Article:

Yang, Xiaorong, Chen, Jia orcid.org/0000-0002-2791-2486, Li, Degui orcid.org/0000-0001-6802-308X et al. (1 more author) (2023) Functional-Coefficient Quantile Regression for Panel Data with Latent Group Structure. Journal of Business and Economic Statistics. ISSN 0735-0015

https://doi.org/10.1080/07350015.2023.2277172

Functional-Coefficient Quantile Regression for Panel Data with Latent Group Structure

Xiaorong Yang[1], Jia Chen[2], Degui Li[3], Runze Li[4*]

[1] School of Statistics and Mathematics and Collaborative Innovation Centre of Statistical Data, Engineering Technology and Application, Zhejiang Gongshang University

[2] Department of Economics and Related Studies, University of York

[3] Department of Mathematics, University of York

[4] Department of Statistics, Pennsylvania State University

Abstract

This paper considers estimating functional-coefficient models in panel quantile regression with individual effects, allowing the cross-sectional and temporal dependence for large panel observations. A latent group structure is imposed on the heterogenous quantile regression models so that the number of nonparametric functional coefficients to be estimated can be reduced considerably. With the preliminary local linear quantile estimates of the subject-specific functional coefficients, a classic agglomerative clustering algorithm is used to estimate the unknown group structure and an easy-to-implement ratio criterion is proposed to determine the group number. The estimated group number and structure are shown to be consistent. Furthermore, a post-grouping local linear smoothing method is introduced to estimate the group-specific functional coefficients, and the relevant asymptotic normal distribution theory is derived with a normalisation rate comparable to that in the literature. The developed methodologies and theory are verified through a simulation study and showcased with an application to house price data from UK local authority districts, which reveals different homogeneity structures at different quantile levels.

Keywords: Cluster analysis; functional-coefficient models; incidental parameter; latent groups; local linear estimation; panel data; quantile regression

*The corresponding author, email address: rzli@psu.edu.

1 Introduction

The quantile regression models and their estimation have received increasing attention since the seminal work by Koenker and Bassett (1978). They have been widely applied in various disciplines including economics, finance, health science and social science. In contrast to classic mean regression, the quantile regression provides a more comprehensive picture in capturing the relationship between the response and explanatory variables, and serves as a robust alternative. Various parametric methods with theoretical treatment and empirical applications have been extensively studied for quantile regression, see Koenker (2005) and Koenker et al. (2017) for comprehensive reviews. Due to wide availability of panel/longitudinal data in many areas, it is natural to extend parametric linear quantile regression from independent data to more general panel data. Subject-specific individual effects are often incorporated in linear quantile panel models to reflect location shift effects on the quantile regression and describe heterogeneity over subjects. The number of these "incidental parameters" diverges as the number of subjects N increases, affecting estimation accuracy of the common quantile regression coefficients (in particular when the number of observations per subject T is fixed). Various quantile estimation and inferential techniques have been proposed in the literature (e.g., Koenker, 2004; Canay, 2011; Kato et al., 2012; Galvao et al., 2013; Galvao and Kato, 2016; Galvao et al., 2020) for large panel data, i.e., both N and T are large. However, the aforementioned literature relies on a pre-specified parametric linear model assumption, which may be too restrictive in quantile regression and is often rejected in practical data analysis. In this paper, we adopt a nonparametric panel modelling approach, allowing data to "speak for themselves" and thus providing more reliable numerical performance in quantile regression than the parametric one.

Nonparametric quantile regression estimation has been systematically studied in the literature for independent cross-sectional data or weakly dependent time series data (e.g., Yu and Jones, 1998; Cai, 2002; Yu and Lu, 2004; Cai and Xu, 2008; Li et al., 2013; Belloni et al., 2019; Li et al., 2021). In recent years, there have been some attempts to study nonparametric quantile regression for panel data with individual effects. For example, Yan and Li (2018) introduce a three-step nonparametric conditional quantile estimation method combining series approximation, first-order difference (to remove incidental parameters) and deconvolution; and Chen (2021) proposes two local-linear-based methods to estimate quantile partial effects and further derives the asymptotic distribution theory under the large panel setting. Among various nonparametric quantile regression models, the functional-coefficient model is one of the most commonly-used frameworks. It is a natural extension of the linear quantile regression model, avoiding the so-called "curse of dimensionality" problem in nonparametric estimation when the number of covariates is large. For the classic independent or time series data setting, the functional-coefficient quantile model and its generalised version have been extensively studied in the literature (e.g., Honda, 2004; Kim, 2007; Cai and Xu, 2008; Wang et al., 2009; Kai et al., 2011; Tang et al., 2013). In particular, the functional-coefficient quantile regression allows the dynamic quantile relationship to vary smoothly over a state variable, and is thus connected to the functional linear quantile regression (e.g., Kato, 2012), but the latter assumes the covariate takes a functional value and bases the estimation methodology on some dimension reduction techniques (such as the functional principal component analysis). For the panel data with subject-specific fixed effects, Su and Hoshino (2016) combine the series approximation and instrumental variable quantile regression (e.g., Chernozhukov and Hansen, 2006) to estimate functional coefficients under a large-T framework, whereas Cai et al. (2018) use a kernel weighted quasi-likelihood to estimate semiparametric functional-coefficient quantile regression under a fixed-T framework.

A typical assumption imposed on nonparametric quantile regression for panel data in the existing papers is that the main nonparametric regression structure (after removing subject-specific location shift effects) is invariant over subjects, indicating that the dynamic relationship between the dependent and explanatory variables is the same for all subjects. However, such an assumption is often too restrictive in panel data studies when subjects involved have very different characteristics. An example is the house price data from UK local authority districts that we consider in Section 5.2. Due to differences in their location, population, and the socio-economic backgrounds of their population, the effects of factors, such as population growth and personal income growth, on house price growth are very unlikely to be homogeneous. As a result, in this paper, we relax the homogenous panel model assumption, allowing the functional-coefficients in nonparametric quantile regression to vary over subjects, see model (2.1).

However, for the heterogenous functional-coefficient panel quantile regression without imposing any structural restriction on the subject-specific coefficients, we can only rely on the sample information from an individual subject to estimate the subject-specific dynamic relationship, which leads to slow convergence of the estimated functional coefficients and unstable numerical performance of the estimates in finite samples. To address this problem, we assume that there exists a latent group structure on the subject-specific functional coefficients at each quantile level. If the underlying group structure is known or can be estimated, more efficient coefficient estimates can be obtained by pooling information belonging to the same group. From an empirical perspective, some panel data studies, such as Phillips and Sul (2007) and Hahn and Moon (2010), have found group structures for the panel models they use. For the UK house price data in Section 5.2, we identify two or five homogeneous groups (depending on the quantile level) for the 335 local authority districts. Such homogeneous groups may exist due to similarities in the characteristics of many districts (e.g., type of district, i.e., urban or rural, and socio-economic background of the majority of the population). Hence, it is both beneficial and reasonable to assume a group structure in some panel studies.

There has been increasing interest on estimating latent group structure in mean regression models for panel data in recent years. For example, Ke et al. (2016) use a binary segmentation technique to identify the latent group structure in linear regression models for panel data, whereas Su et al. (2016) introduce a penalised method via the so-called classifier-LASSO. Vogt and Linton (2017, 2020) propose a kernel-based classification of univariate nonparametric regression functions in panel data, which is further extended by Chen (2019) to estimate the group structure in time-varying coefficient panel data models. Other relevant developments can be found in Bonhomme and Manresa (2015), Ando and Bai (2017), Su et al. (2019), Liu et al. (2020), Wang and Su (2021) and Lian et al. (2021). In contrast, there is sparse literature on quantile regression models for panel data with latent group structures. Chetverikov et al. (2016) study IV panel quantile regression with group-specific coefficients defined as a linear regression with observable group-level covariates, but the "groups" in their paper are essentially the subjects in the context of this paper. Zhang et al. (2019b) propose an L_1-penalised estimation method to identify the group structure on the intercept in linear median regression; Gu and Volgushev (2019) estimate linear quantile regression for panel data with a latent group structure on the subject-specific fixed effects; and Zhang et al. (2019a) introduce an iterative algorithm using an idea similarly to the classic k-means clustering to estimate groups of units in panel data with heterogeneous slope coefficients. These estimation methods and algorithms rely on the parametric linear model assumption in quantile regression and cannot be directly applied to estimate the latent structure in nonparametric panel quantile regression.

In this paper, we aim to consistently estimate the group structure, the group number and the group-specific functional coefficients, all of which are allowed to vary over quantile levels. As there is no prior information on the latent groups, we start with a preliminary local linear quantile estimation of the subject-specific functional coefficients and the incidental parameter, only using the sample information from one subject. Based on the preliminary estimates of the functional coefficients, we compute the distance matrix between the subjects and subsequently use a classic agglomerative clustering algorithm to estimate the unknown group structure for the heterogenous functional coefficients. The resulting estimate is shown to be consistent (once the group number is pre-specified). Then, we introduce a simple ratio criterion to consistently estimate the group number. As the preliminary quantile estimates have rather slow convergence rates, we further propose a post-grouping local linear smoothing method to estimate the group-specific functional coefficients using the consistently estimated group structure, and derive the asymptotic normal distribution theory for the developed estimate with a convergence rate comparable to that in the literature. In the asymptotic analysis, we focus on the large panel setting with both N and T diverging to infinity. The panel observations are allowed to be temporally dependent and cross-sectionally correlated, relaxing the commonly-used cross-sectional independence restriction for panel quantile estimation (e.g., Kato et al., 2012; Cai et al., 2018; Chen, 2021).

