The Canadian Journal of Statistics
Vol. 50, No. 1, 2022, Pages 221–240
La revue canadienne de statistique

Cluster analysis with regression of non-Gaussian functional data on covariates

Jiakun JIANG[1,2], Huazhen LIN[2*], Heng PENG[3], Gang-Zhi FAN[4], and Yi LI[5]

[1]*Center for Statistics and Data Science, Beijing Normal University at Zhuhai, Zhuhai, China*
[2]*Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu, Sichuan, China*
[3]*Department of Mathematics, Hong Kong Baptist University, Hong Kong, China*
[4]*School of Management, Guangzhou University, Guangzhou, China*
[5]*Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.*

Key words and phrases: Cluster analysis; functional data; longitudinal data; semiparametric transformation functional regression; supervised learning.

MSC 2020: Primary 62H30; secondary 62R10.

Abstract: Cluster analysis with functional data often imposes normality assumptions on outcomes and is typically carried out without covariates or supervision. However, nonnormal functional data are frequently encountered in practice, and unsupervised learning, without directly tying covariates to clusters, often makes the resulting clusters less interpretable. To address these issues, we propose a new semiparametric transformation functional regression model, which enables us to cluster nonnormal functional data in the presence of covariates. Our model incorporates several unique features. First, it omits the normality assumptions on the functional response, which adds more flexibility to the modelling. Second, our model allows clusters to have distinct relationships between functional responses and covariates, and thus makes the clusters formed more interpretable. Third, unlike various competing methods, we allow the number of clusters to be unspecified and data-driven. We develop a new method, which combines penalized likelihood and estimating equations, to estimate the number of clusters, regression parameters, and transformation functions simultaneously; we also establish the large-sample properties such as consistency and asymptotic normality. Simulations confirm the utility of our proposed approach. We use our proposed method to analyze Chinese housing market data and garner some interesting findings. *The Canadian Journal of Statistics* 50: 221–240; 2022 © 2021 Statistical Society of Canada

Résumé: En cas de données fonctionnelles, l'analyse par grappes est souvent réalisée sous l'hypothèse de normalité et se fait généralement sans tenir compte de covariables et sans supervision. Mais en pratique, comme il est fréquent que les données fonctionnelles à l'étude ne soient pas gaussiennes, le recours à un apprentissage non supervisé sans un lien direct entre les covariables et les clusters fournit des résultats difficiles à interpréter. Pour remédier à ces problèmes, les auteurs du présent travail proposent un nouveau modèle de régression fonctionnelle de transformation semi-paramétrique (STFR) qui permet de regrouper des données fonctionnelles non normales en présence de covariables. Le modèle proposé intègre plusieurs caractéristiques particulières. Premièrement, en omettant l'hypothèse de normalité de la variable réponse fonctionnelle, il rend la modélisation bien plus flexible. Deuxièmement, en permettant aux relations entre les variables réponses fonctionnelles et les covariables de varier d'un cluster à l'autre, il facilite l'interprétation des clusters construits. Troisièmement, contrairement à diverses méthodes concurrentes, l'approche proposée ne fixe pas le nombre de clusters à l'avance mais adopte davantage un

Additional Supporting Information may be found in the online version of this article at the publisher's website.

* *Corresponding author: linhz@swufe.edu.cn*

choix automatique. La méthode ainsi développée combine les équations d'estimation et la vraisemblance pénalisée pour estimer simultanément le nombre de clusters, les paramètres de régression et les fonctions de transformation. Enfin, en plus d'une étude du comportement asymptotique des estimateurs proposés, dont la convergence et la normalité asymptotiques, les auteurs présentent des simulations et une analyse de données du marché immobilier chinois afin de confirmer les bonnes performances et l'utilité pratique de la méthode proposée. *La revue canadienne de statistique* 50: 221–240; 2022 © 2021 Société statistique du Canada

1. INTRODUCTION

Functional data have been routinely collected in many fields, such as economics, pharmacy, biology, and climatology (Yao, Müller & Wang, 2005a, 2005b; Yao, 2007; Li & Hsing, 2010a, 2010b; Li, Wang & Carroll, 2010; Yao & Müller, 2010; Yao, Fu & Lee, 2010; Horváth & Kokoszka, 2012) and analyses of these data may provide valuable insights for decision makers in these fields. For example, the past two decades have witnessed the skyrocketing housing prices in most cities in China, while the housing markets in a small number of cities have been relatively steady. The sharp market inequality has intrigued scholars and investors (e.g., Zhang et al., 2017; Jia, Wang & Fan 2018), and has sparked interest in understanding how this inequality aligns with the local economy, geography, and demographics, and also which markets are at risk of a "real estate bubble" and which are deemed "healthy." Hence it is crucial to study the change trends in housing prices across cities and to identify the patterns along with local economic conditions and demographic features. Of particular interest is the detection of various types of relationships between these trends and the corresponding macroeconomic factors and the classification of cities or markets accordingly. The results may help identify cities with overheated housing markets.

On the surface, the problem seems to fall into the traditional cluster analysis of functional data, for which various methods are available. For example, the *two-step method* converts the infinite-dimensional clustering problem into a finite-dimensional one and then uses finite-dimensional clustering methods (Abraham et al., 2003; James & Sugar, 2003; Ray & Mallick, 2006; Chiou & Li, 2007; Peng & Müller, 2008; Bouveyron & Jacques, 2011; Samé et al., 2011; Giacofci et al., 2013, Jacques & Preda, 2013, 2014b); *distance-based clustering*, including those of Tarpey & Kinateder (2003), Ferraty & Vieu (2006), Cuesta-Albertos & Fraiman (2007), Tokushige, Yadohisa & Inada (2007), and Ieva et al. (2013), involves clustering that uses the curve data directly; see Jacques & Preda (2014a) for a comprehensive review. However, none of these approaches can classify functional data while accounting for covariates.

Limited research has been published concerning cluster analysis on functional responses with covariates. For example, Titterington, Smith & Makov (1985), Muthén (2001), and McLachlan & Peel (2004) proposed a growth mixture model (GMM), and Nagin (1999) and Nagin (2005) suggested group-based trajectory modelling (GBTM) to identify clusters of individuals based on functional responses as well as covariates. Shi & Wang (2008) developed a mixture of Gaussian process functional regression models to classify relationships between curve responses and covariates. However, these approaches require the functional response to follow a Gaussian distribution, which is violated by our motivating data; see Figure 2. Misleading results may occur when such an assumption is violated. For example, Bauer & Curran (2003) showed that, for nonnormal data, multiple groups can be falsely identified when, in fact, there is only one group. Moreover, as our numerical studies revealed, misspecifications of the polynomial growth curves, which were commonly assumed by these models, may lead to unreliable estimation and classification. Finally, all these methods require the number of clusters to be known a priori, whereas detecting the number of clusters is a centerpiece in cluster analysis. To our knowledge, little research has focused on this topic, and most published work has relied on the Bayesian information criterion (BIC)

(Schwarz, 1978; Jones, Nagin & Roeder, 2001; Nagin, 2005; Shi & Wang, 2008; Andruff et al., 2009), which may involve a marked computational burden. Additionally, large-sample results with the BIC are not available, making it difficult to evaluate its validity for model selection.

We propose a semiparametric transformation functional regression (STFR) model for clustering a functional response with covariates. Our model relaxes the restrictive conditions on the response, the growth curves, and the number of clusters. To introduce the idea, we first note that, for a continuous random variable Y with distribution function F, $\Phi^{-1}(F(Y))$ has a standard normal distribution, where Φ is the standard normal distribution function. This indicates the existence of normal transformation functions, at least in the absence of covariates. Now, with covariates X and an unknown transformation function $H(\cdot)$, we denote by $f_k(H(Y)|X)$ the conditional probability density of the transformed responses in the kth cluster, which we assume is a normal density function given the covariates X. Hence, the conditional distribution of $H(Y)$ given X is a normal mixture. In addition, we also assume that we can specify a semiparametric model with an unknown growth curve. Finally, with penalization on the group probability, we identify clusters that best fit the data. We develop a new method, which combines penalized likelihood and estimating equations, to estimate the number of clusters, regression parameters, and transformation functions simultaneously. We also establish this method's large-sample properties such as consistency and asymptotic normality. We apply our STFR model and the competing methods for clustering to analyze a Chinese housing market dataset. Our model provides a better fit than the competing alternatives, and generates some interesting results that may shed light on the determinants of real estate markets in China.

The remainder of the article is organized as follows. Section 2 introduces our proposed model and the associated method of estimation, and provides a BIC-type procedure for selecting the tuning parameters. Section 3 develops the theoretical properties, including $\sqrt{n}$-consistency, asymptotic normality, and the model selection consistency. In Section 4 we report the results of simulation studies and comparisons with alternative competing clustering methods, while Section 5 concerns our analysis of the Chinese housing market data. We conclude the article with a brief discussion of the possibilities for future research. All the technical proofs may be found in the corresponding Supplementary Material.

