Computational Statistics and Data Analysis 158 (2021) 107192

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Clusterwise functional linear regression models

Ting Li [a], Xinyuan Song [b], Yingying Zhang [c], Hongtu Zhu [d], Zhongyi Zhu [e,*]

[a] School of Statistics and Management, Shanghai University of Finance and Economics, China
[b] Department of Statistics, The Chinese University of Hong Kong, Hong Kong
[c] Academy of Statistics and Interdisciplinary Sciences, East China Normal University, China
[d] Department of Biostatistics, University of North Carolina at Chapel Hill, United States of America
[e] Department of Statistics, Fudan University, China

ARTICLE INFO

Article history:
Received 6 May 2020
Received in revised form 29 January 2021
Accepted 30 January 2021
Available online 14 February 2021

Keywords:
Bayesian information criterion consistency
M-estimation
Subgroup analysis

ABSTRACT

Classical clusterwise linear regression is a useful method for investigating the relationship between scalar predictors and scalar responses with heterogeneous variation of regression patterns for different subgroups of subjects. This paper extends the classical clusterwise linear regression to incorporate multiple functional predictors by representing the functional coefficients in terms of a functional principal component basis. We estimate the functional principal component coefficients based on M-estimation and K-means clustering algorithm, which can classify the data into clusters and estimate clusterwise coefficients simultaneously. One advantage of the proposed method is that it is robust and flexible by adopting a general loss function, which can be broadly applied to mean regression, median regression, quantile regression and robust mean regression. A Bayesian information criterion is proposed to select the unknown number of groups and shown to be consistent in model selection. We also obtain the convergence rate of the set of estimators to the set of true coefficients for all clusters. Simulation studies and real data analysis show that the proposed method is easily implemented, and it consequently improves previous works and also requires much less computing burden than existing methods.

1. Introduction

Functional data analysis (FDA) views data as realizations of random process and has become a growing statistical field in recent years (Ramsay and Silverman, 2005). Regression with functional data is perhaps the most thoroughly researched topic in FDA, and there exists an extensive literature on this topic. Among them, the functional linear regression model and its extensions have received considerable attention. The functional linear regression model characterizes the relationship between a scalar response and random predictor processes based on the underlying assumption that the regression structure is the same for all subjects. However, in some applications, clustered patterns of unobserved heterogeneity are common and the relationships may be different across subgroups of individuals. For example, the ADNI study collects mini-mental state examination (MMSE) scores reflecting cognitive mental status and fractional anisotropy (FA) curves reflecting fiber density, axonal diameter, and myelination in white matter. The effects of white matter on the cognitive function can be different for different subgroups of people since these individuals are usually under different status of disease. Controlling the unobserved heterogeneity in modeling based on their FA curves and MMSE scores is of paramount importance and can be used to identify progression of Alzheimer's Disease.

* Corresponding author.
E-mail address: zhuzy@fudan.edu.cn (Z. Zhu).

https://doi.org/10.1016/j.csda.2021.107192
0167-9473/

To detect potential cluster-level heterogeneity in functional linear models, we consider the clusterwise functional linear regression model to deal with the group-specific regression patterns for functional data. Given a scalar response Y and D smooth random predictor processes $\{X_d(\cdot), d = 1, \ldots, D\}$ on a compact support $\mathcal{T}$ that are square integrable, we have the model as follows:

$$Y_i = \mu_k^0 + \sum_{d=1}^{D} \int_{\mathcal{T}} X_{di}(t)\beta_{dk}^0(t)\mathrm{d}t + \epsilon_i, \quad \text{if the subject } i \text{ belongs to the } k\text{th group } \mathcal{C}_{0k}^{(n)}, \tag{1}$$

where μ_k^0 is the intercept, $\beta_{dk}^0(t)$s are the coefficient functions, $\mathcal{C}_{0k}^{(n)}$ is the index set for the kth group and $k = 1, \ldots, K_0$. Within each group, the regression coefficients are the same, while they may be different across different groups. The true group membership $\mathcal{C}_0^{(n)} = \{\mathcal{C}_{01}^{(n)}, \ldots, \mathcal{C}_{0K_0}^{(n)}\}$ and the number of groups K_0 are often unknown.

Model (1) can be viewed as an extension of the clusterwise linear regression model to incorporate functional predictors. As a useful method for investigating potential cluster-level heterogeneity, clusterwise linear regression involves distinct linear relationships between scalar outcome and scalar predictors across clusters. There exists a huge number of different cluster methods for clusterwise linear regression models with scalar covariates. Späth (1979), Wu and Zen (1999) and Rao et al. (2007) combined K-means algorithm with classical estimation procedures, which enjoy the advantages of conceptual simplicity and computational efficiency. DeSarbo and Cron (1988), Hennig (2000) and McLachlan and Peel (2004) proposed mixture models and adopted EM-algorithm to estimate the parameters by specifying the underlying distribution for each component. These methods are efficient when the underlying distribution is correctly specified, but can be susceptible to model misspecification. Ma and Huang (2016), Ma and Huang (2017) and Zhang et al. (2019) proposed penalization methods through penalizing the pairwise differences of coefficients across subjects. One advantage of the penalization-based methods is that there is no need to pre-specify the number of clusters. However, such convenience comes at the cost of heavy computational burden.

Although many different regression clustering methods for clusterwise linear regression models with scalar covariates have been available, methods for model (1) remains relatively undeveloped in the literature with some exceptions. Preda and Saporta (2005, 2007) proposed the clusterwise partial least square and principal components approaches for perfectly observed functional predictors. However, these methods suffer from three limitations. First, they did not provide any theoretical results about the selection of the number of groups and consistency of the estimators. Second, their least-squares loss function is not robust to non-normal errors. Third, this is rarely the case that the functional covariate is observed continuously and without measurement errors. Usually, functional data are observed intermittently and with errors. Yao et al. (2010) and Ciarleglio and Ogden (2016) extended the classical mixture regression model to the functional mixture model. Likewise, they had the restriction that no consistency result of the criterion is guaranteed for specifying the number of clusters. Moreover, the assumption of a parametric conditional density of the response in the mixture model approach is restrictive and difficult to meet in practice. Hence, there is a need for the development for clusterwise functional linear regression from both practical and theoretical perspectives.

The focus of this paper is to introduce a computational efficient and robust estimation method for model (1) without assumptions on the underlying distribution, and to provide throughout theoretical investigations with respect to both the coefficient estimators and the estimator of the group number. By adopting the functional principal component analysis, we are able to deal with the infinite-dimensional functional coefficients. We propose a general approach to estimate the functional principal component coefficients based on M-estimation and K-means algorithm, which finds the cluster label of each data point, identifies potentially different regression structures simultaneously. To be specific, given an initial group membership of the data, we iteratively cluster functional data into groups which minimize the loss function according to available regression patterns, and then update the regression in each cluster simultaneously until equilibrium is attained. A Bayesian information criterion is proposed to determine the underlying number of clusters. To improve the robustness of the estimation procedure in the presence of heavier-tailed error distributions, we adopt a general loss function which covers mean regression, median regression, quantile regression and robust mean regression.

Compared with the existing literature, we make several major contributions. First, the proposed K-means based algorithm is more robust and more computational convenient compared to existing methods, as evidenced by our simulation studies and analysis of the ADNI dataset. Second, we establish the consistency of the Bayesian information criterion. Third, we derive the consistency rate of the set of estimators in terms of the Hausdorff distance. In particular, the set of estimators for the intercepts enjoys a parametric rate of convergence, whereas the set of estimators for the functional coefficients possesses a nonparametric rate of convergence, which is shown to be optimal in the minimax sense by Hall and Horowitz (2007). To the best of our knowledge, this work is the first to derive the consistency result for model selection by the Bayesian information criterion and the rates of convergence for parameter estimators in the context of clusterwise functional linear models or functional mixture models. In spite of their high importance, these consistency results are relatively rare even for classical clusterwise linear regression models (Müller and Garlipp, 2005). Fourth, the proposed method is applicable to intermittent and noisy trajectories at the price of a more complex theoretical investigation. Meanwhile, it is flexible by choosing different criterion functions according to various interests, and the theoretical results obtained are applicable to a general loss function. Although conceptually similar to Preda and Saporta (2007), their method is restricted to least-squares loss function with perfectly observed functional variables and their paper lacks theoretical results.

The proposed method falls within a semisupervised clustering framework to functional data, which performs unsupervised learning when it clusters data according to their respective unobserved regression structures, and supervised learning when it fits regression patterns to the corresponding data clusters. The unknown group membership and number of clusters pose challenges and distinguish the proposed method from existing M-estimation based methods for classical functional linear models, such as Huang et al. (2014), Shin and Lee (2016), Tang (2017) and Ma et al. (2019). Furthermore, the proposed method is conceptually different from the curve-based clustering or classification methods, such as in Abraham et al. (2003), Wang and Song (2018) and Delaigle et al. (2019). These methods classify trajectories directly, whereas the proposed method clusters the functional processes and the scalar response together according to possibly different regression patterns with an unknown group membership.

The rest of this paper is organized as follows. Detailed estimation method can be found in Section 2. In Section 3, we establish the consistency of the Bayesian information criterion and convergence of the set of estimators. Section 4 demonstrates the practicality of the proposed method through finite sample simulation studies. We apply our method to a real dataset obtained from the ADNI study and to a dataset from an experiment on medfly fecundity in Section 5. All the proofs can be found in the supplementary material.

2. Estimation

In this section, we present the explicit estimation procedure and algorithm, which are easy to implement and do not depend on the knowledge of the conditional density of the response.

