Received: 14 August 2019 | Accepted: 21 July 2020

DOI: 10.1111/biom.13349

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Cluster non-Gaussian functional data

Qingzhi Zhong[1] | **Huazhen Lin**[1] | **Yi Li**[2]

[1] Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu, China 611130

[2] Department of Biostatistics, University of Michigan, Ann Arbor, Michigan

Correspondence
Huazhen Lin, Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu 611130, China.
Email: linhz@swufe.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 11829101, 11931014; National Institutes of Health, Grant/Award Number: R21AG058198

Abstract

Gaussian distributions have been commonly assumed when clustering functional data. When the normality condition fails, biased results will follow. Additional challenges occur as the number of the clusters is often unknown *a priori*. This paper focuses on clustering non-Gaussian functional data without the prior information of the number of clusters. We introduce a semiparametric mixed normal transformation model to accommodate non-Gaussian functional data, and propose a penalized approach to simultaneously estimate the parameters, transformation function, and the number of clusters. The estimators are shown to be consistent and asymptotically normal. The practical utility of the methods is confirmed via simulations as well as an application of the analysis of Alzheimer's disease study. The proposed method yields much less classification error than the existing methods. Data used in preparation of this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative database.

KEYWORDS

clustering analysis, functional principal component analysis, non-Gaussian functional data, nonparametric transformation model, penalized EM algorithm

1 | INTRODUCTION

New technologies allow data to be recorded with high frequency from many research fields, resulting in large volumes of functional data, such as growth curves of infants, patients' blood pressure measured at various time points, daily temperature, and precipitation for consecutive days at national weather stations, term-structured yield curves, and shape representations of body parts. See Ramsay and Silverman (2005), Li and Hsing (2010), Jacques and Preda (2014), Wang *et al.* (2016), and Yao *et al.* (2005) for more details. In this paper, we consider clustering functional data, aiming to identify homogeneous groups of data without using any prior knowledge on the group labels.

As functional data are infinite dimensional, most clustering algorithms project the functional data into a finite-dimensional space, followed by applying a clustering method. For example, Abraham *et al.* (2003), Tarpey and Kinateder (2003), and Suyundykov *et al.* (2010) conducted functional principal component analysis or B-spline expansions, and then detected clusters based on the principal component scores or the coefficients of the B-spline basis, using hierarchical or k-means clustering. Numerous model-based methods have also been developed. For example, Biernacki *et al.* (2000), James and Sugar (2003), and Bouveyron *et al.* (2015) considered the Gaussian mixture model, Liu *et al.* (2003) combined Bayesian clustering and Markov chain Monte Carlo strategies to group functional data, Fröhwirth-Schnatter and Kaufmann (2008) built a clustering algorithm based on time series models, Liu and Yang (2009) developed a coherent framework for simultaneously aligning and clustering functional data, Bouveyron and Jacques (2011) extended a high-dimensional data clustering algorithm to cluster Gaussian functional data, other Gaussian-model based clustering methods included Jacques and Preda (2013, 2014) and Rivera-García *et al.* (2019).

Distance-based clustering methods have sparked much interest. Related works included the L_2 distance-based functional principal component scores (FPCs) developed by Chiou and Li (2007), Peng *et al.* (2008), as well as the weighted L_2 distance designed by Floriello and Vitelli (2017), Ferraty and Vieu (2006), Tarpey and Kinateder (2003) and Tokushige *et al.* (2007) measured dissimilarities between curves using the L_2 distance of their derivatives. Delaigle *et al.* (2019) proposed a modified k-means algorithm for functional data with a given number of clusters.

Most of the aforementioned methods assume, either explicitly or implicitly, the functional data to be Gaussian. When the normality assumption fails, the methods may produce biased results. Particularly, multiple cluster solutions can be falsely identified for homogeneous non-Gaussian functional data (Bauer and Curran, 2003). In practice, non-Gaussian functional data have been commonly observed. For example, the Alzheimer's disease neuroimaging initiative (ADNI) data, which motivated our study, are non-Gaussian (see Figures 7-9 in the Supporting Information). The classification error resulted from the existing Gaussian-based clustering method is 43.1%, while the classification error obtained by applying our proposed method without the Gaussian assumption is merely 2.08%.

In the paper, we propose a semiparametric mixed normal transformation (SMINT) model to group non-Gaussian functional data when the number of clusters is unspecified *a priori*. In the literature, only a few works have been focused on the selection of the number of clusters, and most of them used the Bayesian information criterion (BIC; Schwarz, 1978). However, the BIC method is computationally burdensome, and the large sample model selection results, such as model selection consistency and oracle property, are elusive. We hence propose a penalized approach that selects the number of clusters and estimates all of the parameters and functions simultaneously for non-Gaussian functional data. Our method is interpretable and flexible by allowing unspecified distributions and unknown numbers of clusters. Moreover, our method is computationally feasible. We estimate the mean function and eigenfunctions based on one-dimensional B-splines, instead of directly estimating covariance functions, which is a two-dimensional nonparametric problem (Yao *et al.*, 2005).

The remainder of the paper is organized as follows. In Section 2, we introduce the SMINT model. In Section 3, we propose a combination of penalized likelihood and estimating equations methods to select the number of clusters and estimate the regression parameters and transformation function for each cluster simultaneously. We further propose a BIC-type procedure to select tuning parameters. Section 4 focuses on the theoretical properties, including $n^{1/2}$-consistency and asymptotic normality, and Section 5 reports simulation results and an analysis of the ADNI data. We provide concluding remarks in Section 6 and defer all the proofs to the Supporting Information. The R code for the proposed method is available in the Supporting Information.

2 | SEMIPARAMETRIC MIXED NORMAL TRANSFORMATION MODEL

In classical functional principal component analysis (James *et al.*, 2000; Yao *et al.*, 2005; Jacques and Preda, 2014), the stochastic process $X(t)$ can be written as $X(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t)$, where $\mu(t) = E\{X(t)\}$ is the mean function; $\phi_k(t)$ is the kth orthonormal basis function; and ξ_k is the normal functional principal component scores, which satisfy the following conditions:

(C0) $\phi_k(t)$ is the kth orthonormal eigenfunction of the covariance operator $\Sigma(s,t) = \mathrm{Cov}\{X(s), X(t)\}$, which satisfies $\int \phi_k(t)\phi_j(t)dt = 1$ if $j = k$, and 0 otherwise, and the ξ_k is the normal functional principal component scores with $E(\xi_k) = 0$, $\mathrm{var}(\xi_k) = \lambda_k$, and $\mathrm{cov}(\xi_j, \xi_k) = 0$ if $j \neq k$, with the constraint of $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ and $\sum_{k=1}^{\infty} \lambda_k < \infty$.

As $\sum_{k=1}^{\infty} \lambda_k < \infty$ so the λ_k usually decreases rapidly to 0, the number of included eigenfunctions, K, is usually small or moderate. Hence, we can embed $X(t)$ in a sufficiently flexible but suitable function space with measurement errors, and assume the following model

$$X(t) = \mu(t) + \sum_{k=1}^{K} \xi_k \phi_k(t) + \varepsilon_t, \tag{2.1}$$

where ε_t are errors and independent of ξ_k. Model (2.1) with a fixed K and a Gaussian distribution for ε_t are commonly adopted for longitudinal and functional data analysis (see, eg, James *et al.*, 2000; Yao *et al.*, 2005; Hall *et al.*, 2008). The fixed K may lead to biased estimation and classification, we allow $K \to \infty$ as $n \to \infty$ in the paper.

We propose a model to accommodate non-Gaussian functional data by using an unknown transformation function. Without loss of generality, let $\mathcal{T} = [0,1]$. We assume that the random functions $Y_i(\cdot), i = 1, \ldots, n$ are independent copies of a stochastic process $Y(\cdot)$ on $\mathcal{T}$. For non-Gaussian random variables, Box-Cox power transformations have been routinely used in practice, and nonparametric transformations were proposed for added flexibility (Zhou *et al.*, 2008). Our idea is to suppose the existence of a nonparametric functional operator, denoted by $H(\cdot) : \mathfrak{R} \to \mathfrak{R}$, onto $Y_{i(t)} = Y_i(t)$, such that the following

model holds:

$$H(Y_{i(t)}) = \mu(t) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t) + \varepsilon_{i(t)}, \tag{2.2}$$

where $\mu(t) = E[H\{Y_{i(t)}\}]$ is the mean function of the transformed response, with ξ_{ik}s and ϕ_ks satisfying the conditions in (C0), and measurement errors $\varepsilon_{i(t)}$ are independently and identically distributed as $N(0, \sigma^2)$ and independent of ξ_{ik}. For more flexibility, we do not put any parametric assumptions on $H(\cdot)$, except that it is monotonic function, and allow $K \to \infty$ as $n \to \infty$ (Hall and Hosseini-Nasab, 2006; Lin *et al.*, 2018). We term (2.2) an SMINT model, which includes model (2.1) as a special case with $H(x) = x$.

We cluster functional data based on SMINT, which is to identify different subprocesses underlying observations. To proceed, we introduce a cluster membership indicator by $g \in \{1, \dots, C\}$, with a marginal probability mass π_g satisfying $\sum_{g=1}^{C} \pi_g = 1$. We add g to the subscript of the aforementioned $Y_{i(t)}$ and modify (2.2) in order to model the subprocess $Y_{gi}(\cdot)$ by

$$H(Y_{\text{git}}) = \mu_g(t) + \sum_{k=1}^{K_g} \xi_{\text{gik}}\phi_{gk}(t) + \varepsilon_{\text{git}}, \tag{2.3}$$

where $Y_{\text{git}} = Y_{gi}(t)$, each item is as defined in (2.1), with an added subscript g for the gth subpopulation. The scale and location normalization is needed for H to make the model identifiable. We specify two conditions: $N^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i} H(Y_{ij}) = 0$ and $N^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i} H^2(Y_{ij}) = 1$, where $N = \sum_{i=1}^{n} n_i$, n_i is the number of observation times for curve i and $Y_{ij} = Y_i(t_{ij})$.

