Cluster Analysis with Regression of Non-Gaussian Functional Data on Covariates

Jiakun Jiang[a], Huazhen Lin[a]*, Heng Peng[b], Gang-Zhi Fan[c], and Yi Li[d]
[a] Center of Statistical Research and School of Statistics,
Southwestern University of Finance and Economics, Chengdu, Sichuan, China
[b] Department of Mathematics, Hong Kong Baptist University, Hong Kong
[c] School of Management, Guangzhou University, China
[d] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

Abstract

Cluster analysis with functional data often poses normality assumptions on outcomes, and is typically conducted unsupervisedly without covariates. However, non-normal functional data are frequently encountered in practice, and unsupervised learning, without directly tying covariates to clusters, often makes obtained clusters less interpretable. To address these issues, we propose a new semiparametric transformation functional regression (STFR) model, based on which we cluster non-normal functional data in the presence of covariates. Our model presents several unique features. First, it drops normality assumptions on the functional response, which adds more flexibility to the modeling. Second, our model allows clusters to have distinct relationships between functional responses and covariates, and thus makes the detected clusters more interpretable. Third, as opposed to various competing methods, we allow the number of clusters to be unspecified and data-driven. We develop a new estimation approach, which combines penalized likelihood and estimating equation methods, for estimating the number of clusters, regression parameters and

*Corresponding author. The research was supported by National Natural Science Foundation of China (No.11931014 and 11829101) and Fundamental Research Funds for the Central Universities (No. JBK1806002) of China. We sincerely thank Professor Yuhong Yang for his greatly helpful comment and suggestion.

transformation functions simultaneously, and establish large sample properties such as consistency and asymptotic normality. Simulations confirm the utility of the proposed method. We apply the method to analyze a China housing market dataset and garner some interesting findings.

Keywords: Cluster analysis; Functional data; Longitudinal data; Semiparametric transformation functional regression; Supervised learning

1 Introduction

Functional data have been routinely collected in many fields, such as in economics, pharmacy, biology, and climatology (Horváth and Kokoszka, 2012; Li and Hsing, 2010a,b; Li et al., 2010; Yao, 2007; Yao and Müller, 2010; Yao et al., 2010, 2005a,b), and analyses of these data may provide valuable insight for decision makers in these fields. For example, the past two decades have seen 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 et al. 2018), and has sparked interest in understanding how the inequality aligns with local economy, geography and demographics, and which markets are at the risk of "real estate bubble" and which are deemed "healthy." Hence, it is crucial to study the change trends of the housing prices across cities, and identify the patterns along with with local economic conditions and demographic compositions. Of our particular interest is to detect various types of relationships between these trends and the macroeconomic factors, and to classify cities or markets accordingly. The results may identify cities with over-heated 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 applied the finite-dimensional clustering methods (Abraham et al., 2003; Chiou and Li, 2007; Peng and Müller, 2008; Bouveyron and Jacques, 2011; Giacofci et al., 2013; Jacques and Preda, 2013, 2014b; James and Sugar, 2003; Ray and Mallick, 2006; Samé et al., 2011); *distance-based clustering*, including those of Cuesta-Albertos and Fraiman (2007), Ferraty and Vieu (2006), Ieva et al. (2013), Tarpey and Kinateder (2003) and Tokushige et al. (2007), directly conducts clustering based on the curve data; see Jacques and Preda (2014a) for a comprehensive review. However, none of these approaches could classify functional data while accounting for covariates.

Limited work is available for conducting cluster analysis on functional responses with covariates. For example, McLachlan and Peel (2004), Muthén (2001) and Titterington et al. (1985) proposed a growth mixture model (GMM); Nagin (1999, 2005) suggested group-based trajectory modeling (GBTM) to identify clusters of individuals based on functional responses as well as covariates; Shi and Wang (2008) developed a mixture of Gaussian process functional regression models to classify relationships between curve responses and covariates. However, these approaches required 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 and Curran (2003) showed that, for non-normal data, multiple groups can be falsely identified when, in fact, there is only one group. Moreover, as our numerical studies revealed, mis-specifications of the polynomial growth curves, which were commonly assumed by these models, may lead to unreliable estimation and classification. Finally, these methods required 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 has been done in this topic and most of the work relied on the Bayesian information criterion (BIC) (Nagin, 2005; Schwarz et al.,

1978; Shi and Wang, 2008; Andruff et al., 2009; Jones et al., 2001), which may incur much computational burden. Additionally, large sample results with 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 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, 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 is assumed to be a normal density function given the covariates X. Hence, the conditional distribution of $H(Y)$ given X is a normal mixture, and, on top of it we can further specify a semi-parametric model with an unknown growth curve. Finally, with penalization on the group probability, we identify clusters that can best fit the data. We develop a new estimation approach, which combines penalized likelihood and estimating equation methods, for estimating the number of clusters, regression parameters and transformation functions simultaneously, and establish large sample properties such as consistency and asymptotic normality. We apply the STFR model and the competing methods to analyze a China housing market dataset. Our model provides a better fit than the other competing methods, and generates some interesting results that may shed light on the determinants of the real estate conditions in China.