Considering that for $d = 1, \ldots, D$, $k = 1, \ldots K$, each $\beta_{dk}(t)$ lies in an infinite-dimensional space, dimension reduction is mandatory. We apply the functional principal component analysis and represent the functional coefficients by functional principal component basis functions. Specifically, for $d = 1, \ldots, D$, the functional predictor X_{di} admits $X_{di}(t) = \mu_{xd}(t) + \sum_{m=1}^{\infty} \xi_{dim}\varphi_{dm}(t)$, where $\xi_{dim} = \int_{\mathcal{T}}\{X_{di}(t) - \mu_{xd}(t)\}\varphi_{dm}(t)dt$ is the functional principal component score of X_{di} satisfying $E(\xi_{dim}) = 0$ and $Var(\xi_{dim}) = \lambda_{dm}$ with $\sum \lambda_{dm} < \infty$, and $\{\varphi_{dm}(t)\}_{m=1,2,\ldots}$ are orthonormal eigenfunctions. Hence, the functional parameters β_{dk}s admit the representation $\beta_{dk}(t) = \sum_{m=1}^{\infty} b_{dkm}\varphi_{dm}(t)$ and can be approximated by a finite sum of the leading m_{dn} terms,

$$\beta_{dk}(t) \approx \sum_{m=1}^{m_{dn}} b_{dkm}\varphi_{dm}(t),$$

where m_{dn} can be chosen according to the total variation explained up to a certain threshold.

With the above representation, model (1) becomes

$$Y_i = \mu_k^0 + \sum_{d=1}^{D}\sum_{m=1}^{m_{dn}} b_{dkm}\xi_{dim} + \epsilon_i^*, \quad \text{if the subject } i \text{ belongs to the } k\text{th group.} \tag{2}$$

Denote $\Pi_K^{(n)} = \{C_1^{(n)}, \ldots, C_K^{(n)}\}$ as any possible partition of the n observations for K groups. Estimates of the partition and coefficients μ_k and b_{dkm} are obtained through

$$\min_{\Pi_K^{(n)}} \min_{\mu_k, b_{dkm}} \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in C_k^{(n)}} \phi(y_i - \mu_k - \sum_{d=1}^{D}\sum_{m=1}^{m_{dn}} b_{dkm}\xi_{dim}), \tag{3}$$

where ϕ is a general increasing function of the absolute value of the argument. Interesting special cases are least-squares method $\phi(x) = x^2$, check loss function $\phi(x) = x(\tau - I(x < 0))$ for a given quantile level τ, and Huber's function $\phi_c(x) = \frac{1}{2}x^2 I(|x| \le c) + (cx - \frac{1}{2}c^2)I(|x| > c)$.

In practice, we do not have the entire trajectory $\{X_{di}, d = 1, \ldots, D\}$ but only intermittent noisy measurements

$$X_{dij} = X_{di}(t_{dij}) + e_{dij}, \ i = 1, \ldots, n, \ j = 1, \ldots, \tilde{m}_{di}, \tag{4}$$

where e_{dij}s are independent and identically distributed measurement errors independent of $X_{di}(t_{dij})$, satisfying $E(e_{dij}) = 0$ and $Var(e_{dij}) = \sigma_{xd}^2$. Estimation of the functional principal component scores has been previously studied in the literature. For densely and noisily observed functional data, Hall and Hosseini-Nasab (2006) applied a local linear smoothing to each individual curve and then employ functional principal component analysis to the smoothed curve. Specifically, the discrete observations of $X_{di}(t)$ are pre-smoothed by fitting a local linear regression as follows,

$$(\hat{\theta}_{di0}, \hat{\theta}_{di1}) = \arg\min_{(\theta_{di0}, \theta_{di1})} \sum_{j=1}^{\tilde{m}_{di}}\{X_{dij} - \theta_{di0} - \theta_{di1}(t_{dij} - t)\}^2 K\{(t_{dij} - t)/h_w\},$$

where $K(\cdot)$ is a kernel function and h_w is the bandwidth for the smoothing step. Then the estimate $\hat{X}_{di}(t) = \hat{\theta}_{di0}(t)$ is used instead of the true trajectory $X_{di}(t)$ to construct the covariance, eigen-system and functional principal component scores. For sparse and noisy functional observations, Yao et al. (2005) applied the conditional expectation method, which can tackle both dense and sparse observed observations with measurement errors. The two methods have similar performance

numerically for dense design (Kong et al., 2016a). For computational and theoretical simplicity, we consider the first approach to investigate the theoretical properties, and adopt the method proposed by Yao et al. (2005) to obtain the estimates $\{(\hat{\varphi}_{dm}, \hat{\lambda}_{dm})\}$ and $\{\hat{\xi}_{dim}\}$ in practice.

To derive the estimators, we take advantage of close connection between (3) and the well-known K-means clustering algorithm, and develop a computational efficient algorithm to solve (3). For a fixed K, we are now ready to provide the algorithm to solve (3) following the spirit of K-means type method as follows:

(1) Initially, randomly split the data into K groups $\Pi_K^{(n)} = \{C_1^{(n)}, \ldots, C_K^{(n)}\}$.
(2) For a given partition $\Pi_K^{(n)} = \{C_1^{(n)}, \ldots, C_K^{(n)}\}$, obtain the parameter estimates for the kth group by

$$(\hat{\mu}_k, \hat{b}_{dkm}) = \arg\min_{\mu_k, b_{dkm}} \sum_{i \in C_k^{(n)}} \phi(y_i - \mu_k - \sum_{d=1}^{D} \sum_{m=1}^{m_{dn}} b_{dkm} \hat{\xi}_{dim}).$$

(3) For each subject i, assign subject i to the $\tilde{k}$th group such that $\tilde{k} = \arg\min_k \phi(y_i - \hat{\mu}_k - \sum_{d=1}^{D} \sum_{m=1}^{m_{dn}} \hat{b}_{dkm} \hat{\xi}_{dim})$, $k = 1, \ldots, K$, and update the partition $\Pi_K^{(n)}$ correspondingly.
(4) Repeat the above two steps until that the partition $\Pi_K^{(n)}$ does not change. Obtain the estimates of the functional parameters by $\hat{\beta}_{dk}(t) = \sum_{m=1}^{m_{dn}} \hat{b}_{dkm} \hat{\varphi}_{dm}(t)$ for $k = 1, \ldots, K$, $d = 1, \ldots, D$.

The proposed algorithm is computationally efficient by alternating between the "assignment" and "update" steps. Specifically, each subject is assigned to the group k whose residual is the smallest. In the " update" step, the parameters are estimated through M-estimation. It can be a linear regression, a quantile regression or a robust regression due to various interests. The proposed estimation procedure can be robust to non-normal and heavy-tailed errors by choosing the median regression or the robust regression. Because the loss function is non-decreasing as the increase of iteration numbers, numerical convergence is very fast.

To select the number of groups, we define the Bayesian information criterion for a partition $\Pi_K^{(n)}$ as follows:

$$\text{BIC}(\Pi_K^{(n)}) = \log\Big(\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k^{(n)}} \phi(y_i - \hat{\mu}_k - \sum_{d=1}^{D} \sum_{m=1}^{m_{dn}} \hat{b}_{dkm} \hat{\xi}_{dim})\Big) + \frac{q(K) A_n}{n} \Big(\sum_{d=1}^{D} m_{dn} + 1\Big),$$

where $q(K)$ is an increasing function of K and A_n is a sequence related to n. For example, one may take $q(K) = K$ and $A_n = \log\log n$. Note that the term $(\sum_{d=1}^{D} m_{dn} + 1)$ is the number of parameters involved in the regression structure within groups. Then, we can obtain the estimate of the underlying number of groups K_0 by

$$\hat{K}_n = \arg\min_K \min_{\Pi_K^{(n)}} \text{BIC}(\Pi_K^{(n)}).$$

To validate the use of the BIC, we show that the proposed criterion can specify the true number of groups with probability approaching one as the sample size increases to infinity in next section.

Notably, Preda and Saporta (2007) considered a special case of the proposed method, but they neither took the noisy observations of the functional process into account, nor did they study the theoretical results of the estimators.

3. Theoretical results

In this section, we present the consistency of the proposed BIC and the consistency of the set of estimators. Denote $\|\cdot\|$ as the Euclidean norm and $\|\cdot\|_{L^2}$ as the L^2 norm.

Without loss of generality, we fix $D = 1$ and drop the subscript d for simplicity. An extension to the case of $D > 1$ is straightforward. Recall that $C_0^{(n)} = \{C_{01}^{(n)}, \ldots, C_{0K_0}^{(n)}\}$ is the true group membership such that for $i \in C_{0k}^{(n)}$, $k = 1, \ldots, K_0$, we have

$$Y_i = \mu_k^0 + \int_{\mathcal{T}} X_i(t) \beta_k^0(t) \mathrm{d}t + \epsilon_i.$$

For $k = 1, \ldots, K_0$, each true functional coefficient admits the expansion $\beta_k^0(t) = \sum_{m=1}^{\infty} b_{km}^0 \varphi_m(t) \approx \sum_{m=1}^{m_n} b_{km}^0 \varphi_m(t)$, where m_n is the number of functional principal component basis used, and $\{b_{km}^0\}_{m=1}^{\infty}$ are the true Fourier coefficients. Let $n_k = |C_{0k}^{(n)}|$ be the number of observations in the kth group. Before investigating the theoretical results, we need to make the following assumptions.

Assumption 1. The population comprises $K_0 < \infty$ subpopulations with proportions $\pi_1, \ldots, \pi_{K_0}$ with $\pi_i > a_0 > 0$ and $\sum_{i=1}^{K_0} \pi_i = 1$, where a_0 is a constant.

Assumption 2. The function $\phi(\cdot)$ is convex and $E[\phi(\epsilon_j)]$ is finite for $j \in \mathcal{C}_{0k}^{(n)}$ and $k = 1, \ldots, K_0$. For any β, μ and observations in $\mathcal{C}_{0k}^{(n)}$, $\liminf_{n_k \to \infty} \sum_{j \in \mathcal{C}_{0k}^{(n)}} E[\phi(\epsilon_j - \mu - \int X_j(t)\beta(t)dt) - \phi(\epsilon_j)]/n_k \geq g\big((\|\beta\|_{L^2}^2 + \mu^2)^{1/2}\big)$, where $g(\cdot)$ is a nonnegative convex function and strictly convex monotonic in a right neighborhood of 0.