Even for Gaussian functional data, most existing methods have assumed the same eigenfunctions ϕ_{gk} across groups (Bouveyron and Jacques, 2011; Serban and Jiang, 2012). The assumption was made for computational feasibility because the cluster-specified covariance function is not available with unknown clusters. But, with this assumption, a large number of eigenfunctions are needed for estimation accuracy. In contrast, our model allows cluster-specific eigenfunctions, as a result, the functional curve in each group can be represented by fewer eigenfunctions, resulting in more concise, informative, and interpretable clusters.

3 | ESTIMATION

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{i,n_i})'$ represent the measurements on individual i over n_i evaluation points, denoted by $\mathbf{t}_i = (t_{i1}, t_{i2}, \dots, t_{i,n_i})'$. Without loss of generality, we assume the t_{ij}'s are scaled within [0,1]. Let $f(\cdot)$ denote the density function of the random vector, $\mathbf{H}(\mathbf{Y}_i) = \{H(Y_{i1}), H(Y_{i2}), \dots, H(Y_{i,n_i})\}'$. Under (2.3), $f(\cdot)$ has the following form:

$$f\{\mathbf{H}(\mathbf{Y}_i)\} = \sum_{g=1}^{C} \pi_g f_g\{\mathbf{H}(\mathbf{Y}_i)\}, \quad \sum_{g=1}^{C} \pi_g = 1, \tag{3.4}$$

where $f_g\{\mathbf{H}(\mathbf{Y}_i)\} = (2\pi)^{-n_i/2}|\mathbf{\Delta}_g(\mathbf{t}_i)|^{-1/2}\exp[-\frac{1}{2}\{\mathbf{H}(\mathbf{Y}_i) - \boldsymbol{\mu}_{gi}\}'\mathbf{\Delta}_g(\mathbf{t}_i)^{-1}\{\mathbf{H}(\mathbf{Y}_i) - \boldsymbol{\mu}_{gi}\}]$ is the density function of $\mathbf{H}(\mathbf{Y}_i)$ if individual i belongs to the gth cluster, $\boldsymbol{\mu}_{gi} = \mu_g(\mathbf{t}_i) \doteq \{\mu_g(t_{i1}), \dots, \mu_g(t_{i,n_i})\}'$, $\mathbf{\Delta}_g(\mathbf{t}_i) = \mathbf{\Phi}_g(\mathbf{t}_i)\mathbf{\Lambda}_g\mathbf{\Phi}_g(\mathbf{t}_i)' + \sigma_g^2\mathbf{I}_{n_i}$, $\mathbf{\Phi}_g(\mathbf{t}_i) = \{\boldsymbol{\phi}_g(t_{i1}), \dots, \boldsymbol{\phi}_g(t_{i,n_i})\}'$, $\boldsymbol{\phi}_g(t) = \{\phi_{g1}(t), \dots, \phi_{g,K_g}(t)\}'$, $\mathbf{\Lambda}_g = \text{diag}(\lambda_{g1}, \dots, \lambda_{g,K_g})$ and $\mathbf{I}_{n_i}$ is the $n_i \times n_i$ identity matrix. Then the covariance function of the transformed response is $\Sigma_g(s,t) = \boldsymbol{\phi}_g(t)'\mathbf{\Lambda}_g\boldsymbol{\phi}_g(s) + \sigma_g^2 I(s=t)$ for the gth cluster. Let $\mathbf{\Omega} = (\mathbf{\Lambda}', \boldsymbol{\sigma}^{2'}, \boldsymbol{\pi}', \boldsymbol{\mu}', \boldsymbol{\phi}')'$ with $\mathbf{\Lambda} = (\lambda_{gk}, g = 1, \dots, C, k = 1, \dots, K_g)'$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_C^2)'$, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_C)'$, $\boldsymbol{\mu} = (\mu_g, g = 1, \dots, C)'$ and $\boldsymbol{\phi} = (\phi_{gk}, g = 1, \dots, C, k = 1, \dots, K_g)'$.

Our goal is to simultaneously estimate the number of clusters C, the transformation function H, the mean function μ_g, and the covariance function $\Sigma_g(s,t)$ of each cluster via estimating $\mathbf{\Omega}$, which includes finite-dimensional parameters and infinite-dimensional functions. Different from the existing method that directly estimating $\Sigma_g(s,t)$ by two-dimensional nonparametric technique (Yao *et al.*, 2005; Cai and Yuan, 2010), we estimate it via its eigenfunctions using univariate splines, hence effectively increase the convergence rate of the estimator of $\Sigma_g(s,t)$ from $n^{-1/3}$ to $n^{-2/5}$.

We start with modeling $\mu_g(\cdot)$ and $\phi_{gk}(\cdot)$. Denote by

$$\mathcal{G} = \{g(\cdot) : |g^{(q_1)}(t_1) - g^{(q_1)}(t_2)| \le c_0|t_1 - t_2|^{q_2}, \text{ for any } 0 \le t_1, t_2 \le 1\}, \tag{3.5}$$

where q_1 is a nonnegative integer, $q_2 \in (0, 1]$, $r = q_1 + q_2 \ge 2$, and c_0 is a generic constant. The smoothness assumption (3.5) is often used in nonparametric estimation. With the assumption of $\mu_g(\cdot), \phi_{gk}(\cdot) \in \mathcal{G}$, we approximate $\mu_g(\cdot)$ and $\phi_{gk}(\cdot)$ by

$$\mu_{ng}(t) = \sum_{j=1}^{q_n} \alpha_{gj}b_j(t) = \boldsymbol{\alpha}_g'\mathbf{B}_n(t), \tag{3.6}$$

$$\phi_{ngk}(t) = \sum_{j=1}^{q_n} \beta_{gkj}b_j(t) = \boldsymbol{\beta}_{gk}'\mathbf{B}_n(t), \tag{3.7}$$

respectively, where $\mathbf{B}_n(\cdot) = \{b_1(\cdot), \dots, b_{q_n}(\cdot)\}'$ is an orthogonal set of spline basis functions of order $r+1$ with knots

$0 = \zeta_0 < \zeta_1 < \cdots < \zeta_{M_n} < \zeta_{M_n+1} = 1$, satisfying $\max(\zeta_j - \zeta_{j-1} : j = 1, \ldots, M_n) = O(n^{-\upsilon})$, where $q_n = M_n + r + 1$, and M_n is the integer part of n^{υ} with $0 < \upsilon < 0.5$. Substituting (3.6) and (3.7) into (3.4), we obtain the log-likelihood,

$$L_n(\boldsymbol{\Omega}_n; H) = \sum_{i=1}^{n} \log \left\{ \sum_{g=1}^{C} \pi_g f_{gi}(\boldsymbol{\Omega}_n; H) \right\}, \tag{3.8}$$

subject to $N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n_i} H(Y_{ij}) = 0$ and $N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n_i} H^2(Y_{ij}) = 1$,

$$\sum_{g=1}^{C} \pi_g = 1, \quad \int_t \mathbf{B}_n(t)\mathbf{B}_n(t)'dt = \mathbf{I}_{q_n}, \quad \boldsymbol{\beta}_g \boldsymbol{\beta}_g' = \mathbf{I}_{K_g}, \quad \text{for } g = 1, \ldots, C, \tag{3.9}$$

and the first nonzero element of each row of $\boldsymbol{\beta}_g$ to be positive, where $\boldsymbol{\beta}_g = (\boldsymbol{\beta}_{g1}, \ldots, \boldsymbol{\beta}_{g,K_g})'$,

$$f_{gi}(\boldsymbol{\Omega}_n; H) = \frac{(2\pi)^{-n_i/2}}{|\boldsymbol{\Delta}_{gi}(\boldsymbol{\Omega}_n)|^{1/2}} \times \exp\left[-\frac{1}{2} \left\{ \mathbf{H}(\mathbf{Y}_i) - \mathbf{B}_{ni}\boldsymbol{\alpha}_g \right\}' \boldsymbol{\Delta}_{gi}(\boldsymbol{\Omega}_n)^{-1} \left\{ \mathbf{H}(\mathbf{Y}_i) - \mathbf{B}_{ni}\boldsymbol{\alpha}_g \right\} \right],$$

$\mathbf{B}_{ni} = \{\mathbf{B}_n(t_{i1}), \ldots, \mathbf{B}_n(t_{i,n_i})\}'$, $\boldsymbol{\Delta}_{gi}(\boldsymbol{\Omega}_n) = \mathbf{B}_{ni}\boldsymbol{\beta}_g' \boldsymbol{\Lambda}_g \boldsymbol{\beta}_g \mathbf{B}_{ni}' + \sigma_g^2 \mathbf{I}_{n_i}$, and $\boldsymbol{\Omega}_n = \{\lambda_{gk}, \sigma_g^2, \pi_g, \boldsymbol{\alpha}_g, \boldsymbol{\beta}_{gk}, k = 1, \ldots, K_g, g = 1, \ldots, C\}$. (3.9) is orthogonality constraints on the eigenfunctions. To adhere to the orthonormal constraints on the B-spline, we denote by $\mathbf{A} = \int_t \widetilde{\mathbf{B}}_n(t)\widetilde{\mathbf{B}}_n(t)'dt$, where $\widetilde{\mathbf{B}}_n(t)$ is the cubic B-spline basis functions (Schumaker, 2007). As $\mathbf{A}$ is a symmetric matrix, we have $\int_t \{\mathbf{A}^{-\frac{1}{2}}\widetilde{\mathbf{B}}_n(t)\}\{\mathbf{A}^{-\frac{1}{2}}\widetilde{\mathbf{B}}_n(t)\}'dt = \mathbf{I}_{q_n}$, and $\mathbf{B}_n(t) = \mathbf{A}^{-\frac{1}{2}}\widetilde{\mathbf{B}}_n(t)$ satisfies the orthonormal constraints. Throughout the paper, n in the subscript represents the corresponding parameters and functions when using splines to approximate $\mu_g(\cdot)$ and $\phi_{gk}(\cdot)$.