The remainder of the paper is organized as follows. Section 2 introduces the proposed model and the estimation method, and provides a BIC-type procedure to select tuning parameters; Section 3 develops the theoretical properties, including $\sqrt{n}$-consistency, asymp-

totic normality, and the model selection consistency; Section 4 presents simulations and comparisons with the other competing methods and Section 5 analyzes the China housing market data in Section 5. We conclude the paper with a brief discussion about the future research in Section 6. All of the technical proofs are relegated to the Supplementary Material.

2 Model and Estimation

2.1 Model and objective function

Let $(Y_i(t), X_i(t)), i = 1, \cdots, n$ be independent realizations of $(Y(t), X(t))$, where $t \in [t_0, t_1]$ with $0 \leq t_0 < t_1 < \infty$ being two fixed constants, and $X(t)$ is a $p-$dimensional covariate that may be time-dependent. Instead of observing the full trajectories $Y_i(\cdot)$ and $X_{i1}(\cdot), \cdots, X_{ip}(\cdot)$, 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 $X_{i1}(\cdot), \cdots, X_{ip}(\cdot)$ at time points $\mathbf{t}_i = (t_{i1} \cdots, t_{i,n_i})$ from $[t_0, t_1]$. For notational simplicity and without loss of generality, we hereafter assume this bounded set is $[0, 1]$, and the measurement time t_{ij} randomly distributed on $[0, 1]$. We denote by $Y_{ij} = Y_i(t_{ij})$ and $\mathbf{X}_{ij} = (X_{i1}(t_{ij}), \cdots, X_{ip}(t_{ij}))', j = 1, \cdots, n_i$. Then $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \cdots, Y_{i,n_i})'$ and $\mathbf{X}_i = (\mathbf{X}_{i1}, \mathbf{X}_{i2}, \cdots, \mathbf{X}_{i,n_i})'$ represent the sequences of measurements on individual i over n_i time points. For a non-random function H, let $f(H(\mathbf{Y}_i)|\mathbf{X}_i, \mathbf{t}_i)$ denote the conditional probability density of the transformed responses $H(\mathbf{Y}_i) = (H(Y_{i1}), H(Y_{i2}), \cdots, H(Y_{i,n_i}))'$ 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 mix-

ture of K normal densities, i.e.,

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

with $\pi_k \geq 0$, where π_k is the marginal probability of an individual belonging 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 $g_{ik} = (g_k(t_{i1}), g_k(t_{i2}), \cdots, g_k(t_{i,n_i}))'$, $g_k(\cdot)$ is an unknown smooth function, $\boldsymbol{\beta}_k$ is a parameter vector with dimension p, $\mathbf{X}_i$ is a matrix of covariates with dimension $n_i \times p$. We assume the covariance matrix $\Delta_i(\boldsymbol{\gamma}_k)$ following some parametric structures, but we do not require certain structure of the covariance matrix structure. Particularly, the covariance matrix can be a linear combination of various covariance matrixes of certain structure. Model (2.1) 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 modeling proposed by Nagin (1999, 2005) as a special case. In this paper, we propose to estimate K, along with $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)'$, the growth curve $g_k(\cdot)$, and the transformation function $H(\cdot)$ simultaneously.

Denote

$$\mathcal{G} = \{g(\cdot) : |g^{(q_1)}(t_1) - g^{(q_1)}(t_2)| \leq c_0|t_1 - t_2|^{q_2}, \text{ for any } 0 \leq t_1, t_2 \leq 1\}, \tag{2.1}$$

where q_1 is a non-negative integer, $q_2 \in (0, 1]$, $r = q_1 + q_2 \geq 2$, and $c_0 > 0$ is a constant. The smoothness assumption (2.1) is often used in non-parametric curve estimation. With the assumption $g_k \in \mathcal{G}$ for $k = 1, \cdots, K$, we approximate $g_k(\cdot)$ by $g_{nk}(t) = \boldsymbol{\alpha}_k' B_n(t)$ for $k = 1, \cdots, K$, where $B_n(\cdot) = \{b_1(\cdot), \cdots, b_{q_n}(\cdot)\}'$ 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 $\max(t_j - t_{j-1} : j = 1, \cdots, M_n) = O(n^{-v})$. 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 of

$$H(\mathbf{Y}_i)|\mathbf{X}_i, \mathbf{t}_i \sim N(\mathbf{B}_n(\mathbf{t}_i)\boldsymbol{\alpha}_k + \mathbf{X}_i\boldsymbol{\beta}_k, \Delta_i(\boldsymbol{\gamma}_k)), \tag{2.2}$$

where $\mathbf{B}_n(\mathbf{t}_i) = (B_n(t_{i1}), \cdots, B_n(t_{i,n_i}))'$. We denote $\mu_{nik} = \mathbf{B}_n(\mathbf{t}_i)\boldsymbol{\alpha}_k + \mathbf{X}_i\boldsymbol{\beta}_k$ and $f_{nk}(H(\mathbf{Y}_i)|\mathbf{X}_i, \mathbf{t}_i) = (2\pi)^{-n_i/2}|\Delta_i(\boldsymbol{\gamma}_k)|^{-1/2} \exp\left[-\frac{1}{2}\{H(\mathbf{Y}_i) - \mu_{nik}\}' \Delta_i(\boldsymbol{\gamma}_k)^{-1} \{H(\mathbf{Y}_i) - \mu_{nik}\}\right]$. We can draw inference based on the logarithmic likelihood function,

$$L_n(\boldsymbol{\theta}_n; H) = \sum_{i=1}^{n} \log\left\{\sum_{k=1}^{K} \pi_k f_{nk}(H(\mathbf{Y}_i)|\mathbf{X}_i, \mathbf{t}_i)\right\}, \tag{2.3}$$

