbol{\beta}})+\sum_{1\le i<j\le n}P\left(|\mu_i-\mu_j|,\lambda_1\right)+n\sum_{j=1}^{p_n}P\left(\|\boldsymbol{\beta}_j\|_2,\lambda_2\right),$$

where

$$L_n(\boldsymbol{\mu},\tilde{\boldsymbol{\beta}})=\frac{1}{2}\sum_{i=1}^{n}\left\{y_i-\sum_{j=1}^{p_n}\sum_{k=1}^{s_n}\left(\hat{\xi}_{ijk}v_{jk}^{-1/2}\right)\tilde{\beta}_{jk}-\mu_i\right\}^2=\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p_n}\sum_{k=1}^{s_n}\hat{\xi}_{ijk}\beta_{jk}-\mu_i\right)^2.$$

Now we show that $\check{\zeta}=\left(\hat{\mu}^{o\top},\check{\beta}^{o\top}\right)^\top$ with $\hat{\mu}^o=\boldsymbol{Z}\hat{\alpha}^o$ is a local minimizer of the proposed penalized objective function $Q_n(\boldsymbol{\mu},\tilde{\boldsymbol{\beta}})$ with probability approaching one. Denote $T_i^*(\zeta)=\partial\left\{1/2\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p_n}\sum_{k=1}^{s_n}\hat{\xi}_{ijk}\beta_{jk}-\mu_i\right)^2\right\}\Big/\partial\mu_i$ and $\boldsymbol{T}_j(\zeta)=\partial\left\{1/2\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p_n}\sum_{k=1}^{s_n}\hat{\xi}_{ijk}\beta_{jk}-\mu_i\right)^2\right\}\Big/\partial\boldsymbol{\beta}_j$. Inspired by the proof of Theorem 1 in Jeon et al. [11], the goal is to show that $\check{\zeta}$ satisfies the following conditions with probability tending to one, provided that conditions (C1)-(C7) and (CA1)-(CA2) hold. That is

$$\min_{j\in\mathcal{A}}\|\hat{\boldsymbol{\beta}}_j^o\|_2>a_\lambda\lambda_2,\quad \max_{j\in\mathcal{A}^c}\left\|\boldsymbol{T}_j(\check{\zeta})\right\|_2\le\lambda_2 n,\quad \sum_{i\in g_k^0}T_i^*(\check{\zeta})=0\ ,\ 1\le k\le K, \tag{B.20}$$

$$\min_{i\in g_k^0,j\in g_l^0,1\le k<l\le K}\left|\hat{u}_i^0-\hat{u}_j^0\right|>a_\lambda\lambda_1, \tag{B.21}$$

$$\max_{i\in g_k^0,1\le k\le K}\left|T_i^*(\check{\zeta})\right|/\left(|g_k^0|-1\right)\le\lambda_1. \tag{B.22}$$

Obviously, (B.20) holds by the definition of $\check{\zeta}$. Next, we prove (B.21). Under condition (C6), we have $\left\|\hat{\alpha}^0-\alpha^0\right\|_2=O_p\left(\alpha_n\right)=o_p\left(\lambda_1\right)$. As

$$\min_{i\in g_k^0,j\in g_l^0,1\le k<l\le K}\left|\hat{u}_i^0-\hat{u}_j^0\right|/\lambda_1\ge\min_{1\le k<l\le K}\left|\alpha_k^0-\alpha_l^0\right|/\lambda_1-2\left\|\hat{\alpha}^0-\alpha^0\right\|_2/\lambda_1,$$

we have $\Pr\left(\min_{i\in g_k^0,j\in g_l^0,1\le k<l\le K}\left|\hat{u}_i^0-\hat{u}_j^0\right|\le a_\lambda\lambda_1\right)\to 0$ as $n\to\infty$.