As C is not given *a priori* and to make the model inclusive, we often start with a relative large C, which, however, may lead to many nuisance parameters and cause large variation of estimates. This necessitates developing a more formal procedure for estimating C. To proceed, we first note that if π_g is found to be 0, the gth cluster is not necessary and can be deleted from the model. Hence, the selection of the clusters corresponds to the selection of nonzero elements of $\{\pi_g, g = 1, \ldots, C\}$. However, we may not directly penalize on $\{\pi_g, g = 1, \ldots, C\}$ to achieve model selection. To see that, we consider the complete data for individual i as $\mathbf{D}_i = \{\mathbf{Y}_i, \boldsymbol{\delta}_i\}$, where $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{iC})'$ with $\delta_{ig} = 1$ if $\mathbf{Y}_i$ arises from the gth cluster, otherwise $\delta_{ig} = 0$. The expected complete-data log-likelihood function is

$$\sum_{i=1}^{n} \sum_{g=1}^{C} \left(b_{ig} \left[\log(\pi_g) + \log\{f_{gi}(\boldsymbol{\Omega}_n; H)\} \right] \right) \tag{3.10}$$

where $b_{ig} = E\{\delta_{ig}|\mathbf{Y}_i\}$. As (3.10) contains $\log(\pi_g)$ with an unbounded derivative when π_g is close zero, setting L_p penalties directly on π_g will not return an exact zero solution for π_g. Instead, we propose to penalize on $\log\{\pi_g\}$ in order to achieve the sparsity for $(\pi_g, 1 \le g \le C)'$. However, large π_g might be overly shrunk to 0 as we penalize on $\log\{\pi_g\}$. Therefore, following Huang *et al.* (2017), we propose the following penalized likelihood,

$$Q_n(\boldsymbol{\Omega}_n; H) = L_n(\boldsymbol{\Omega}_n; H) - n\lambda \sum_{g=1}^{C} \log \left\{ \frac{\epsilon + \pi_g}{\epsilon} \right\}, \tag{3.11}$$

where ϵ is a very small positive number, say 10^{-6} or $o\{n^{-1/2}(\log n)^{-1}\}$ (Huang *et al.*, 2017). Then it is natural to define the penalized log-likelihood estimator

$$(\widehat{\boldsymbol{\Omega}}_n, \widehat{H}_n) = \arg\max_{\boldsymbol{\Omega}_n, H} Q_n(\boldsymbol{\Omega}_n; H), \tag{3.12}$$

based on which we show that there is a nonzero probability of estimating some π_g's to be exactly zero and achieving automatic cluster selection. Our procedure naturally integrates the steps of cluster selection and parameter estimation, which makes computation feasible.

The penalized likelihood function $Q_n(\boldsymbol{\Omega}_n; H)$ involves the infinite-dimensional function $H(\cdot)$ and mixture distribution, so a direct maximization is not feasible. We resort to a two-stage approach. Particularly, we estimate $\boldsymbol{\Omega}_n$ by maximizing the penalized pseudo-likelihood, which is implemented by a penalized expectation maximization (EM) algorithm based on $Q_n(\boldsymbol{\Omega}_n; H)$ with H replaced by its estimated value with (3.20) described in Section 3.2. We repeat the procedure until convergence.

3.1 | Penalized expectation maximization algorithm

As the penalized likelihood function involves both finite- and infinite-dimensional parameters, we resort to a two-stage iterative algorithm. We estimate the parameter $\boldsymbol{\Omega}_n$ by a penalized EM algorithm, then we use a series of estimating equations to estimate the transformation function H.

We first consider the penalized maximum likelihood estimator for $\boldsymbol{\Omega}_n$ given H, and propose a penalized EM algorithm (Dempster *et al.*, 1977), which was originally designed for handling missing data. In our setting, we treat

$\boldsymbol{\delta}_i$ as the missing data and view the complete data for individual i as $\mathbf{D}_i = \{\mathbf{Y}_i, \boldsymbol{\delta}_i\}$. Thus, the penalized complete-data log-likelihood function is

$$Q_c(\boldsymbol{\Omega}_n; H) = \log \mathcal{L}_c(\boldsymbol{\Omega}_n; H) - n\lambda \sum_{g=1}^{C} \log\left\{\frac{\epsilon + \pi_g}{\epsilon}\right\}, \tag{3.13}$$

where $\log \mathcal{L}_c(\boldsymbol{\Omega}_n; H) \propto \sum_{i=1}^{n}\sum_{g=1}^{C}(\delta_{ig}[\log(\pi_g) + \log\{f_{gi}(\boldsymbol{\Omega}_n; H)\}])$.

In the maximization step, we maximize the conditional expectation of $Q_c(\boldsymbol{\Omega}_n; H)$ given the observed data. Differentiating $E\{Q_c(\boldsymbol{\Omega}_n; H)|\mathbf{Y}_i, i = 1, \ldots, n\}$ with respect to $\boldsymbol{\Omega}_n$ and setting the derivatives to zero lead to

$$\sum_{i=1}^{n}\frac{E(\delta_{ig}|\mathbf{Y}_i)}{\pi_g} - \sum_{i=1}^{n}\frac{E(\delta_{i1}|\mathbf{Y}_i)}{1-\sum_{j=2}^{C}\pi_j} + \frac{n\lambda}{\epsilon + 1 - \sum_{j=2}^{C}\pi_j} - \frac{n\lambda}{\epsilon + \pi_g} = 0, \quad g \geq 2, \tag{3.14}$$

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_{gk} = &\left\{\sum_{i=1}^{n} E(\delta_{ig}|\mathbf{Y}_i)\mathbf{B}_{ni}'\boldsymbol{\Delta}_{gi}(\boldsymbol{\Omega}_n)^{-1}\mathbf{B}_{ni}\right\}^{-1} \\ &\times \sum_{i=1}^{n} E(\delta_{ig}|\mathbf{Y}_i)\mathbf{B}_{ni}'\boldsymbol{\Delta}_{gi}(\boldsymbol{\Omega}_n)^{-1}\left\{\mathbf{H}(\mathbf{Y}_i) - \mathbf{B}_{ni}\boldsymbol{\alpha}_g\right\}^{\otimes 2}\boldsymbol{\Delta}_{gi}(\boldsymbol{\Omega}_n)^{-1}\mathbf{B}_{ni}\boldsymbol{\beta}_{gk},\end{aligned} \tag{3.15}$$

$$\lambda_{gk} = \frac{-\sum_{i=1}^{n} E(\delta_{ig}|\mathbf{Y}_i)\boldsymbol{\beta}_{gk}'\mathbf{B}_{ni}'\boldsymbol{\Delta}_{gi}(\boldsymbol{\Omega}_n)^{-1}\mathbf{R}_{gi,-k}(\boldsymbol{\Omega}_n)\boldsymbol{\Delta}_{gi}(\boldsymbol{\Omega}_n)^{-1}\mathbf{B}_{ni}\boldsymbol{\beta}_{gk}}{\sum_{i=1}^{n} E(\delta_{ig}|\mathbf{Y}_i)\boldsymbol{\beta}_{gk}'\mathbf{B}_{ni}'\boldsymbol{\Delta}_{gi}(\boldsymbol{\Omega}_n)^{-1}\mathbf{B}_{ni}\boldsymbol{\beta}_{gk}\boldsymbol{\beta}_{gk}'\mathbf{B}_{ni}'\boldsymbol{\Delta}_{gi}(\boldsymbol{\Omega}_n)^{-1}\mathbf{B}_{ni}\boldsymbol{\beta}_{gk}}, \tag{3.16}$$

$$\sigma_g^2 = \frac{-\sum_{i=1}^{n} E(\delta_{ig}|\mathbf{Y}_i)\mathrm{tr}\left(\boldsymbol{\Delta}_{gi}(\boldsymbol{\Omega}_n)^{-2}\left[\mathbf{B}_{ni}\sum_{k=1}^{K_g}\lambda_{gk}\boldsymbol{\beta}_{gk}\boldsymbol{\beta}_{gk}'\mathbf{B}_{ni}' - \left\{\mathbf{H}(\mathbf{Y}_i) - \mathbf{B}_{ni}\boldsymbol{\alpha}_g\right\}^{\otimes 2}\right]\right)}{\sum_{i=1}^{n} E(\delta_{ig}|\mathbf{Y}_i)\mathrm{tr}\left\{\boldsymbol{\Delta}_{gi}(\boldsymbol{\Omega}_n)^{-2}\right\}}, \tag{3.17}$$

$$\boldsymbol{\alpha}_g = \left\{\sum_{i=1}^{n} E(\delta_{ig}|\mathbf{Y}_i)\mathbf{G}_{gi}(\boldsymbol{\Omega}_n)\right\}^{-1}\sum_{i=1}^{n} E(\delta_{ig}|\mathbf{Y}_i)\mathbf{B}_{ni}'\boldsymbol{\Delta}_{gi}(\boldsymbol{\Omega}_n)^{-1}\mathbf{H}(\mathbf{Y}_i), \tag{3.18}$$

for $k = 1, \ldots, K_g$ and $g = 1, \ldots, C$, where $\mathbf{G}_{gi}(\boldsymbol{\Omega}_n) = \mathbf{B}_{ni}'\boldsymbol{\Delta}_{gi}(\boldsymbol{\Omega}_n)^{-1}\mathbf{B}_{ni}$, $\mathbf{R}_{gi,-k}(\boldsymbol{\Omega}_n) = \mathbf{B}_{ni}\sum_{r=1,r\neq k}^{K_g}\lambda_{gr}\boldsymbol{\beta}_{gr}\boldsymbol{\beta}_{gr}'\mathbf{B}_{ni}' + \sigma_g^2\mathbf{I}_{n_i} - \{\mathbf{H}(\mathbf{Y}_i) - \mathbf{B}_{ni}\boldsymbol{\alpha}_g\}^{\otimes 2}$. Given a small ϵ such that $\frac{1}{\pi_j + \epsilon} \approx \frac{1}{\pi_j}$ for all j, we obtain an approximate solution for (3.14),

$$\tilde{\pi}_g = \max\left\{0, \frac{1}{1 - C\lambda}\left[\frac{1}{n}\sum_{i=1}^{n} E(\delta_{ig}|\mathbf{Y}_i) - \lambda\right]\right\}. \tag{3.19}$$