Assumption 3. Let $\psi(\cdot)$ be the subgradient of $\phi(\cdot)$ and $\mathcal{U}$ be the set of discontinuity points of $\psi(\cdot)$. The distribution function F_i of ϵ_i is unimodal and satisfies $F_i(\mathcal{U}) = 0$, Furthermore, $E[\psi(\epsilon_i)] = 0$ and $E[\psi^2(\epsilon_i)] < \infty$.

Assumption 4. There exist some positive constants c_1, c_2, c_3, and c_4, such that as $u \to 0$, $E[\psi(\epsilon_i + u)] = c_1 u + O(u^2)$. Also, $E\big([\psi(\epsilon_i + u) - \psi(\epsilon_i)]^2\big) \leq c_2|u|$, and $|\psi(v + u) - \psi(v)| \leq c_3$ for any $|u| \leq c_4$ and $v \in \mathbb{R}$.

Assumption 5. The eigenvalues $\{\lambda_m\}$ satisfy $c_5^{-1}m^{-a} \leq \lambda_m \leq c_5 m^{-a}$ and $\lambda_m - \lambda_{m+1} \geq c_5^{-1}m^{-a-1}$ for $m \geq 1$. Also, the true Fourier coefficients $|b_{km}^0| \leq c_6 m^{-b}$, where $a > 1$, $b > a/2 + 1$, and c_5 and c_6 are some positive constants.

Assumption 1 requires that the sample sizes of the subgroups are comparable, implying that $a_0 n \leq n_k \leq n$ for $k = 1, \ldots, K_0$. Assumption 2 is parallel to the conditions in Rao et al. (2007) for the loss function. Assumptions 3–4 are imposed on the subgradient of the loss function, which is commonly encountered in the literature of M-estimation (Rao et al., 2007; Wu and Zen, 1999; Tang, 2017; He and Shi, 1996). Assumption 5 ensures the identifiability of eigenfunctions and smoothness of the functional coefficients.

Similar to Kong et al. (2016b), we need the following assumptions to guarantee the asymptotic equivalence between the estimators obtained from $\hat{X}_i$ and those obtained from the true X_i. Recall that $\tilde{m}_i$ is the number of observations for X_i. Denote that $\tilde{m} = \inf_{i=1,\ldots,n} \tilde{m}_i$.

Assumption 6. For any $C > 0$, there exist an $\epsilon > 0$ such that $\sup_{s \in \mathcal{T}}\{E|X(s)|^C\} < \infty$, and $\sup_{s,t \in \mathcal{T}}\{E[(|s-t|^{-\epsilon}|X(s) - X(t)|)^C]\} < \infty$.

Assumption 7. X is twice continuously differentiable on $\mathcal{T}$ with probability 1, $E(X(t)) = 0$ and $\int E(X^{(2)}(t))^4\, dt < \infty$, where $X^{(2)}(t)$ denotes the second derivative of $X(t)$.

Assumption 8. The observation points $\{t_{ij}, j = 1, \ldots, \tilde{m}_i\}$ are deterministic and ordered increasingly for $i = 1, \ldots, n$. There exist densities g_i uniformly smooth over i, satisfying $\int_0^1 g_i(t)dt = 1$ and $0 < c_1 < \inf_i\{\inf_{t \in \mathcal{T}} g_i(t)\} < \sup_i\{\sup_{t \in \mathcal{T}} g_i(t)\} < c_2 < \infty$. The t_{ij}s are generated according to $t_{ij} = G_i^{-1}\{j/(m_i + 1)\}$, where G_i^{-1} is the inverse of $G_{ij} = \int_{-\infty}^t g_i(s)ds$. The kernel density function is smooth and compactly supported.

Assumption 9. $\sup_i \sup\{t_{i(j+1)} - t_{ij}, j = 1, \ldots, \tilde{m}_i\} = O(\tilde{m}^{-1})$, $h_w \sim \tilde{m}^{-1/5}$, $\tilde{m}n^{-5/4} \to \infty$.

With the preparations above, we are able to derive the theoretical properties of the criterion and the estimators.

Theorem 1. *Suppose the conditions in Assumptions 1–9 are satisfied, if $A_n \to \infty$ and $A_n m_n/n \to 0$ as $n \to \infty$, then $P(\hat{K}_n = K_0) \to 1$.*

Theorem 1 establishes the consistency of the BIC, which is new for clusterwise functional linear models and functional mixture models. It guarantees that the proposed criterion can select the true number of clusters with probability approaching one. Once the true number of clusters is given, we can obtain the consistency of the set of estimators.

Denote $\mathcal{A}^0 = \{\mu_1^0, \ldots, \mu_{K_0}^0\}$ and $\mathcal{B}^0 = \{\beta_1^0, \ldots, \beta_{K_0}^0\}$ as the sets of the true intercepts and functional coefficients, respectively. Let $\hat{\mathcal{A}} = \{\hat{\mu}_1, \ldots, \hat{\mu}_{K_0}\}$ and $\hat{\mathcal{B}} = \{\hat{\beta}_1, \ldots, \hat{\beta}_{K_0}\}$, where $\hat{\beta}_k(t) = \sum_{m=1}^{m_n} \hat{b}_{km}\hat{\varphi}_m(t)$, be the corresponding sets of estimators when the true number of clusters K_0 is given. The following theorem gives the rates of convergence for the two sets of estimators in terms of the Hausdorff metric.

Theorem 2. *If the conditions in Theorem 1 hold, and $m_n \sim n^{1/(a+2b)}$ as $n \to \infty$, when the true number of groups K_0 is given, we have*

$$H(\hat{\mathcal{A}}, \mathcal{A}^0) = O_p(n^{-1/2}) \quad \text{and} \quad H(\hat{\mathcal{B}}, \mathcal{B}^0) = O_p(n^{-(2b-1)/(2a+4b)}),$$

where the Hausdorff distances are $H(\hat{\mathcal{A}}, \mathcal{A}^0) = \sup_{\mu_1 \in \hat{\mathcal{A}}} \inf_{\mu_2 \in \mathcal{A}^0} |\mu_1 - \mu_2|$ and $H(\hat{\mathcal{B}}, \mathcal{B}^0) = \sup_{\beta_1 \in \hat{\mathcal{B}}} \inf_{\beta_2 \in \mathcal{B}^0} \|\beta_1 - \beta_2\|_{L^2}$.

Theorem 2 shows that the proposed method can consistently estimate the set of true coefficients. Specifically, the estimator of the set of true intercepts enjoys a parametric rate of convergence, whereas the estimator of the set of true functional coefficients enjoys the nonparametric rate of convergence $n^{-(2b-1)/(2a+4b)}$, which is the same as that of Hall and Horowitz (2007) and optimal in the minimax sense. Yao et al. (2010) also derived consistent parameter estimation by imposing the conditional density, but they did not investigate the convergence rate. It is also worth mentioning that the results in Theorems 1 and 2 are applicable to a general loss function, which brings great flexibility in practical applications.

Similar to the disadvantage of K-means type methods, the estimated group membership may not converge to the population group membership, thereby leading to a nonzero group misclassification probability. An intuition is that when the error term takes value from negative infinity to positive infinity, we cannot separate the clusters perfectly according to the distance between the observation value and the sample mean (or even the population mean). In other words, the cluster boundaries are not very precisely located. However, given that observations near the cluster boundaries effectively contribute to the means of observations in both clusters (Pollard, 1982), we can consistently estimate the parameters.

4. Simulation studies

We conduct simulation studies to illustrate the empirical performance of the proposed method. We choose least-squares loss function $\phi(x) = x^2$, least absolute deviation loss function $\phi(x) = |x|$ and Huber loss function $\phi_c(x) = \frac{1}{2}x^2 I(|x| \le c) + (cx - \frac{1}{2}c^2)I(|x| > c)$ with the commonly used $c = 1.345$. Denote the three choices as LS, LAD and HL, respectively. For comparison, we also include the functional mixture regression (FMR) of Yao et al. (2010) and the clusterwise partial least squares regression (PLS) of Preda and Saporta (2005). Following Yu et al. (2016), the number of PLS basis is obtained by minimizing the Akaike information criterion. The number of functional principal scores of the functional processes is truncated by the threshold of 90% of overall variation, and the number of groups is obtained by minimizing the proposed BIC. Following Zhang et al. (2019), we choose $q(K) = K$ and $A_n = 10 \log\log n$ for least squares-loss function and Huber loss function, and $A_n = 5 \log\log n$ for least absolute deviation loss function.

To evaluate the performance of the above methods, we calculate several frequently used external validity measures, namely, the rand index, adjusted rand index, Jaccard index, and the purity function. Denote the true positive (TP) to be the number of pairs of subjects from the same cluster and assigned to the same cluster, the true negative (TN) to be the number of pairs of subjects from different clusters and assigned to different clusters, the false positive (FP) to be the number of pairs of subjects from different clusters but assigned to the same cluster, and the false negative (FN) to be the number of pairs of subjects from the same cluster but assigned to different clusters. The rand index and Jaccard index are defined by

$$\text{RI} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad \text{and} \quad \text{Jaccard} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}.$$

However, rand index tends to be large even under random partitions. The adjusted rand index (Hubert and Arabie, 1985; Zhu et al., 2018) corrects this problem and is calculated as follows:

$$\text{ARI} = \frac{\text{RI} - E(\text{RI})}{\max(\text{RI}) - \min(\text{RI})}.$$

The purity function is defined by

$$\text{Purity} = \frac{1}{n}\sum_{j=1}^{K} \max_{1 \le k \le K_0} |\mathcal{C}_{0k}^{(n)} \cap \mathcal{C}_j^{(n)}|,$$

where $\{\mathcal{C}_{01}^{(n)}, \ldots, \mathcal{C}_{0K_0}^{(n)}\}$ are the index sets of the true classes, $\{\mathcal{C}_1^{(n)}, \ldots, \mathcal{C}_K^{(n)}\}$ are the estimated index sets, and $|\mathcal{C}_{0k}^{(n)} \cap \mathcal{C}_j^{(n)}|$ is the number of samples in cluster j that belongs to original class k. For these external measures, a higher value indicates a better agreement between the selected and the true group memberships.