We apply the proposed method to the house price data from UK local authority districts over the period Q1/1997–Q4/2016 and discover different group structures at different quantiles. At the lower quartile and median, we find more homogeneity in the effects of population and income growth on house price growth across districts, while at the upper quartile, more groups (i.e., five) are identified. By allowing the group structure to vary with the quantile level, we uncover a clearer picture about the relationship between population and income growth and house price growth across the distribution of house price growth.

The rest of the paper is organised as follows. Section 2 introduces the model and latent group structure. Section 3 describes the clustering algorithm and the ratio criterion for estimating the latent structure, and the post-grouping local linear quantile estimation. The technical assumptions and main asymptotic properties are provided in Section 4. Section 5 reports both the simulation and empirical studies. Section 6 concludes the paper. Proofs of the main theorems and technical lemmas, extensions of the developed methods and theory, and additional simulation and empirical results are available in a supplement.

2 Model structure

Suppose that we collect the panel random observations $(Y_{it}, \mathbf{X}_{it})$, $i = 1, \cdots, N$, $t = 1, \cdots, T$, and time series random observations Z_t, $t = 1, \cdots, T$, where Y_{it} and Z_t are univariate and $\mathbf{X}_{it}$ is d-dimensional. Let α_i be a subject-specific effect which may be correlated with X_{it} and Z_t. At a given quantile level $0 < \tau < 1$, the conditional quantile function for the i-th subject has the following functional-coefficient regression form:

$$Q_{\tau,i}\left(Y_{it}|\mathbf{X}_{it}, Z_t, \alpha_i\right) = \mathbf{X}_{it}^{\intercal}\boldsymbol{\beta}_{\tau,i}(Z_t) + \alpha_{\tau,i}, \tag{2.1}$$

where $\boldsymbol{\beta}_{\tau,i}(\cdot)$ is a d-dimensional vector of subject-specific functional coefficients. Both $\boldsymbol{\beta}_{\tau,i}(\cdot)$ and $\alpha_{\tau,i}$ are allowed to depend on the quantile level τ. It is worth stressing that model (2.1) is different from the random-coefficient quantile regression model (Koenker and Xiao, 2006), see the discussion in Appendix C.1 of the supplement. The functional coefficients $\boldsymbol{\beta}_{\tau,i}(Z_t)$ capture smooth changes of the dynamic quantile relationship (over Z_t) between Y_{it} and $\mathbf{X}_{it}$ at a fixed quantile level. Without loss of generality, we assume Z_t has a compact support $[0,1]$[1]. In practical applications, we may replace the random index variable Z_t in (2.1) by the fixed scaled time, t/T, or a variable Z_{it} that varies over both i and t, which would lead to the following functional-coefficient quantile regressions:

$$Q_{\tau,i}\left(Y_{it}|\mathbf{X}_{it},\alpha_i\right)=\mathbf{X}_{it}^{\intercal}\boldsymbol{\beta}_{\tau,i}(t/T)+\alpha_{\tau,i}, \quad \text{or} \quad Q_{\tau,i}\left(Y_{it}|\mathbf{X}_{it},Z_{it},\alpha_i\right)=\mathbf{X}_{it}^{\intercal}\boldsymbol{\beta}_{\tau,i}(Z_{it})+\alpha_{\tau,i}. \tag{2.2}$$

With slight modification, the methodology and theory to be developed in Sections 3 and 4 are still applicable to the above two model variants. Models in (2.1) and (2.2) can be seen as an extension of the functional-coefficient/time-varying panel data models studied by Li et al. (2011), Chen (2019), Su et al. (2019) and Phillips and Wang (2022) from mean regression to quantile regression.

In this paper, we further assume that there exists a partition of the index set $\{1,2,,\cdots,N\}$, denoted by $\mathcal{G}_\tau=\{\mathcal{G}_1^\tau,\mathcal{G}_2^\tau,\cdots,\mathcal{G}_{R_{\tau,0}}^\tau\}$ such that

$$\mathcal{G}_j^\tau\cap\mathcal{G}_k^\tau=\emptyset \ \text{ for } 1\leqslant j\neq k\leqslant R_{\tau,0}, \ \text{ and } \ \boldsymbol{\beta}_{\tau,i}(\cdot)=\boldsymbol{\gamma}_{\tau,j}(\cdot) \ \text{ for } i\in\mathcal{G}_j^\tau, \tag{2.3}$$

where $\boldsymbol{\gamma}_{\tau,j}(\cdot)$ denotes a d-dimensional vector of group-specific functional coefficients that may also depend on τ. Neither the group membership nor the group number is known a priori. Combining (2.1) and (2.3), we readily have that

$$Q_{\tau,i}\left(Y_{it}|\mathbf{X}_{it},Z_t,\alpha_i\right)=\mathbf{X}_{it}^{\intercal}\boldsymbol{\gamma}_{\tau,j}(Z_t)+\alpha_{\tau,i}, \ \ i\in\mathcal{G}_j^\tau, \ \ j=1,\cdots,R_{\tau,0}. \tag{2.4}$$

Note that the total number of unknown functional coefficients in (2.4) is $dR_{\tau,0}$, which is much smaller than dN, the number of heterogenous functional coefficients in (2.1).

Two remarks are in order here. First, the group membership $\mathcal{G}_\tau$ and the group number $R_{\tau,0}$ are allowed to vary over τ, which implies that the latent group structure can

[1]In the main model, we assume that Z_t is a continuous random variable with a density function satisfying Assumption 2(ii). This ensures that the maximum distance between two consecutive observations of Z_t is of order $O_P(\log T/T)$, indicating that there is a large number of observations in any neighborhood of $z\in[0,1]$ when T is sufficiently large. This is crucial for the kernel-based nonparametric estimation methodology and theory to be developed in the subsequent sections.

change over quantile levels. This makes the proposed quantile regression model framework more flexible and applicable than the mean regression one for practical research. Second, although we assume a group structure on the functional coefficients, $\boldsymbol{\beta}_{\tau,i}(\cdot)$, no group structure is imposed on the individual effects $\alpha_{\tau,i}$. This means that the subjects belonging to the same group are still allowed some degree of heterogeneity, as represented by their individual specific effects, albeit having the same functional slope coefficients.

The main interest of this paper lies in the estimation of $\mathcal{G}_\tau$, $R_{\tau,0}$ and $\boldsymbol{\gamma}_{\tau,j}(\cdot)$, $j = 1, \cdots, R_{\tau,0}$. For notational simplicity, we write $\boldsymbol{\beta}_{\tau,i}(\cdot) = \boldsymbol{\beta}_i(\cdot) = [\beta_{i,1}(\cdot), \cdots, \beta_{i,d}(\cdot)]^\intercal$, $\boldsymbol{\gamma}_{\tau,j}(\cdot) = \boldsymbol{\gamma}_j(\cdot) = [\gamma_{j,1}(\cdot), \cdots, \gamma_{j,d}(\cdot)]^\intercal$, $\alpha_{\tau,i} = \alpha_i$, $R_{\tau,0} = R_0$ and $\mathcal{G}_\tau = \mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_{R_0}\}$, suppressing their dependence on τ.

3 Estimation methodology

3.1 Preliminary local linear estimation and clustering algorithm

As mentioned in the introductory section, model (2.1) is a semiparametric functional-coefficient quantile model by treating α_i as an incidental parameter for fixed i. Assume that the unknown functional coefficients have continuous second-order derivatives. For $z \in [0, 1]$, with the sample information from the i-th subject, we define

$$\sum_{t=1}^{T} \rho_\tau \left(Y_{it} - \mathbf{X}_{it}^\intercal \mathbf{b}_1 - a_1 - (Z_t - z)\mathbf{X}_{it}^\intercal \mathbf{b}_2 - (Z_t - z)a_2 \right) K_h(Z_t - z), \tag{3.1}$$

where $\rho_\tau(\cdot)$ is the quantile check function defined by $\rho_\tau(z) = z[\tau - I(z \leqslant 0)]$ with $I(\mathcal{A})$ being the indicator function of the event $\mathcal{A}$, $K_h(u) = K(u/h)$, $K(\cdot)$ is a kernel function and h is a bandwidth. The local linear estimates $\widehat{\boldsymbol{\beta}}_i(z), \widehat{\boldsymbol{\beta}}_i'(z), \widehat{\alpha}_i(z), \widehat{\alpha}_i'(z)$ are obtained as the solution to minimise the objective function in (3.1).

Let $\boldsymbol{\Delta}$ be an $N \times N$ distance matrix among the true functional coefficients $\boldsymbol{\beta}_j(\cdot)$, $j = 1, \cdots, N$. The diagonal elements of $\boldsymbol{\Delta}$ are zeros, whereas the off-diagonal elements $\Delta(j, k)$, $1 \leqslant j \neq k \leqslant d$, are defined by

$$\Delta(j, k) = \int_0^1 \left\| \boldsymbol{\beta}_j(z) - \boldsymbol{\beta}_k(z) \right\| f_Z(z) dz,$$

where $f_Z(\cdot)$ is the density function of Z_t and $\|\cdot\|$ denotes the Euclidean norm. With the preliminary local linear quantile estimates, we have the following estimate of $\Delta(j,k)$[2]:

$$\widehat{\Delta}(j,k) = \frac{1}{T}\sum_{t=1}^{T}\left\|\widehat{\boldsymbol{\beta}}_j(Z_t) - \widehat{\boldsymbol{\beta}}_k(Z_t)\right\|.$$

With $\widehat{\Delta}(j,k)$, we obtain $\widehat{\boldsymbol{\Delta}}$, an $N\times N$ estimated distance matrix of $\boldsymbol{\Delta}$. The (j,k)-entry of $\widehat{\boldsymbol{\Delta}}$ is $\widehat{\Delta}(j,k)$ and the diagonal elements of $\widehat{\boldsymbol{\Delta}}$ are zeros. Using the estimated distance matrix, we may apply the agglomerative clustering method which has been widely used in the literature of cluster analysis (e.g., Everitt et al., 2011; Rencher and Christensen, 2012). Recently, such a method, combined with the kernel-based smoothing technique, has been applied to estimate the homogeneity/group structure in nonparametric mean regression models (e.g., Chen, 2019; Vogt and Linton, 2020; Chen et al., 2021). However, so far as we know, there is virtually no work on applying the kernel-based agglomerative clustering method to quantile regression models with a latent group structure. We next introduce the clustering algorithm when the group number is assumed to be R.