2. THE MODEL AND ITS ESTIMATION

2.1. Model and Objective Function

Let $\big(Y_i(t), X_i(t)\big), i = 1, \ldots, n$ be independent realizations of $(Y(t), X(t))$, where $t \in \big[t_0, t_1\big]$ with $0 \le t_0 < t_1 < \infty$ being two fixed constants, and $X(t)$ is a p-dimensional covariate vector which includes time-dependent as well as time-independent components. Instead of observing the full trajectories $Y_i(\cdot)$ and $X_{i1}(\cdot), \ldots, X_{ip}(\cdot)$, we measure them at sparse and irregular time points. To adequately describe the subject-specific time points underlying the measurements, we assume there are n_i measurements for $Y_i(\cdot)$ and $X_{i1}(\cdot), \ldots, X_{ip}(\cdot)$ at time points $\mathbf{t}_i = \big(t_{i1}, \ldots, t_{i,n_i}\big)$ from $\big[t_0, t_1\big]$. For notational simplicity and without loss of generality, we hereafter assume this bounded set is $[0, 1]$, and the measurement time t_{ij} is randomly distributed on $[0, 1]$. Let $Y_{ij} = Y_i\big(t_{ij}\big)$ and $\mathbf{X}_{ij} = \big(X_{i1}\big(t_{ij}\big), \ldots, X_{ip}\big(t_{ij}\big)\big)', j = 1, \ldots, n_i$; then $\mathbf{Y}_i = \big(Y_{i1}, Y_{i2}, \ldots, Y_{i,n_i}\big)'$ and $\mathbf{X}_i = \big(\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{i,n_i}\big)'$ represent the sequences of measurements on individual i over n_i time points. For a nonrandom function H, let $f(H(\mathbf{Y}_i)|\mathbf{X}_i, \mathbf{t}_i)$ denote the conditional probability density of the transformed responses $H\big(\mathbf{Y}_i\big) = \big(H\big(Y_{i1}\big), H\big(Y_{i2}\big), \ldots, H\big(Y_{i,n_i}\big)\big)'$ given time-dependent covariates $\mathbf{X}_i$ and time points $\mathbf{t}_i$. We assume that $H(\cdot)$ is chosen such that $H(\mathbf{Y}_i)$ given $\mathbf{X}_i, \mathbf{t}_i$ follows a mixture of K normal densities, i.e.,

$$f\big(H\big(\mathbf{Y}_i\big)|\mathbf{X}_i, \mathbf{t}_i\big) = \sum_{k=1}^{K} \pi_k f_k\big(H(\mathbf{Y}_i)|\mathbf{X}_i, \mathbf{t}_i\big), \quad \sum_{k=1}^{K} \pi_k = 1 \tag{1}$$

with $\pi_k \geq 0$, where π_k is the marginal probability that an individual belongs to cluster k and $f_k(H(\mathbf{Y}_i)|\mathbf{X}_i, \mathbf{t}_i)$ is the multivariate normal density function with mean μ_{ik} and covariance matrix $\Delta_i(\boldsymbol{\gamma}_k)$, where $\boldsymbol{\gamma}_k$ is an m-dimensional parameter vector. We assume $\mu_{ik} = g_{ik} + \mathbf{X}_i\boldsymbol{\beta}_k$, where $g_{ik} = \left(g_k\left(t_{i1}\right), g_k\left(t_{i2}\right), \ldots, g_k\left(t_{i,n_i}\right)\right)'$, $g_k(\cdot)$ is an unknown smooth function, $\boldsymbol{\beta}_k$ is a parameter vector with dimension p, and $\mathbf{X}_i$ is a matrix of covariates with dimension $n_i \times p$. We assume the covariance matrix $\Delta_i(\boldsymbol{\gamma}_k)$ follows some parametric structure, but we do not require any particular structure, such as AR(1) or blockwise diagonal, on the covariance matrix. In particular, the covariance matrix can be a linear combination of various covariance matrices with certain structures. The model specified in Equation (1) generalizes the existing models. When $H(x) = x$, $g_k(\cdot)$ is specified, and the true number of clusters is known, our model includes the group-based trajectory modelling proposed by Nagin (1999, 2005) as a special case. In this article, we propose to estimate K, along with $\boldsymbol{\pi} = \left(\pi_1, \ldots, \pi_K\right)'$, the growth curve $g_k(\cdot)$, and the transformation function $H(\cdot)$ simultaneously.

Denote

$$\mathcal{G} = \{g(\cdot) : |g^{(q_1)}(t_1) - g^{(q_1)}(t_2)| \leq c_0|t_1 - t_2|^{q_2}, \text{ for any } 0 \leq t_1, t_2 \leq 1\}, \tag{2}$$

where q_1 is a nonnegative integer, $q_2 \in (0, 1]$, $r = q_1 + q_2 \geq 2$, and $c_0 > 0$ is a constant. The smoothness assumption identified in Equation (2) is often used in nonparametric curve estimation. With the assumption $g_k \in \mathcal{G}$ for $k = 1, \ldots, K$, we approximate $g_k(\cdot)$ by $g_{nk}(t) = \boldsymbol{\alpha}_k' B_n(t)$ for $k = 1, \ldots, K$, where $B_n(\cdot) = \left\{b_1(\cdot), \ldots, b_{q_n}(\cdot)\right\}'$ is a set of B-spline basis functions of order $r + 1$ with knots $0 = t_0 < t_1 < \cdots < t_{M_n} = 1$, satisfying $\max\left(t_j - t_{j-1} : j = 1, \ldots, M_n\right) = O(n^{-v})$. Here, $q_n = M_n + r + 1$, and M_n is the integer part of n^v with $0 < v < 0.5$; see Schumaker (2007). The resulting model for cluster k has the form

$$H\left(\mathbf{Y}_i\right)|\mathbf{X}_i, \mathbf{t}_i \sim N(\mathbf{B}_n\left(\mathbf{t}_i\right)\boldsymbol{\alpha}_k + \mathbf{X}_i\boldsymbol{\beta}_k, \Delta_i\left(\boldsymbol{\gamma}_k\right)), \tag{3}$$

where $\mathbf{B}_n\left(\mathbf{t}_i\right) = \left(B_n\left(t_{i1}\right), \ldots, B_n\left(t_{i,n_i}\right)\right)'$. We denote $\mu_{nik} = \mathbf{B}_n\left(\mathbf{t}_i\right)\boldsymbol{\alpha}_k + \mathbf{X}_i\boldsymbol{\beta}_k$ and

$$f_{nk}\left(H\left(\mathbf{Y}_i\right)|\mathbf{X}_i, \mathbf{t}_i\right) = (2\pi)^{-n_i/2}|\Delta_i\left(\boldsymbol{\gamma}_k\right)|^{-1/2} \times \exp\left[-\frac{1}{2}\left\{H\left(\mathbf{Y}_i\right) - \mu_{nik}\right\}'\Delta_i\left(\boldsymbol{\gamma}_k\right)^{-1}\left\{H\left(\mathbf{Y}_i\right) - \mu_{nik}\right\}\right].$$

We can base our inference on the logarithmic likelihood function

$$L_n\left(\boldsymbol{\theta}_n; H\right) = \sum_{i=1}^{n} \log\left\{\sum_{k=1}^{K} \pi_k f_{nk}\left(H\left(\mathbf{Y}_i\right)|\mathbf{X}_i, \mathbf{t}_i\right)\right\}, \tag{4}$$

with $\boldsymbol{\theta}_n = \left\{\boldsymbol{\beta}_k, \boldsymbol{\alpha}_k, \boldsymbol{\gamma}_k, \pi_k, k = 1, \ldots, K\right\}$ and $\sum_{k=1}^{K} \pi_k = 1$.

Since the number of clusters is unknown, we begin with a bigger model that has the number of clusters $K \geq K_0$ with K_0 being the true number of clusters. This implies that some clusters are redundant or can be merged. As $\pi_k = 0$ indicates that the kth cluster is not necessary and can be deleted from the model, cluster detection corresponds to the selection of nonzero $\{\pi_k, k = 1, \ldots, K\}$, which, however, cannot be achieved by directly penalizing $\left(\pi_k, k = 1, \ldots, K\right)'$. To see that, we denote $\delta_{ik} = 1$ if $\mathbf{Y}_i$ arises from the kth cluster, and $\delta_{ik} = 0$ otherwise, and denote $\delta_i = \left(\delta_{i1}, \ldots, \delta_{iK}\right)'$; then the complete

data for individual i is $D_i = \{\mathbf{Y}_i, \delta_i, \mathbf{X}_i\}$. The expected complete-data log-likelihood function is

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\left(b_{ik}\left[\log\left(\pi_k\right) + \log\left\{f_{nk}\left(H\left(\mathbf{Y}_i\right)|\mathbf{X}_i, \mathbf{t}_i\right)\right\}\right]\right), \tag{5}$$

where $b_{ik} = E\{\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i\}$. Since Equation (5) contains $\log(\pi_k)$, which has a gradient that increases quickly when π_k is close to zero, the L_p-type penalties cannot directly set small values of π_k to zero and and hence we need to consider imposing a penalty on $\log\{\pi_k\}$ for sparsity. Following Huang, Peng & Zhang (2017), we consider the penalized log-likelihood function

$$Q_n\left(\boldsymbol{\theta}_n; H\right) = L_n\left(\boldsymbol{\theta}_n; H\right) - n\lambda\sum_{k=1}^{K}\log\left\{\frac{\epsilon + \pi_k}{\epsilon}\right\}, \tag{6}$$

where $\epsilon > 0$ is small, say 10^{-6} or $o\{n^{-1/2}(\log n)^{-1}\}$ (Huang, Peng & Zhang, 2017). With this, we can show that there is a positive probability of some estimated values of π_k equalling zero exactly, resulting in the estimation of the number of clusters.