Example 1. Samples are generated from

$$Y_i = \int X_i(t)\beta_k(t)dt + \sigma(X_i)\epsilon_i, \ \ k = 1, 2, \quad \text{if the subject } i \text{ belongs to the } k\text{th group.}$$

Similar to Yao et al. (2010), the functional predictor has the form $X_i(t) = \xi_{i1}\varphi_1(t) + \xi_{i2}\varphi_2(t)$, where $\xi_{i1} \sim N(0, 1)$, $\xi_{i2} \sim N(0, 4)$ and $\varphi_1(t) = \sin(\pi t/10)/\sqrt{5}$, $\varphi_2(t) = \sin(2\pi t/10)/\sqrt{5}$. We assume that observation X_{ij} is the realization of $X_i(t)$ at 100 evenly spaced points $\{t_{ij}, t_{ij} \in [0, 10]\}$ with i.i.d. error $e_{ij} \sim N(0, 0.2^2)$. The functional coefficients are set to be $\beta_1(t) = \varphi_1(t) + \varphi_2(t)$ for the first $n/2$ subjects with $n = 200$, and $\beta_2(t) = \varphi_1(t) - \varphi_2(t)$ for the rest of the samples. We consider $\sigma(X_i) = 1$ and $\sigma(X_i) = 0.5\xi_{i1} + 0.4\xi_{i2}$, which correspond to homoscedastic and heteroscedastic cases, respectively. Each ϵ_i is generated from $0.2N(0, 1)$, $0.2t(3)$ or $0.2(\chi^2(3) - 3)$. For the initial partition, we randomly assign the samples into groups with equal size. We repeat 200 times in each scenario, using R (version 3.3.2) on a Dell desktop computer (equipped with Intel(R) Core(TM) i5-4690S CPU@ 3.20 GHz, 8 GB RAM). On the basis of 200 replications, the means of the aforementioned metrics and other summary statistics are calculated and reported.

Table 1 presents results for the homoscedastic cases. We also calculate the mean squared error for the functional coefficient by using $\text{MSE}_\beta = n^{-1}\sum_{i=1}^{n}\int[\hat{\beta}_i(t) - \beta(t)]^2 dt$. For normal errors, the proposed methods perform similarly to the method of Yao et al. (2010), but better than the method of Preda and Saporta (2005). However, for models with heavy-tailed and asymmetric errors, the proposed method based on the least absolute deviation consistently outperforms the other methods and produces higher percentage of times for identifying the true number of groups, closer-to-truth

Table 1
Results for the homoscedastic cases in Example 1.

		Correct_p	$\hat{K}$	RI	ARI	Purity	Jaccard	MisError	MSE_β	Time (s)	Iteration
N(0, 1)	FMR	0.995	2.005	0.847	0.693	0.916	0.736	0.084	0.330	0.475	13.720
	PLS	0.980	2.025	0.799	0.597	0.886	0.665	0.115	0.459	56.591	7.465
	LS	0.980	2.020	0.858	0.716	0.924	0.752	0.080	0.299	0.100	5.700
	LAD	1.000	2.000	0.859	0.718	0.924	0.754	0.076	0.300	0.090	4.195
	HL	1.000	2.000	0.863	0.727	0.926	0.760	0.074	0.290	0.135	4.680
t(3)	FMR	0.750	2.265	0.808	0.615	0.895	0.675	0.115	0.414	0.366	21.920
	PLS	0.980	2.010	0.768	0.535	0.865	0.624	0.136	0.545	56.230	7.530
	LS	0.975	2.030	0.825	0.650	0.904	0.702	0.099	0.388	0.114	5.835
	LAD	1.000	2.000	0.827	0.655	0.904	0.706	0.096	0.376	0.091	4.325
	HL	0.990	2.010	0.825	0.650	0.903	0.702	0.098	0.381	0.135	5.140
$\chi^2(3)$	FMR	0.775	2.225	0.736	0.472	0.847	0.579	0.165	0.603	0.397	24.210
	PLS	0.970	1.990	0.717	0.435	0.827	0.563	0.175	1.025	53.820	7.720
	LS	0.975	2.025	0.751	0.503	0.855	0.601	0.149	0.585	0.118	6.330
	LAD	1.000	2.000	0.754	0.507	0.856	0.605	0.144	0.573	0.101	5.300
	HL	1.000	2.000	0.758	0.517	0.859	0.611	0.141	0.561	0.141	5.720

Table 2
Results for the heterscedastic cases in Example 1.

		Correct_p	$\hat{K}$	RI	ARI	Purity	Jaccard	MisError	MSE_β	Time (s)	Iteration
N(0, 1)	FMR	0.410	2.850	0.796	0.591	0.914	0.639	0.177	0.339	0.319	39.430
	PLS	0.975	2.030	0.812	0.624	0.895	0.684	0.107	0.427	58.150	8.080
	LS	0.990	2.010	0.868	0.735	0.929	0.766	0.073	0.279	0.106	5.725
	LAD	0.995	2.005	0.866	0.733	0.928	0.765	0.073	0.282	0.095	4.300
	HL	0.990	2.010	0.866	0.732	0.928	0.764	0.074	0.283	0.148	5.000
t(3)	FMR	0.255	3.095	0.778	0.556	0.898	0.614	0.179	0.404	0.282	40.160
	PLS	0.980	2.020	0.797	0.595	0.885	0.664	0.116	0.472	58.250	8.435
	LS	0.960	2.045	0.835	0.671	0.910	0.717	0.094	0.362	0.114	6.510
	LAD	1.000	2.000	0.839	0.679	0.912	0.724	0.088	0.346	0.102	4.450
	HL	0.995	2.005	0.841	0.682	0.913	0.726	0.088	0.342	0.150	5.270
$\chi^2(3)$	FMR	0.040	3.540	0.720	0.439	0.870	0.514	0.263	0.535	0.160	43.380
	PLS	0.970	1.980	0.776	0.553	0.867	0.641	0.134	0.596	32.300	8.250
	LS	0.865	2.135	0.778	0.556	0.875	0.632	0.138	0.517	0.072	7.515
	LAD	1.000	2.000	0.795	0.591	0.884	0.661	0.116	0.464	0.068	5.555
	HL	0.975	2.025	0.788	0.576	0.880	0.649	0.123	0.482	0.169	6.520

$\hat{K}$, larger rand index, adjusted rand index, Jaccard index and purity, and smaller misclassification error. The results for the heteroscedastic cases are summarized in Table 2. The proposed methods outperform the methods of Yao et al. (2010) and Preda and Saporta (2005) in all aspects, and the proposed method based on the least absolute deviation performs better than those based on least squares and Huber loss. Tables 1–2 also show that the proposed methods are computationally more efficient than the other two methods with less computation times.

Furthermore, we conduct a sensitive study for A_n of the proposed methods for different loss functions. We choose $A_n = c \log\log n$ and $c \in \{1, 2, \ldots, 25\}$.Fig. 1 gives the RI and MSE of the three methods under the heterscedastic cases for different errors. It shows that there exist stable regions of A_n for the three methods where *RI* or *MSE* behaves similarly for different A_ns. Meanwhile, we can see that $c = 10$ is suitable for LS and HL, and $c = 5$ is suitable for LAD.

Example 2. In this example, we consider different functional coefficients and intercepts across groups. We generate samples from

$$Y_i = \mu_k + \int X_i(t)\beta_k(t)dt + \sigma(X_i)\epsilon_i, \quad k = 1, 2, 3.$$

Different from Example 1 that the functional variables and the functional coefficients can be expressed in the first few eigenfunctions, here we consider $X_i(t) = \sum_{\ell=1}^{50} \xi_{i\ell}\varphi_\ell(t)$, where $\varphi_{2\ell-1}(t) = \sqrt{2}\cos((2\ell-1)\pi t)$ and $\varphi_{2\ell}(t) = \sqrt{2}\sin((2l-1)\pi t)$ for $\ell = 1, \ldots, 25$ and $\xi_{i\ell} \sim N(0, 16\ell^{-2})$. Also, the observation X_{ij} is the realization of $X_i(t)$ at 100 evenly spaced points $\{t_{ij}, t_{ij} \in [0, 1]\}$ with i.i.d. error $e_{ij} \sim N(0, 0.2^2)$. We set $\beta_1(t) = \sum_{\ell=1}^{50} b_{1\ell}\varphi_\ell(t)$ for the first $n/3$ subjects, $\beta_2(t) = \sum_{\ell=1}^{50} b_{2\ell}\varphi_\ell(t)$ for the next $n/3$ subjects, and $\beta_3(t) = \sum_{\ell=1}^{50} b_{3\ell}\varphi_\ell(t)$ for the last $n/3$ subjects. Specifically, $b_{11} = 1, b_{12} = 0.8, b_{13} = 0.6, b_{14} = 0.5$, and $b_{21} = 1, b_{22} = -0.8, b_{23} = -0.6, b_{24} = -0.5$, and $b_{31} = -1, b_{32} = -0.8, b_{33} = -0.6, b_{34} = -0.5$ and $b_{1\ell} = b_{2\ell} = b_{3\ell} = 8(\ell-2)^{-4}$ for $\ell = 5, \ldots, 50$. Sample size $n = 300$ and a total of 200 replications are considered. The intercepts are chosen to be $\mu_1 = 3$, $\mu_2 = -3$ and $\mu_3 = 0$. We have $\sigma(X_i) = 1$ for