Some $\tilde{\pi}_g$ may be shrunk to zero, in which case, we only need to renormalize $\tilde{\pi}_g$ by enforcing $\sum_{g=1}^{C}\tilde{\pi}_g = 1$ after the EM algorithm converges. Denote the estimate of $\boldsymbol{\Omega}_n$ from the rth step by $\widetilde{\boldsymbol{\Omega}}_n$. We update $\pi_g^{(r-1)}$ from step $r-1$ by $\pi_g^{(r)} = \widetilde{\pi}_g / \sum_{j=1}^{C}\widetilde{\pi}_j$ and further perform a QR decomposition on $\widetilde{\boldsymbol{\beta}}_g = (\widetilde{\boldsymbol{\beta}}_{g1}, \ldots, \widetilde{\boldsymbol{\beta}}_{g,K_g})'$ obtained from (3.15) to get $\widetilde{\boldsymbol{\beta}}_g = \mathbf{QR}$ and update $\boldsymbol{\beta}_g^{(r-1)}$ from step $r-1$ by $\boldsymbol{\beta}_g^{(r)} = \mathbf{Q}'$. It is easy to see that $\boldsymbol{\beta}_g^{(r)}\boldsymbol{\beta}_g^{(r)\prime}$ is an identity matrix. We estimate $\boldsymbol{\Omega}_n$ by repeatedly using equations (3.15)-(3.19) until convergence. At each step, $\boldsymbol{\Omega}_n$ on the right side of the equations is replaced by its most updated value. When computing the conditional mean δ_{ig} given $\mathbf{Y}_i$, $E(\delta_{ig}|\mathbf{Y}_i) = \frac{f_g\{\mathbf{H}(\mathbf{Y}_i)\}\pi_g}{\sum_{j=1}^{C} f_j\{\mathbf{H}(\mathbf{Y}_i)\}\pi_j}$, we also replace all the unknown parameters and functions with their estimates from the previous step.

3.2 | Estimation of the transformation function

For any given $y = Y_{ij}$, $j = 1, \ldots, n_i$ and $i = 1, \ldots, n$, we have

$$\begin{aligned}Pr(Y_{ij} \leq y) &= Pr\{H(Y_{ij}) \leq H(y)\} \\ &= \sum_{g=1}^{C}\pi_g Pr\{H(Y_{ij}) \leq H(y)|\delta_{ig} = 1\} \\ &= \sum_{g=1}^{C}\pi_g\Phi\left\{\frac{H(y) - \mathbf{B}_n(t_{ij})'\boldsymbol{\alpha}_g}{\sqrt{\mathbf{B}_n(t_{ij})'\boldsymbol{\beta}_g'\boldsymbol{\Lambda}_g\boldsymbol{\beta}_g\mathbf{B}_n(t_{ij}) + \sigma_g^2}}\right\},\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution of the standard normal variable. For each y in the support of Y_{ij}, we estimate $H(y)$ by solving

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[I(Y_{ij}\le y)-\sum_{g=1}^{C}\pi_g\Phi\times\left\{\frac{H(y)-\mathbf{B}_n(t_{ij})'\boldsymbol{\alpha}_g}{\sqrt{\mathbf{B}_n(t_{ij})'\boldsymbol{\beta}_g'\boldsymbol{\Lambda}_g\boldsymbol{\beta}_g\mathbf{B}_n(t_{ij})+\sigma_g^2}}\right\}\right]=0. \quad (3.20)$$

For each y, the estimating equation implicitly defines a value of $\theta\equiv H(y)$. Clearly $\sum_{i=1}^{n}\sum_{j=1}^{n_i}I\{Y_{ij}\le y\}$ is a nondecreasing function of y. Therefore, the second term in the estimating equation (3.20) is a nondecreasing function of y. As Φ is an increasing function of its argument, we must have θ nondecreasing in y. By the same reasoning, we can see that θ is piecewise constant, with nonzero jumps at the data values of y.

Remark 1. The computational cost for $H(\cdot)$ is very limited. Coupled with the closed-form estimator for Ω_n at each step, the implementation and computation of the proposed method are simple. Unlike a traditional nonparametric approach to estimate the transformation function (Horowitz, 1996), our approach does not involve nonparametric smoothing and thus does not suffer from smoothing-related problems, for example, selection of a smoothing parameter.

Denote the resulting estimate for $H(Y_{ij})$ from the rth step by $H_{nr}(Y_{ij})$. Then we update $H^{(r-1)}(Y_{ij})$ by $H^{(r)}(Y_{ij})=\{H_{nr}(Y_{ij})-\overline{H}_{nr}\}/\mathrm{sd}\{H_{nr}\}$ for identification, where $\bar{H}_{nr}$ and $\mathrm{sd}\{H_{nr}\}$ is the empirical mean and standard deviation of $H_{nr}(Y_{ij})$ over i, j.

3.3 | Selection of hyper-parameters

The method requires to tune three parameters: the truncation parameter K_g, the penalty parameter λ, and the number of interior knots M_n. For standard LASSO and SCAD penalty functions, Wang *et al.* (2007) showed that the BIC yields model selection consistency, we hence propose to select K_g, M_n, and λ and by maximizing

$$BIC(K_g,M_n,\lambda)=L_n(\boldsymbol{\Omega}_n;H)-\frac{1}{2}DF(\boldsymbol{\Omega}_n)\log\left(\sum_{i=1}^{n}n_i\right), \quad (3.21)$$

where $L_n(\boldsymbol{\Omega}_n;H)$ is as defined in (3.8) and $DF(\boldsymbol{\Omega}_n)$ is the number of parameters $\boldsymbol{\Omega}_n$. As reported in Figure S.4(a)-(d) in the Supporting Information, we have examined the performance of criterion (3.21) in selecting K_g, M_n, and λ, which show that the optimal K_g, M_n, and λ are nearly independent, suggesting K_g, M_n, and λ can be separately chosen. Furthermore, we can see the proposed method is not sensitive to the choice of M_n and hence a rough selection for M_n is enough. For smooth and either monotonic or unimodal functions, $2-6$ knots seem quite adequate and that is what we recommend. The simulations and data analysis suggested the BIC criterion (3.21) performs well.

4 | LARGE SAMPLE PROPERTIES

Let $\widehat{\boldsymbol{\Omega}}_n=\{\widehat{\lambda}_{gk},\widehat{\sigma}_g^2,\widehat{\pi}_g,\widehat{\boldsymbol{\alpha}}_g,\widehat{\boldsymbol{\beta}}_{gk},k=1,\dots,K_g,g=1,\dots,\widehat{C}\}$ and $\widehat{H}_n$ be the estimator of $\boldsymbol{\Omega}$ and H derived above. The mean function $\mu_g(t)$ and covariance functions $\Sigma_g(s,t)$ can be estimated by $\widehat{\mu}_g(t)=\widehat{\boldsymbol{\alpha}}_g'\mathbf{B}_n(t)$ and $\widehat{\Sigma}_g(s,t)=\sum_{k=1}^{K_g}\widehat{\phi}_{gk}(t)\widehat{\lambda}_{gk}\widehat{\phi}_{gk}(s)+\widehat{\sigma}_g^2I(s=t)$, respectively, where $\widehat{\phi}_{gk}(t)=\widehat{\boldsymbol{\beta}}_{gk}'\mathbf{B}_n(t)$. In this section, we focus on the theoretical properties, including $n^{1/2}$-consistency and asymptotic normality.

Denote the Euclidean norm and the L^2 norm by $\|\cdot\|$ and $\|\cdot\|_2$, respectively, and the parametric space $\boldsymbol{\Omega}^*=\{\boldsymbol{\Omega}=(\boldsymbol{\Lambda}',\boldsymbol{\sigma}^{2\prime},\boldsymbol{\pi}',\boldsymbol{\mu}',\boldsymbol{\phi}')'\in R_+^{\sum_{g=1}^{C}K_g}\otimes R_+^C\otimes[0,1]^C\otimes\mathcal{G}^C\otimes\mathcal{G}^{\sum_{g=1}^{C}K_g}\}$ with $R_+=(0,\infty)$. Denote $\boldsymbol{\Theta}=(\boldsymbol{\pi}',\boldsymbol{\mu}',\boldsymbol{\Sigma}')'$ with $\boldsymbol{\Sigma}=(\Sigma_1,\dots,\Sigma_C)'$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and covariance functions of interest, and $\boldsymbol{\pi}$ is the probabilities. We define the norm between $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ by

$$d(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2)=\left(\|\boldsymbol{\pi}_1-\boldsymbol{\pi}_2\|^2+\sum_{g=1}^{C}\|\mu_{g,1}-\mu_{g,2}\|_2^2+\sum_{g=1}^{C}\|\Sigma_{g,1}-\Sigma_{g,2}\|_2^2\right)^{1/2}.$$

Throughout the paper, 0 in the subscript represents the true values of corresponding parameters and functions. Without loss of generality, we assume that $\pi_{1,0}\ge\pi_{2,0}\ge\cdots\ge\pi_{C_0,0}>0$, $\sum_{g=1}^{C_0}\pi_{g,0}=1$, and $\pi_{C_0+1,0}=\cdots=\pi_{C,0}=0$. We set the following conditions.

(C1) $\max_{1\le j\le M_n}(\zeta_j-\zeta_{j-1})=O(n^{-\upsilon})$ with $0<\upsilon<1/2$. Moreover, $\max_{1\le j\le M_n}(\zeta_j-\zeta_{j-1})/\min_{1\le j\le M_n}(\zeta_j-\zeta_{j-1})$ is bounded.

(C2) $\boldsymbol{\Omega}_0$ is an interior point of $\boldsymbol{\Omega}^*$ and $\mu_{g0}\in\mathcal{G},\phi_{gk0}\in\mathcal{G}$ for $1\le g\le C, 1\le k\le K_g$.

(C3) There exists $[y_1,y_2]$ so that $\frac{1}{\sum_{i=1}^{n}n_i}\sum_{i=1}^{n}\sum_{j=1}^{n_i}I(Y_{ij}\notin[y_1,y_2])=o_p(n^{-1/2})$.

(C4) The transformation function $H(y)$ is strictly increasing and its first derivative is continuous over $y\in[y_1,y_2]$.