Since $Q_n\left(\boldsymbol{\theta}_n; H\right)$ involves the infinite-dimensional function $H(\cdot)$ and hence a direct maximization is infeasible, we resort to a two-stage approach. We first use a series of estimating equations to estimate the transformation function H. We then estimate $\boldsymbol{\theta}_n$ by maximizing $Q_n\left(\boldsymbol{\theta}_n; H\right)$ with H replaced by its current estimate. We repeat the procedure until convergence or the number of iterations exceeds 100. This iterative estimator may converge to a local minimizer since the objective function is nonconvex. These local minimizers may differ from each other. Multiple initial values are recommended so that the optimum value can be identified. Since the choice of initial values plays a vital role in nonconvex optimization, we have designed a strategy for selecting the initial values. Specifically, by noting that H is monotonic increasing, we use the Box–Cox transformation (with a ρ) to select an initial value for $H(\cdot)$. With a fixed large K, a specified H, and a B-spline approximation of the nonparametric mean function $g_k(t)$, fitting this semiparametric model has been reduced to a linear regression problem without the penalty term on π_k, enabling us to apply a common mixture regression statistical software, such as the R package *flexmix* (Leisch, 2004). In this way, we can easily obtain the initial values as desired. For the choice of ρ, we take a value that maximizes the resulting log-likelihood function. Our simulation studies suggest that this strategy will work well in practice.

2.2. A Penalized EM Algorithm for $\boldsymbol{\theta}_n$ Given H

To estimate $\boldsymbol{\theta}_n$ given H, we propose to use a penalized expectation–maximization (EM) algorithm (Dempster, Laird & Rubin, 1977). Since the complete data for individual i is $D_i = \{\mathbf{Y}_i, \delta_i, \mathbf{X}_i\}$, the penalized complete-data log-likelihood function is

$$Q_c\left(\boldsymbol{\theta}_n; H\right) = \log\mathcal{L}_c\left(\boldsymbol{\theta}_n; H\right) - n\lambda\sum_{k=1}^{K}\log\left\{\frac{\epsilon + \pi_k}{\epsilon}\right\}, \tag{7}$$

where

$$\log\mathcal{L}_c\left(\boldsymbol{\theta}_n; H\right) \propto \sum_{i=1}^{n}\sum_{k=1}^{K}\left(\delta_{ik}\left[\log\left(\pi_k\right) + \log\left\{f_{nk}\left(H\left(\mathbf{Y}_i\right)|\mathbf{X}_i, \mathbf{t}_i\right)\right\}\right]\right). \tag{8}$$

We estimate $\boldsymbol{\theta}_n$ by maximizing $E\{\mathcal{Q}_c(\boldsymbol{\theta}_n; H)|\mathbf{Y}_i, \mathbf{X}_i, , \mathbf{t}_i, i = 1, \ldots, n\}$ with respect to $\boldsymbol{\theta}_n$. To proceed, we differentiate $E\{\mathcal{Q}_c(\boldsymbol{\theta}_n; H)|\mathbf{Y}_i, \mathbf{X}_i, , \mathbf{t}_i, i = 1, \ldots, n\}$ with respect to $\boldsymbol{\theta}_n$ and set the derivatives to zero, thereby obtaining the following estimation equations:

$$\sum_{i=1}^{n} \frac{E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i)}{\pi_k} - \sum_{i=1}^{n} \frac{E(\delta_{i1}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i)}{1 - \sum_{j=2}^{K} \pi_j} + \frac{n\lambda}{\epsilon + 1 - \sum_{j=2}^{K} \pi_j} - \frac{n\lambda}{\epsilon + \pi_k} = 0,$$

$$\text{for } k = 2, \ldots, K, \tag{9}$$

$$\sum_{i=1}^{n} E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i) \left\{ \text{tr}\left(\Delta_i(\boldsymbol{\gamma}_k)^{-1} \frac{\partial \Delta_i(\boldsymbol{\gamma}_k)}{\partial \gamma_{kj}} \Delta_i(\boldsymbol{\gamma}_k)^{-1} \left[\Delta_i(\boldsymbol{\gamma}_k) - \{H(\mathbf{Y}_i) - \mu_{nik}\}^{\otimes 2} \right] \right) \right\} = 0,$$

$$\text{for } k = 1, \ldots, K, \ j = 1, \ldots, m, \tag{10}$$

$$\begin{aligned} \boldsymbol{\alpha}_k = & \left\{ \sum_{i=1}^{n} E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i) \mathbf{B}_n(\mathbf{t}_i)' \Delta_i(\boldsymbol{\gamma}_k)^{-1} \mathbf{B}_n(\mathbf{t}_i) \right\}^{-1} \\ & \times \sum_{i=1}^{n} E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i) \mathbf{B}_n(\mathbf{t}_i)' \Delta_i(\boldsymbol{\gamma}_k)^{-1} \{H(\mathbf{Y}_i) - \mathbf{X}_i \boldsymbol{\beta}_k\}, \end{aligned} \tag{11}$$

$$\begin{aligned} \boldsymbol{\beta}_k = & \left\{ \sum_{i=1}^{n} E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i) \mathbf{X}_i' \Delta_i(\boldsymbol{\gamma}_k)^{-1} \mathbf{X}_i \right\}^{-1} \\ & \times \sum_{i=1}^{n} E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i) \mathbf{X}_i' \Delta_i(\boldsymbol{\gamma}_k)^{-1} \{H(\mathbf{Y}_i) - \mathbf{B}_n(\mathbf{t}_i) \boldsymbol{\alpha}_k\}, \end{aligned} \tag{12}$$

with $\sum_{k=1}^{K} \|\boldsymbol{\alpha}_k\| = c_0$ for identifiability; γ_{kj} is the jth component of $\boldsymbol{\gamma}_k$, and $a^{\otimes 2} = aa'$. Given that ϵ is so small that $\frac{1}{\pi_j + \epsilon} \approx \frac{1}{\pi_j}$ for any π_j, we arrive at an approximating solution of the estimating equations identified in Equation (9), namely

$$\widehat{\pi}_k = \max\left\{ 0, \frac{1}{1 - K\lambda} \left[\frac{1}{n} \sum_{i=1}^{n} E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i) - \lambda \right] \right\}. \tag{13}$$

Some $\widehat{\pi}_k$ may be shrunk to zero and the constraint $\sum_{k=1}^{K} \widehat{\pi}_k = 1$ may not be satisfied. However, this result neither decreases the likelihood function nor affects the estimate of the posterior probability $E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i)$ in the E-step or the update of π_k in the M-step. In this particular case, we normalize $\widehat{\pi}_k$ by enforcing $\sum_{k=1}^{K} \widehat{\pi}_k = 1$ after the EM algorithm converges. Then, we estimate $\boldsymbol{\theta}_n$ by repeatedly using Equations (10)–(13) until $\boldsymbol{\theta}_n$ converges. For each step, the values of π_k, $\boldsymbol{\alpha}_k$, and $\boldsymbol{\beta}_k$ on the left-hand side of the equations are replaced by the iterative values from the previous step, and $\boldsymbol{\gamma}_k$ is estimated by Newton–Raphson iteration using Equation (10). To estimate $\boldsymbol{\theta}_n$, we compute

$$E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i) = \frac{f_{nk}(H(\mathbf{Y}_i)|\mathbf{X}_i, \mathbf{t}_i)\pi_k}{\sum_{j=1}^{K} f_{nj}(H(\mathbf{Y}_i)|\mathbf{X}_i, \mathbf{t}_i)\pi_j}. \tag{14}$$

At the rth step, $E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i)$ is estimated by the left-hand side of Equation (14), with the unknown parameters and functions replaced by the estimators from the previous iteration.

2.3. Estimation of H Given θ_n

For any given y, we have

$$\begin{aligned} Pr\left(Y_{ij} \le y|\mathbf{X}_i, \mathbf{t}_i\right) &= Pr\left(H\left(Y_{ij}\right) \le H(y)|\mathbf{X}_i, \mathbf{t}_i\right) \\ &= \sum_{k=1}^{K} \pi_k Pr\left(H\left(Y_{ij}\right) \le H(y)|\delta_{ik} = 1, \mathbf{X}_i, \mathbf{t}_i\right) \\ &= \sum_{k=1}^{K} \pi_k \Phi\left\{\frac{H(y) - \boldsymbol{\alpha}_k' B_n\left(t_{ij}\right) - \mathbf{X}_{ij}'\boldsymbol{\beta}_k}{\sqrt{\sigma_{kj}}}\right\}, \end{aligned}$$

where σ_{kj} denotes the element (j, j) of $\Delta_i(\boldsymbol{\gamma}_k)$. We estimate $H(y)$ by solving

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[I\left(Y_{ij} \le y\right) - \sum_{k=1}^{K}\pi_k\Phi\left\{\frac{H(y) - \boldsymbol{\alpha}_k' B_n\left(t_{ij}\right) - \mathbf{X}_{ij}'\boldsymbol{\beta}_k}{\sqrt{\sigma_{kj}}}\right\}\right] = 0, \tag{15}$$

for any given y in the support of Y_{ij}. Specifically, let $v_1, \ldots, v_{s_n}$ denote the distinct points of Y_{ij}, $i = 1, 2, \ldots, n, j = 1, 2, \ldots, n_i$. Given $y = v_s, s = 1, \ldots, s_n$, we estimate $H(y)$ by solving the equation

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[I\left(Y_{ij} \le y\right) - \sum_{k=1}^{K}\pi_k\Phi\left\{\frac{\theta - \boldsymbol{\alpha}_k' B_n(t_{ij}) - \mathbf{X}_{ij}'\boldsymbol{\beta}_k}{\sqrt{\sigma_{kj}}}\right\}\right] = 0, \tag{16}$$

for θ. Equation (16) says that the estimator $\widehat{H}(y)$ is a nondecreasing step function with jumps that occur only at the observed values of Y_{ij}. Varying y among $\left\{v_1, \ldots, v_{s_n}\right\}$ and repeating the estimation procedure for each y, we obtain the whole curve estimator of $H(\cdot)$. We use the Newton–Raphson algorithm to solve Equation (16), with a moderate computational cost. Coupled with the closed-form estimator for $\boldsymbol{\theta}_n$ at each step, the implementation of our proposed method is straightforward. Unlike traditional nonparametric approaches (Horowitz, 1996), our approach does not involve nonparametric smoothing or need to select smoothing parameters.