Now we prove (B.22). It suffices to show that

$$\Pr\left\{\max_{i\in g_k^0,1\le k\le K}\left|T_i^*(\check{\zeta})\right|>\left(|g_k^0|-1\right)\lambda_1\right\}\to 0. \tag{B.23}$$

For $k \in \{1, \dots, K\}$, we have $T_i^*(\zeta) = -\left(\Gamma_2^i + \epsilon_i\right)$, where Γ_2^i is the ith component of Γ_2. Hence, it follows that

$$\begin{aligned}\Pr\left\{\max_{i \in g_k^0, 1 \le k \le K} \left|T_i^*(\breve{\zeta})\right| > \left(|g_k^0| - 1\right)\lambda_1\right\} \le \Pr&\left\{\max_{i \in g_k^0, 1 \le k \le K} \left|\Gamma_2^i\right| > \left(|g_{\min}^0| - 1\right)\lambda_1/2\right\} \\ &+ \Pr\left\{\max_{i \in g_k^0, 1 \le k \le K} |\epsilon_i| > \left(|g_{\min}^0| - 1\right)\lambda_1/2\right\}.\end{aligned}$$

From the previous derivations, we have $\|\Gamma_2\|_2^2 = O_p\left(r_n^2\right)$. Then we have

$$\Pr\left\{\max_{i \in g_k^0, 1 \le k \le K} \left|\Gamma_2^i\right| > \left(|g_{\min}^0| - 1\right)\lambda_1/2\right\} \le \Pr\left\{\|\Gamma_2\|_2 > \left(|g_{\min}^0| - 1\right)\lambda_1/2\right\} \to 0 \tag{B.24}$$

under condition (C6). Besides, we have

$$\Pr\left\{|\epsilon_i| > \left(|g_{\min}^0| - 1\right)\lambda_1/2\right\} \le 2\exp\left\{-C_3\lambda_1^2\left(|g_{\min}^0| - 1\right)^2/4\right\}.$$

Then

$$\begin{aligned}\Pr\left\{\max_{i \in g_k^0, 1 \le k \le K} |\epsilon_i| > \left(|g_{\min}^0| - 1\right)\lambda_1/2\right\} &\le 2n\exp\left\{-C_3\lambda_1^2\left(|g_{\min}^0| - 1\right)^2/4\right\} \\ &= 2\exp\left[\log n\left\{1 - C_3\lambda_1^2\left(|g_{\min}^0| - 1\right)^2/(4\log n)\right\}\right] \to 0\end{aligned} \tag{B.25}$$

under condition (C6). By (B.24) and (B.25), (B.23) holds.

This completes the proof. □

Appendix C. Technical assumptions and theorem for a homogeneity model

When the true model is homogeneous given as Eq. (1) with $\mu_1 = \cdots = \mu_n = \mu = \alpha$ and $K = 1$, we also reparameterize by writing $\tilde{\beta}_{jk} = v_{jk}^{1/2}\beta_{jk}$ and introduce the following conditions.

(C8) The smoothing parameters s_n and q_n satisfy $q_n^2 s_n^{a+4}/n \to 0$, $s_n^{2a+2}/n \to 0$, $s_n^{2b-1}/n \to \infty$.

(C9) Tuning parameters λ_1 and λ_2 satisfy: (i) $\lambda_1 = o(1)$, $\left(q_n s_n + q_n^2\right)n^{-1} = o\left(\lambda_1^2\right)$; (ii) $\lambda_2 = o(1)$, $\min_{j \in \mathcal{A}} \|\beta_j^0\|_2/\lambda_2 \to \infty$, $\max\left\{s_n^a\left(q_n s_n + q_n^2\right)n^{-1}, s_n^3 n^{-1}, s_n \log\left(p_n s_n\right)n^{-1}\right\} = o\left(\lambda_2^2\right)$.

(C10) Define $\boldsymbol{U}^* = \begin{pmatrix}1 & 0 \\ 0 & \mathrm{E}(\tilde{\boldsymbol{N}}_i\tilde{\boldsymbol{N}}_i^\top)\end{pmatrix}$, where $\tilde{\boldsymbol{N}}_i = \left(\xi_{i11}v_{11}^{-1/2}, \dots, \xi_{iq_n s_n}v_{q_n s_n}^{-1/2}\right)^\top$ is $(q_n s_n) \times 1$. $0 < C_4^* \le \lambda_{\min}(\boldsymbol{U}^*) \le \lambda_{\max}(\boldsymbol{U}^*) \le C_5^* < \infty$.