(C5) $K_g = O(n^e)$ for $1 \le g \le C$, with $0 \le e < \min(1-\upsilon, 2r\upsilon)$.
(C6) $\sum_{k=1}^{K_g} \lambda_{gk0} < \infty$ for $1 \le g \le C$.
(C7) The matrix $E\{S(\theta_0)S(\theta_0)'\}$ is finite and positive definite, where θ_0 is the true value of $\theta = (\sigma^{2'}, \pi')'$, and $S(\theta_0)$ is defined in the proof of Theorem 3.

Condition (C1) is common in the spline smoothing (Chen *et al.*, 2017), and Condition (C2) is often assumed in semiparametric analyses (Chen and Tong, 2010; Ma *et al.*, 2015). Condition (C3) is needed to avoid the tail problem (Lin *et al.*, 2012). Condition (C4) is a regular assumption for the transformation function (Horowitz, 1996; Zhou *et al.*, 2008). In practice, K_g is small and Condition (C5) is easy to be satisfied. Condition (C6) is needed to avoid unbounded covariance (Hall *et al.*, 2007). Condition (C7) is to ensure the existence of the asymptotic covariance matrix (Ma *et al.*, 2015; Chen *et al.*, 2017). We state below Theorems 1-3, which indicate the proposed estimates are consistent and asymptotically normal, forming the basis for statistical inference. We defer the proofs to the Supporting Information .

Theorem 1. *Under Conditions (C1)-(C6), $\lambda\sqrt{n} \to 0$, $\lambda\sqrt{n}\log(n) \to \infty$, and $\epsilon = o\{\frac{1}{\sqrt{n}\log(n)}\}$, we have $\widehat{C} \to C_0$ with probability tending to 1.*

Theorem 2. *With Conditions (C1)-(C6), $\lambda\sqrt{n} \to 0$, and $\epsilon = o\{\frac{1}{\sqrt{n}\log(n)}\}$, we have*

$$\widehat{H}_n(y) \to H_0(y) \text{ uniformly over } y \in [y_1, y_2] \text{ and}$$

$$d(\widehat{\boldsymbol{\Theta}}_n, \boldsymbol{\Theta}_0) = O_p\{n^{-\min(\frac{1-\upsilon-e}{2}, r\upsilon-\frac{e}{2})}\},$$

as $n \to \infty$, r is defined in (3.5), and $\upsilon \in (0, 0.5]$ is given to determine the number of knots for the spline basis $B_n(\cdot)$.

As e increases, the number of eigenfunctions increases and the parameter space becomes larger, causing the error $d(\widehat{\boldsymbol{\Theta}}_n, \boldsymbol{\Theta}_0)$ to increase. When $e = 0$, corresponding to a fixed number of eigenfunctions, $d(\widehat{\boldsymbol{\Theta}}_n, \boldsymbol{\Theta}_0) = O_p\{n^{-\min(\frac{1-\upsilon}{2}, r\upsilon)}\}$. In this case, taking $\upsilon = 1/(2r+1)$, we have $d(\widehat{\boldsymbol{\Theta}}_n, \boldsymbol{\Theta}_0) = O_p\{n^{-r/(2r+1)}\}$, which is the optimal rate for the estimation of univariate nonparametric functions (Stone, 1980).

Theorem 3. *Assume that Conditions (C1)-(C7) hold with $r \ge 2$, $\frac{1}{4r} < \upsilon < \frac{1}{2}$, $\lambda\sqrt{n} \to 0$, and $\epsilon = o\{\frac{1}{\sqrt{n}\log(n)}\}$. As $n \to \infty$, the estimator $\widehat{\theta}$ for finite parameters θ satisfies*

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to N\{0, \mathbf{I}^{-1}(\theta_0)\},$$

where $\mathbf{I}(\theta_0) = E\{S(\theta_0)S(\theta_0)'\}$ is defined in the Supporting Information .

5 | SIMULATIONS

With unspecified transformation functions, we expect our method to be robust and flexible. To investigate the trade-off between the added robustness and the efficiency, we compare the proposed method with the model-based and distance-based clustering methods. The former includes the method with a correctly specified transformation function (CT), the method without transformation (WoT), and the FunFEM method (Bouveyron *et al.*, 2015). We take the DHP method recently proposed by Delaigle *et al.* (2019) as a representation of distance-based clustering methods. We also examine the sensitivity of the proposed method to the initial value of the number of groups and the performance of criterion (3.21) in selecting K_g, M_n, and λ. As our method requires the measurement error of the transformed responses to be normally distributed, we further investigate the sensitivity of the proposed method to the normality assumption. We assess the performance in terms of bias, standard deviation (sd), and root mean squared error (RMSE), defined by $\text{bias} = [\frac{1}{n_{\text{grid}}}\sum_{i=1}^{n_{\text{grid}}}\{E\widehat{f}(t_i) - f(t_i)\}^2]^{1/2}$, $\text{sd} = [\frac{1}{n_{\text{grid}}}\sum_{i=1}^{n_{\text{grid}}} E\{\widehat{f}(t_i) - E\widehat{f}(t_i)\}^2]^{1/2}$, and $\text{RMSE} = [\text{bias}^2 + \text{sd}^2]^{1/2}$ for any estimation $\widehat{f}(\cdot)$ of $f(\cdot)$, with t_i $(i = 1, \ldots, n_{\text{grid}})$ being the grid evaluation points and $E\widehat{f}(t_i)$ being approximated by its sample mean, based on 300 replications and $n_{\text{grid}} = 300$. To evaluate the classification accuracy, we use the measurements of purity function (PF) and adjusted Rand index (ARI), commonly used in the classification literature (Delaigle *et al.*, 2019). In the following simulations, we use the cubic B-spline, choose the number of interior knots M_n via BIC in (3.21), and place the knots at the M_n quantiles of the observation times.

5.1 | Performance of estimation

To investigate the efficiency and robustness, we compare the proposed method with the CT and WoT in terms of bias, sd, and RMSE.

Simulation 1. We generate $n = 400$ samples from a three-component population with mixing probabilities $\pi = (1/3, 1/3, 1/3)'$. The data in the gth cluster are generated from $H(Y_{\text{git}}) = \mu_g(t) + \sum_{k=1}^{2}\xi_{\text{gik}}\phi_{gk}(t) + \varepsilon_{\text{git}}$, $g = 1, 2, 3$, where $\mu_1(t) = t + \sin(\pi t), \mu_2(t) = \exp(t)$, $\mu_3(t) = 2t^2 + 2$; $\phi_{11}(t) = \sqrt{2}\cos(\pi t), \phi_{12}(t) = \sqrt{2}\sin(\pi t)$, $\phi_{21}(t) = \sqrt{2}\cos(2\pi t), \quad \phi_{22}(t) = \sqrt{2}\cos(\pi t), \quad \phi_{31}(t) =$

TABLE 1 Results for Case 1 of Simulation 1

	Proposed ($C = 7$)		CT ($C = 3$)		CT ($C = 7$)		WoT ($C = 3$)	
	bias (sd)	RMSE	bias (sd)	RMSE	bias (sd)	RMSE	bias (sd)	RMSE
π_1	0.003 (0.034)	0.034	0.000 (0.032)	0.032	0.001 (0.034)	0.034	0.017 (0.095)	0.096
π_2	0.009 (0.039)	0.039	0.002 (0.034)	0.034	0.003 (0.039)	0.039	0.031 (0.112)	0.116
π_3	0.004 (0.038)	0.038	0.001 (0.031)	0.031	0.001 (0.037)	0.037	0.048 (0.113)	0.123
σ_1^2	0.001 (0.009)	0.009	0.001 (0.005)	0.005	0.001 (0.006)	0.006	0.044 (0.044)	0.062
σ_2^2	0.001 (0.014)	0.014	0.001 (0.012)	0.012	0.000 (0.014)	0.014	0.010 (0.054)	0.055
σ_3^2	0.006 (0.030)	0.031	0.001 (0.011)	0.011	0.002 (0.013)	0.013	0.070 (0.052)	0.087
λ_{11}	0.030 (0.145)	0.148	0.008 (0.121)	0.121	0.009 (0.125)	0.125	0.526 (0.648)	0.835
λ_{12}	0.011 (0.036)	0.038	0.003 (0.033)	0.033	0.006 (0.034)	0.034	0.059 (0.427)	0.431
λ_{21}	0.037 (0.140)	0.145	0.022 (0.137)	0.139	0.023 (0.137)	0.138	0.301 (0.609)	0.680
λ_{22}	0.001 (0.011)	0.011	0.000 (0.010)	0.010	0.000 (0.011)	0.011	0.146 (0.354)	0.382
λ_{31}	0.038 (0.149)	0.154	0.030 (0.114)	0.118	0.038 (0.139)	0.144	0.143 (0.485)	0.506
λ_{32}	0.017 (0.056)	0.058	0.013 (0.050)	0.052	0.015 (0.054)	0.056	0.164 (0.350)	0.386
$\mu_1(\cdot)$	0.020 (0.130)	0.132	0.007 (0.103)	0.103	0.014 (0.112)	0.113	0.729 (0.448)	0.856
$\mu_2(\cdot)$	0.033 (0.180)	0.183	0.013 (0.143)	0.144	0.019 (0.166)	0.167	0.753 (0.506)	0.907
$\mu_3(\cdot)$	0.049 (0.184)	0.191	0.013 (0.140)	0.141	0.040 (0.163)	0.167	1.300 (0.518)	1.399
$\phi_{11}(\cdot)$	0.020 (0.070)	0.073	0.004 (0.065)	0.065	0.005 (0.067)	0.067	0.534 (0.569)	0.780
$\phi_{12}(\cdot)$	0.013 (0.083)	0.084	0.006 (0.080)	0.080	0.007 (0.081)	0.081	0.421 (0.620)	0.750
$\phi_{21}(\cdot)$	0.037 (0.109)	0.115	0.022 (0.086)	0.089	0.026 (0.092)	0.096	1.262 (0.723)	1.455
$\phi_{22}(\cdot)$	0.014 (0.115)	0.116	0.007 (0.099)	0.099	0.009 (0.100)	0.100	1.147 (0.888)	1.451
$\phi_{31}(\cdot)$	0.035 (0.106)	0.112	0.033 (0.098)	0.103	0.035 (0.100)	0.106	1.154 (0.551)	1.279
$\phi_{32}(\cdot)$	0.023 (0.107)	0.110	0.013 (0.052)	0.054	0.017 (0.082)	0.084	1.135 (0.777)	1.376
$\Sigma_1(\cdot,\cdot)$	0.061 (0.267)	0.274	0.013 (0.146)	0.146	0.014 (0.149)	0.150	0.748 (0.784)	1.083
$\Sigma_2(\cdot,\cdot)$	0.072 (0.271)	0.280	0.061 (0.236)	0.244	0.063 (0.269)	0.276	1.130 (0.837)	1.406
$\Sigma_3(\cdot,\cdot)$	0.073 (0.293)	0.302	0.063 (0.260)	0.267	0.066 (0.268)	0.276	1.088 (1.087)	1.538
♯cluster	0.070 (0.354)	0.361	–	–	0.056 (0.212)	0.219	–	-

[a]*Note*. "–" not available.