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

As 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 $(\pi_k, k = 1, \ldots, K)'$. To see that, we denote $\delta_{ik} = 1$ if $\mathbf{Y}_i$ arises from the kth cluster, and $\delta_{ik} = 0$ otherwise, and denote $\delta_i = (\delta_{i1}, \cdots, \delta_{iK})'$, 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

$$\sum_{i=1}^{n}\sum_{k=1}^{K} \left(b_{ik}\left[\log(\pi_k) + \log\{f_{nk}(H(\mathbf{Y}_i)|\mathbf{X}_i, \mathbf{t}_i)\}\right]\right), \tag{2.4}$$

where $b_{ik} = E\{\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i\}$. As (2.4) contains $\log(\pi_k)$, whose gradient grows fast when π_k is close to zero, the L_p type penalties can not directly set small π_k to zero and we need to instead consider a penalty on $\log\{\pi_k\}$ for sparsity. Following Huang et al. (2017), we consider the following penalized likelihood,

$$Q_n(\boldsymbol{\theta}_n; H) = L_n(\boldsymbol{\theta}_n; H) - n\lambda \sum_{k=1}^{K} \log\left\{\frac{\epsilon + \pi_k}{\epsilon}\right\}, \tag{2.5}$$

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

As $Q_n(\boldsymbol{\theta}_n; H)$ involves the infinite-dimensional function $H(\cdot)$, a direct maximization is infeasible, we resort to a two-stage approach. We first use a series of estimating equations to estimate the transformation functions H. We then estimate $\boldsymbol{\theta}_n$ by maximizing $Q_n(\boldsymbol{\theta}_n; H)$ with H replaced by its current estimate. We repeat the procedure until convergence or the number of iteration larger than 100. The iterative estimator may converge to a local minimizer since the objective function is non-convex. These local minimizers may differ from each other. Multiple initial values are recommended so that the optimum value can be identified. In numerical studies and real data analysis, noting that H is a monotonic increasing function, we take Box-Cox transformation with ρ as the initial value of $H(\cdot)$. Given H, with B-spline approximation of nonparametric mean function $g_k(t)$, estimating the semiparametric model returned to the linear regression model framework. Therefore we can quite easily apply the common mixture regression statistical software, such as R package *flexmix*(Leisch, 2004), to get the initial value of regression parameters. We take ρ from previously given values which maximizing the resulting likelihood function. In our simulation, resulting in about 95% of runs converging, which suggest the choice works well.

2.2 Penalized EM algorithm for $\boldsymbol{\theta}_n$ given H

To estimate $\boldsymbol{\theta}_n$ given H, we propose a penalized expectation maximization (EM) algorithm (Dempster et al., 1977). With the complete data for an individual i being $D_i = \{\mathbf{Y}_i, \delta_i, \mathbf{X}_i\}$,

the penalized complete-data log-likelihood function is

$$\mathcal{Q}_c(\boldsymbol{\theta}_n; H) = \log \mathcal{L}_c(\boldsymbol{\theta}_n; H) - n\lambda \sum_{k=1}^{K} \log\left\{\frac{\epsilon + \pi_k}{\epsilon}\right\}, \tag{2.6}$$

where

$$\log \mathcal{L}_c(\boldsymbol{\theta}_n; H) \propto \sum_{i=1}^{n}\sum_{k=1}^{K} \left(\delta_{ik}\left[\log(\pi_k) + \log\{f_{nk}(H(\mathbf{Y}_i)|\mathbf{X}_i, \mathbf{t}_i)\}\right]\right). \tag{2.7}$$

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

$$\sum_{i=1}^{n} \frac{E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i)}{\pi_k} - \sum_{i=1}^{n} \frac{E(\delta_{i1}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i)}{1 - \sum_{j=2}^{K}\pi_j} + \frac{n\lambda}{\epsilon + 1 - \sum_{j=2}^{K}\pi_j} - \frac{n\lambda}{\epsilon + \pi_k} = 0,$$
$$\text{for } k = 2, \cdots, K, \tag{2.8}$$

$$\sum_{i=1}^{n} E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i)\left\{\text{tr}\left(\Delta_i(\boldsymbol{\gamma}_k)^{-1}\frac{\partial \Delta_i(\boldsymbol{\gamma}_k)}{\partial \gamma_{kj}}\Delta_i(\boldsymbol{\gamma}_k)^{-1}\left[\Delta_i(\boldsymbol{\gamma}_k) - \{H(\mathbf{Y}_i) - \mu_{nik}\}^{\otimes 2}\right]\right)\right\} = 0,$$
$$\text{for } k = 1, \cdots, K,\ j = 1, \cdots, m, \tag{2.9}$$

$$\boldsymbol{\alpha}_k = \left\{\sum_{i=1}^{n} E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i)\mathbf{B}_n(\mathbf{t}_i)'\Delta_i(\boldsymbol{\gamma}_k)^{-1}\mathbf{B}_n(\mathbf{t}_i)\right\}^{-1}$$
$$\times \sum_{i=1}^{n} E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i)\mathbf{B}_n(\mathbf{t}_i)'\Delta_i(\boldsymbol{\gamma}_k)^{-1}\left\{H(\mathbf{Y}_i) - \mathbf{X}_i\boldsymbol{\beta}_k\right\}, \tag{2.10}$$