1. Start with N clusters, each of which corresponds to one of the N subjects. Search for the smallest off-diagonal element in $\widehat{\boldsymbol{\Delta}}$ which is the smallest distance estimate.

2. Merge the two clusters with the smallest distance. Consequently, the cluster number reduces from N to $N-1$. Update the estimated distance matrix for the $N-1$ clusters. Here the distance between two clusters $\mathcal{A}_1$ and $\mathcal{A}_2$ is calculated via the complete linkage, i.e., compute the farthest distance between an element in $\mathcal{A}_1$ and that in $\mathcal{A}_2$.

3. Repeat the previous steps with the updated distance matrix, and stop the algorithm when the number of clusters reaches R.

Let $\widehat{\mathcal{G}}_{1|R},\cdots,\widehat{\mathcal{G}}_{R|R}$ be the estimated clusters for a given group number R. If the true number of groups, R_0, is known a priori, we denote the estimated groups as $\widehat{\mathcal{G}}_r = \widehat{\mathcal{G}}_{r|R_0}$, $r = 1,\cdots,R_0$, whose consistency property is given in Theorem 4.1.

[2]When the index variable is Z_{it} which varies over i and t, we estimate the functional coefficients at a set of equidistant grid points $u_1, u_2, \cdots, u_m$, where m is a sufficiently large positive integer. Then, we define $\widehat{\Delta}(j,k) = \frac{1}{m}\sum_{t=1}^{m}\|\widehat{\boldsymbol{\beta}}_j(u_t) - \widehat{\boldsymbol{\beta}}_k(u_t)\|$.

3.2 Estimation of the group number

We next introduce a ratio criterion to consistently estimate the group number R_0. For a given number R, with the estimated groups $\widehat{\mathcal{G}}_{r|R}$ defined in Section 3.1, we may pool the estimated functional coefficients $\widehat{\boldsymbol{\beta}}_j(\cdot)$, $j \in \widehat{\mathcal{G}}_{r|R}$, and obtain the following estimate:

$$\widehat{\boldsymbol{\beta}}_{r|R}(z) = \frac{1}{\left|\widehat{\mathcal{G}}_{r|R}\right|} \sum_{j \in \widehat{\mathcal{G}}_{r|R}} \widehat{\boldsymbol{\beta}}_j(z), \quad r = 1, \cdots, R,$$

where $|\mathcal{A}|$ denotes the cardinality of a set $\mathcal{A}$. Then, we calculate the average deviation for $\widehat{\boldsymbol{\beta}}_j(\cdot)$ if the group number is assumed to be R:

$$\mathsf{D}(R) = \frac{1}{TR} \sum_{r=1}^{R} \frac{1}{\left|\widehat{\mathcal{G}}_{r|R}\right|} \sum_{j \in \widehat{\mathcal{G}}_{r|R}} \sum_{t=1}^{T} \left\| \widehat{\boldsymbol{\beta}}_j(Z_t) - \widehat{\boldsymbol{\beta}}_{r|R}(Z_t) \right\|. \tag{3.2}$$

It follows from Theorem 4.1 that the functional-coefficient quantile panel regression model is either correctly- or over-fitted (with probability tending to one) when $R \geqslant R_0$, and $\mathsf{D}(R)$ is thus convergent to zero. On the other hand, the model is under-fitted when $R < R_0$, and at least two groups are falsely merged. Consequently $\mathsf{D}(R)$ is strictly larger than a positive constant (using Assumption 5(ii) in Section 4.1). Hence, it is sensible to determine R_0 via the following simple ratio criterion:

$$\widehat{R} = \arg\min_{1 \leqslant R \leqslant \overline{R}} \frac{\mathsf{D}(R)}{\mathsf{D}(R-1)}, \tag{3.3}$$

where $\overline{R}$ is a pre-specified positive integer larger than R_0, and we set $\mathsf{D}(1)/\mathsf{D}(0) = 1$, $\mathsf{D}(R) = 0$ if $\mathsf{D}(R)$ is smaller than ω_{NT}, a threshold satisfying some mild restrictions, and define $0/0 \equiv 1$. A similar ratio criterion is also used by Lam and Yao (2012) and Ahn and Horenstein (2013) to find the number of latent factors in approximate factor models, and by Li et al. (2020) to determine the dimension of dominant sub-space in functional time series. Theorem 4.2 in Section 4.2 below shows that $\widehat{R}$ is a consistent estimate of R_0. With $\widehat{R}$, we may extract the estimated groups $\widetilde{\mathcal{G}} = \{\widetilde{\mathcal{G}}_1, \widetilde{\mathcal{G}}_2, \cdots, \widetilde{\mathcal{G}}_{\widehat{R}}\}$ by terminating the agglomerative clustering algorithm when $R = \widehat{R}$.

3.3 Post-grouping local linear estimation

Note that the preliminary functional coefficient estimates defined in Section 3.1 only make use of the sample information from one subject, resulting in a relatively slow uniform convergence rate, see Lemma A.2 in Appendix A of the supplement. The numerical performance of these estimates may be unstable in finite samples in particular when T is not sufficiently large. With the consistent estimates of the group number and membership constructed in Sections 3.1 and 3.2, we next propose a post-grouping local linear quantile estimation method for the group-specific functional coefficients $\boldsymbol{\gamma}_j(\cdot)$, $j = 1, \cdots, R_0$, improving the convergence rate of the preliminary functional coefficient estimates. From Corollary 4.1 to be given in Section 4.2, for any $j = 1, \cdots, R_0$, there exists $1 \leqslant j_* \leqslant \widehat{R}$ such that $\mathsf{P}\left(\mathcal{G}_j = \widetilde{\mathcal{G}}_{j_*}\right) \to 1$. Without loss of generality, we let $j_* = j$ in the rest of the section. Define the post-grouping local linear weighted objective function:

$$\sum_{i \in \widetilde{\mathcal{G}}_j} \sum_{t=1}^{T} \rho_\tau \left(Y_{it} - \mathbf{X}_{it}^\intercal \mathbf{b}_1 - a_{i1} - (Z_t - z)\mathbf{X}_{it}^\intercal \mathbf{b}_2 - (Z_t - z)a_{i2}\right) K_{h_1}(Z_t - z), \tag{3.4}$$

where $K_{h_1}(z) = K(z/h_1)$, $K(\cdot)$ is the kernel function and h_1 is a bandwidth which may be different from h used in the preliminary local linear estimation. The post-grouping local linear estimates $\widetilde{\boldsymbol{\gamma}}_j(z), \widetilde{\boldsymbol{\gamma}}_j'(z), \widetilde{\alpha}_i(z), \widetilde{\alpha}_i'(z)$, $i \in \widetilde{\mathcal{G}}_j$, are obtained as a solution to minimise the objective function in (3.4). As the fixed effects α_i are treated as "nuisance parameters", our primary interest lies in $\widetilde{\boldsymbol{\gamma}}_j(z)$ whose asymptotic distribution theory will be derived in Section 4.2 below.

4 Main asymptotic theory

4.1 Technical assumptions

For $i = 1, \cdots, N$, we let

$$\boldsymbol{\Omega}_i(z) = f_Z(z) \cdot \mathsf{E}\left[f_{ie}(0|\mathbf{X}_{it}, Z_t) \begin{pmatrix} 1 \\ \mathbf{X}_{it} \end{pmatrix} \left(1, \mathbf{X}_{it}^\intercal\right) \middle| Z_t = z \right], \tag{4.1}$$

where $f_Z(\cdot)$ is the density of Z_t and $f_{ie}(\cdot|\mathbf{x}, z)$ is the conditional density of $e_{it} = Y_{it} - \mathbf{X}_{it}^\intercal \boldsymbol{\beta}_i(Z_t) - \alpha_i$ given $\mathbf{X}_{it} = \mathbf{x}$ and $Z_t = z$. Assumptions 1–5 below are sufficient to prove the consistency properties for the estimated group membership and number.

Assumption 1. *For each* i*, the process* $\{(Y_{it}, \mathbf{X}_{it}, Z_t)\}$ *is stationary and* α*-mixing dependent with the mixing coefficient* $\alpha_i(\cdot)$ *satisfying* $\max_{1 \leqslant i \leqslant N} \alpha_i(s) \asymp \rho^s$, $0 < \rho < 1$.