Theorem 2. *Under conditions (C1)-(C4), (C8)-(C10) and (CA1)–(CA2), there exists a local minimizer $\left(\hat{\mu}^\top, \breve{\beta}^\top\right)^\top$ of objective function $Q_n(\mu, \tilde{\beta})$ satisfying*

(i) $\Pr\left(\hat{\mu}_i = \hat{\mu}_j, \forall i, j\right) \to 1$, *i.e.*, $\Pr\left(\hat{K} = 1\right) \to 1$,

(ii) $\Pr\left\{\hat{b}_{q_n+1}(t) = \cdots = \hat{b}_{p_n}(t) = 0\right\} \to 1$,

(iii) $\left\|\left(\hat{\mu}^\top, \breve{\beta}^\top\right)^\top - \left(\mu^{0\top}, \tilde{\beta}^{0\top}\right)^\top\right\|_2 = O_p\left\{\left(q_n s_n + q_n^2\right)^{1/2} n^{-1/2}\right\}$.

Theorem 2 shows that homogeneity and sparsity can be recovered with a high probability. Similar to Theorem 1, we can conclude that $\left\|\hat{b}_j(t) - b_j(t)\right\|^2 = O_p\left\{s_n^a\left(q_n s_n + q_n^2\right)n^{-1}\right\}$. Next, we give a brief sketch of the proof of Theorem 2.

Proof of Theorem 2. When the true subgroup memberships of samples are known, that is, $\boldsymbol{Z} = \mathbf{1}_n$ is known, the oracle estimators for $\boldsymbol{\mu}$ and $\tilde{\boldsymbol{\beta}}$ are:

$$\left(\hat{\mu}^o, \breve{\beta}^o\right) = \underset{\mu \in \mathcal{M},\ \tilde{\beta} \in \mathbb{R}^{p_n s_n}}{\arg\min}\ L_n(\mu, \tilde{\beta}) + n\sum_{j=1}^{p_n} P\left(\|\beta_j\|_2, \lambda_2\right),$$

where

$$L_n(\mu, \tilde{\beta}) = \frac{1}{2}\sum_{i=1}^n\left\{y_i - \sum_{j=1}^{p_n}\sum_{k=1}^{s_n}\left(\hat{\xi}_{ijk}v_{jk}^{-1/2}\right)\tilde{\beta}_{jk} - \mu_i\right\}^2,$$

and $\mathcal{M}$ is the subspace of $\mathbb{R}^n$ defined as

$$\mathcal{M} = \left\{\mu \in \mathbb{R}^n : \mu_1 = \cdots = \mu_n\right\}.$$

Correspondingly, the oracle estimators for the common intercepts α and $\tilde{\boldsymbol{\beta}}$ are

$$\left(\hat{\alpha}^o, \check{\boldsymbol{\beta}}^o\right) = \underset{\alpha\in\mathbb{R},\ \tilde{\boldsymbol{\beta}}\in\mathbb{R}^{p_n s_n}}{\arg\min}\ L_n(\mathbf{1}_n\alpha, \tilde{\boldsymbol{\beta}}) + n\sum_{j=1}^{p_n} P\left(\|\boldsymbol{\beta}_j\|_2, \lambda_2\right),$$

with $\hat{\boldsymbol{\mu}}^o = \mathbf{1}_n\hat{\alpha}^o$. Denote $\tilde{\eta} = \left(\alpha, \tilde{\boldsymbol{\beta}}^\top\right)^\top$ and $Q_n^o(\tilde{\eta}) = L_n(\mathbf{1}_n\alpha, \tilde{\boldsymbol{\beta}}) + n\sum_{j=1}^{p_n} P\left(\|\boldsymbol{\beta}_j\|_2, \lambda_2\right)$.