$$\boldsymbol{\beta}_k = \left\{\sum_{i=1}^{n} E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i)\mathbf{X}_i'\Delta_i(\boldsymbol{\gamma}_k)^{-1}\mathbf{X}_i\right\}^{-1}$$
$$\times \sum_{i=1}^{n} E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i)\mathbf{X}_i'\Delta_i(\boldsymbol{\gamma}_k)^{-1}\left\{H(\mathbf{Y}_i) - \mathbf{B}_n(\mathbf{t}_i)\boldsymbol{\alpha}_k\right\}, \tag{2.11}$$

with $\sum_{k=1}^{K}\|\boldsymbol{\alpha}_k\| = c_0$ for identifiability, and γ_{kj} being the j-th 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 for (2.8),

$$\widehat{\pi}_k = \max\left\{0, \frac{1}{1-K\lambda}\left[\frac{1}{n}\sum_{i=1}^{n} E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i) - \lambda\right]\right\}. \tag{2.12}$$

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 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 case, we normalize π_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 (2.9), (2.10), (2.11) and (2.12) until $\boldsymbol{\theta}_n$ converges. For each step, 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 the Newton-Raphson iteration using (2.9). To estimate $\boldsymbol{\theta}_n$, we compute

$$E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i) = \frac{f_{nk}(H(\mathbf{Y}_i)|\mathbf{X}_i, \mathbf{t}_i)\pi_k}{\sum_{j=1}^{K} f_{nj}(H(\mathbf{Y}_i)|\mathbf{X}_i, \mathbf{t}_i)\pi_j}. \tag{2.13}$$

At the rth step, $E(\delta_{ik}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{t}_i)$ is estimated by the left-hand side of (2.13) with the unknown parameters and functions replaced by the estimators from the $(r-1)$th step.

2.3 Estimation of H given $\boldsymbol{\theta}_n$

For any given y, we have

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

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

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[I\left(Y_{ij}\leq y\right)-\sum_{k=1}^{K}\pi_k\Phi\left\{\frac{H(y)-\boldsymbol{\alpha}_k'B_n(t_{ij})-\mathbf{X}_{ij}'\boldsymbol{\beta}_k}{\sqrt{\sigma_{kj}}}\right\}\right]=0, \tag{2.14}$$

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

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[I\left(Y_{ij}\leq y\right)-\sum_{k=1}^{K}\pi_k\Phi\left\{\frac{\theta-\boldsymbol{\alpha}_k'B_n(t_{ij})-\mathbf{X}_{ij}'\boldsymbol{\beta}_k}{\sqrt{\sigma_{kj}}}\right\}\right]=0. \tag{2.15}$$

Equation (2.15) entails that the estimator $\widehat{H}(y)$ is a nondecreasing step function with jumps only at the observed Y_{ij}. Varying y among $\{v_1,\cdots,v_{s_n}\}$ and repeating the estimation procedure for each y, we obtain the whole curve estimator of $H(\cdot)$. In practice, the function $H(\cdot)$ may not be estimated well in the area of extreme observations because of sparsity. It may be necessary to specify a parametric form of $H(\cdot)$ for the area of extreme observations. In general, the parametric form for the area of extreme observations can be inducted according to the estimated function in the interiors of the observations. In the simulations and the real data example, we assume that the linear form for $H(\cdot)$ at the tail of observations.

We use the Newton-Raphson algorithm to solve the equation, with a moderate computational cost. Coupled with the closed-form estimator for $\boldsymbol{\theta}_n$ at each step, the implementation of the 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

Estimation of 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 proposed algorithm is not sensitive to the choice of the number of knots, which is consistent with those found in the literature, such as Winsberg and Ramsay (1981) who wrote: "The curve tends to be insensitive to knot choice provided there are no discontinuities and the curve is smooth in the sense of a small bound on the modulus of its second derivative". For smooth functions, 3-6 knots seem quite adequate and that is we recommend. We consider a BIC-based procedure to select λ, which yields model selection consistency for linear regression models (Wang et al., 2007). Specifically, we choose λ by maximizing

$$BIC(\lambda) = \log L_n(\boldsymbol{\theta}_n; H) - \frac{1}{2} DF_\lambda \log\left(\sum_{i=1}^{n} n_i\right), \tag{2.16}$$

where DF_λ is the generalized degree of freedom, which can be consistently estimated by the number of nonzero parameters; see Zhang et al. (2010) for generalized linear models. In our numerical studies, we select λ using grid search, which seems 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 $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K), \boldsymbol{\pi} = (\pi_1, \cdots, \pi_K), \boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_K), \mathbf{g} = (g_1, \cdots, g_K)$, $\|\cdot\|$ 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 $\boldsymbol{\theta}$, K_0 be the true number of clusters. Without loss of generality, the first K_0 components of $\boldsymbol{\pi}_0$ are non-zero with $\sum_{k=1}^{K} \pi_{k0} = 1$ and $\pi_{10} \geq \pi_{20} \geq \cdots \geq \pi_{K0} > 0$ for identifiability. Theorems 1 to 3 summarize the large sample properties under the following regularity conditions; the proofs are deferred to the Supplementary Material.