$\sqrt{2}\cos(2\pi t)$, $\phi_{32}(t) = \sqrt{2}\sin(\pi t)$, $\xi_{\text{gik}} \sim N(0, \lambda_{gk})$ with $\lambda_{11} = 1, \lambda_{12} = 0.25$, $\lambda_{21} = 1.1, \lambda_{22} = 0.2$, $\lambda_{31} = 0.9, \lambda_{32} = 0.15$, and $\varepsilon_{\text{git}} \sim N(0, \sigma_g^2)$ with $\sigma_1^2 = 0.1, \sigma_2^2 = 0.15, \sigma_3^2 = 0.2$. The errors ε_{git} for $g = 1, 2, 3$ and $i = 1, \ldots n$ are independently and identically distributed over time t. For each subject, the number of observations is randomly drawn from a discrete uniform distribution on $\{8, 9, 10, 11, 12\}$ and the observation times are sampled from $U(0, 1)$. We take $H(y) = 3\log(y)$ and $10(\sqrt{y} - 1)$ for Cases 1 and 2, respectively.

Based on 300 repetitions, Table 1 and Table 6 (in the Supporting Information) summarize the results obtained by using the proposed method with $C = 7$, the CT method with $C = 3, 7$, and the WoT method with $C = 3$ for Cases 1 and 2. We take $K_g = 2$, $M_n = 2$, and $\lambda = 0.05$ for all of the three methods. We implement the CT and the WoT estimators by using the proposed algorithm with the transformation function H specified by the true transformation function and the misspecified transformation $H(y) = y$, respectively. The CT method with $C = 3$ is served as the gold standard, the CT method with $C = 7$ is used to investigate the effect of the selection of the number of clusters, and the WoT method with $C = 3$ is used to investigate the effect of misspecification of the transformation function. In Table 1 and Table 6, we also present ♯cluster, the estimated number of clusters.

Tables 1 and 6 reveal the WoT method produces large biases and variances, with biases often overwhelming variances, suggesting that the misspecification of the transformation function leads to severely biased and unstable estimates. In contrast, the proposed method is unbiased with a variance close to that of the CT estimator, suggesting that our method is robust with little loss of efficiency. Moreover, the proposed method consistently selects the number of clusters.

Figures 3 and 4 (both in the Supporting Information) display the averaged estimates of the transformation functions, mean functions, and eigenfunctions, along with their empirical 95% pointwise confidence intervals based on the 300 simulated datasets. In all the cases, the

average estimates of the functions match well with the true functions, with confidence intervals of reasonable width.

To investigate the sensitivity of the proposed procedure to the initial number of clusters, we compare the results with the initial values of $C = 7, 10, 15$ for Case 1. Figure 5 in the Supporting Information illustrates the estimation error $\text{Err}(f) = [\frac{1}{n_{\text{grid}}} \sum_{i=1}^{n_{\text{grid}}} \{\widehat{f}(t_i) - f(t_i)\}^2]^{1/2}$ over 300 simulated datasets. Figure 5 shows that the proposed estimates are nearly identical with different initial numbers of clusters, supporting the conjecture of robustness. It further hints that our method may be as efficient as the method with a known number of clusters, which is the oracle property.

Figures 6(a)-(d) in the Supporting Information report the performance of criterion (3.21) in selecting K_g, M_n, and λ under Case 1 of Simulation 1. The candidates of K_g, M_n, and λ are {1, 2, 3, 4}, {1, 2, 3, 4, 5}, and {0.01, 0.03, 0.05, 0.07}, respectively. The largest BIC is achieved when $K_g = 2$, which is the true value. These figures reveal that the optimal K_g, M_n, and λ are nearly independent, suggesting K_g, M_n, and λ can be separately chosen. Furthermore, we can see the proposed method is not sensitive to the choice of M_n. Figure 6(e) in the Supporting Information, the barplot of estimated number of clusters based on the strategy and BIC criterion (3.21) for Case 1 of Simulation 1, shows that the proposed method can correctly identify the number of clusters.

Simulation 2. To match the real data of the ADNI study where the function data are a restricted density, we generate data similarly as Case 1 of Simulation 1, except that $\boldsymbol{\pi} = (0.229, 0.277, 0.494)'$, the transformation function, the mean functions, the eigenfunctions, the eigenvalues, and error variances are taken as the estimators from the real data analysis of the ADNI study in Section 6. Table 7 in the Supporting Information shows that the proposed method is unbiased with a variance close to that of the CT estimator, and has much less RMSE than the WoT method, and can correctly select the number of clusters.

Simulation 3. To assess the sensitivity of our proposed method to the assumption of Gaussian errors, we generate data similarly as Case 1 of Simulation 1, except that we generate ε_{git} from a mixed distribution with each component being the centralized and scaled gamma distribution $\sigma_g \times \{\text{Gamma}(\tau, 1) - \tau\}/\sqrt{\tau}$ with $\sigma_1^2 = 0.1, \sigma_2^2 = 0.15, \sigma_3^2 = 0.2$, which approaches the normal distribution as τ increases. We take $\tau = 1, 5, 10, 100$. Table 8 (in the Supporting Information) presents the bias, sd, and RMSE for the parameters and the nonparametric functions when $C = 7$. When $\tau \geq 10$ and both skewness and excess kurtosis are less than 1, the proposed estimators are nearly unbiased. When both skewness and excess kurtosis approximate 1, the proposed estimators are acceptable, although the estimators are moderately biased. Taken altogether, these results suggest the robustness toward the Gaussian error assumption.

5.2 | Performance of classification

To assess the classification accuracy, we compare the proposed method with the FunFEM (Bouveyron *et al.*, 2015) and the DHP method proposed by Delaigle *et al.* (2019), based on the PF and ARI. The larger PF and ARI, the better the clustering. The DHP method requires a specification of C. For fair comparisons, we always take C to be the true number of clusters when required. We perform the DHP method with the Haar basis (DHP_{HA}), the Daubechies DB2 wavelet basis (DHP_{DB}), and the principal component basis (DHP_{PC}). We consider two settings.

Simulation 4. The setting is the same as Setting (a) in Delaigle *et al.* (2019). The data in the gth cluster are generated from $Y_{\text{git}} = \sum_{k=1}^{40} (\lambda_k^{1/2} \xi_{\text{gik}} + \gamma_{kg}) \phi_k(t)$, on a grid of 128 equispaced time points in [0,1], where $\lambda_k = k^{-2}, \phi_k(t) = \sqrt{2}\sin(k\pi t)$, and $\xi_{\text{gik}} \sim N(0,1)$ for $k = 1, \ldots, 40$ and $g = 1, 2$; $(\gamma_{11}, \gamma_{21}, \gamma_{31}, \gamma_{41}, \gamma_{51}, \gamma_{61}) = (0, -0.3, 0.6, -0.3, 0.6, -0.3)$, $(\gamma_{12}, \gamma_{22}, \gamma_{32}, \gamma_{42}, \gamma_{52}, \gamma_{62}) = (0, -0.45, 0.45, -0.09, 0.84, 0.6)$, and $\gamma_{kg} = 0$ for $g = 1, 2$ and $k > 6$. We generate 100 replications, each with sample size $n = 200$, where half of the data come from the first cluster and the remaining half come from the second cluster.

Simulation 5. The setting is the same as Bouveyron *et al.* (2015) except that we consider two clusters. A total of $n = 100$ curves with equal mixing proportions are generated from $Y_{\text{git}} = U_i + (1 - U_i) h_g(t) + \varepsilon_{i(t)}$ on a grid of 101 equispaced time points in [1,21] for $g = 1, 2$, where U_i is uniformly distributed on [0,1], $\varepsilon_{i(t)} \sim N(0, 0.5)$, $h_1(t) = 6 - |t - 7|$ and $h_2(t) = 6 - |t - 15|$. We generate 100 replications.

We take $(C, K_g, M_n, \lambda) = (7, 6, 6, 0.08)$ and $(C, K_g, M_n, \lambda) = (7, 2, 6, 0.15)$ for Simulations 4 and 5, respectively, for the proposed method. Table 2 displays the averages of PF and ARI using the proposed method, the FunFEM, and the DHP methods for Simulations 1, 4, and 5. Table 2 presents that the proposed method yields larger PF and ARI than the FunFEM and DHP methods when the assumptions required by our method are satisfied, and produces comparable or slightly better results than FunFEM and DHP when the assumptions follow those specified in Bouveyron *et al.* (2015) and Delaigle *et al.* (2019), respectively.