2.4. Selection of the Tuning Parameter λ

Estimating K relates to the selection of the tuning parameter λ and the number of interior knots M_n. Through our simulation studies, we found that our proposed algorithm is not sensitive to the choice of the number of knots, which is consistent with observations in the literature made by other investigators; see Winsberg & Ramsay (1981). For smooth functions, three to six knots seemed adequate, as we later recommend. We consider a BIC-based procedure to select λ, which yields model selection consistency for linear regression models (Wang, Li & Tsai, 2007). Specifically, we choose λ by maximizing

$$\mathrm{BIC}(\lambda) = \log L_n\left(\boldsymbol{\theta}_n; H\right) - \frac{1}{2}DF_\lambda \log\left(\sum_{i=1}^{n} n_i\right), \tag{17}$$

where DF_λ is the generalized degree of freedom, which can be consistently estimated by the number of nonzero parameters; see Zhang, Li & Tsai (2010) for corresponding results involving generalized linear models. In our numerical studies, we selected λ using a grid search, which seemed to work well.

3. LARGE-SAMPLE PROPERTIES

Denote the estimators of $\boldsymbol{\theta}_n$ and H by $\widehat{\boldsymbol{\theta}}_n$ and $\widehat{H}_n$, respectively. Also define $\|f\|_\infty = \sup_t |f(t)|$, $Pf = \int f(x)dP(x)$, and $P_n f = \frac{1}{n}\sum_{i=1}^n f(x_i)$ for any function f, and for $c_0 > 0$

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\gamma}, \mathbf{g}) \in R^{Kp} \otimes [0,1]^K \otimes R^{K\times m} \otimes \mathcal{G}^K, \|\boldsymbol{\beta}\| + \|\boldsymbol{\pi}\| + \|\boldsymbol{\gamma}\| \le c_0\},$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K), \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K), \boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K), \mathbf{g} = (g_1, \ldots, g_K)$, and $\|\cdot\|$ is the Euclidean norm. Furthermore, we define a distance metric

$$d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \Big(\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|^2 + \|\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2\|^2 + \|\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2\|^2 + \sum_{k=1}^K \|g_{k,1} - g_{k,2}\|_2^2\Big)^{1/2},$$

where $\|g_{k,1} - g_{k,2}\|_2^2 = \int_0^1 \{g_{k,1}(t) - g_{k,2}(t)\}^2 dt$. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\pi}_0, \boldsymbol{\gamma}_0, \mathbf{g}_0)$ be the true value of $\boldsymbol{\theta}$, and K_0 be the true number of clusters. Without loss of generality, we suppose the first K_0 components of $\boldsymbol{\pi}_0$ are nonzero with $\sum_{k=1}^K \pi_{k0} = 1$ and $\pi_{10} \ge \pi_{20} \ge \cdots \ge \pi_{K0} > 0$ for identifiability. Theorems 1–3 summarize the large-sample properties under the following regularity conditions; the proofs may be found in the Supplementary Material.

(A1) $\{X(t), t \in (0,1)\}$ is bounded.

(A2) $g_{k0} \in \mathcal{G}, k = 1, \ldots, K_0$ and $\boldsymbol{\theta}_0$ is an interior point of $\boldsymbol{\Theta}$.

(A3) There exists $[\underline{y}, \bar{y}]$ such that $\frac{1}{N}\sum_{i=1}^n \sum_{j=1}^{n_i} I\big(Y_{ij} \notin [\underline{y}, \bar{y}]\big) = o_p\big(n^{-1/2}\big)$, where $N = \sum_{i=1}^n n_i$.

(A4) The transformation function $H(y)$ is strictly increasing with a continuous first derivative over $y \in [\underline{y}, \bar{y}]$, and satisfies a restriction $H(a) = b$ for constants a and $b \neq 0$.

(A5) $\Delta/\delta \le c_0$ uniformly in n, where $\delta = \min_{1\le i\le M_n} |t_i - t_{i-1}|$, $\Delta = \max_{1\le i\le M_n} |t_i - t_{i-1}| = O(n^{-\upsilon})$.

Condition (A1), for mathematical convenience, is commonly used in the nonparametric and semiparametric literature (Fan & Gijbels, 1996; Horowitz, 1996, 2001; Zhang, Li & Xia, 2015). The boundedness assumption for the regressor $X(t)$ is technical to simplify the proofs and may be relaxed to allow bounded high-order moments. (A2) is commonly assumed in the semiparametric literature (Chen & Tong, 2010). Condition (A3) is used to avoid the tail problem, which is also required by Lin, Zhou & Li (2012), while Condition (A4) is a common requirement for the transformation function (Zhou, Lin & Johnson, 2008). Condition (A5) is often assumed for spline analysis (Lu, Zhang & Huang, 2009).

Theorem 1. *Under Conditions (A1)–(A5),* $\lambda\sqrt{n} \to 0$, $\lambda\sqrt{n}\log n \to \infty$, *and* $\epsilon = o\left(\frac{1}{\sqrt{n}\log(n)}\right)$, *the estimated number of components* $\widehat{s}_n \to K_0$ *with probability tending to* 1.

Theorem 2. *Under Conditions (A1)–(A5),* $\lambda\sqrt{n} \to 0$, *and* $\epsilon = o\left(\frac{1}{\sqrt{n}\log(n)}\right)$,

$$\widehat{H}_n(y) \xrightarrow{a.s.} H_0(y) \quad \textit{uniformly over } y \in [\underline{y}, \bar{y}],$$

$$d\big(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0\big) = O_p\big(n^{-\min\left(\frac{1-\upsilon}{2}, r\upsilon\right)}\big),$$

where r *is a smooth parameter defined in Equation (2), and* $0 < \upsilon < 0.5$ *is given for determining the spline basis* $B_n(\cdot)$.

The choice of $\upsilon = 1/2r + 1$ yields the optimal rate of convergence $n^{r\upsilon}$ for the nonparametric function (Stone, 1980).

Theorem 3. *Under Conditions (A1)–(A5) with* $r \geq 2$ *and* $\frac{1}{4r} < v < \frac{1}{2}$, $\sqrt{n}\lambda \to 0$ *and* $\epsilon = o\left(\frac{1}{\sqrt{n}\log(n)}\right)$,

$$\sqrt{n}\left(\widehat{\Upsilon} - \Upsilon_0\right) \to N\left(0, I^{-1}\left(\Upsilon_0\right)\right),$$

where $\Upsilon = \left\{\boldsymbol{\beta}_k, \boldsymbol{\gamma}_k, \pi_k, k = 1, \ldots, K_0\right\}$, Υ_0 *is the true value of* Υ, *and* $I^{-1}\left(\Upsilon_0\right)$ *is defined in the Supplementary Material.*

4. SIMULATION STUDY

Since our proposed method allows the transformation function as well as the distribution of the functional data to be unknown, we investigated whether our approach is more robust than alternative existing parametric or semiparametric procedures that need to specify the distributions of the response curves, and, if so, whether the robustness of our approach comes at the expense of reduced efficiency. We compare our method with the following: (i) the model with correct transformation (CT), where the transformation function is correctly specified and the growth curve is estimated by B-splines, and (ii) the untransformed model (WOT), with the growth curves estimated by B-splines. The CT and WOT methods are used to evaluate the efficiency and robustness of our proposed method, respectively. We also assess the accuracy of cluster selection. Finally, as our method assumes a Gaussian distribution for the transformed responses within each cluster, we investigate the sensitivity of our method to departures from this assumption. We used the criteria of bias, standard error (SE), and root-mean-square error (RMSE), defined by

$$bias = \left[\frac{1}{n_{grid}}\sum_{i=1}^{n_{grid}}\left\{E\widehat{g}\left(t_i\right) - g\left(t_i\right)\right\}^2\right]^{1/2}, \quad \text{SE} = \left[\frac{1}{n_{grid}}\sum_{i=1}^{n_{grid}} E\{\widehat{g}\left(t_i\right) - E\widehat{g}\left(t_i\right)\}^2\right]^{1/2},$$

and $\text{RMSE} = \left[bias^2 + \text{SE}^2\right]^{1/2}$, where t_i $\left(i = 1, \ldots, n_{grid}\right)$ denote the grid points on which $g(\cdot)$ is estimated. For each parameter configuration detailed below, we generated $N = 200, 400$ independent datasets, and used the cubic B-spline approximation with the number of knots $K_n = n^{1/3}$, the knots were placed at the K_n-quantiles of the observation times, and $n_{grid} = 200$. We approximated $E\widehat{g}(t_i)$ by the sample mean based on these N simulated datasets.