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

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

(A3) There exists $[\underline{y}, \bar{y}]$ such that $\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i} I(Y_{ij} \notin [\underline{y}, \bar{y}]) = o_p(n^{-1/2})$, 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) $\Delta/\delta \leq c_0$ uniformly in n, where $\delta = \min_{1 \leq i \leq M_n} |t_i - t_{i-1}|$, $\Delta = \max_{1 \leq i \leq M_n} |t_i - t_{i-1}| = O(n^{-\upsilon})$.

The assumption (A1), for mathematical convenience, is widely used in the nonparametrics literature (Zhang et al., 2015; Horowitz, 1996, 2001; Fan and Gijbels, 1996). The boundedness assumption for the regressor $X(t)$ is technical to simplify the proofs and may be relaxed to high-order moments being bounded. (A2) is commonly assumed in the semiparametrics literature (Chen and Tong, 2010). Conditions (A3) is used to avoid the

tail problem, which is also required by Lin et al. (2012). (A4) is a common condition for the transformation function (Zhou et al., 2008). Condition (A5) is often assumed for spline analysis (Lu et al., 2009).

Theorem 3.1. 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 one.

Theorem 3.2. Under Conditions (A1)-(A5), $\lambda\sqrt{n} \to 0$, and $\epsilon = o\left(\frac{1}{\sqrt{n}\log(n)}\right)$,

$$\widehat{H}_n(y) \xrightarrow{a.s.} H_0(y) \quad \text{uniformly} \quad \text{over } y \in [\underline{y}, \overline{y}],$$
$$d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_p(n^{-\min(\frac{1-\upsilon}{2}, r\upsilon)}),$$

where r is a smooth parameter defined in (2.1), and $0 < \upsilon < 0.5$ is given for determining the spline basis $B_n(\cdot)$.

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

Theorem 3.3. Under Conditions (A1)-(A5) with $r \geq 2$ and $\frac{1}{4r} < \upsilon < \frac{1}{2}$, $\sqrt{n}\lambda \to 0$ and $\epsilon = o\left(\frac{1}{\sqrt{n}\log(n)}\right)$,

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

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

4 Simulation

As our method allows the transformation function as well as the distribution of the functional data to be unknown, we investigate whether it is more robust than the existing

parametric or semi-parametric 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: (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) theuntransformed model (WOT), with the growth curves estimated by B-splines. The CT and WOT methods are used to evaluate the efficiency and robustness of the proposed method, respectively. We also assess the accuracy of cluster selection. Finally, as our method assumes a Gaussian distribution on the transformed responses within each cluster, we investigate the sensitivity of our method to the departure of this assumption. We use the criteria of bias, standard error (SE), and root-mean-square error (RMSE), defined by

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

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

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

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

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 $g_3(t) = -0.5t^2 + 0.5$; $\epsilon_i(t)$ is 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 consider 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 present the biases, empirical SEs, and RMSEs for the proposed method with the initial number of clusters $K^{(0)} = 7$, the CT and WOT estimators for Cases 1 and 2, respectively. When implementing the CT and WOT methods, the number of clusters is correctly specified as $K = K_0 = 3$, while our method starts with a larger number of clusters than the true one and provides the estimated $\sharp$cluster, the number of clusters. Table 1 indicates that the WOT estimates have large biases and variances (with biases even dominating the corresponding SEs), suggesting that mis-specifications of transformation functions may lead to biased and unstable estimates of the regression parameters and the growth curves. In contrast, our method yields estimates close to the true values, with biases and variances close to those for the CT estimates. The comparison with the WOT and CT estimators suggests that our procedure is robust with little loss of efficiency, and, moreover, our method can select 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, along with the 95% pointwise confidence intervals, shows that the estimates, on average, are very close to the true functions.

As our method requires an initial number of clusters, we further apply the proposed method with different initial number of clusters $K^{(0)} = 7, 14,$ and 21, respectively, for Case 2. Table 3 shows that the proposed estimates are almost the same with different initial numbers of clusters, suggesting the robustness to the initial specification of the number of

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		*			*	

*$K^{(0)}$ in the proposed method is the initial value for the number of clusters, K_0 is the true number of clusters.

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		*			*	

*$K^{(0)}$ in the proposed method is the initial value for the number of clusters, K_0 is the true number of clusters.

clusters.

Simulation 2. Our method assumes a Gaussian distribution on the transformed responses. To investigate the robustness of our method to this assumption, we generate data using the similar settings as in Case 2 of Simulation 1, except that $\epsilon_i(t)$ is generated from a mixed distribution with each component being the centralized and scaled gamma distribution $\sigma \times (Gamma(\tau, 1) - \tau)/\sqrt{\tau}$, and the correlation is constructed through a normal copula function. Taking $\tau = 5, 10, 50$, Table 4 presents the results with $K^{(0)} = 7$.

A useful rule to evaluate the severity of bias, as suggested by Olsen and Schafer (2001), is to check whether the standardized bias (bias over SE) exceeds 0.4. When $\tau \geq 10$, 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.