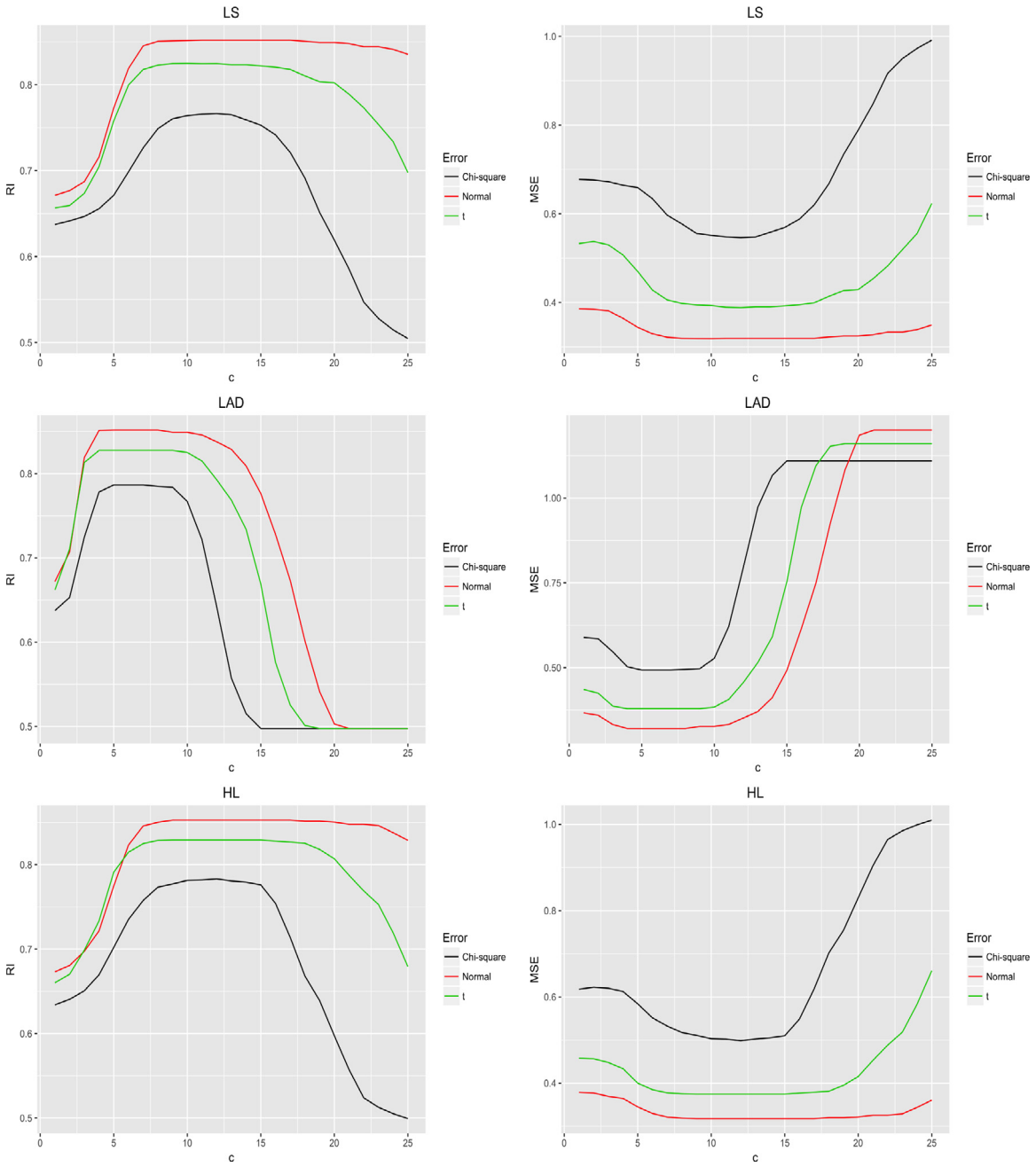

Fig. 1. RI (left) and MSE (right) of the proposed methods with different $A_n = c \log\log n$ for the heterscedastic cases with top for LS, middle for LAD and bottom for HL. The red line corresponds to the normal errors, and the green and black lines correspond to errors from t and χ^2 distributions, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

homoscedastic cases, and $\sigma(X_i) = \int_0^1 X_i(t)\sigma(t)dt$ with $\sigma(t) = \sum_{\ell=1}^{50} \sigma_\ell \varphi_\ell(t)$, where $\sigma_1 = 0.5, \sigma_2 = 0.4, \sigma_3 = 0.2, \sigma_4 = 0.1$ and $\sigma_\ell = 4(\ell - 2)^{-4}$ for $\ell = 5, \ldots, 50$. Except for mean squared errors for the functional coefficients, we also calculate mean squared errors for the intercept term, defined as $\text{MSE}_\mu = n^{-1} \sum_{i=1}^{n} (\hat{\mu}_i - \mu_i)^2$.

Tables 3–4 present results for the homoscedastic and heterscedastic cases. Our methods outperform the methods of Yao et al. (2010) and Preda and Saporta (2005) in almost all the scenarios, and the proposed method based on the least absolute deviation and Huber loss function have advantage over other methods in the heterscedastic cases.

Table 3
Results for the homoscedastic cases in Example 2.

		Correct_p	$\hat{K}$	RI	ARI	Purity	Jaccard	MisError	MSE_μ	MSE_β	Time (s)	Iteration
N(0, 1)	FMR	1.000	3.000	0.960	0.910	0.969	0.887	0.031	0.556	0.216	0.332	18.125
	PLS	0.590	2.520	0.709	0.419	0.703	0.489	0.233	5.454	2.731	363.271	33.025
	LS	0.845	3.155	0.956	0.900	0.967	0.875	0.039	0.585	0.247	0.232	11.970
	LAD	0.960	3.040	0.959	0.907	0.969	0.883	0.033	0.566	0.225	0.185	7.050
	HL	0.955	3.045	0.959	0.907	0.969	0.884	0.033	0.559	0.222	0.558	8.560
t(3)	FMR	0.960	3.040	0.946	0.878	0.959	0.850	0.043	0.714	0.283	0.284	20.075
	PLS	0.510	2.400	0.678	0.369	0.668	0.459	0.256	5.701	3.182	366.963	32.320
	LS	0.865	3.135	0.945	0.874	0.958	0.845	0.048	0.731	0.319	0.251	11.710
	LAD	0.965	3.035	0.948	0.882	0.959	0.854	0.042	0.700	0.284	0.193	7.305
	HL	0.995	3.005	0.948	0.883	0.960	0.855	0.041	0.697	0.274	0.574	9.445
$\chi^2(3)$	FMR	0.825	3.190	0.915	0.807	0.938	0.772	0.077	1.064	0.423	0.323	26.780
	PLS	0.485	2.440	0.681	0.362	0.668	0.448	0.246	5.895	3.238	363.432	34.555
	LS	0.930	3.070	0.921	0.820	0.938	0.786	0.066	1.056	0.435	0.253	12.645
	LAD	0.965	3.005	0.918	0.817	0.934	0.785	0.064	1.129	0.451	0.218	8.370
	HL	1.000	3.000	0.924	0.829	0.940	0.796	0.060	1.015	0.402	0.613	9.985

Table 4
Results for the heterscedastic cases in Example 2.

		Correct_p	$\hat{K}$	RI	ARI	Purity	Jaccard	MisError	MSE_μ	MSE_β	Time (s)	Iteration
N(0, 1)	FMR	0.715	3.305	0.938	0.858	0.960	0.827	0.066	0.829	0.273	0.283	27.850
	PLS	0.530	2.455	0.699	0.404	0.690	0.480	0.234	5.147	2.955	359.755	32.060
	LS	0.870	3.130	0.948	0.881	0.961	0.854	0.046	0.836	0.302	0.240	12.200
	LAD	0.960	3.030	0.951	0.890	0.962	0.865	0.038	0.810	0.261	0.193	7.485
	HL	1.000	3.000	0.953	0.894	0.964	0.868	0.036	0.780	0.249	0.607	9.380
t(3)	FMR	0.550	3.480	0.912	0.799	0.938	0.761	0.089	1.221	0.452	0.300	32.385
	PLS	0.525	2.455	0.689	0.383	0.681	0.463	0.243	5.594	3.115	356.640	31.170
	LS	0.975	2.975	0.924	0.834	0.937	0.807	0.061	1.180	0.390	0.263	12.065
	LAD	0.985	2.985	0.933	0.850	0.946	0.822	0.051	1.098	0.355	0.213	8.400
	HL	0.935	3.065	0.935	0.852	0.949	0.821	0.054	1.060	0.337	0.387	10.375
$\chi^2(3)$	FMR	0.105	4.130	0.866	0.686	0.904	0.644	0.159	1.856	0.726	0.251	43.955
	PLS	0.480	3.815	0.716	0.333	0.699	0.379	0.361	9.648	2.903	234.344	40.245
	LS	0.630	2.260	0.690	0.488	0.703	0.588	0.297	3.278	1.147	0.182	8.775
	LAD	0.985	2.985	0.901	0.779	0.920	0.745	0.078	1.633	0.515	0.154	9.160
	HL	0.915	3.085	0.901	0.776	0.921	0.740	0.083	1.627	0.534	0.442	11.755

Example 3. In this example, we consider two functional variables. Samples are generated from

$$Y_i = \mu_k + \int X_{1i}(t)\beta_{1k}(t)dt + \int X_{2i}(t)\beta_{2k}(t)dt + \sigma(X_{1i}, X_{2i})\epsilon_i, \ \ k = 1, 2, 3,$$

where μ_ks are the same as those in Example 2. For $l = 1, 2$, $X_{li}(t) = \xi_{li1}\varphi_1(t) + \xi_{li2}\varphi_2(t)$, where $\xi_{li1} \sim N(0, 1)$, $\xi_{li2} \sim N(0, 4)$, and $\varphi_1(t)$ and $\varphi_2(t)$ are the same as those in Example 1. The functional coefficients are set to be $\beta_{11}(t) = \varphi_1(t) + \varphi_2(t)$, $\beta_{21}(t) = \varphi_1(t) + 0.5\varphi_2(t)$ for the first $n/3$ subjects, $\beta_{12}(t) = \varphi_1(t) - \varphi_2(t)$, $\beta_{22}(t) = \varphi_1(t) - 0.5\varphi_2(t)$ for the next $n/3$ subjects, and $\beta_{13}(t) = -\varphi_1(t) - \varphi_2(t)$, $\beta_{23}(t) = -0.5\varphi_1(t) - \varphi_2(t)$ for the last $n/3$ subjects. Likewise, $\sigma(X_{1i}, X_{2i}) = 1$ and $\sigma(X_{1i}, X_{2i}) = \sum_{l=1}^{2}(0.5\xi_{li1} + 0.4\xi_{li2})$ denote the homoscedastic and heteroscedastic cases, respectively.