The proof includes two steps. In Step 1, we establish properties of the oracle estimators $\check{\eta}^o = \left(\hat{\alpha}^o, \check{\boldsymbol{\beta}}^{o\top}\right)^\top$. The proof follows the same arguments as the proof of Theorem 1 by letting $\boldsymbol{Z} = \mathbf{1}_n$ and $|g^0_{\min}| = n$. Thus we omit it. In Step 2, we show that $\check{\xi} = \left(\hat{\boldsymbol{\mu}}^{o\top}, \check{\boldsymbol{\beta}}^{o\top}\right)^\top$ with $\hat{\boldsymbol{\mu}}^o = \mathbf{1}_n\hat{\alpha}^o$ is a local minimizer of the proposed penalized objective function $Q_n(\mu, \tilde{\boldsymbol{\beta}})$ with probability approaching one. It follows similar procedures as the proof of Theorem 1 with details below. The goal is to show that $\check{\xi}$ satisfies the following additional conditions with probability tending to 1, under conditions (C1)-(C4), (C8)-(C10) and (CA1)-(CA2). We note that:

$$\min_{j\in\mathcal{A}} \|\hat{\boldsymbol{\beta}}^o_j\|_2 > a_\lambda\lambda_2, \quad \max_{j\in\mathcal{A}^c} \left\|\boldsymbol{T}_j(\check{\xi})\right\|_2 \le \lambda_2 n, \quad \sum_{i=1}^{n} T_i^*(\check{\xi}) = 0, \quad \max_i \left|T_i^*(\check{\xi})\right| /(n-1) \le \lambda_1.$$

The proof follows the same arguments as the proof of Theorem 1. We omit details here. □

References

[1] H. Cardot, F. Ferraty, A. Mas, Testing hypotheses in the functional linear model, Scand. J. Stat. 30 (1) (2003) 241–255.

[2] J. Chen, Z. Chen, Extended Bayesian information criteria for model selection with large model spaces, Biometrika 95 (3) (2008) 759–771.

[3] J.S. Chou, Y. Tai, L.J. Chang, Predicting the development cost of TFT-LCD manufacturing equipment with artificial intelligence models, Int. J. Prod. Econ. 128 (1) (2010) 339–350.

[4] A. Cuevas, M. Febrero, R. Fraiman, Linear functional regression: the case of fixed design and functional response, Canad. J. Statist. 30 (2) (2002) 285–300.

[5] Y. Fan, G.M. James, P. Radchenko, Functional additive regression, Ann. Statist. 43 (5) (2015) 2296–2325.

[6] P. Hall, J.L. Horowitz, Methodology and convergence rates for functional linear regression, Ann. Statist. 35 (1) (2007) 70–91.

[7] C. Happ, S. Greven, Multivariate functional principal component analysis for data observed on different (dimensional) domains, J. Amer. Statist. Assoc. 113 (522) (2018) 649–659.

[8] C.Y. Hsu, C.F. Chien, K.Y. Lin, C.Y. Chien, Data mining for yield enhancement in TFT-LCD manufacturing: an empirical study, J. Chin. Inst. Ind. Eng. 27 (2) (2010) 140–156.

[9] X. Hu, J. Huang, L. Liu, D. Sun, X. Zhao, Subgroup analysis in the heterogeneous Cox model, Stat. Med. 40 (3) (2021) 739–757.

[10] J. Huang, P. Breheny, S. Ma, A selective review of group selection in high-dimensional models, Statist. Sci. 27 (4) (2012) 481–499.

[11] J. Jeon, S. Kwon, H. Choi, Homogeneity detection for the high-dimensional generalized linear model, Comput. Statist. Data Anal. 114 (2017) 61–74.

[12] D. Kong, K. Xue, F. Yao, H.H. Zhang, Partially functional linear regression in high dimensions, Biometrika 103 (1) (2016) 147–159.

[13] J. Lei, Adaptive global testing for functional linear models, J. Amer. Statist. Assoc. 109 (506) (2014) 624–634.

[14] Y. Li, T. Hsing, On rates of convergence in functional linear regression, J. Multivariate Anal. 98 (9) (2007) 1782–1804.

[15] H. Lian, Shrinkage estimation and selection for multiple functional regression, Statist. Sinica 23 (1) (2013) 51–74.

[16] L. Liu, M. Gordon, J.P. Miller, M. Kass, L. Lin, S. Ma, L. Liu, Capturing heterogeneity in repeated measures data by fusion penalty, Stat. Med. 40 (8) (2021) 1901–1916.

[17] L. Liu, L. Lin, Subgroup analysis for heterogeneous additive partially linear models and its application to car sales data, Comput. Statist. Data Anal. 138 (2019) 239–259.