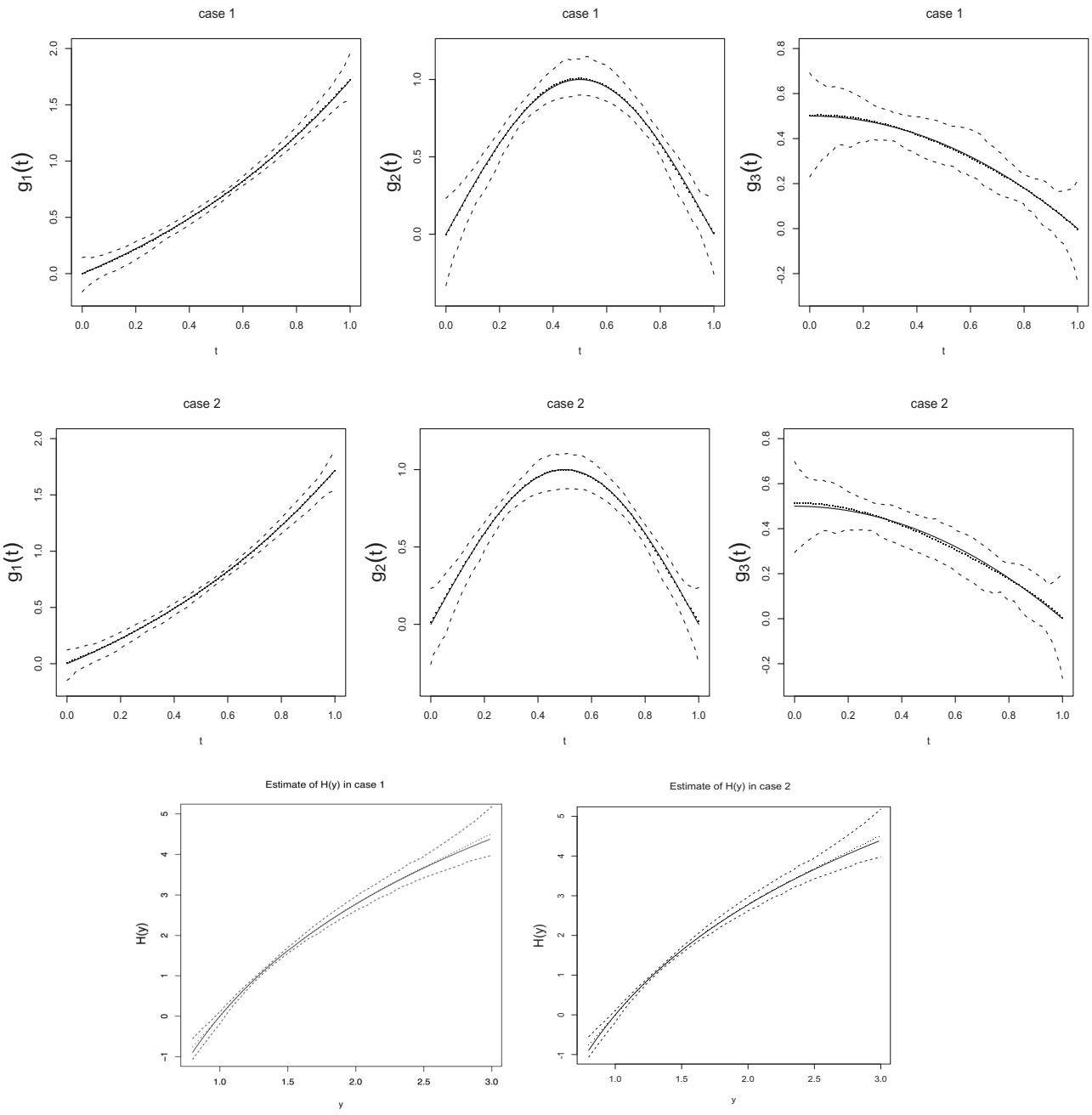

Figure 1: The estimated growth curves and transformation function for Cases 1 and 2 of Simulation 1 (solid: true function; dashed: 95% confidential limit; dotted: average of the estimated growth curve).

Simulation 3. We investigate the case of time-dependent covariate. We generate data using the similar settings as in Case 2 of Simulation 1, except that $X_{ij} = X_i(t_{ij}) \sim U[0, 1]$ is time-dependent. Table 5 presents the results with taking initial number of clusters $K^{(0)} = 7$. Similar conclusions with those from Simulation 1 can be obtained.

5 Analysis of the China Housing Market (2007-2014)

Rising housing prices in most of the Chinese cities between 2007-2014 had led to a public outcry over the seriously overheating markets in these regions, while the real estate markets in a small number of cities had been stable in the same time period (Zhang et al., 2017). From the perspective of policy making as well as personal investment, it is of substantial interest to study how such inequality linked to local economy, geography and demographics, and which markets behaved more similarly compared to the others. Previous studies on similar topics often made restrictive conditions, e.g. linear relationships, homogeneity and normal assumptions (Guo and Li, 2011; Burdekin and Tao, 2014) on the relationships between the change trends of housing prices and local economic and demographic conditions. However, these assumptions may not be satisfied. For example, the normal assumption may be problematic as exhibited in Figure 2. Ren et al. (2012) and Zhang et al. (2017) did relax these conditions, but only conducted classifications without considering covariates.