Simulation 1. We generated observations from a three-component mixture model with $\pi_1 = \pi_2 = \pi_3 = 1/3$. The data in cluster k were obtained using

$$H\left(Y_i\left(t_{ij}\right)\right) = g_k\left(t_{ij}\right) + X_i\beta_k + \epsilon_i\left(t_{ij}\right),$$

for $k = 1, 2, 3$, where X_i is generated from $U(0, 1)$ with coefficients $\beta_1 = 1, \beta_2 = 2$ and $\beta_3 = 3$; $g_1(t) = \exp(t) - 1, g_2(t) = \sin(\pi t)$, and $g_3(t) = -0.5t^2 + 0.5$; $\epsilon_i(t)$ denotes a Gaussian process with mean zero and a covariance function $\text{cov}(\epsilon_i(t_1), \epsilon_i(t_2)) = \sigma_k^2 \times \rho_k^{|t_1 - t_2|}$ with $\sigma_1^2 = 0.1, \rho_1 = 0.3, \sigma_2^2 = 0.15, \rho_2 = 0.35, \sigma_3^2 = 0.2, \rho_3 = 0.4$. For each individual i, $n_i = 5$, and the observation time t_{ij} was sampled from a uniform distribution on $U(0, 1)$. We considered two transformations: the logarithm transformation $H(y) = 4\log(y)$ (Case 1), and the Box–Cox transformation $H(y) = (y^{0.5} - 1)/0.5$ (Case 2).

Tables 1 and 2 report the biases, empirical SEs, and RMSEs for our proposed method with the initial number of clusters $K^{(0)} = 7$, as well as the observed results for the CT and WOT

TABLE 1: Performance of the proposed method, CT, and WOT for Case 1 in Simulation 1.

	Proposed ($K^{(0)} = 7$)			CT ($K = K_0 = 3$)			WOT ($K = K_0 = 3$)		
	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
					Case 1				
π_1	0.003	0.026	0.026	0.002	0.024	0.024	0.039	0.043	0.058
π_2	0.002	0.079	0.079	0.005	0.039	0.040	0.217	0.073	0.229
π_3	0.000	0.082	0.082	0.002	0.040	0.040	0.256	0.064	0.264
β_1	0.006	0.055	0.056	0.006	0.047	0.047	0.246	0.150	0.288
β_2	0.006	0.066	0.066	0.001	0.058	0.058	0.022	0.139	0.141
β_3	0.000	0.070	0.070	0.007	0.054	0.055	0.268	0.165	0.315
ρ_1	0.009	0.045	0.046	0.005	0.046	0.046	0.026	0.070	0.075
ρ_2	0.008	0.063	0.064	0.006	0.050	0.050	0.079	0.173	0.190
ρ_3	0.012	0.054	0.055	0.008	0.046	0.046	0.070	0.038	0.079
σ_1^2	0.000	0.012	0.012	0.002	0.010	0.010	0.191	0.111	0.221
σ_2^2	0.002	0.015	0.015	0.004	0.017	0.017	0.263	0.044	0.267
σ_3^2	0.001	0.033	0.033	0.006	0.017	0.018	0.153	0.017	0.154
$g_1(t)$	0.002	0.042	0.042	0.001	0.029	0.029	0.059	0.059	0.084
$g_2(t)$	0.004	0.061	0.062	0.002	0.042	0.042	0.117	0.286	0.310
$g_3(t)$	0.003	0.055	0.055	0.005	0.048	0.048	0.123	0.049	0.133
♯cluster	0	0	0		[a]			[a]	

[a] $K^{(0)}$ in the proposed method is the initial value for the number of clusters; K_0 is the true number of clusters.

estimators for Cases 1 and 2, respectively. When implementing the CT and WOT methods, the number of clusters was correctly specified as $K = K_0 = 3$, whereas our proposed method was initialized with a larger number of clusters than the true value, and yielded #cluster, the estimated number of clusters. Tables 1 and 2 indicate that the WOT resulted in large biases and variances (with biases even dominating the corresponding SEs), suggesting that misspecifying of transformation functions may lead to biased and unstable estimates of the regression parameters and the growth curves. In contrast, our method does not require the specification of transformation functions and hence avoids the bias and instability that results when transformation functions are misspecified. Furthermore, our proposed method yielded estimates with corresponding estimated biases and variances that were close to those values obtained when the transformation functions were correctly specified. This result suggests that our proposed method achieves robust results with little loss of efficiency. Moreover, our method is able to estimate the number of clusters accurately.

Figure 1, which displays the average estimates of the transformation function and the growth curves based on the 200 simulations, together with the corresponding 95% pointwise confidence intervals, shows that the estimates, on average, are very close to the true functions.

Since our proposed method requires an initial number of clusters, we also investigated its behaviour using different initial numbers of clusters $K^{(0)} = 7$, 14, and 21, respectively, for Case 2.

TABLE 2: Performance of the proposed method, CT, and WOT for Case 2 in Simulation 1.

	Proposed ($K^{(0)} = 7$)			CT ($K = K_0 = 3$)			WOT ($K = K_0 = 3$)		
	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
					Case 2				
π_1	0.004	0.026	0.026	0.002	0.024	0.024	0.205	0.079	0.220
π_2	0.006	0.083	0.083	0.004	0.040	0.040	0.122	0.108	0.164
π_3	0.011	0.085	0.085	0.002	0.040	0.040	0.082	0.092	0.124
β_1	0.012	0.053	0.054	0.005	0.047	0.047	0.246	0.150	0.288
β_2	0.003	0.067	0.067	0.001	0.058	0.058	0.022	0.139	0.141
β_3	0.008	0.070	0.070	0.007	0.054	0.055	0.268	0.165	0.315
ρ_1	0.010	0.045	0.046	0.005	0.046	0.046	0.297	0.039	0.299
ρ_2	0.007	0.058	0.058	0.005	0.050	0.050	0.000	0.135	0.135
ρ_3	0.014	0.057	0.059	0.009	0.046	0.046	0.060	0.085	0.104
σ_1^2	0.001	0.012	0.012	0.002	0.010	0.010	0.077	0.051	0.093
σ_2^2	0.003	0.015	0.015	0.004	0.017	0.017	0.234	0.045	0.238
σ_3^2	0.001	0.036	0.036	0.006	0.017	0.018	0.271	0.026	0.272
$g_1(t)$	0.002	0.043	0.043	0.001	0.029	0.029	0.132	0.067	0.148
$g_2(t)$	0.002	0.061	0.061	0.002	0.042	0.042	0.067	0.164	0.177
$g_3(t)$	0.005	0.055	0.055	0.005	0.048	0.048	0.049	0.104	0.115
♯cluster	0	0	0		[a]			[a]	

[a] $K^{(0)}$ in the proposed method is the initial value for the number of clusters; K_0 is the true number of clusters.

Table 3 shows that despite the differing initial values of $K^{(0)}$, the resulting estimates were almost the same, suggesting that our proposed approach appears to be robust to the initial specification of the number of clusters.

Simulation 2. Our proposed method assumes a Gaussian distribution for the transformed responses. To investigate the robustness of our method to this particular assumption, we generated data using parameter settings that were similar to those used in Case 2 of Simulation 1, except that the $\epsilon_i(t)$ were generated from a mixed distribution, with each component being the centralized and scaled gamma distribution $\sigma \times (Gamma(\tau, 1) - \tau)/\sqrt{\tau}$; also, the correlation was introduced via a normal copula function. Taking $\tau = 5, 10, 50$, Table 4 reports the observed results for $K^{(0)} = 7$.

A useful rule to evaluate the severity of bias, as suggested by Olsen & Schafer (2001), involves checking whether the standardized bias (bias over SE) exceeds 0.4. When $\tau \geq 10$, both the skewness and the excess kurtosis were less than 1, and our proposed estimators were nearly unbiased. When both the skewness and the excess kurtosis were approximately 1, our proposed estimators yielded observed results that were moderately biased but nonetheless acceptable.