TABLE 2 The averages of PF and ARI for Simulations 1, 4, and 5

	Proposed	DHP_{HA}	DHP_{DB}	DHP_{PC}	FunFEM
	Case 1 of Simulation 1				
PF	0.924	0.599	0.522	0.587	0.557
ARI	0.812	0.242	0.118	0.227	0.177
	Case 2 of Simulation 1				
PF	0.937	0.622	0.528	0.628	0.577
ARI	0.836	0.275	0.126	0.277	0.211
	Simulation 4				
PF	0.934	0.850	0.857	0.915	0.548
ARI	0.793	0.661	0.552	0.808	0.009
	Simulation 5				
PF	0.927	0.824	0.860	0.827	0.896
ARI	0.664	0.424	0.532	0.433	0.638

6 | ANALYSIS OF THE ADNI STUDY

Alzheimer's disease is an irreversible and the most common form of dementia, and can result in the loss of thinking, memory, and language skills. It is of substantial interest to unravel the complex brain changes involved in the onset and progression of Alzheimer's disease. The effort helps develop effective therapies targeting specific progression mechanisms in order to stop or prevent the actual underlying cause of the disease. In particular, the volume of hippocampus, which is the brain region that is associated with memory loss and disorientation, has been found to be associated with the cognitive function. We explore using the volume of hippocampus to distinguish patients with different levels of cognitive impairment. Specifically, we propose to use the density function of the volumes of hippocampus, obtained from various sampling locations, as a basis for grouping patients with cognitive impairment (AD), mild cognitive impairment (MCI, an early stage of AD) and cognitively normal (CN). The dataset includes 768 participants enrolled in ADNI1 (Mueller *et al.*, 2005), the first phase of ADNI study, a large cohort study designed to prevent and treat Alzheimer's disease. Each patient's record consists of the density for each of the observed 501 equispaced sampling volumes, which are in the interval of [−255,255]. Among the 768 patients, 172 subjects are diagnosed with AD, 378 MCI, and 218 CN.

To proceed, denote by $Y(t)$ the density function of the log of the Jacobian volume of the hippocampus (denoted by t), which is to be used as the functional response in the analysis. The density curves for all the subjects are plotted in Figure 1. Figures 7-9 in the Supporting Information for three groups, respectively, display the histogram for Y-values (Y = density) given x-value (x = log Jacobian volume) at 16 points that are uniformly distributed over the

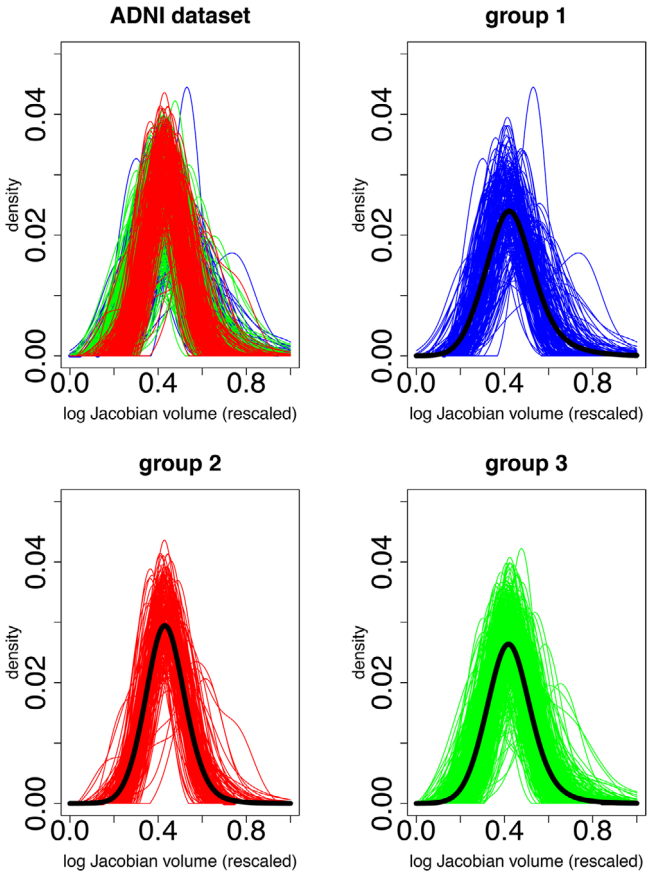

FIGURE 1 ADNI dataset (top left) and separate plots for each group with mean function (Black thick line). This figure appears in color in the electronic version of this article, and any mention of color refers to that version

support of x. It is obvious that the values of density are not normally distributed. We scale the log Jacobian volume into [0,1] before analysis. The density functions are all nonnegative and each integrates to 1, we fit the transformed density functions by the proposed estimation procedure. We conduct unsupervised learning of the data based on model (2.2), without using the known labeling of AD, MCI, or CN. Then we compare the resulting estimators with the known grouping information to examine the clustering performance.

We also compare the proposed method with the untransformed method, that is, WoT with $H(y) = y$. We adopt the cubic B-spline with interior knots chosen by (3.21). To reduce the computational burden, we first select M_n, λ, and a common K_g for all clusters by the BIC criterion (3.21), which are $(M_n, \lambda, K_g) = (2, 0.003, 2)$ and (2,0.002,1) for the proposed method and the WoT, respectively. With the selected (M_n, λ, K_g), both methods identify the number of clusters as 3. Table 3 displays the resulting estimates, their sd, and an ad hoc P-values (based on bootstrap resamples). The sd and P-values are estimated based on 200 bootstrap resamples, where the number of 200 is determined

TABLE 3 The resulting estimates for ADNI data

	WoT		Proposed	
	Estimate (sd)	***P*-value**	**Estimate (sd)**	***P*-value**
π_1	0.321 (0.176)	0.068	0.229 (0.027)	0.000
π_2	0.329 (0.187)	0.079	0.277 (0.023)	0.000
π_3	0.351 (0.180)	0.051	0.494 (0.032)	0.000
σ_1^2	2.031×10^{-5} (2.628×10^{-5})	0.440	0.306 (0.090)	0.001
σ_2^2	1.888×10^{-5} (2.490×10^{-5})	0.448	0.302 (0.088)	0.001
σ_3^2	2.475×10^{-5} (2.778×10^{-5})	0.373	0.018 (0.131)	0.891
λ_{11}	– (–)	–	0.163 (0.075)	0.030
λ_{31}	– (–)	–	0.308 (0.027)	0.000
λ_{32}	– (–)	–	0.160 (0.013)	0.000

[a]*Note*. "–" not available.

TABLE 4 The PF and ARI for ADNI data

	Proposed	DHP_{HA}	DHP_{DB}	DHP_{PC}	**FunFEM**	**WoT**
PF	0.962	0.516	0.451	0.509	0.451	0.569
ARI	0.899	0.042	0.459	0.025	0.034	0.097

by monitoring the stability of the sd. The *P*-value for testing the parameters equalling 0 is obtained as the frequency that the replicates fall in the region (estimator $\pm$ 1.96 $\times$ sd). As the estimates of $\lambda_{12}, \lambda_{21}, \lambda_{22}$ are not significant by using the proposed method, we set $K_1 = 1, K_2 = 0, K_3 = 2$ for our proposed method. Figure 2 displays the estimated mean and transformation functions, eigenfunctions, and their corresponding 95% point-wise confidence limits.

Table 3 reveals that the estimates in the WoT method are all nonsignificant at the level of 0.05, which does not seem reasonable. In contrast, the proposed method yields mostly significant estimates, and Figure 2 shows that the estimated transformation function appears deviating from a linear function, which has a negative intercept causing $\widehat{\mu}_1$ and $\widehat{\mu}_3$ to be negative. Figure 2 shows that the shapes of the mean functions of different clusters are almost same, but the mean function of cluster 2 is higher than those in the remaining two clusters, and cluster 3 is second. The first eigenfunctions of clusters 1 and 3 are nearly the same, but clusters 1, 2, and 3 extracted one, zero and two principal components, respectively. That is, the proposed method can detect different clusters with the ADNI data. With Table 5, we see that group 2 (CN) is cognitively normal and $K_2 = 0$ implies that the variations over volumes and across subjects are simply due to randomness. These findings are consistent with what researchers expect of the three cognitive groups displayed in Figure 1, which shows that the pointwise variance of group 2 is smaller than that of groups 1 and 3 and the mean functions of these 3 groups are almost the same.

To check the clustering performance, we compare PF and ARI among the DHP_{HA}, DHP_{DB}, DHP_{PC}, FunFEM, and WoT ($C = 3$) methods in Table 4. Table 4 reports that the proposed method has larger PF and ARI than the DHP and FunFEM methods, suggesting that the proposed method performs better than the DHP and FunFEM methods in clustering. Furthermore, we estimate the class label for each individual using the Bayes' optimal allocation rule, $g_i = \arg\max_g \frac{f_g\{\mathbf{H}(\mathbf{Y}_i)\}\pi_g}{\sum_{j=1}^{C} f_j\{\mathbf{H}(\mathbf{Y}_i)\}\pi_j}$, and classify 186, 219, and 363 subjects to clusters 1, 2, and 3, respectively, by the proposed method, and 200, 159, and 409 to clusters 1, 2, and 3 by the WoT method. Figure 2 reveals that the mean functions are unimodal for all three groups, with an

TABLE 5 The confusion matrix for the estimated clusters of the ADNI1 patients using the proposed method

		Estimated			
		Cluster 1 (AD)	**Cluster 2 (CN)**	**Cluster 3 (MCI)**	**Total**
True	Cluster 1 (AD)	172	0	0	172
	Cluster 2 (CN)	1	217	0	218
	Cluster 3 (MCI)	13	2	363	378
	All	186	219	363	768