We apply the proposed method to cluster 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 area in China. The house price-to-income ratio is often used as a measurement for housing valuation and affordability (Wu et al., 2012). For each city, our data include house prices ($PRICE_t$), average monthly income ($INCOME_t$), real estate investment (INV_t), resident population

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

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

τ	5			10			50			∞		
	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

"$\tau = \infty$" represents a normal distribution

Table 5: Resulting estimators of the proposed method with $K^{(0)} = 7$ 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	0.001	0.019	-0.020	-0.004	0.000	-0.005	-0.013	0.016	-0.003	-0.009	-0.004	-0.022	0.007	0.011	0.011
SE	0.030	0.065	0.064	0.047	0.048	0.044	0.042	0.050	0.057	0.013	0.016	0.028	0.035	0.048	0.047
RMSE	0.030	0.067	0.067	0.047	0.048	0.044	0.044	0.053	0.057	0.016	0.016	0.035	0.035	0.049	0.049

size (POP_t), and total GDP(GDP_t). A total of 1,230 observations are included in the data with n_i varying from 1 to 8. The data are extracted from the official website of the National Bureau of Statistics of China (www.stats.gov.cn).

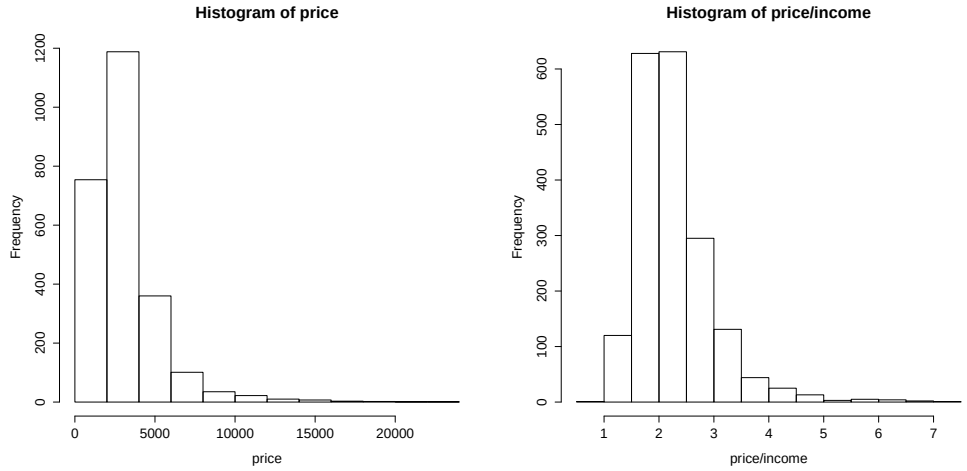

Figure 2: Histogram of the price and price/income, left is for price and right is for price/income.

First, we rescale the time range to $[0, 1]$. Following the housing demand-supply theory (DiPasquale and Wheaton, 1992), we take the housing price-to-income ratio $Y_i(t) = PRICE_t/INCOME_t$ in city i as dependent variable, and use 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. We use the growth rates or the change rates of economic

data, in lieu of the original values, to be the predictors in the model as they can better capture the dynamics of GDP, population and investment over years, and facilitate horizontal and vertical comparisons across different markets.

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

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

for $k = 1, \cdots, K$, 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 inital number of clusters $K^{(0)} = 7$ and $K^{(0)} = 10$ yield the same results, we take $K^{(0)} = 7$. We adopt the cubic B-spline approximation, with the number and locations of the interior knots taken based on the strategy in Section 4. The tuning parameter $\lambda = 1/200$ is selected by minimizing the BIC defined in (2.16). Our method identifies two clusters in 252 cities, with probabilities of 0.947 and 0.053 corresponding to clusters 1 and 2, respectively. The estimated coefficients and standard error (SE) are reported in Table 6. The SE is based on 200 bootstrap samples, where 200 is adopted by monitoring the stability of the SE. The estimation of 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 benefit from using the proposed method, we split data into 4 nearly equal-sized parts in which 3 parts are taken as training data and remain as validation dataset. We compare the proposed method with the logarithm transformation and Box-Cox transformation by the out-of-sample prediction error (PE) $= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}|\widehat{H}^{-1}_{-d(i)}(W_{ij}) - Y_{ij}|/|C_i|$, where $W_{ij} = \sum_{k=1}^{K} I(\widehat{\delta}_{ik}^{-d(i)} = 1)\{\widehat{g}_k^{-d(i)}(t_{ij}) + \mathbf{X}_i(t_{ij})'\widehat{\boldsymbol{\beta}}_k^{-d(i)}\}$ and $C_i = \max_j(Y_{ij}, j = 1, 2, \cdots, n_i), d = 1, \cdots, 4$. The estimate $\widehat{H}_{-d(i)}, \widehat{g}_k^{-d(i)}, \widehat{\boldsymbol{\beta}}_k^{-d(i)}, \widehat{\delta}_{ik}^{-d(i)}$ are computed with removing dth part of the data in which the ith sample belongs to. We perform the Box-Cox

transformation with $\lambda = 0.2, 0.25, 0.5, 1, 2, 3$. Table 7 suggests that the proposed method has a lower PE than both logarithm transformation and Box-Cox transformation.

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

Table 6 suggests that the impacts of covariates are similar in the two clusters, with more significant effects in cluster 2. All the regression coefficients are found 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 of resident population is also significantly negative in cluster 2, but is less significant 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 significant effect on the house price-to-income ratios in both clusters. These findings provide valuable insight into the housing market conditions in China.