Simulation 3. We investigated the case of a time-dependent covariate. We generated the data using parameter settings that were similar to those used previously in Case 1 of Simulation 1,

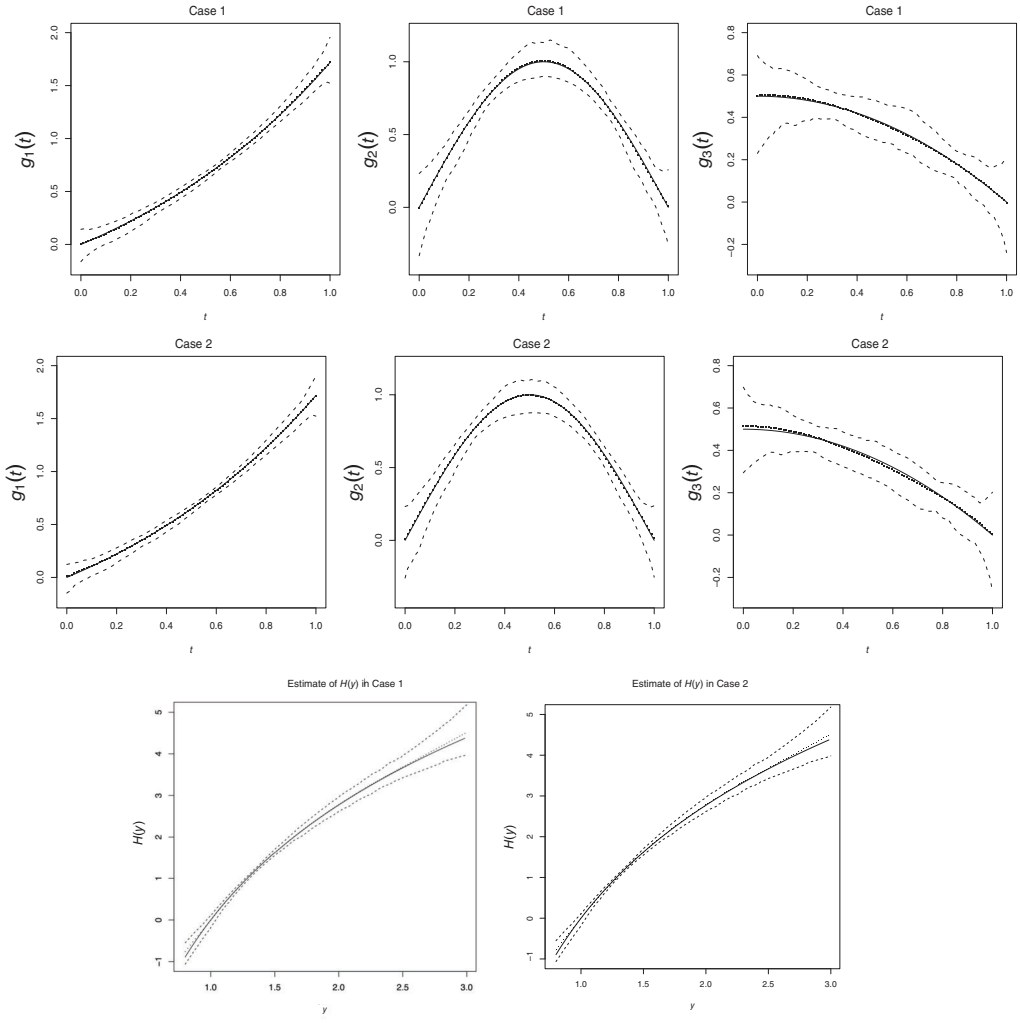

FIGURE 1: Estimated growth curves and transformation function for Cases 1 and 2 of Simulation 1 (solid: true function; dashed: 95% confidential limit; dotted: average of the estimated growth curve).

except that we generated $X_i(t)$ from Brownian motion. Table 5 summarizes the observed results that were obtained using an initial number of clusters $K^{(0)} = 7$. Our conclusions paralleled those derived from the results of Simulation 1.

5. ANALYSIS OF THE CHINESE HOUSING MARKET (2007–2014)

Rising housing prices in most of the Chinese cities between 2007 and 2014 had led to a public outcry over the seriously overheating markets in these regions, while the corresponding real estate markets in a small number of cities were stable in the same time period (Zhang et al., 2017). From the perspective of public policy as well as personal investment, it is therefore of substantial interest to study how such inequality may be linked to local economy, geography, and demographics, and which particular markets were more alike compared to other markets. Previous studies in a similar vein often made restrictive conditions, e.g., linear relationships,

TABLE 3: Performance of the proposed method, the CT, and the WOT for Case 2 in Simulation 1.

	$K^{(0)} = 7$			$K^{(0)} = 14$			$K^{(0)} = 21$		
	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
π_1	0.004	0.026	0.026	0.004	0.026	0.026	0.003	0.026	0.026
π_2	0.006	0.083	0.083	0.006	0.082	0.083	0.014	0.072	0.074
π_3	0.011	0.085	0.085	0.002	0.084	0.084	0.011	0.075	0.075
β_1	0.012	0.053	0.054	0.011	0.053	0.054	0.010	0.053	0.054
β_2	0.003	0.067	0.067	0.003	0.067	0.067	0.007	0.063	0.064
β_3	0.008	0.070	0.070	0.008	0.070	0.070	0.003	0.066	0.066
ρ_1	0.010	0.045	0.046	0.009	0.045	0.047	0.008	0.045	0.046
ρ_2	0.007	0.058	0.058	0.006	0.056	0.057	0.003	0.053	0.053
ρ_3	0.014	0.057	0.059	0.014	0.057	0.059	0.016	0.053	0.053
σ_1^2	0.001	0.012	0.012	0.001	0.012	0.012	0.002	0.012	0.012
σ_2^2	0.003	0.015	0.015	0.002	0.015	0.015	0.002	0.014	0.014
σ_3^2	0.001	0.036	0.036	0.002	0.036	0.036	0.008	0.032	0.033
$g_1(t)$	0.002	0.043	0.043	0.003	0.043	0.043	0.001	0.041	0.041
$g_2(t)$	0.002	0.061	0.061	0.002	0.061	0.061	0.003	0.056	0.056
$g_3(t)$	0.005	0.055	0.055	0.005	0.055	0.055	0.010	0.053	0.054
♯cluster	0	0	0	0	0	0	0	0	0

homogeneity, and normality assumptions (Guo & Li, 2011; Burdekin & Tao, 2014) on the relationships between the change trends in housing prices and local economic and demographic conditions. However, these various assumptions may not be satisfied. For example, the normal assumption may be problematic, as Figure 2 seems to suggest. Ren, Xiong & Yuan (2012) and Zhang et al. (2017) did relax these conditions, but only attempted to classify the data without considering covariates.

We used our proposed method to cluster the housing markets, after controlling for local economic levels and demographics, based on the average housing price-to-income ratios from 2007 to 2014 in a total of 252 cities, which cover most of the urban areas of China. The house price-to-income ratio is often used as an indicator of housing valuation and affordability (Wu, Gyourko & Deng, 2012). For each city, our data included house prices ($PRICE_t$), average monthly income ($INCOME_t$), real estate investment (INV_t), resident population size (POP_t), and total GDP (GDP_t). A total of 1230 observations were included in the dataset with n_i varying from 1 to 8. The data were extracted from the official website of the National Bureau of Statistics of China (www.stats.gov.cn).

First, we rescaled the time range to $[0, 1]$. Following housing demand–supply theory (DiPasquale & Wheaton, 1992), we used the housing price-to-income ratio $Y_i(t) = PRICE_t/INCOME_t$ in city i as the dependent variable, and the rates of growth $GR(t) = (GDP_t - GDP_{t-1})/GDP_{t-1}$, $PR(t) = (POP_t - POP_{t-1})/POP_{t-1}$, and $IR(t) = (INV_t - INV_{t-1})/INV_{t-1}$ as predictors. In particular, we used the growth rates or the change rates of economic data, in lieu of the original values, as predictors in the model since they can better capture the dynamics of GDP,

TABLE 4: Resulting estimators of the proposed method with $K^{(0)} = 7$ for Simulation 2.

	5			10			50			∞		
τ	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
π_1	0.008	0.027	0.028	0.007	0.027	0.028	0.004	0.027	0.028	0.004	0.026	0.026
π_2	0.037	0.091	0.098	0.028	0.088	0.092	0.007	0.084	0.085	0.006	0.083	0.083
π_3	0.045	0.090	0.100	0.035	0.087	0.094	0.012	0.084	0.085	0.011	0.085	0.085
ρ_1	0.015	0.049	0.052	0.007	0.049	0.050	0.004	0.045	0.046	0.010	0.045	0.046
ρ_2	0.024	0.073	0.077	0.028	0.066	0.072	0.015	0.057	0.059	0.007	0.058	0.058
ρ_3	0.007	0.058	0.058	0.003	0.059	0.059	0.006	0.055	0.056	0.014	0.057	0.059
β_1	0.042	0.059	0.072	0.030	0.058	0.065	0.011	0.059	0.060	0.012	0.053	0.054
β_2	0.021	0.067	0.071	0.009	0.063	0.064	0.002	0.062	0.063	0.003	0.067	0.067
β_3	0.063	0.082	0.103	0.040	0.076	0.086	0.013	0.074	0.075	0.008	0.070	0.070
σ_1^2	0.002	0.012	0.012	0.000	0.012	0.012	0.001	0.011	0.012	0.001	0.012	0.012
σ_2^2	0.023	0.026	0.035	0.016	0.019	0.025	0.008	0.016	0.018	0.003	0.015	0.015
σ_3^2	0.028	0.046	0.054	0.021	0.040	0.046	0.007	0.037	0.038	0.001	0.036	0.036
$g_1(t)$	0.013	0.040	0.042	0.010	0.042	0.043	0.003	0.042	0.042	0.002	0.043	0.043
$g_2(t)$	0.008	0.058	0.059	0.012	0.057	0.058	0.006	0.057	0.058	0.002	0.061	0.061
$g_3(t)$	0.031	0.060	0.068	0.021	0.056	0.060	0.006	0.054	0.054	0.005	0.055	0.055

Note: "$\tau = \infty$" represents a normal distribution.