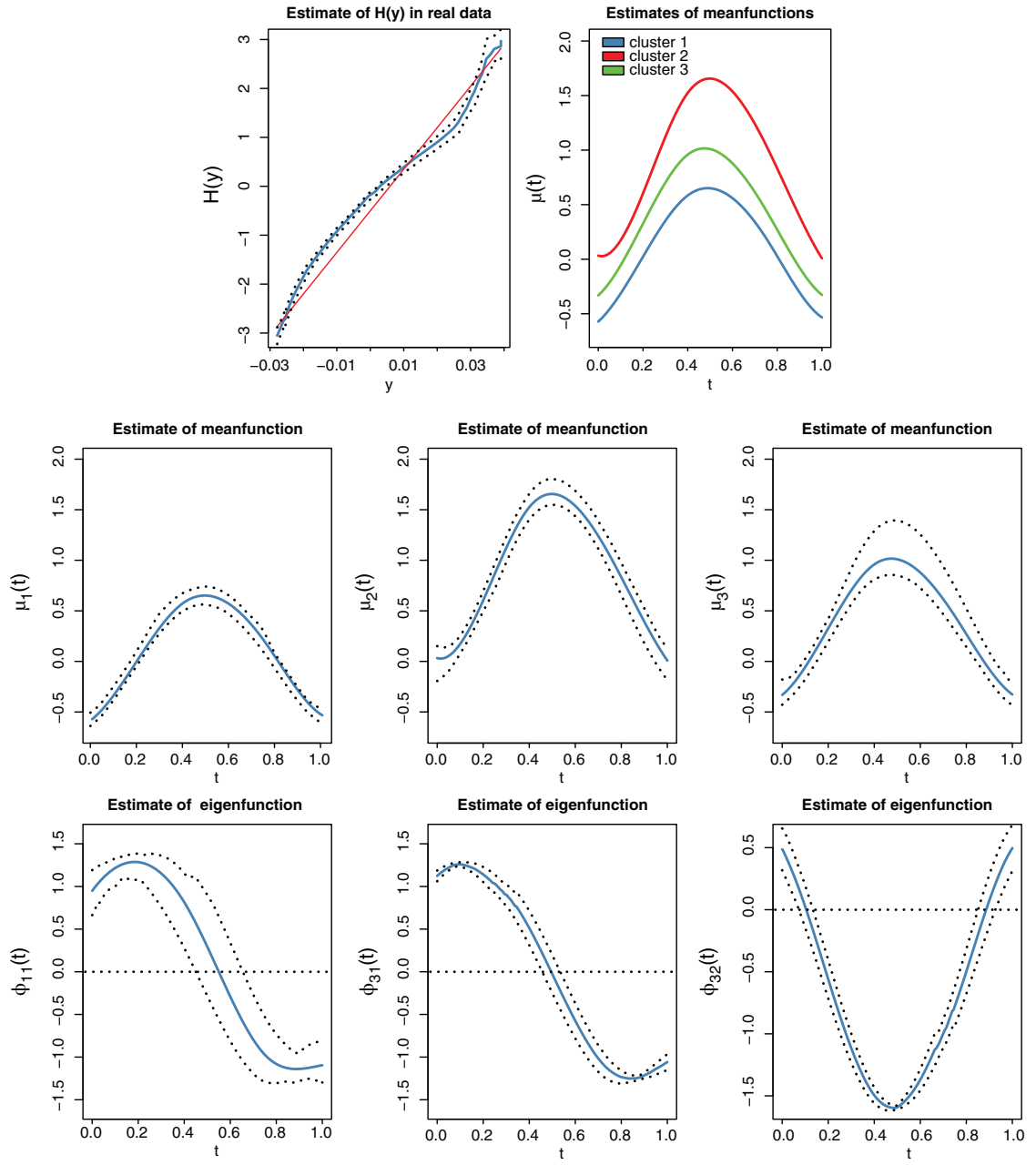

FIGURE 2 The estimated transformation function and mean functions (top), and eigenfunctions (bottom) for ADNI data (dotted—95% confidence limit; solid—estimated function). This figure appears in color in the electronic version of this article, and any mention of color refers to that version

obvious ordering. The mean function of cluster 2 is on top of those for clusters 1 and 3. Cluster 3 comes second, and cluster 1 is the last. As AD patients tend to have low hippocampal volumes, indicating a high level of cognitive impairment, we label clusters 1, 2, and 3 as AD, CN, and MCI, respectively. Table 5 displays the comparisons of the estimated and true clusters. The classification error by applying the WoT method is 43.1%, whereas the classification error by applying our proposed method without the Gaussian assumption is merely 2.08%.

7 | CONCLUSION

We have proposed an SMINT model to cluster non-Gaussian functional data, and used functional principal components for dimension reduction. We have utilized cubic B-spline approximation for eigenfunctions to avoid computing cluster-specific covariance functions, and allowed eigenfunctions to differ across clusters. We have developed an computationally efficient algorithm to estimate the unknown finite- and infinite-dimensional param-

eters. The proposed method has some appealing features: (a) the model is flexible as both the distribution of the response and the number of clusters are unspecified, and (b) the estimates are robust, efficient, consistent, and asymptotically normal.

Although focused on univariate functional data, our method can be extended to accommodate multivariate functional data. It is also possible to extend our method to accommodate external covariates. However, we envision that the theory and implementation may be more complicated, which warrants further study.

ACKNOWLEDGMENTS

We thank the editor, the AE, and two anonymous referees for their insightful suggestions that have helped substantially improve the manuscript. This research is partially supported by National Natural Science Foundation of China (Nos. 11931014 and 11829101), Fundamental Research Funds for the Central Universities (No. JBK1806002) of China and the National Institutes of Health (R21AG058198). Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

ORCID

Huazhen Lin https://orcid.org/0000-0002-4890-9550
Yi Li https://orcid.org/0000-0003-1720-2760

REFERENCES

Abraham, C., Cornillon, P.-A., Matzner-Løber, E. and Molinari, N. (2003) Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics*, 30, 581–595.

Bauer, D.J. and Curran, P.J. (2003) Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychological Methods*, 8, 338.

Biernacki, C., Celeux, G. and Govaert, G. (2000) Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 719–725.

Bouveyron, C., Côme, E. and Jacques, J. (2015) The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics*, 9, 1726–1760.

Bouveyron, C. and Jacques, J. (2011) Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification*, 5, 281–300.

Cai, T. and Yuan, M. (2010) *Nonparametric covariance function estimation for functional and longitudinal data*. University of Pennsylvania and Georgia Institute of Technology.

Chen, X., Hu, T. and Sun, J. (2017) Sieve maximum likelihood estimation for the proportional hazards model under informative censoring. *Computational Statistics & Data Analysis*, 112, 224–234.

Chen, K. and Tong, X. (2010) Varying coefficient transformation models with censored data. *Biometrika*, 97, 969–976.

Chiou, J.-M. and Li, P.-L. (2007) Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 679–699.

Delaigle, A., Hall, P. and Pham, T. (2019) Clustering functional data into groups by using projections. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81, 271–304.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39, 1–22.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. New York, NY: Springer Science & Business Media.

Floriello, D. and Vitelli, V. (2017) Sparse clustering of functional data. *Journal of Multivariate Analysis*, 154, 1–18.

Fröhwirth-Schnatter, S. and Kaufmann, S. (2008) Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*, 26, 78–89.

Hall, P. and Horowitz, J.L. (2007) Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35, 70–91.

Hall, P. and Hosseini-Nasab, M. (2006) On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 109–126.

Hall, P., Müller, H.-G. and Yao, F. (2008) Modelling sparse generalized longitudinal observations with latent Gaussian processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 703–723.

Horowitz, J.L. (1996) Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica*, 64, 103–137.

Huang, T., Peng, H. and Zhang, K. (2017) Model selection for gaussian mixture models. *Statistica Sinica*, 27, 147–169.

Jacques, J. and Preda, C. (2013) Funclust: a curves clustering method using functional random variables density approximation. *Neurocomputing*, 112, 164–171.

Jacques, J. and Preda, C. (2014) Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8, 231–255.

James, G.M., Hastie, T.J. and Sugar, C.A. (2000) Principal component models for sparse functional data. *Biometrika*, 87, 587–602.

James, G.M. and Sugar, C.A. (2003) Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98, 397–408.

Li, Y. and Hsing, T. (2010) Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, 38, 3321–3351.

Lin, Z., Müller, H.-G. and Yao, F. (2018) Mixture inner product spaces and their application to functional data analysis. *The Annals of Statistics*, 46, 370–400.

Lin, H., Zhou, X.-H. and Li, G. (2012) A direct semiparametric receiver operating characteristic curve regression with unknown link and baseline functions. *Statistica Sinica*, 22, 1427–1456.

Liu, X. and Yang, M.C. (2009) Simultaneous curve registration and clustering for functional data. *Computational Statistics & Data Analysis*, 53, 1361–1376.

Liu, J.S., Zhang, J.L., Palumbo, M.J. and Lawrence, C.E. (2003) Bayesian clustering with variable and transformation selections. *Bayesian Statistics*, 7, 249–275.

Ma, L., Hu, T. and Sun, J. (2015) Sieve maximum likelihood regression analysis of dependent current status data. *Biometrika*, 102, 731–738.

Mueller, S., Weiner, M., Thal, L., Petersen, R., Jack, C., Jagust, W., Trojanowski, J., Toga, A. and Beckett, L. (2005) The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15, 869–77.

Peng, J. and Müller, H.-G. (2008) Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *The Annals of Applied Statistics*, 2, 1056–1077.

Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*. Berlin: Springer.

Rivera-García, D., García-Escudero, L.A., Mayo-Iscar, A. and Ortega, J. (2019) Robust clustering for functional data based on trimming and constraints. *Advances in Data Analysis and Classification*, 13, 201–225.

Schumaker, L. (2007). *Spline Functions: Basic Theory*. Cambridge: Cambridge University Press.

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.

Serban, N. and Jiang, H. (2012) Multilevel functional clustering analysis. *Biometrics*, 68, 805–814.

Stone, C.J. (1980) Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 1348–1360.

Suyundykov, R., Puechmorel, S. and Ferré, L. (2010) Multivariate functional data clusterization by PCA in Sobolev space using wavelets. *42èmes Journées de Statistique*.

Tarpey, T. and Kinateder, K.K. (2003) Clustering functional data. *Journal of Classification*, 20, 93–114.

Tokushige, S., Yadohisa, H. and Inada, K. (2007) Crisp and fuzzy k-means clustering algorithms for multivariate functional data. *Computational Statistics*, 22, 1–16.

Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2016) Functional data analysis. *Annual Review of Statistics and its Application*, 3, 257–295.

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

Yao, F., Müller, H.-G. and Wang, J.-L. (2005) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100, 577–590.

Zhou, X.-H., Lin, H. and Johnson, E. (2008) Non-parametric heteroscedastic transformation regression models for skewed data with an application to health care costs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 1029–1047.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures, the implemented R code referenced in Sections 4-6 are available with this paper at the Biometrics website on Wiley Online Library. An R code implementing the proposed methods also is available at https://github.com/zhzhqz/SMINT.

How to cite this article: Zhong Q, Lin H, Li Y. Cluster non-Gaussian functional data. *Biometrics*. 2021;77:852–865. https://doi.org/10.1111/biom.13349