Tables 5–6 present results for the homoscedastic and heterscedastic cases. Similar to Examples 1 and 2, the proposed methods outperform the methods of Yao et al. (2010) and Preda and Saporta (2005). Meanwhile, the proposed methods based on the least absolute deviation and Huber loss are more robust to heavy-tailed and heteroscedastic errors compared with the other two methods.

5. Real data analysis

5.1. Analysis of the ADNI data

Alzheimer's disease (AD) is a chronic neurodegenerative disease that causes brain cells to degenerate and die. As the most common cause of dementia, AD is associated with a continuous decline in thinking as well as behavioral and social skills, which eventually disrupts a person's ability to function independently. Hence, it is of great interest to discover or validate prognostic biomarkers that may identify subjects at great risk for future cognitive decline and investigate the effects of various biomarkers on the conversion from cognitive normal (CN) to AD.

Table 5
Results for the homoscedastic cases in Example 3.

		Correct_p	$\hat{K}$	RI	ARI	Purity	Jaccard	MisError	MSE_μ	MSE_{β_1}	MSE_{β_2}	Time (s)	Iteration
N(0, 1)	FMR	1.000	3.000	0.962	0.914	0.971	0.892	0.029	0.324	0.169	0.098	0.225	20.570
	PLS	0.725	3.305	0.843	0.642	0.854	0.627	0.165	2.079	0.718	0.421	104.134	21.210
	LS	0.820	3.205	0.956	0.901	0.968	0.876	0.041	0.360	0.187	0.113	0.167	12.890
	LAD	0.925	3.080	0.960	0.909	0.969	0.885	0.034	0.343	0.181	0.105	0.125	7.755
	HL	0.905	3.105	0.959	0.907	0.969	0.884	0.036	0.342	0.178	0.102	0.339	10.035
t(3)	FMR	0.955	3.045	0.943	0.871	0.956	0.842	0.046	0.482	0.253	0.146	0.209	22.450
	PLS	0.665	2.810	0.796	0.574	0.797	0.593	0.163	2.458	0.919	0.521	104.713	17.645
	LS	0.865	3.135	0.942	0.868	0.955	0.838	0.051	0.502	0.256	0.157	0.181	13.640
	LAD	0.855	3.075	0.936	0.857	0.946	0.831	0.052	0.590	0.311	0.181	0.136	8.210
	HL	1.000	3.000	0.947	0.881	0.959	0.853	0.041	0.457	0.237	0.136	0.353	10.215
$\chi^2(3)$	FMR	0.780	3.240	0.909	0.792	0.932	0.756	0.086	0.752	0.397	0.227	0.227	29.920
	PLS	0.690	3.345	0.805	0.552	0.812	0.549	0.213	2.815	0.898	0.525	101.794	20.490
	LS	0.705	3.320	0.902	0.775	0.921	0.739	0.094	0.872	0.475	0.276	0.191	15.195
	LAD	0.915	3.075	0.915	0.807	0.932	0.773	0.072	0.755	0.394	0.238	0.155	9.510
	HL	0.870	3.135	0.913	0.803	0.931	0.768	0.075	0.752	0.399	0.235	0.395	11.965

Table 6
Results for the heterscedastic cases in Example 3.

		Correct_p	$\hat{K}$	RI	ARI	Purity	Jaccard	MisError	MSE_μ	MSE_{β_1}	MSE_{β_2}	Time (s)	Iteration
N(0, 1)	FMR	0.635	3.385	0.944	0.872	0.968	0.841	0.067	0.380	0.189	0.108	0.196	32.595
	PLS	0.725	3.310	0.842	0.638	0.853	0.623	0.166	2.187	0.731	0.419	105.565	20.990
	LS	0.840	3.150	0.954	0.897	0.966	0.872	0.041	0.411	0.198	0.118	0.163	13.365
	LAD	0.875	3.095	0.956	0.901	0.964	0.880	0.037	0.413	0.212	0.120	0.129	8.165
	HL	1.000	3.000	0.963	0.916	0.971	0.894	0.029	0.335	0.163	0.093	0.345	9.935
t(3)	FMR	0.550	3.465	0.922	0.821	0.947	0.785	0.084	0.648	0.308	0.182	0.221	36.145
	PLS	0.620	3.400	0.820	0.587	0.831	0.577	0.194	2.610	0.832	0.490	103.683	21.730
	LS	0.765	3.250	0.935	0.851	0.950	0.818	0.060	0.620	0.306	0.177	0.182	13.635
	LAD	0.905	3.100	0.946	0.876	0.958	0.848	0.046	0.516	0.239	0.141	0.149	8.615
	HL	0.905	3.100	0.944	0.872	0.956	0.843	0.048	0.533	0.246	0.143	0.387	10.635
$\chi^2(3)$	FMR	0.110	4.055	0.879	0.718	0.914	0.675	0.144	1.100	0.512	0.325	0.283	50.205
	PLS	0.655	3.380	0.790	0.519	0.798	0.519	0.228	3.250	0.981	0.578	103.275	20.400
	LS	0.775	3.245	0.900	0.772	0.919	0.736	0.092	1.042	0.472	0.288	0.202	15.885
	LAD	0.890	3.110	0.912	0.801	0.929	0.766	0.075	0.918	0.406	0.233	0.167	10.470
	HL	0.900	3.080	0.910	0.798	0.927	0.764	0.075	0.938	0.407	0.237	0.456	12.955

We applied model (1) to the diffusion-weighted imaging dataset collected by the ADNI study to illustrate the empirical utility of our proposed method. The diffusion-weighted imaging data were processed by using TBSS-ENIGMA pipeline (Smith et al., 2006), which included eddy current correction, masking, tensor calculation, creation of FA images and quality controls. We adopted linear registration to register each of the FA images to the Enigma FA template at $1 \times 1 \times 1$ mm spatial resolution. We performed quality controls again after registration and exclude subjects with bad registration. Next, we applied nonlinear registration to align the linearly registered FA images to the ENIGMA FA template and mask the registered FA with a template mask.

The aim of this data analysis is to examine the grouped effects of FA curves along cingulum and body of corpus callosum skeletons on cognitive performance. We treated the MMSE score as the response because it has been widely used to assess cognitive mental status, with low scores indicating impairment. The dataset consists of $n = 139$ subjects (93 CN controls and 49 AD patients) and the FA curves along the cingulum and body of corpus callosum are recorded at 100 locations. Except for the proposed methods based on least-squares, least absolute deviation, and Huber loss function, we also implemented the functional mixture regression method (Yao et al., 2010) and the partial least squares method (Preda and Saporta, 2005) for comparison. The number of functional components was selected, such that the percentage of variance explained was 90%. We used the disease status as a benchmark for the "true groups" to assess the performance of the preceding methods. However, the disease status was treated as unknown throughout the data analysis for illustration.

The five methods all selected two groups based on the values of BIC. Hence, the BIC correctly select true number of groups. Regarding CN and AD people as the true groups, the rand index, adjusted rand index, Jaccard index, purity and misclassification error were calculated and reported in Table 7. To better understand the interoperability of the methods, we also calculated the value of R^2. Apparently, the proposed clustering methods based on M-estimation consistently outperform the methods of Yao et al. (2010) and Preda and Saporta (2005).

Fig. 2 presents the boxplots of MMSE scores for the true groups and groups obtained by using all the methods. The estimates of the two functional coefficients are depicted in Fig. 3. The five methods produce similar estimation results for the body of corpus callosum functional coefficient and share similar trends for the estimates of the cingulum functional coefficient. However, the estimate of the cingulum functional coefficient for the PLS method exhibits higher oscillations.

Table 7
Results for the ADNI data.

	RI	ARI	Purity	Jaccard	MisError	R^2
FMR	0.722	0.444	0.834	0.581	0.165	0.694
PLS	0.865	0.729	0.928	0.777	0.071	0.797
LS	0.903	0.806	0.949	0.836	0.050	0.807
LAD	0.903	0.806	0.949	0.836	0.050	0.793
HL	0.903	0.806	0.949	0.836	0.050	0.806