TABLE 5: Resulting estimators of the proposed method with X generated from Brownian motion for Simulation 3.

	π_1	π_2	π_3	ρ_1	ρ_2	ρ_3	β_1	β_2	β_3	σ_1^2	σ_2^2	σ_3^2	$g_1(t)$	$g_2(t)$	$g_3(t)$
Bias	0.003	0.027	−0.030	0.002	−0.002	−0.018	0.034	0.077	−0.111	−0.002	−0.006	−0.033	0.004	0.010	0.023
SE	0.031	0.065	0.065	0.049	0.048	0.059	0.098	0.133	0.150	0.023	0.024	0.037	0.049	0.055	0.054
RMSE	0.031	0.070	0.072	0.049	0.048	0.061	0.104	0.153	0.187	0.023	0.024	0.049	0.050	0.056	0.059

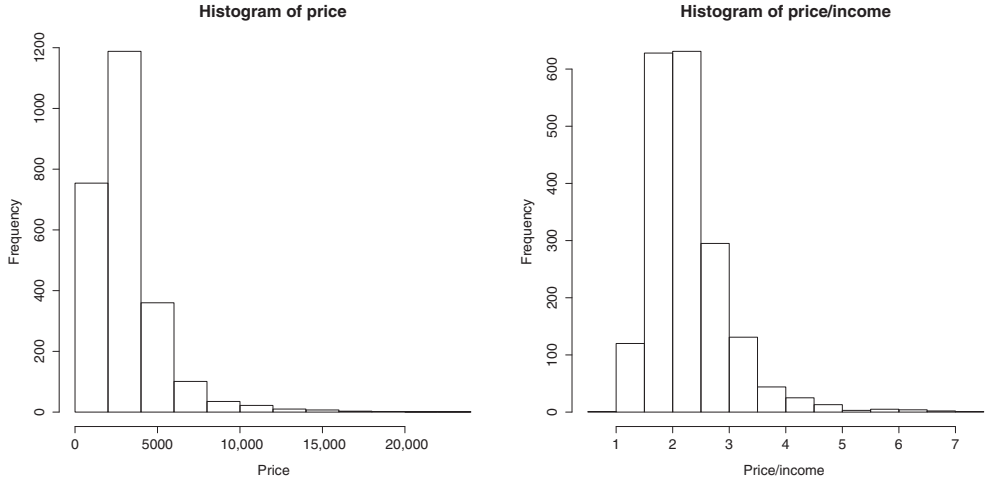

FIGURE 2: Histogram of the price (left) and price/income (right).

 DOI: 10.1002/cjs.11680

TABLE 6: Estimated coefficients of parameters for data of China Housing Market.

	Cluster 1			Cluster 2		
	Est.	SE	*P*-Value	Est.	SE	*P*-Value
π	0.953	0.013	0	0.047	0.013	0.0002
$GR(t-1)$	−0.017	0.008	0.0335	−0.620	0.095	0
$PR(t-1)$	−0.015	0.009	0.0955	−0.316	0.072	1e−05
$IR(t-1)$	−0.031	0.070	0.6578	−0.040	0.074	0.5888

population, and investment over time, and also facilitate horizontal and vertical comparisons across different markets.

Allowing impacts to have a 1-year lag, we regressed $Y_i(t)$ on $GR_i(t-1), PR_i(t-1)$, and $IR_i(t-1)$ with

$$H\big(Y_i(t)\big) = g_k(t) + \mathbf{X}_i(t-1)'\boldsymbol{\beta}_k + \epsilon_i(t),$$

for $k = 1, \ldots, K$, where $\mathbf{X}_i(t-1) = (GR_i(t-1), PR_i(t-1), IR_i(t-1))'$, and $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \beta_{k3})'$. As an initial number of clusters, the values $K^{(0)} = 7$ and $K^{(0)} = 10$ yielded the same results, so we used $K^{(0)} = 7$. We adopted the cubic B-spline approximation, with the number and locations of the interior knots chosen based on the strategy outlined in Section 4. The tuning parameter $\lambda = 1/200$ was selected by minimizing the BIC defined in Equation (17). Our method identified two clusters in 252 cities, with estimated probabilities of 0.947 and 0.053 corresponding to clusters 1 and 2, respectively. The estimated coefficients and corresponding estimated standard errors (SE) are reported in Table 6. The SE is based on 200 bootstrap samples, where 200 was adopted by monitoring the stability of the SE. The estimates of the transformation function H and growth curves g_1, g_2 are displayed in Figure 3, which reveals that the transformation curve resembles a logarithm transformation, and that g_1, g_2 have different change trends. In order to demonstrate how we benefitted from using our proposed method, we split the original data into four subgroups of roughly equal size. Three of these groups were used as training data, and the remaining quarter served as a validation dataset. We compared our proposed method with the logarithm transformation and Box−Cox transformation by using the out-of-sample prediction error (PE) $= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\big|\widehat{H}^{-1}_{-d(i)}\big(W_{ij}\big) - Y_{ij}\big|/\big|C_i\big|$, where $W_{ij} = \sum_{k=1}^{K} I\big(\widehat{\delta}^{-d(i)}_{ik} = 1\big)\Big\{\widehat{g}^{-d(i)}_{k}\big(t_{ij}\big) + \mathbf{X}_i\big(t_{ij}\big)'\widehat{\boldsymbol{\beta}}^{-d(i)}_{k}\Big\}$ and $C_i = \max_j\big(Y_{ij}, j = 1, 2, \ldots, n_i\big), d = 1, \ldots, 4$. The estimates $\widehat{H}_{-d(i)}, \widehat{g}^{-d(i)}_{k}, \widehat{\boldsymbol{\beta}}^{-d(i)}_{k}, \widehat{\delta}^{-d(i)}_{ik}$ were computed by omitting the dth record in the data to which the ith sample belongs. We then evaluated the Box−Cox transformation with $\lambda = 0.2, 0.25, 0.5, 1, 2, 3$. Table 7 suggests that our fitted model has a smaller PE than both the logarithm and the Box−Cox transformations.

Table 6 suggests that the effects of the covariates are similar in the two clusters, though perhaps more significant in cluster 2. All the regression coefficients were estimated to be negative, which seems to reflect the actual situation in China from 2007 to 2014. In particular, the growth rate of GDP has a significant effect on the housing price-to-income ratios in both clusters. The effect of the growth rate associated with the size of the resident population was more marked in cluster 2 than in cluster 1. These results suggest that the positive growth rate of GDP or POP actually reduces the housing price-to-income ratios, whereas the growth rate of real estate investment (INT) has no apparent effect on the housing price-to-income ratios in either cluster. These findings

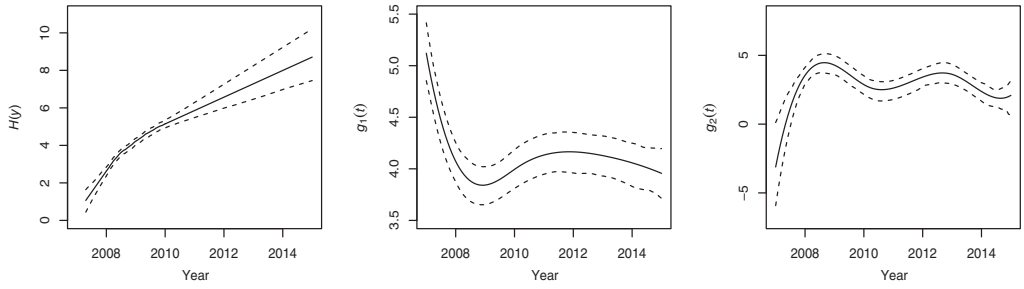

FIGURE 3: Estimates of the transformation H and mean functions g_1, g_2 (solid: average of the estimated function; dashed: 95% confidential limit).

TABLE 7: Prediction error.

Proposed	log	$\lambda = 0.2$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
0.162	0.186	0.196	0.202	0.243	0.357	0.538	0.620

provide valuable insight into the housing market conditions in China during the observation period 2007–2014.

Figure 3 reveals that the two groups of cities exhibit different change patterns in housing price-to-income ratios. In cluster 1, the ratios sharply decreased from 2007 to 2009, and then stabilized, whereas in cluster 2 the ratio slightly increased from 2007 to 2009, and became stable thereafter. In 2009, the central government introduced a series of regulations to manage the housing markets, which explains the stability after 2009. The decline from 2007 to 2009 that occurred in cluster 1 was due to the global financial crisis, whereas the housing markets in cities belonging to cluster 2 were relatively healthy and withstood the impact of the financial crisis by maintaining a steady rate of growth from 2007 to 2009.

To shed more light on these two clusters, we estimated the probability of each city belonging to each cluster based on Equation (14); then, using the majority voting rule, we assigned each city to cluster 1 or 2. Figure 4, which depicts the time series of housing price-to-income ratios for each city, shows that most cities with higher housing price-to-income ratios belong to cluster 1, while the rest form cluster 2. This is consistent with the observation that most of the China cities were deemed overheated between 2007 and 2014. Since the price-to-income ratios often serve as an important indicator for detecting market bubbles, our results highlight the need to distinguish these two clusters when formulating and implementing housing regulation policies.