Assumption 2. *(i) The conditional density function* $f_{ie}(\cdot|\mathbf{x}, z)$ *is continuous and has a bounded first-order derivative. In addition,* $f_{ie}(0|\mathbf{x}, z)$ *is continuous with respect to z and satisfies that*

$$0 < \underline{c}_e \leqslant \min_{1\leqslant i\leqslant N} \inf_{\mathbf{x},z} f_{ie}(0|\mathbf{x}, z) \leqslant \max_{1\leqslant i\leqslant N} \sup_{\mathbf{x},z} f_{ie}(0|\mathbf{x}, z) \leqslant \overline{c}_e < \infty,$$

where $\underline{c}_e$ *and* $\overline{c}_e$ *are two positive constants.*

(ii) The density function $f_Z(\cdot)$ *has continuous first-order derivative, and is bounded away from zero and infinity.*

Assumption 3. *(i) The matrix* $\mathbf{\Omega}_i(z)$ *defined in (4.1) is continuous (with respect to z) and positive definite with all the eigenvalues bounded away from zero and infinity uniformly over* $z \in [0, 1]$ *and* $1 \leqslant i \leqslant N$*. Furthermore,*

$$\max_{1\leqslant i\leqslant N} \mathsf{E}\left[\|\mathbf{X}_{it}\|^{\kappa+\epsilon}|Z_t\right] < \infty \ \ \text{a.s.}, \ \ 4 < \kappa < \infty, \ \ \epsilon > 0. \tag{4.2}$$

(ii) The subject-specific coefficient functions $\boldsymbol{\beta}_i(\cdot)$ *are twice continuously differentiable. In addition, there exists a positive constant* c_β *such that*

$$\max_{1\leqslant i\leqslant N} \sup_{0\leqslant z\leqslant 1} \left\|\boldsymbol{\beta}_i'(z)\right\| + \max_{1\leqslant i\leqslant N} \sup_{0\leqslant z\leqslant 1} \left\|\boldsymbol{\beta}_i''(z)\right\| \leqslant c_\beta,$$

where $\boldsymbol{\beta}_i'(z)$ *and* $\boldsymbol{\beta}_i''(z)$ *are the first-order and second-order derivatives of* $\boldsymbol{\beta}_i(z)$*, respectively. Similar conditions also hold for the group-specific coefficient functions* $\boldsymbol{\gamma}_j(\cdot)$.

Assumption 4. *(i)* $K(\cdot)$ *is a bounded, Lipschitz continuous and symmetric probability density function with a compact support* $[-1, 1]$.

(ii) The bandwidth h *satisfies that*

$$h^5 = o\left(\frac{\log(T \vee N)}{T}\right), \ \ \frac{Th}{(NT)^{4/\kappa}\log^5(N \vee T)} \to \infty, \tag{4.3}$$

where κ *is defined in Assumption 3(i). In addition, define* $\xi_{NT}^2 = \frac{\log(N\vee T)}{Th}$*, then it holds that*

$$\xi_{NT}^2 = o\left(\zeta_{NT}^2\right), \ \ \text{where} \ \ \zeta_{NT} = \min_{1\leqslant j\neq k\leqslant R_0} \int_0^1 \left\|\boldsymbol{\gamma}_j(z) - \boldsymbol{\gamma}_k(z)\right\| f_Z(z)dz. \tag{4.4}$$

Assumption 5. *(i) Let the group number* R_0 *be fixed and there exist* $0 < \underline{c}_g \leqslant \overline{c}_g < 1$ *such that*

$$\underline{c}_g N \leqslant \min_{1\leqslant r\leqslant R_0} |\mathcal{G}_r| \leqslant \max_{1\leqslant r\leqslant R_0} |\mathcal{G}_r| \leqslant \overline{c}_g N.$$

(ii) For any k *different (true) groups* $\mathcal{G}_{r_1}, \cdots, \mathcal{G}_{r_k}$*, when they are falsely merged, define*

$$\overline{\boldsymbol{\gamma}}_*(z) = \frac{1}{|\mathcal{G}_{r_1} \bigcup \cdots \bigcup \mathcal{G}_{r_k}|} \left[|\mathcal{G}_{r_1}| \boldsymbol{\gamma}_{r_1}(z) + \cdots + |\mathcal{G}_{r_k}| \boldsymbol{\gamma}_{r_k}(z) \right].$$

There exist $j \in \{r_1, \cdots, r_k\}$ *and a positive constant* c_* *such that*

$$\int_0^1 \left\| \overline{\boldsymbol{\gamma}}_*(z) - \boldsymbol{\gamma}_j(z) \right\| f_Z(z) dz > c_*.$$

(iii) The threshold parameter ω_{NT}*, used in the ratio criterion in Section 3.2, satisfies* $\omega_{NT} = o(1)$ *and* $\xi_{NT} = o(\omega_{NT})$*, where* ξ_{NT} *is defined in (4.4).*

Remark 4.1. Most of the above regularity conditions are mild and justifiable. Assumption 1 shows that the panel data are temporally dependent over t. The α-mixing dependence is one of the weakest mixing dependence conditions, which is satisfied for some commonly-used time series models (such as a vector ARMA process). The smoothness conditions on the (conditional) density functions and functional coefficients in Assumptions 2 and 3 are needed due to application of the local linear smoothing technique to estimate the unknown functions in quantile regression (e.g., Cai and Xu, 2008). The relatively strong moment condition in Assumption 3(i) is crucial to derive the uniform Bahadur representation and uniform consistency for the local linear quantile estimates, see Lemmas A.1 and A.2 in Appendix A of the supplement. It is worthwhile to point out that when κ is larger (indicating a stronger moment condition on $\mathbf{X}_i$), we may relax the bandwidth restriction and allow N to diverge at a faster polynomial rate of T. Letting $h \propto T^{-1/5}$ and κ be sufficiently large, we may show that the two conditions in (4.3) are satisfied. Similar to Assumption 4(iii) in Chen (2019) and Assumption 4(ii) in Chen et al. (2021), the condition (4.4) indicates that the minimum Euclidean distance between distinct coefficient functions is allowed to converge to zero. When $\zeta_{NT} > \underline{c}_* > 0$, (4.4) would be automatically satisfied. Assumption 5 is mainly used to derive the consistency for the group number estimate stated in Theorem 4.2. Assumption 5(i) shows that the latent groups have similar sizes (with the same divergence rate), whereas Assumption 5(ii) is crucial to prove that $D(R)$ defined in (3.2) would be strictly larger than a positive constant when the model is under-fitted (i.e., $R < R_0$). In fact, Assumption 5(ii) can be verified by using Assumption 5(i) and assuming $\zeta_{NT} > \underline{c}_* > 0$.

With the latent structure (2.3), we write $e_{it} = Y_{it} - \mathbf{X}_{it}^\intercal \boldsymbol{\gamma}_j(Z_t) - \alpha_i$ for $i \in \mathcal{G}_j$ and let

$$b_{it}(z) = \mathbf{X}_{it}^\intercal \left[\boldsymbol{\beta}_i(Z_t) - \boldsymbol{\beta}_i(z) - \boldsymbol{\beta}_i'(z)(Z_t - z)\right] = \mathbf{X}_{it}^\intercal \left[\boldsymbol{\gamma}_j(Z_t) - \boldsymbol{\gamma}_j(z) - \boldsymbol{\gamma}_j'(z)(Z_t - z)\right].$$

Re-write $\boldsymbol{\Omega}_i(z)$ defined in (4.1) in the block-matrix form:

$$\boldsymbol{\Omega}_i(z) = \begin{pmatrix} \omega_i^\alpha(z) & \boldsymbol{\Omega}_i^{\alpha\gamma}(z) \\ \boldsymbol{\Omega}_i^{\gamma\alpha}(z) & \boldsymbol{\Omega}_i^{\gamma}(z) \end{pmatrix}, \tag{4.5}$$

where $\omega_i^\alpha(z)$ is univariate and $\boldsymbol{\Omega}_i^{\gamma}(z)$ is a $d \times d$ matrix. For $j = 1, \cdots, R_0$, we define

$$\boldsymbol{\Omega}(z; \mathcal{G}_j) = \frac{1}{|\mathcal{G}_j|} \sum_{i \in \mathcal{G}_j} \left[\boldsymbol{\Omega}_i^{\gamma}(z) - \boldsymbol{\Omega}_i^{\gamma\alpha}(z)\boldsymbol{\Omega}_i^{\alpha\gamma}(z)/\omega_i^\alpha(z)\right], \tag{4.6}$$

$$\Gamma_{t0}(\mathcal{G}_j) = \sum_{i \in \mathcal{G}_j} \eta_{it}(z), \quad \boldsymbol{\Gamma}_{t1}(\mathcal{G}_j) = \sum_{i \in \mathcal{G}_j} \eta_{it}(z) \left[\mathbf{X}_{it} - \boldsymbol{\Omega}_i^{\gamma\alpha}(z)/\omega_i^\alpha(z)\right], \tag{4.7}$$

where $\eta_{it}(z) = \tau - I\left(e_{it} \leqslant -b_{it}(z)\right)$. To derive the asymptotic distribution theory for the post-grouping local linear quantile estimation, we need some additional conditions.

Assumption 6. *(i) The joint process* $\{(\mathbb{Y}_t, \mathbb{X}_t, Z_t)\}$ *is stationary and* α*-mixing dependent with the mixing coefficient satisfying* $\alpha(s) \asymp \rho^s$*, where* ρ *is defined as in Assumption 1,* $\mathbb{Y}_t = \{Y_{it} : i = 1, 2, \cdots\}$ *and* $\mathbb{X}_t = \{\mathbf{X}_{it} : i = 1, 2, \cdots\}$.

(ii) The bandwidth condition in (4.3) holds when h *is replaced by* h_1*, the bandwidth used in the post-grouping local linear estimation. For any* $j = 1, \cdots, R_0$,

$$N_j = |\mathcal{G}_j| = o\left((Th_1)^{1/2}/(\log T)^{3/2}\right).$$

(iii) For $j = 1, \cdots, R_0$, $\boldsymbol{\Omega}(z; \mathcal{G}_j)$ *defined in (4.6) is positive definite with all the eigenvalues bounded away from zero and infinity.*

Assumption 7. *(i) The joint density function of* (Z_t, Z_s) *exists and is bounded for any* $t \neq s$.