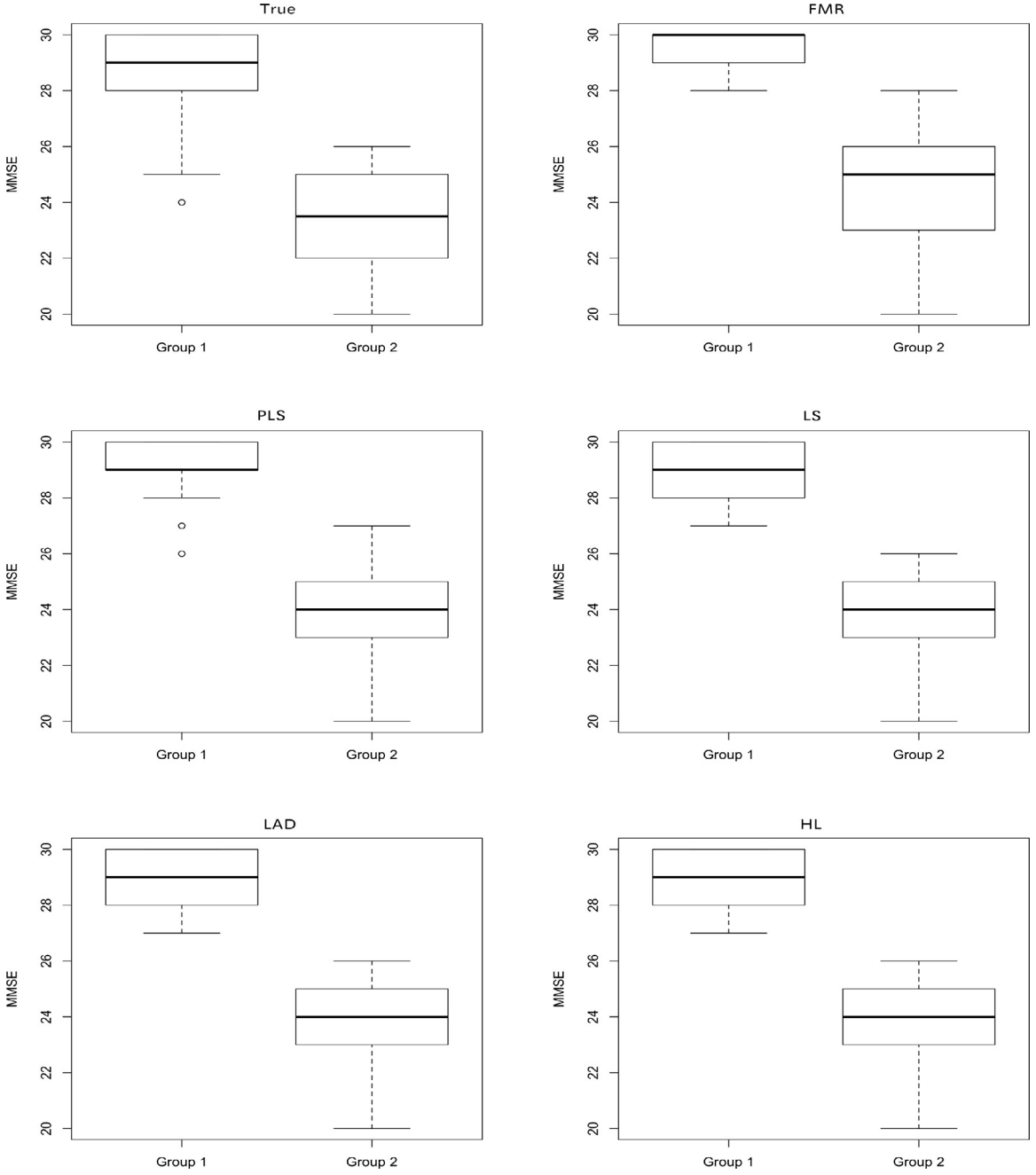

Fig. 2. Boxplots of the MMSE scores for the true, PLS, LAD, FMR, LS and HL (from top to bottom, from left to right).

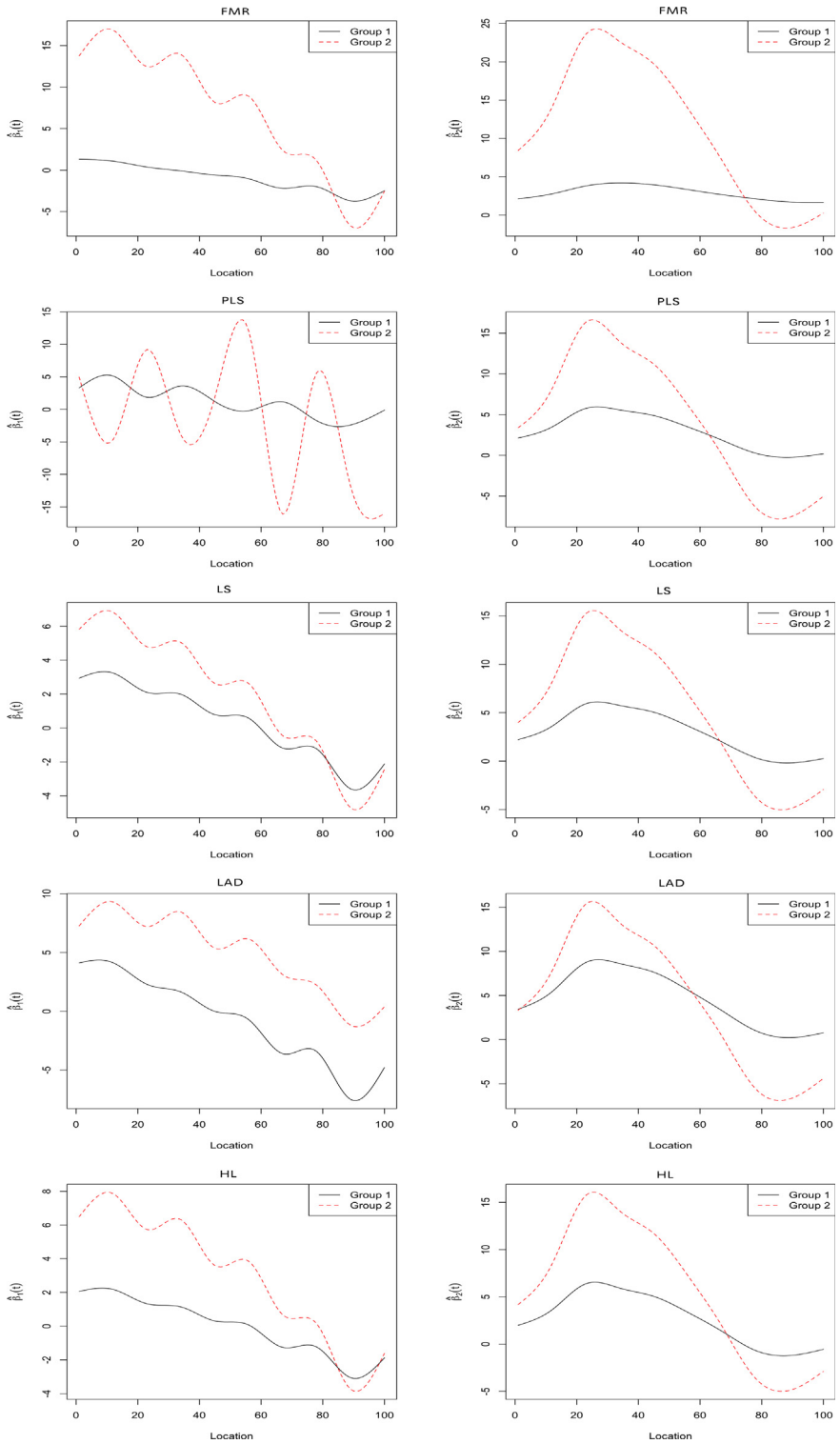

Fig. 3. Estimates of the functional coefficients with the left for cingulum and the right for body of corpus callosum.

Most interestingly, the FA curves along cingulum and body of corpus callosum affect cognitive ability in different manners for the two groups. The effects of the FA curves along cingulum and body of corpus callosum on cognitive ability are greater

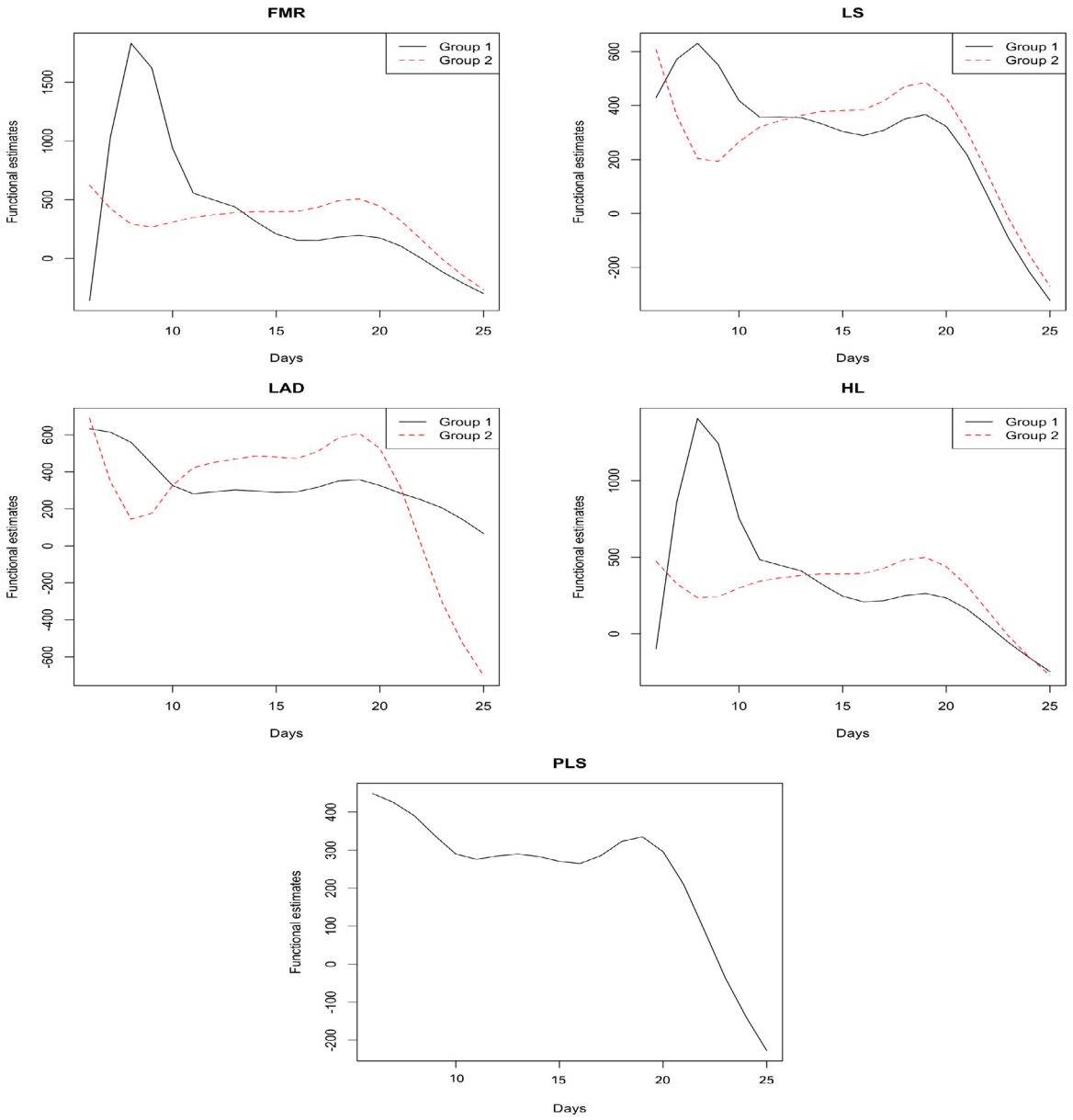

Fig. 4. Estimates of the functional coefficients for the medfly data, where group 1 corresponds to those medflies with a larger mean of lifetime.

for people with worse cognitive function than for people with better one. For subjects in the bad cognitive function group, high FA values along cingulum and body of corpus callosum are associated with high MMSE scores at most of the locations. This conclusion coincides with the findings of Nir et al. (2013) and Zhang et al. (2009) that the higher FA values of cingulum and body of corpus callosum typically indicate better cognitive performance.