It is generally believed that such housing regulation policies mainly affected the housing prices in the major cities, such as Beijing, Shanghai, Guangzhou, Shenzhen, Chongqing, Chengdu, Hangzhou, and Nanjing. However, after 2009, the government began to adopt real estate policies based on local conditions, which seems to agree with the findings of this study. Indeed, our estimated cluster 2 includes the major cities (Guangzhou, Shenzhen, Hangzhou, and Nanjing) as well as some rapidly developing cities (Xiamen, Ningbo, Fuzhou, Dongguan, Foshan, Zhuhai, Haikou, Sanya, Dalian, Lishui, Shaoxing, Taizhou, Wenzhou, and Zhoushan), whose GDP usually grew faster during 2007–2014 than that of the cities estimated to belong to cluster 1. In general, GDP growth was associated with increasing incomes in these cities, while the corresponding housing prices grew relatively slowly during the same period. As a result, a

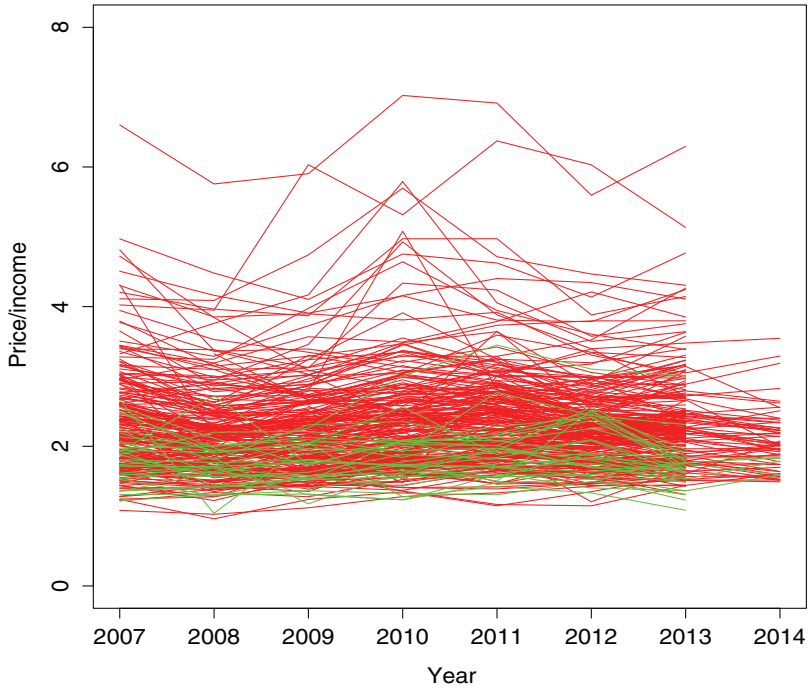

FIGURE 4: Categorization of 252 Chinese cities into two clusters: red curves indicate cluster 1 and green curves cluster 2.

marked decline in the housing price-to-income ratios occurred in cluster 2, which is reflected in the magnitude of β for $GR(t-1)$, which we estimated to be greater for cluster 2 than for cluster 1.

6. DISCUSSION

We have proposed an STFR model to cluster non-Gaussian functional data. Our proposed method can simultaneously estimate the unknown cluster number, transformation function, growth curves, and regression parameters. Via theoretical and numerical studies, we have shown that our proposed method performs well in selecting the number of clusters and in estimating the various unknown parameters and functions.

There are several open questions. Though our methods have focused on continuous responses, they can be extended to accommodate discrete responses as well. We envision such an extension to be nontrivial. It is also possible to extend our methods to cope with high-dimensional covariates by using a suitable penalty, but such an extension will require the development of new theory in order to provide performance guarantees. These research questions warrant further investigation.

ACKNOWLEDGEMENTS

This research was supported by the National Natural Science Foundation of China (Nos. 11931014 and 11829101) and the Fundamental Research Funds for the Central Universities (No. JBK1806002) of China. We sincerely thank Professor Yuhong Yang for his very helpful comments and suggestions. We are grateful to the editor, the review editor, and two anonymous reviewers for their constructive comments and suggestions that led to an improved article.

REFERENCES

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

Andruff, H., Carraro, N., Thompson, A., Gaudreau, P., & Louvet, B. (2009). Latent class growth modelling: A tutorial. *Tutorials in Quantitative Methods for Psychology*, 5, 11–24.

Bauer, D. J. & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, 8, 338–363.

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

Burdekin, R. C. & Tao, R. (2014). Chinese real estate market performance: Stock market linkages, liquidity pressures, and inflationary effects. *Chinese Economy*, 47, 5–26.

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

Chiou, J.-M. & 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.

Cuesta-Albertos, J. A. & Fraiman, R. (2007). Impartial trimmed *k*-means for functional data. *Computational Statistics & Data Analysis*, 51, 4864–4877.

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

DiPasquale, D. & Wheaton, W. C. (1992). The markets for real estate assets and space: A conceptual framework. *Real Estate Economics*, 20, 181–198.

Fan, J. & Gijbels, I. (1996). Local polynomial modelling and its applications. *Monographs on Statistics and Applied Probability*, Vol. 66, CRC Press, Boca Raton, FL.

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

Giacofci, M., Lambert-Lacroix, S., Marot, G., & Picard, F. (2013). Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics*, 69, 31–40.

Guo, S. & Li, C. (2011). Excess liquidity, housing price booms and policy challenges in China. *China & World Economy*, 19, 76–91.

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

Horowitz, J. L. (2001). Nonparametric estimation of a generalized additive model with an unknown link function. *Econometrica*, 69, 499–513.

Horváth, L. & Kokoszka, P. (2012). *Inference for Functional Data with Applications*, Vol. 200. Springer Science & Business Media, Berlin.

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

Ieva, F., Paganoni, A. M., Pigoli, D., & Vitelli, V. (2013). Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62, 401–418.

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

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

Jacques, J. & Preda, C. (2014b). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, 71, 92–106.

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

Jia, S., Wang, Y., & Fan, G.-Z. (2018). Home-purchase limits and housing prices: Evidence from China. *The Journal of Real Estate Finance and Economics*, 56, 386–409.

Jones, B. L., Nagin, D. S., & Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods & Research*, 29, 374–393.

Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11, 1–18.

Li, Y. & Hsing, T. (2010a). Deciding the dimension of effective dimension reduction space for functional and high-dimensional data. *The Annals of Statistics*, 38, 3028–3062.

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

Li, Y., Wang, N., & Carroll, R. J. (2010). Generalized functional linear models with semiparametric single-index interactions. *Journal of the American Statistical Association*, 105, 621–633.

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

Lu, M., Zhang, Y., & Huang, J. (2009). Semiparametric estimation methods for panel count data using monotone B-splines. *Journal of the American Statistical Association*, 104, 1060–1070.

McLachlan, G. & Peel, D. (2004). *Finite Mixture Models*. John Wiley & Sons, New York.

Muthén, B. O. (2001). Latent variable mixture modeling. *New Developments and Techniques in Structural Equation Modeling*, Psychology Press, New York.

Nagin, D. S. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods*, 4, 139–157.

Nagin, D. S. (2005). *Group-based Modeling of Development*. Harvard University Press, Cambridge, MA.

Olsen, M. K. & Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, 96, 730–745.

Peng, J. & 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.

Ray, S. & Mallick, B. (2006). Functional clustering by Bayesian wavelet methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 305–332.

Ren, Y., Xiong, C., & Yuan, Y. (2012). House price bubbles in China. *China Economic Review*, 23, 786–800.

Samé, A., Chamroukhi, F., Govaert, G., & Aknin, P. (2011). Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, 5, 301–321.

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

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

Shi, J. & Wang, B. (2008). Curve prediction and clustering with mixtures of Gaussian process functional regression models. *Statistics and Computing*, 18, 267–283.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8, 1348–1360.

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

Titterington, D. M., Smith, A. F., & Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

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

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

Winsberg, S. & Ramsay, J. O. (1981). Analysis of pairwise preference data using integrated B-splines. *Psychometrika*, 46, 171–186.

Wu, J., Gyourko, J., & Deng, Y. (2012). Evaluating conditions in major Chinese housing markets. *Regional Science and Urban Economics*, 42, 531–543.

Yao, F. (2007). Functional principal component analysis for longitudinal and survival data. *Statistica Sinica*, 17, 965–983.

Yao, F., Fu, Y., & Lee, T. C. (2010). Functional mixture regression. *Biostatistics*, 12, 341–353.

Yao, F. & Müller, H. G. (2010). Functional quadratic regression. *Biometrika*, 97, 49–64.

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

Yao, F., Müller, H.-G., & Wang, J.-L. (2005b). Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33, 2873–2903.

Zhang, D., Liu, Z., Fan, G. Z., & Horsewood, N. (2017). Price bubbles and policy interventions in the Chinese housing market. *Journal of Housing and the Built Environment*, 32, 133–155.

Zhang, W., Li, D., & Xia, Y. (2015). Estimation in generalised varying-coefficient models with unspecified link functions. *Journal of Econometrics*, 187, 238–255.

Zhang, Y., Li, R., & Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105, 312–323.

Zhou, X. H., Lin, H., & 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.

Received 19 September 2020
Accepted 14 October 2021