(ii) There exists $\iota > 4$ *such that*

$$\mathsf{E}\left[|\Gamma_{t0}(\mathcal{G}_j) - \mathsf{E}\left[\Gamma_{t0}(\mathcal{G}_j)\right]|^\iota\right] = O\left(N_j^{\iota/2}\right), \quad \mathsf{E}\left[\|\boldsymbol{\Gamma}_{t1}(\mathcal{G}_j) - \mathsf{E}\left[\boldsymbol{\Gamma}_{t1}(\mathcal{G}_j)\right]\|^\iota\right] = O\left(N_j^{\iota/2}\right).$$

(iii) There exists a $d \times d$ *matrix* $\boldsymbol{\Lambda}(z; \mathcal{G}_j)$ *such that, as* $N_j \to \infty$,

$$\frac{1}{N_j}\mathsf{E}\left[\left(\boldsymbol{\Gamma}_{t1}(\mathcal{G}_j) - \mathsf{E}\left[\boldsymbol{\Gamma}_{t1}(\mathcal{G}_j)\right]\right)\left(\boldsymbol{\Gamma}_{t1}(\mathcal{G}_j) - \mathsf{E}\left[\boldsymbol{\Gamma}_{t1}(\mathcal{G}_j)\right]\right)^\intercal | Z_t = z\right] \to \boldsymbol{\Lambda}(z; \mathcal{G}_j).$$

Remark 4.2. Assumptions 6(i) and 7(ii)(iii) show that the panel observations are allowed to be temporally correlated over t and cross-sectionally dependent over i. The high-level conditions would be satisfied when Y_{it}, $\mathbf{X}_{it}$ and Z_t are independent over i and t. The condition $N_j = o\left((Th_1)^{1/2}/(\log T)^{3/2}\right)$ in Assumption 6(ii) indicates that the number of subjects (in each group) needs to be much smaller than the time series length in order to derive the limit distribution theory in Theorem 4.3 with root-($N_j T h_1$) convergence. Analogous restrictions can also be found in Kato et al. (2012), Galvao and Kato (2016) and Chen (2021), and more comments will be given in Remark 4.4.

4.2 Asymptotic properties

We start with the consistency property for the group membership estimate when R_0 is pre-specified.

Theorem 4.1. *Suppose that Assumptions 1–4 are satisfied and the group number R_0 is known a priori. Then we have*

$$\mathsf{P}\left(\left\{\widehat{\mathcal{G}}_r,\ r = 1, \cdots, R_0\right\} = \left\{\mathcal{G}_r,\ r = 1, \cdots, R_0\right\}\right) \to 1,\ \ T \to \infty. \tag{4.8}$$

Remark 4.3. The consistency result (4.8) is similar to Theorem 3.1 in Vogt and Linton (2017), Theorem 1 in Chen (2019) and Theorem 4.1(a) in Vogt and Linton (2020), all of which study nonparametric mean regression for panel data with a latent group structure. By the clustering algorithm, to achieve the consistency property, it is sufficient to show that $\max_{1\leqslant j,k\leqslant N}\left|\widehat{\Delta}(j,k) - \Delta(j,k)\right|$ is of order smaller than the minimum distance between true group-specific functional coefficients. This can be proved by using the uniform consistency result for the preliminary local linear quantile estimation (see Lemma A.2 in Appendix A) and Assumption 4(ii). We do not need to impose any restriction on the cross-sectional dependence structure for panel random observations in this theorem.

The following theorem shows that the simple ratio criterion proposed in Section 3.2 consistently estimates the group number R_0.

Theorem 4.2. *Suppose that Assumptions 1–5 are satisfied. Then*

$$\mathsf{P}\left(\widehat{R} = R_0\right) \to 1,\ \ T \to \infty. \tag{4.9}$$

Combining Theorems 4.1 and 4.2, we readily have the following corollary on consistency of the group membership estimate when R_0 is unknown.

Corollary 4.1. *Suppose that the assumptions in Theorem 4.2 are satisfied. Then*

$$\mathsf{P}\left(\widetilde{\mathcal{G}} = \mathcal{G}\right) \to 1, \ \ T \to \infty, \tag{4.10}$$

where $\widetilde{\mathcal{G}} = \{\widetilde{\mathcal{G}}_1, \widetilde{\mathcal{G}}_2, \cdots, \widetilde{\mathcal{G}}_{\widehat{R}}\}$ *and* $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_R\}$.

We finally turn to the asymptotic normal distribution theory for the post-grouping estimate $\widetilde{\boldsymbol{\gamma}}_j(z)$ of the group-specific functional coefficient. It follows from Corollary 4.1 that there exists $1 \leqslant j_* \leqslant \widehat{R}$ such that $\mathsf{P}\left(\mathcal{G}_j = \widetilde{\mathcal{G}}_{j_*}\right) \to 1$. Without loss of generality, we let $j_* = j$, and define $\mu_k = \int z^k K(z) dz$ and $\nu_k = \int z^k K^2(z) dz$ for $k = 0, 1, 2, \cdots$.

Theorem 4.3. *Suppose that Assumptions 1–7 are satisfied. For z, an interior point of* $[0, 1]$, *and* $j = 1, \cdots, R_0$, *we have*

$$\sqrt{N_j T h_1}\left[\widetilde{\boldsymbol{\gamma}}_j(z) - \boldsymbol{\gamma}_j(z) - \mathbf{B}_j(z)\right] \xrightarrow{d} \mathsf{N}\left(\mathbf{0}_d, \ \boldsymbol{\Sigma}(z; \mathcal{G}_j)\right), \tag{4.11}$$

where $\mathbf{B}_j(z) = \frac{1}{2} h_1^2 \boldsymbol{\gamma}_j''(z) \mu_2$ *with* $\boldsymbol{\gamma}_j''(z)$ *being the second-order derivative of* $\boldsymbol{\gamma}_j(z)$, *and* $\boldsymbol{\Sigma}(z; \mathcal{G}_j) = [\boldsymbol{\Omega}(z; \mathcal{G}_j)]^{-1} [\nu_0 \boldsymbol{\Lambda}(z; \mathcal{G}_j)] [\boldsymbol{\Omega}(z; \mathcal{G}_j)]^{-1}$ *with* $\boldsymbol{\Omega}(z; \mathcal{G}_j)$ *defined in (4.6) and* $\boldsymbol{\Lambda}(z; \mathcal{G}_j)$ *defined in Assumption 7(iii).*

Remark 4.4. (i) Theorem 4.3 can be seen as an extension of the asymptotic normality given in Theorem 3.2 of Kato et al. (2012) from linear quantile regression to functional-coefficient quantile regression. To obtain the root-($N_j T h_1$) convergence, we need to asymptotically remove the influence of the nuisance parameters α_i, $i \in \mathcal{G}_j$, and impose a somehow restrictive condition on the divergence rate of N_j. Specifically, we assume that $N_j = o\left((Th_1)^{1/2}/(\log T)^{3/2}\right)$ in Assumption 6(ii), analogous to the condition $N = o\left(T^{1/2}/(\log T)^{3/2}\right)$ in Kato et al. (2012) if Th_1 is treated as the *effective* sample size for each subject in the kernel-based estimation. Galvao et al. (2020) show that the latter restriction may be relaxed to $N = o\left(T/(\log T)^2\right)$ in the context of linear panel quantile regression with fixed effects. This improvement is achieved by more precisely computing the orders of the remainder terms in the Bahadur representation of the quantile regression estimation. In particular, they show that the main remainder term can be approximated by a cross-sectional average of independent random elements under the cross-sectional independence restriction. Extension of this technique in the asymptotic proofs to our more general setting is non-trivial since we allow the panel observations to be cross-sectionally correlated. However, we conjecture that the restriction of $N_j = o\left((Th_1)^{1/2}/(\log T)^{3/2}\right)$ may be similarly relaxed by imposing some additional high-level conditions and handling the remainder terms of the Bahadur representation more carefully. This will be left in our future research.

(ii) Theorem 4.3 can be used to conduct point-wise statistical inference on the group-specific functional coefficients. For this, we have to estimate the bias and asymptotic variance matrix in (4.11), both of which contain some unknown quantities. Appendix C.3 in the supplement introduces nonparametric methods to estimate these quantities and subsequently obtains the bias and variance matrix estimate denoted by $\widetilde{\mathbf{B}}_j(z)$ and $\widetilde{\mathbf{\Sigma}}(z;\mathcal{G}_j)$, respectively. Let u_l be a d-dimensional vector with the l-th element being one and the others being zeros. For $\alpha \in (0,1)$, the $100(1-\alpha)\%$ confidence interval of $\gamma_{j,l}(z) = u_l^\intercal \boldsymbol{\gamma}_j(z)$ is constructed as

$$\left[\widetilde{\gamma}_{j,l}(z) - u_l^\intercal \widetilde{\mathbf{B}}_j(z) - c_{1-\alpha/2}\left(\frac{u_l^\intercal \widetilde{\mathbf{\Sigma}}(z;\mathcal{G}_j)u_l}{N_jTh_1}\right)^{1/2}, \widetilde{\gamma}_{j,l}(z) - u_l^\intercal \widetilde{\mathbf{B}}_j(z) + c_{1-\alpha/2}\left(\frac{u_l^\intercal \widetilde{\mathbf{\Sigma}}(z;\mathcal{G}_j)u_l}{N_jTh_1}\right)^{1/2}\right],$$

where $\widetilde{\gamma}_{j,l}(z) = u_l^\intercal \widetilde{\boldsymbol{\gamma}}_j(z)$ and $c_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. In the empirical study, we ignore the bias term by appropriate undersmoothing to simplify our construction of the confidence intervals. Alternatively, the estimation bias term can be removed via the so-called jackknife correction.