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

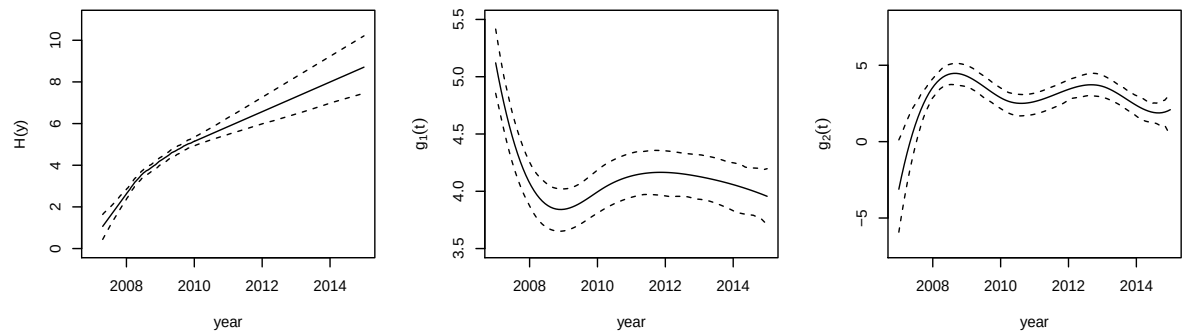

Figure 3: Estimates of the transformation H and mean functions g_1, g_2 (solid-average of the estimated function; dashed-95% confidential limit).

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

To shed more light on these two clusters, we estimate the probability of each city belonging to each cluster based on (2.13) and, by using the majority voting rule, we assign 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 observations that most of the China cities were deemed overheated between 2007 and 2014. As 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 making and implementing housing regulation policies.

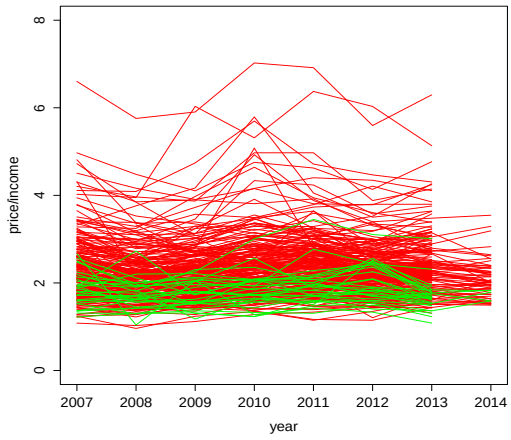

Figure 4: Categorize 252 Chinese cities into two clusters: red curves indicate cluster 1 and green curves cluster 2.

It is generally believed that the 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 seemed to be in agreement with the findings of this study. Indeed, our obtained 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 those in Cluster 1. The quicker growth of GDP has greatly increased the income levels in these cities, while their housing prices grew relatively slowly during the period. As a

result, there is a significant decline in the housing price-to-income ratios in cluster 2 which is reflected by that the magnitude of β for $GR(t-1)$ is much larger for Cluster 2.

6 Concluding remarks

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

There are several open questions. Though focused on continuous responses, our methods can be extended to accommodate discrete responses. We envision such an extension to be nontrivial. It is also possible to extend our methods to handle high-dimensional covariates by using a suitable penalty, but new theory needs to be developed to provide performance guarantees. These warrant future studies.

References

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

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

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

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

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

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

Chiou, J.-M. and P.-L. Li (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. and R. Fraiman (2007). Impartial trimmed k-means for functional data. *Computational Statistics & Data Analysis 51*, 4864–4877.

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

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

Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, Volume 66. CRC Press.

Ferraty, F. and P. Vieu (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media.

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

Guo, S. and C. Li (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.

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

Horváth, L. and P. Kokoszka (2012). *Inference for functional data with applications*, Volume 200. Springer Science & Business Media.

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

Ieva, F., A. M. Paganoni, D. Pigoli, and V. Vitelli (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. and C. Preda (2013). Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing 112*, 164–171.

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

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

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

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

Jones, B. L., D. S. Nagin, and K. Roeder (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*(8), 1–18.

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

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

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

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

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

McLachlan, G. and D. Peel (2004). *Finite mixture models*. John Wiley & Sons.

Muthén, B. O. (2001). Latent variable mixture modeling. In *New developments and techniques in structural equation modeling*. Psychology Press.

Nagin, D. S. (1999). Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychological methods 4*, 139.

Nagin, D. S. (2005). *Group-based modeling of development*. Harvard University Press.

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

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

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

Samé, A., F. Chamroukhi, G. Govaert, and P. Aknin (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.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics 6*, 461–464.

Shi, J. and B. Wang (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. and K. K. Kinateder (2003). Clustering functional data. *Journal of classification 20*, 093–114.

Titterington, D. M., A. F. Smith, and U. E. Makov (1985). *Statistical analysis of finite mixture distributions*. Wiley,.

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

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

Winsberg, S. and J. O. Ramsay (1981). Analysis of pairwise preference data using integrated b-splines. *Psychometrika 46*(2), 171–186.

Wu, J., J. Gyourko, and Y. Deng (2012). Evaluating conditions in major chinese housing markets. *Regional Science and Urban Economics 42*(3), 531–543.

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

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

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

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

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

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

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

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

Zhou, X.-H., H. Lin, and E. Johnson (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.