[18] W. Lu, G. Qin, Z. Zhu, D. Tu, Multiply robust subgroup identification for longitudinal data with dropouts via median regression, J. Multivariate Anal. 181 (2021) 104691.

[19] Y. Lv, X. Zhu, Z. Zhu, A. Qu, Nonparametric cluster analysis on multiple outcomes of longitudinal data, Statist. Sinica 30 (4) (2020) 1829–1856.

[20] S. Ma, J. Huang, A concave pairwise fusion approach to subgroup analysis, J. Amer. Statist. Assoc. 112 (517) (2017) 410–423.

[21] S. Ma, J. Huang, Z. Zhang, M. Liu, Exploration of heterogeneous treatment effects via concave fusion, Int. J. Biostat. 16 (1) (2019) 20180026.

[22] H. Ma, T. Li, H. Zhu, Z. Zhu, Quantile regression for functional partially linear model in ultra-high dimensions, Comput. Statist. Data Anal. 129 (2019) 135–147.

[23] J. Ramsay, B. Silverman, Functional Data Analysis, Springer, New York, 2005.

[24] J. Shen, X. He, Inference for subgroup analysis with a structured logistic-normal mixture model, J. Amer. Statist. Assoc. 110 (509) (2015) 303–312.

[25] Y.C. Su, M.H. Hung, F.T. Cheng, Y.T. Chen, A processing quality prognostics scheme for plasma sputtering in TFT-LCD manufacturing, IEEE Trans. Semicond. Manuf. 19 (2) (2006) 183–194.

[26] K. Tsai, S. Wang, Multi-site available-to-promise modeling for assemble-to-order manufacturing: an illustration on TFT-LCD manufacturing, Int. J. Prod. Econ. 117 (1) (2009) 174–184.

[27] Y. Ukai, TFT-LCD manufacturing technology—current status and future prospect–, in: 2007 International Workshop on Physics of Semiconductor Devices, IEEE, 2007, pp. 29–34.

[28] X. Wang, Clustering of longitudinal curves via a penalized method and em algorithm, 2019, p. 1910.11258, arXiv: Methodology arXiv.

[29] W. Wang, Y. Sun, H.J. Wang, Latent group detection in functional partially linear regression models, Biometrics (2021) 1–12, http://dx.doi.org/10.1111/biom.13557.

[30] X. Wang, Z. Zhu, H.H. Zhang, Spatial automatic subgroup analysis for areal data with repeated measures, 2019, p. 1906.01853, arXiv: Methodology arXiv.
[31] S. Wei, M.R. Kosorok, Latent supervised learning, J. Amer. Statist. Assoc. 108 (503) (2013) 957–970.
[32] X. Yan, G. Yin, X. Zhao, Subgroup analysis in censored linear regression, Statist. Sinica 31 (2) (2021) 1027–1054.
[33] X. Yang, X. Yan, J. Huang, High-dimensional integrative analysis with homogeneity and sparsity recovery, J. Multivariate Anal. 174 (2019) 104529.
[34] F. Yao, H.G. Müller, J.L. Wang, Functional data analysis for sparse longitudinal data, J. Amer. Statist. Assoc. 100 (470) (2005) 577–590.
[35] F. Yao, H.G. Müller, J.L. Wang, Functional linear regression analysis for longitudinal data, Ann. Statist. 33 (6) (2005) 2873–2903.
[36] Y. Zhang, H.J. Wang, Z. Zhu, Robust subgroup identification, Statist. Sinica 29 (4) (2019) 1873–1889.
[37] X. Zhang, Y. Wu, L. Wang, R. Li, Variable selection for support vector machines in moderately high dimensions, J. R. Stat. Soc. Ser. B Stat. Methodol. 78 (1) (2016) 53.
[38] X. Zhu, A. Qu, Cluster analysis of longitudinal profiles with subgroups, Electron. J. Stat. 12 (1) (2018) 171–193.
[39] H. Zhu, M. Vannucci, D.D. Cox, A Bayesian hierarchical model for classification with selection of functional predictors, Biometrics 66 (2) (2010) 463–473.