5 Numerical studies

5.1 Monte-Carolo simulation

Data is generated via the following functional-coefficient quantile regression:

$$Y_{it} = \mathbf{X}_{it}^\intercal \boldsymbol{\beta}_i(Z_t) + \alpha_i + e_{it}, \quad i = 1, \cdots, N, \quad t = 1, \cdots, T, \tag{5.1}$$

where Z_t, $t = 1, \cdots T$, are independently drawn from the uniform distribution $U[0,1]$, $\mathbf{X}_{it} = (X_{it,1}, X_{it,2})^\intercal$, $i = 1, \cdots, N, t = 1, \cdots, T$, are independently drawn from a bivariate normal distribution with zero means, unit variances, and a correlation coefficient of 1/2, $\alpha_i = \left(\bar{X}_{i,1}^2 + \bar{X}_{i,2}^2\right)/5$ with $\bar{X}_{i,k} = \frac{1}{T}\sum_{t=1}^{T} X_{it,k}$, and the idiosyncratic errors e_{it} are independently generated from one of the following distributions: $N(0,1)$, $t(5)$, and $0.4\left[\chi^2(3) - 3\right]$, and are independent of Z_t and $\mathbf{X}_{it}$. The results for the $N(0,1)$ errors can be used as the benchmark for assessing how the proposed method performs when the errors are heavy tailed (with $t(5)$ distribution) or asymmetrically distributed (with $0.4\left[\chi^2(3)-3\right]$ distribution, where the scaling factor 0.4 is used to give a comparable error variance to $N(0,1)$).

As in Chen (2019) and Su et al. (2019), the heterogenous functional coefficients $\boldsymbol{\beta}_i(\cdot) = [\beta_{i,1}(\cdot), \beta_{i,2}(\cdot)]^\intercal$ satisfy the following group structure:

$$\beta_{i,1}(z) = \begin{cases} \gamma_{1,1}(z) = 3F(z; 0.5, 0.1) & \text{if } i \in \mathcal{G}_1, \\ \gamma_{2,1}(z) = 3[2z - 6z^2 + 4z^3 + F(z; 0.7, 0.05)] & \text{if } i \in \mathcal{G}_2, \\ \gamma_{3,1}(z) = 3[4z - 8z^2 + 4z^3 + F(z; 0.6, 0.05)] & \text{if } i \in \mathcal{G}_3, \end{cases}$$

and

$$\beta_{i,2}(z) = \begin{cases} \gamma_{1,2}(z) = 3[2z - 4z^2 + 2z^3 + F(z; 0.6, 0.1)] & \text{if } i \in \mathcal{G}_1, \\ \gamma_{2,2}(z) = 3[z - 3z^2 + 2z^3 + F(z; 0.7, 0.04)] & \text{if } i \in \mathcal{G}_2, \\ \gamma_{3,2}(z) = 3[0.5z - 0.5z^2 + F(z; 0.4, 0.07)] & \text{if } i \in \mathcal{G}_3, \end{cases}$$

where $F(z;\xi,\eta) = 1/(1+\exp[-(z-\xi)/\eta])$, $\mathcal{G}_1 = \{1, 2, \cdots, N_1\}$, $\mathcal{G}_2 = \{N_1 + 1, \cdots, N_1 + N_2\}$, and $\mathcal{G}_3 = \{N_1 + N_2 + 1, \cdots, N_1 + N_2 + N_3\}$ with $N_1 = \lfloor 0.3N \rfloor$, $N_2 = \lfloor 0.3N \rfloor$ and $N_3 = N - N_1 - N_2$.

The sample size is $N = 50, 100$ and $T = 50, 100$, and the number of replications is $M = 200$. We consider three quantile levels, $\tau = 0.25, 0.50$ and 0.75, and use the Gaussian kernel, $K(u) = e^{-u^2/2}/\sqrt{2\pi}$, in the local linear estimation. The bandwidth is selected via the leave-one-out cross-validation method. To gauge the performance of the ratio criterion (3.3) in estimating the number of groups, we set $\overline{R} = 5$ and report the percentage of replications where each integer (between 1 and $\overline{R}$) is chosen. To understand the accuracy of the estimated groups (and their membership) from the agglomerative clustering algorithm in Section 3.1, we consider two measures: purity and the normalised mutual information (NMI) of the estimated groups $\widetilde{\mathfrak{G}} = \left\{\widetilde{\mathcal{G}}_1, \cdots, \widetilde{\mathcal{G}}_{\widehat{R}}\right\}$ with the true groups

$\mathcal{G} = \{\mathcal{G}_1, \cdots, \mathcal{G}_R\}$, both of which are classic criteria of clustering quality and are defined respectively as

$$\text{Purity}(\widetilde{\mathcal{G}}, \mathcal{G}) = \frac{1}{N} \sum_{k=1}^{\widehat{R}} \max_{1 \leqslant j \leqslant R_0} \left| \widetilde{\mathcal{G}}_k \cap \mathcal{G}_j \right|, \qquad \text{NMI}(\widetilde{\mathcal{G}}, \mathcal{G}) = 2 \cdot \frac{I(\widetilde{\mathcal{G}}, \mathcal{G})}{H(\widetilde{\mathcal{G}}) + H(\mathcal{G})},$$

where $I(\widetilde{\mathcal{G}}, \mathcal{G})$ is the mutual information between $\widetilde{\mathcal{G}}$ and $\mathcal{G}$ defined as

$$I(\widetilde{\mathcal{G}}, \mathcal{G}) = \sum_{k=1}^{\widehat{R}} \sum_{j=1}^{R_0} \frac{\left| \widetilde{\mathcal{G}}_k \cap \mathcal{G}_j \right|}{N} \cdot \log_2 \left(\frac{N \left| \widetilde{\mathcal{G}}_k \cap \mathcal{G}_j \right|}{\left| \widetilde{\mathcal{G}}_k \right| \cdot |\mathcal{G}_j|} \right),$$

$H(\widetilde{\mathcal{G}})$ is the entropy of $\widetilde{\mathcal{G}}$ defined as

$$H(\widetilde{\mathcal{G}}) = -\sum_{k=1}^{\widehat{R}} \frac{\left| \widetilde{\mathcal{G}}_k \right|}{N} \log_2 \left(\frac{\left| \widetilde{\mathcal{G}}_k \right|}{N} \right),$$

and $H(\mathcal{G})$ is defined analogously. The closer the values of NMI and purity are to 1, the more accurate the estimated groups are to the true ones. We also look at the estimation accuracy of the functional coefficients for both the preliminary local linear quantile estimator defined in Section 3.1 and the post-grouping local linear quantile estimator defined in Section 3.3. For this we compute the average root mean squared errors (RMSE) defined as

$$\text{RMSE}(\widehat{\boldsymbol{\beta}}) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{T} \sum_{t=1}^{T} \left\| \widehat{\boldsymbol{\beta}}_i(Z_t) - \boldsymbol{\beta}_i(Z_t) \right\|_2^2 \right]^{1/2},$$

for an estimate, $\widehat{\boldsymbol{\beta}}(\cdot) = \left(\widehat{\boldsymbol{\beta}}_1(\cdot), \cdots, \widehat{\boldsymbol{\beta}}_N(\cdot)\right)^\intercal$, of the true functional coefficients $\boldsymbol{\beta}(\cdot) = \left(\boldsymbol{\beta}_1(\cdot), \cdots, \boldsymbol{\beta}_N(\cdot)\right)^\intercal$. As a benchmark, we also compute the RMSE of the oracle estimator, which assumes that the true group structure is known a priori and pools data belonging to each group to obtain group-specific estimates of the functional coefficients. The results for DGP 1 are reported in Table 1 (for $\mathsf{N}(0,1)$ errors), Table 2 (for $t(5)$ errors), and Table 3 (for $0.4\left[\chi^2(3) - 3\right]$ errors).

Table 1 shows that for DGP 1, when the errors follow the $\mathsf{N}(0,1)$ distribution, for the smallest sample size considered (i.e., $N = 50$ and $T = 50$), the ratio criterion (3.3) picks the correct number of groups in about 75% of the replications at $\tau = 0.50$ and about 60% of the replications at $\tau = 0.25$ and 0.75. These percentages increase markedly as T

increases and usually as N increases, but to a lesser extent. When $T = 100$, the correct R_0 is chosen in 100% of the replications at $\tau = 0.50$ and more than 85% of the replications at $\tau = 0.25$ and 0.75. This is consistent with the theoretical result in Theorem 4.2. On the other hand, the NMI values are above 0.77 and purities above 0.91 at all quantiles when $T = 50$, and they increase to above 0.95 when T increases to 100, verifying the consistency results in Theorem 4.1 and Corollary 4.1 for the agglomerative clustering algorithm. The lower block of Table 1 shows that by pooling data belonging to the same group in the estimation, the post-grouping local linear estimator cuts the RMSE of the preliminary local linear estimator by more than 30%, and its RMSE values are not far from those of the oracle estimator. When T increases to 100 and the groups are accurately estimated, the RMSEs of the post-grouping estimator are very close to the oracle estimator. Similar findings can be drawn from Tables 2 and 3, where the errors in DGP 1 follow the heavier tailed $t(5)$ distribution and the asymmetric $0.4\left[\chi^2(3) - 3\right]$ distribution, respectively. The results for $t(5)$ errors are in general worse than those of $N(0, 1)$ and $0.4[\chi^2(3) - 3]$ errors, which may be due to the fact that the $t(5)$ distribution has a larger variance than the other two. For $0.4[\chi^2(3) - 3]$ errors, the results are better than those of $N(0, 1)$ at $\tau = 0.25$, worse than $N(0, 1)$ at $\tau = 0.75$ and comparable at median $\tau = 0.50$. This may be because at $\tau = 0.25$, the $0.4\left[\chi^2(3) - 3\right]$ distribution has more data points than the $N(0, 1)$ distribution and hence, the preliminary functional coefficients estimation and latent group estimation are more accurate. At $\tau = 0.75$, the reverse is true.