5.2. Analysis of the medfly data

In this subsection, we applied the proposed methods and the competing methods to the data from an experiment on medfly fecundity (Carey et al., 1998). The experiment consisted of 1000 female medflies for which daily egg production was recorded. Evolution is closely related to reproductive success, and the connection between the evolution of aging and reproduction is intriguing. Investigating the relationship between reproductive and aging patterns has attracted more and more research interests in recent years. We aimed to determine the dependence of longevity of the medflies on their early fertility process and inspect whether the regression relationship varies due to some unknown mechanism.

Table 8
Values of R^2 for the five methods for the medfly data.

	FMR	PLS	LS	LAD	HL
R^2	0.740	0.773	0.914	0.881	0.920

Because egg counts of most of the medflies were 0 before day 6, we selected $n = 490$ medflies that were fertile more than 15 days from the sixth day and also survived beyond. The trajectories of the number of daily eggs during the early life period from day 6 to day 25 were treated as the functional predictors, while the lifetime served as the response. Similar to Yao et al. (2010), we took a log-transformed of egg counts to achieve homogeneity.

To identify possible relationship changes in early life reproductive trajectories that tend to influence longevity, we implemented the proposed methods based on least-squares, least absolute deviation, and Huber loss function, and two competing methods, the functional mixture regression method (Yao et al., 2010) and the partial least squares method (Preda and Saporta, 2005). The proposed methods and FMR suggested 2 groups with different regression structures in terms of the value of BIC. However, PLS suggested there is no obvious grouping effect and the regression structure is the same across all the subjects. Different from the analysis of the ADNI data in Section 5.1, the relevant classes are not known for this dataset. Hence, we only calculated the values of R^2 for the above five methods in Table 8, which indicates an obvious gain in interoperability of the proposed methods compared to FMR and PLS.

Fig. 4 presents the functional estimates of the five methods. Observing Fig. 4 reveals that a higher level of fertility in the late period seems to shorten lifespan, while a rise of egg production in the early period helps to prolong the lifespan. Specifically, for flies belonging to group 1 with a larger mean of longevity, the impact of the egg production on the lifespan is larger than the other group. These findings may help to recognize distinct underlying mechanisms relating to longevity and early fertility.

6. Discussion

The clusterwise functional linear regression model offers a flexible yet parsimonious approach to deal with the unobserved grouped patterns of heterogeneity. To cover a wide range of estimators, we considered M-estimation for clusterwise functional linear regression models with multiple functional predictors. The infinite-dimensional functional coefficients were represented by the functional principal component basis. A Bayesian information criterion was proposed to select the number of groups and shown to be consistent in specifying the true number of groups. Meanwhile, we showed that the set of estimators to be consistent when the true number of groups is given. Through the analysis of the ADNI data, we showed that the proposed method is a valuable statistical tool for detecting heterogeneous regression patterns between the FA values along cingulum and BCC and the cognitive function for different groups of people.

There are several directions for future study. A useful area for improvement would be to allow some of the regression coefficients to be the same for all subjects. Another interesting consideration for future research would be to accommodate functional response. Furthermore, considering different regimes of functional data in theoretical studies is very interesting. However, additional technical challenges arise in figuring out the impact of the eigen-system construction under different cases on the rates of the estimates. The detailed research will be pursued in the future study.

Acknowledgments

Xinyuan Song's research was partially supported by GRF grants 14301918 and 14302519 from the Research Grant Council of the Hong Kong Special Administrative Region. Zhongyi Zhu's research was partially supported by NFSC 12071087 and NFSC 11731011. Ting Li's work was partially supported by the Fundamental Research Funds for the Central Universities 2020110918.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2021.107192. The supplementary material contains additional simulation results and all the technical proofs of Theorem 1 and Theorem 2.

References

Abraham, C., Cornillon, P.-A., Matzner-Løber, E., Molinari, N., 2003. Unsupervised curve clustering using b-splines. Scand. J. Stat. 30, 581–595.
Carey, J.R., Liedo, P., Müller, J.-L., Chiou, J.-M., 1998. Relationship of age patterns of fecundity to mortality, longevity, and lifetime reproduction in a large cohort of mediterranean fruit fly females. J. Gerontol. Ser. A 53, B245–B251.
Ciarleglio, A., Ogden, R.T., 2016. Wavelet-based scalar-on-function finite mixture regression models. Comput. Statist. Data Anal. 93, 86–96.
Delaigle, A., Hall, P., Pham, T., 2019. Clustering functional data into groups by using projections. J. R. Stat. Soc. Ser. B Stat. Methodol. 81, 271–304.
DeSarbo, W.S., Cron, W.L., 1988. A maximum likelihood methodology for clusterwise linear regression. J. Classification 5, 249–282.
Hall, P., Horowitz, J.L., 2007. Methodology and convergence rates for functional linear regression. Ann. Statist. 35, 70–91.
Hall, P., Hosseini-Nasab, M., 2006. On properties of functional principal components analysis. J. R. Stat. Soc. Ser. B Stat. Methodol. 68, 109–126.
He, X., Shi, P., 1996. Bivariate tensor-product b-splines in a partly linear model. J. Multivariate Anal. 58, 162–181.

Hennig, C., 2000. Identifiablity of models for clusterwise linear regression. J. Classification 17, 273–296.
Huang, L., Wang, H., Zheng, A., 2014. The m-estimator for functional linear regression model. Statist. Probab. Lett. 88, 165–173.
Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classification 2, 193–218.
Kong, D., Staicu, A.-M., Maity, A., 2016a. Classical testing in functional linear models. J. Nonparametr. Stat. 28, 813–838.
Kong, D., Xue, K., Yao, F., Zhang, H.H., 2016b. Partially functional linear regression in high dimensions. Biometrika 103, 147–159.
Ma, S., Huang, J., 2016. Estimating subgroup-specific treatment effects via concave fusion. arXiv preprint arXiv:1607.03717.
Ma, S., Huang, J., 2017. A concave pairwise fusion approach to subgroup analysis. J. Amer. Statist. Assoc. 112, 410–423.
Ma, H., Li, T., Zhu, H., Zhu, Z., 2019. Quantile regression for functional partially linear model in ultra-high dimensions. Comput. Statist. Data Anal. 129, 135–147.
McLachlan, G., Peel, D., 2004. Finite Mixture Models. John Wiley & Sons.
Müller, C.H., Garlipp, T., 2005. Simple consistent cluster methods based on redescending m-estimators with an application to edge identification in images. J. Multivariate Anal. 92, 359–385.
Nir, T.M., Jahanshad, N., Villalon-Reina, J.E., Toga, A.W., Jack, C.R., Weiner, M.W., Thompson, P.M., ADNI, 2013. Effectiveness of regional dti measures in distinguishing alzheimer's disease, mci, and normal aging. NeuroImage: Clin. 3, 180–195.
Pollard, D., 1982. A central limit theorem for *k*-means clustering. Ann. Probab. 10, 919–926.
Preda, C., Saporta, G., 2005. Clusterwise pls regression on a stochastic process. Comput. Statist. Data Anal. 49, 99–108.
Preda, C., Saporta, G., 2007. Pcr and pls for clusterwise regression on functional data. In: Selected Contributions in Data Analysis and Classification. Springer, pp. 589–598.
Ramsay, J.O., Silverman, B.W., 2005. Functioanl Data Analysis, second ed. Springer.
Rao, C.R., Wu, Y., Shao, Q., 2007. An m-estimation-based procedure for determining the number of regression models in regression clustering. In: Advances in Decision Sciences.
Shin, H., Lee, S., 2016. An rkhs approach to robust functional linear regression. Statist. Sinica 26, 255–272.
Smith, S.M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T.E., Mackay, C.E., Watkins, K.E., Ciccarelli, O., Cader, M.Z., Matthews, P.M., et al., 2006. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. Neuroimage 31, 1487–1505.
Späth, H., 1979. Algorithm 39 clusterwise linear regression. Computing 22, 367–373.
Tang, Q., 2017. M-estimation for functional linear regression. Comm. Statist. Theory Methods 46, 3782–3800.
Wang, G., Song, X., 2018. Functional sufficient dimension reduction for functional data classification. J. Classification 35, 250–272.
Wu, Y., Zen, M.-M., 1999. A strongly consistent information criterion for linear model selection based on m-estimation. Probab. Theory Related Fields 113, 599–625.
Yao, F., Fu, Y., Lee, T.C., 2010. Functional mixture regression. Biostatistics 12, 341–353.
Yao, F., Müller, H.-G., Wang, J.-L., 2005. Functional data analysis for sparse longitudinal data. J. Amer. Statist. Assoc. 100, 577–590.
Yu, D., Kong, L., Mizera, I., 2016. Partial functional linear quantile regression for neuroimaging data analysis. Neurocomputing 195, 74–87.
Zhang, Y., Schuff, N., Du, A.-T., Rosen, H.J., Kramer, J.H., Gorno-Tempini, M.L., Miller, B.L., Weiner, M.W., 2009. White matter damage in frontotemporal dementia and alzheimer's disease measured by diffusion mri. Brain 132, 2579–2592.
Zhang, Y., Wang, H.J., Zhu, Z., 2019. Robust subgroup identification. Statist. Sinica 29, 1873–1889.
Zhu, X., Qu, A., et al., 2018. Cluster analysis of longitudinal profiles with subgroups. Electron. J. Stat. 12, 171–193.