In the supplement, we consider additional simulation study with quantile-dependent functional coefficients and group structure and obtain similar results, see Tables D.1–D.3 in Appendix D.2.

5.2 Empirical analysis

To further illustrate the applicability and usefulness of our methods, we next consider a panel house price growth model for UK local authority districts (LADs). Similarly to Chen et al. (2022), we use quarterly house price data over the period Q1/1997 –

Table 1: Simulation results with $N(0,1)$ idiosyncratic errors

		Percentage of replications each $\widehat{R}$ value is selected by the ratio criterion with true $R_0 = 3$ for $\tau = 0.25, 0.50, 0.75$									
T \ N		N = 50					N = 100				
		$\widehat{R}=1$	$\widehat{R}=2$	$\widehat{R}=3$	$\widehat{R}=4$	$\widehat{R}=5$	$\widehat{R}=1$	$\widehat{R}=2$	$\widehat{R}=3$	$\widehat{R}=4$	$\widehat{R}=5$
50	$\tau = 0.25$	0.0%	2.5%	63.0%	27.5%	7.0%	0.0%	32.5%	65.0%	2.5%	0.0%
	$\tau = 0.50$	0.0%	2.5%	74.5%	19.0%	4.0%	0.0%	2.0%	92.5%	4.5%	1.0%
	$\tau = 0.75$	0.0%	42.5%	56.5%	1.0%	0.0%	0.0%	4.0%	89.0%	7.0%	0.0%
100	$\tau = 0.25$	0.0%	0.0%	96.5%	3.5%	0.0%	0.0%	0.0%	99.5%	0.5%	0.0%
	$\tau = 0.50$	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
	$\tau = 0.75$	0.0%	15.0%	85.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%

		Average NMI and purity (with standard deviation in parentheses)			
T \ N		N = 50		N = 100	
		NMI	Purity	NMI	Purity
50	$\tau = 0.25$	0.8363 (0.1305)	0.9174 (0.0868)	0.7807 (0.1466)	0.9556 (0.0552)
	$\tau = 0.50$	0.8923 (0.1084)	0.9508 (0.0644)	0.8994 (0.0956)	0.9646 (0.0468)
	$\tau = 0.75$	0.7766 (0.1416)	0.9647 (0.0502)	0.8347 (0.1385)	0.9452 (0.0575)
100	$\tau = 0.25$	0.9915 (0.0356)	0.9953 (0.0202)	0.9938 (0.0257)	0.9973 (0.0156)
	$\tau = 0.50$	0.9986 (0.0099)	0.9996 (0.0028)	0.9986 (0.0082)	0.9997 (0.0021)
	$\tau = 0.75$	0.9522 (0.1021)	0.9987 (0.0053)	0.9921 (0.0328)	0.9975 (0.0122)

		Average RMSE of $\beta(\cdot)$ estimates (with standard deviation in parentheses)					
T \ N		N = 50			N = 100		
		Oracle	Preliminary	Post-grouping	Oracle	Preliminary	Post-grouping
50	$\tau = 0.25$	0.4211 (0.0213)	0.7085 (0.0280)	0.4770 (0.0653)	0.3877 (0.0181)	0.7105 (0.0235)	0.4910 (0.0906)
	$\tau = 0.50$	0.3783 (0.0195)	0.6608 (0.0282)	0.4190 (0.0602)	0.3693 (0.0170)	0.6646 (0.0204)	0.4025 (0.0497)
	$\tau = 0.75$	0.3866 (0.0191)	0.7038 (0.0287)	0.5057 (0.0924)	0.4003 (0.0176)	0.7114 (0.0214)	0.4598 (0.0727)
100	$\tau = 0.25$	0.3028 (0.0124)	0.5213 (0.0166)	0.3056 (0.0205)	0.2742 (0.0091)	0.5160 (0.0129)	0.2762 (0.0146)
	$\tau = 0.50$	0.2407 (0.0131)	0.4805 (0.0170)	0.2410 (0.0135)	0.2291 (0.0090)	0.4783 (0.0129)	0.2294 (0.0090)
	$\tau = 0.75$	0.2675 (0.0134)	0.5171 (0.0170)	0.3016 (0.0792)	0.2753 (0.0093)	0.5170 (0.0128)	0.2780 (0.0186)

Table 2: Simulation results with $t(5)$ idiosyncratic errors

T \ N		Percentage of replications each $\widehat{R}$ value is selected by the ratio criterion with true $R_0 = 3$ for $\tau = 0.25, 0.50, 0.75$									
		N = 50					N = 100				
		$\widehat{R}=1$	$\widehat{R}=2$	$\widehat{R}=3$	$\widehat{R}=4$	$\widehat{R}=5$	$\widehat{R}=1$	$\widehat{R}=2$	$\widehat{R}=3$	$\widehat{R}=4$	$\widehat{R}=5$
50	$\tau=0.25$	0.0%	7.5%	67.0%	21.0%	4.5%	0.5%	4.5%	59.0%	28.0%	8.0%
	$\tau=0.50$	0.0%	23.0%	70.0%	6.5%	0.5%	0.0%	17.5%	78.0%	4.5%	0.0%
	$\tau=0.75$	0.0%	15.5%	71.5%	10.0%	1.0%	0.0%	15.5%	71.5%	10.5%	2.5%
100	$\tau=0.25$	0.0%	0.5%	85.5%	12.0%	2.0%	0.0%	0.0%	94.5%	5.0%	0.5%
	$\tau=0.50$	0.0%	0.0%	95.5%	4.5%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
	$\tau=0.75$	0.0%	0.0%	91.5%	6.0%	2.5%	0.0%	0.0%	98.5%	1.5%	0.0%

T \ N		Average NMI and purity (with standard deviation in parentheses)			
		N = 50		N = 100	
		NMI	Purity	NMI	Purity
50	$\tau=0.25$	0.7272 (0.1414)	0.8885 (0.0897)	0.6954 (0.1689)	0.8770 (0.0943)
	$\tau=0.50$	0.7940 (0.1641)	0.9497 (0.0618)	0.8026 (0.1364)	0.9494 (0.0589)
	$\tau=0.75$	0.7292 (0.1555)	0.9026 (0.0857)	0.6839 (0.1474)	0.9018 (0.0835)
100	$\tau=0.25$	0.9601 (0.0892)	0.9786 (0.0483)	0.9783 (0.0430)	0.9915 (0.0227)
	$\tau=0.50$	0.9922 (0.0263)	0.9949 (0.0220)	0.9946 (0.0177)	0.9986 (0.0054)
	$\tau=0.75$	0.9690 (0.0617)	0.9850 (0.0382)	0.9768 (0.0505)	0.9926 (0.0218)

T \ N		Average RMSE of $\beta(\cdot)$ estimates (with standard deviation in parentheses)					
		N = 50			N = 100		
		Oracle	Preliminary	Post-grouping	Oracle	Preliminary	Post-grouping
50	$\tau=0.25$	0.4269 (0.0202)	0.7903 (0.0358)	0.5386 (0.0783)	0.4259 (0.0176)	0.7914 (0.0249)	0.5401 (0.0928)
	$\tau=0.50$	0.3999 (0.0207)	0.7178 (0.0308)	0.4980 (0.0967)	0.3776 (0.0182)	0.7195 (0.0223)	0.4646 (0.0850)
	$\tau=0.75$	0.4187 (0.0221)	0.7935 (0.0381)	0.5340 (0.0864)	0.4095 (0.0174)	0.7949 (0.0266)	0.5462 (0.0922)
100	$\tau=0.25$	0.3206 (0.0139)	0.5819 (0.0217)	0.33421 (0.0471)	0.3117 (0.0104)	0.5804 (0.0161)	0.3184 (0.0259)
	$\tau=0.50$	0.2753 (0.0130)	0.5185 (0.0189)	0.2778 (0.0156)	0.2673 (0.0098)	0.5183 (0.0148)	0.2686 (0.0110)
	$\tau=0.75$	0.3204 (0.0143)	0.5825 (0.0206)	0.3324 (0.0377)	0.2955 (0.0102)	0.5810 (0.0165)	0.3037 (0.0305)

Table 3: Simulation results with $0.4 * [\chi^2(3) - 3]$ idiosyncratic errors

		Percentage of replications each $\widehat{R}$ value is selected by the ratio criterion with true $R_0 = 3$ for $\tau = 0.25, 0.50, 0.75$									
T \ N		N = 50					N = 100				
		$\widehat{R} = 1$	$\widehat{R} = 2$	$\widehat{R} = 3$	$\widehat{R} = 4$	$\widehat{R} = 5$	$\widehat{R} = 1$	$\widehat{R} = 2$	$\widehat{R} = 3$	$\widehat{R} = 4$	$\widehat{R} = 5$
50	$\tau = 0.25$	0.0%	0.0%	95.5%	4.5%	0.0%	0.0%	0.5%	97.5%	2.0%	0.0%
	$\tau = 0.50$	0.0%	1.0%	69.0%	24.5%	5.5%	0.0%	0.5%	96.0%	3.5%	0.0%
	$\tau = 0.75$	0.0%	11.0%	66.5%	18.5%	4.0%	0.0%	22.5%	70.5%	6.5%	0.5%
100	$\tau = 0.25$	0.0%	0.5%	99.5%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
	$\tau = 0.50$	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
	$\tau = 0.75$	0.0%	0.0%	98.0%	2.0%	0.0%	0.0%	0.0%	98.0%	1.5%	0.5%

		Average NMI and purity (with standard deviation in parentheses)			
T \ N		N = 50		N = 100	
		NMI	Purity	NMI	Purity
50	$\tau = 0.25$	0.9766 (0.0574)	0.9927 (0.0182)	0.9748 (0.0691)	0.9945 (0.0187)
	$\tau = 0.50$	0.9382 (0.0789)	0.9608 (0.0595)	0.9477 (0.0831)	0.9827 (0.0354)
	$\tau = 0.75$	0.7418 (0.1690)	0.9010 (0.0873)	0.7469 (0.1573)	0.9342 (0.0735)
100	$\tau = 0.25$	0.9986 (0.0199)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	$\tau = 0.50$	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	$\tau = 0.75$	0.9811 (0.0502)	0.9933 (0.0195)	0.9755 (0.0554)	0.9904 (0.0299)

		Average RMSE of $\beta(\cdot)$ estimates (with standard deviation in parentheses)					
T \ N		N = 50			N = 100		
		Oracle	Preliminary	Post-grouping	Oracle	Preliminary	Post-grouping
50	$\tau = 0.25$	0.3259 (0.0191)	0.5420 (0.0267)	0.3363 (0.0436)	0.3212 (0.0171)	0.5420 (0.0201)	0.3334 (0.0470)
	$\tau = 0.50$	0.3526 (0.0189)	0.6041 (0.0278)	0.3736 (0.0448)	0.3300 (0.0162)	0.6007 (0.0223)	0.3498 (0.0498)
	$\tau = 0.75$	0.4172 (0.0220)	0.7685 (0.0380)	0.5251 (0.0935)	0.3944 (0.0158)	0.7638 (0.0238)	0.5039 (0.0944)
100	$\tau = 0.25$	0.1953 (0.0089)	0.3531 (0.0136)	0.1964 (0.0204)	0.1911 (0.0072)	0.3534 (0.0113)	0.1911 (0.0072)
	$\tau = 0.50$	0.2109 (0.0103)	0.4241 (0.0175)	0.2109 (0.0103)	0.2027 (0.0094)	0.4210 (0.0134)	0.2027 (0.0094)
	$\tau = 0.75$	0.3119 (0.0136)	0.5704 (0.0192)	0.3188 (0.0290)	0.3005 (0.0098)	0.5686 (0.0162)	0.3093 (0.0302)

Q4/2016[3], downloaded from the UK Office of National Statistics (ONS) website: https://www.ons.gov.uk/. Growth rates in population and the nominal per capita personal income are used as the explanatory variables. Their data at the individual LAD level are available at the annual rate on the ONS website[4], and we construct quarterly data for the two variables using the interpolation method in Denton (1971) and Dagum and Cholette (2006). For the index variable, we use quarterly inflation rates, which are available at the country level at https://www.bls.gov/cpi/data.htm. This allows for interaction between inflation and the explanatory variables and for the effects of the explanatory variables on house prices to vary with inflation. Data for all the variables have been de-seasoned and de-trended[5].

Assume the following panel quantile regression:

$$Q_{\tau,i}(hp_{it}|pop_{it}, inc_{i,t-1}, inf_t, \alpha_i) = \alpha_i + \beta_{i,1}(inf_t)pop_{it} + \beta_{i,2}(inf_t)inc_{i,t-1}, \qquad (5.2)$$

for $i = 1, \cdots, 335$ and $t = 2, \cdots, 80$, where hp_{it} is the house price growth rate (in %) for the i-th LAD in the t-th quarter, pop_{it} is the population growth (in %), $inc_{i,t-1}$ is the growth in per capita personal income (in %), inf_t is the inflation rate (in %), and α_i is the fixed effect. Here inc is lagged by one time period due to the likely lagged effect of income growth on house price.

While the model (5.2) offers great flexibility by allowing the coefficients to vary across LADs and quantiles as well as with inflation, there may exist some homogeneity groups of LADs at each quantile level, where the coefficients are homogeneous within each group (while still varying with inflation) but heterogeneous across groups. That is, at a given quantile level τ, there may exist a partition of the index set $\{1, 2, , \cdots, 335\}$, denoted as $\{\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_{R_0}\}$, such that

$$\boldsymbol{\beta}_i(\cdot) := \begin{pmatrix} \beta_{i,1}(\cdot) \\ \beta_{i,2}(\cdot) \end{pmatrix} = \boldsymbol{\gamma}_j(\cdot) := \begin{pmatrix} \gamma_{j,1}(\cdot) \\ \gamma_{j,2}(\cdot) \end{pmatrix} \text{ for } i \in \mathcal{G}_j, \quad i = 1, \cdots, 335, \ j = 1, \cdots, R_0.$$

[3]Due to data unavailability, we consider LADs only in England and Wales.

[4]Four LADs from England and Wales - Aylesbury Vale, Gloucester, Norwich, and Powys, are excluded due to their outlying values. This gives a total of 335 LADs for the subsequent analysis.

[5]More data details can be found in Chen et al. (2022).

We consider 3 quantile levels, i.e., $\tau = 0.25, 0.50, 0.75$, and at each quantile, we then use the methods proposed in Section 3 to estimate the number of latent groups as well as the group membership. The results are summarised in Table 4. The post-grouping local linear estimates of the group-specific functional coefficients, together with their 95% confidence intervals, are plotted in Figures 1–3 for $\tau = 0.25, 0.50, 0.75$, respectively. The confidence intervals are computed using the plug-in estimates of the asymptotic bias and variance matrix derived in Theorem 4.3 (see Appendix C.3 in the supplement for detail). The estimated groups at each quantile level, projected onto a choropleth map, are shown in Figures 4–6.

At both $\tau = 0.25$ and $\tau = 0.50$, two groups of LADs are identified, although the membership of the groups is not exactly the same[6]. Furthermore, Figures 1 and 2 show that the coefficient functions for each group have similar patterns at these two quantiles, especially for that of population growth. The coefficients are mostly positive, consistent with the economic theory that growth in population and income leads to growth in demand for housing and hence, a rise in house price. At $\tau = 0.25$, the effects of population and income growth in general decrease as inflation increases. This decreasing trend is more marked for the effect of population growth in Group 1 LADs (88% of which are non-metropolitan districts or unitary authorities), but much less so in Group 2 LADs (which include 88% of the London boroughs and 89% of the metropolitan districts). For example, when inflation rate is -0.5%, for each 1% increase in population growth, house price growth is increased by around 8%. But when inflation rate is 0.5%, a 1% increase in population growth leads only to a 2% increase in house price growth. At $\tau = 0.50$, we observe a similar trend. We can also find, from Figures 1 and 2, that at $\tau = 0.25$ and $\tau = 0.50$, the effect of population growth on house price growth dominates that of income growth for Group 1 LADs, which are mainly non-metropolitan or unitary districts. At $\tau = 0.75$, where five groups are identified, we see some different trends for different groups (see Figure 3). While most values of the coefficient functions are posi-

[6]The membership of the two groups at $\tau = 0.50$ is similar to that of the two groups at $\tau = 0.25$ with a large number of overlapping member LADs (e.g. for Group 1, there are 45 overlapping LADs and for Group 2, there are 247 overlapping LADs).

Table 4: Estimated latent groups for house price data from 335 UK LADs

Quantile	Number of groups	Cardinality of each group
$\tau = 0.25$	$\widehat{R} = 2$	69, 266
$\tau = 0.50$	$\widehat{R} = 2$	64, 271
$\tau = 0.75$	$\widehat{R} = 5$	4, 59, 45, 204, 23

tive, there are negative values for some groups in some subintervals. For example, the coefficient function of population growth for Group 1 exhibits a downward sloping pattern with increasing inflation and is negative when inflation is above -0.7%. A closer examination of the member LADs reveals that three out of the four LADs of this group are urban districts with major or minor conurbation, where immigration are more likely to occur. If population growth is driven by migration inflows, its effect on house price growth might be ambiguous and sometimes even negative (Sá, 2015). Chen et al. (2022) also find negative effects of population growth on house price growth in some LADs.

For comparison, we also conduct a grouping analysis based on the functional-coefficient mean regression for the same variables. The same ratio criterion and HAC algorithm are used for choosing the number of groups and estimating the group membership. The obtained results are similar to those from the quantile analysis at $\tau = 0.50$: two groups are found with similar, although not exactly the same, membership to those from the quantile regression at $\tau = 0.50$. Results from the mean regression are more susceptible to the influence of outliers. More detail about the mean regression grouping results can be found in Appendix D.1 of the online supplement.

The above analysis reveals that at $\tau = 0.25$ and $\tau = 0.50$, there is more homogeneity in the effects of population and income growth on house price growth across LADs, and in general these effects are positive and decrease as inflation increases. At $\tau = 0.75$, more heterogeneity is observed, and for some identified groups the effects of population and income growth are negative for some values of inflation. It also appears that at $\tau = 0.75$, the effect of income growth is larger than at lower quantiles. This empirical application demonstrates the benefit of the quantile grouping analysis: it can shed more light on the impact of population and income growth on house price growth across LADs than a mean regression analysis or a mean regression based grouping analysis does. The

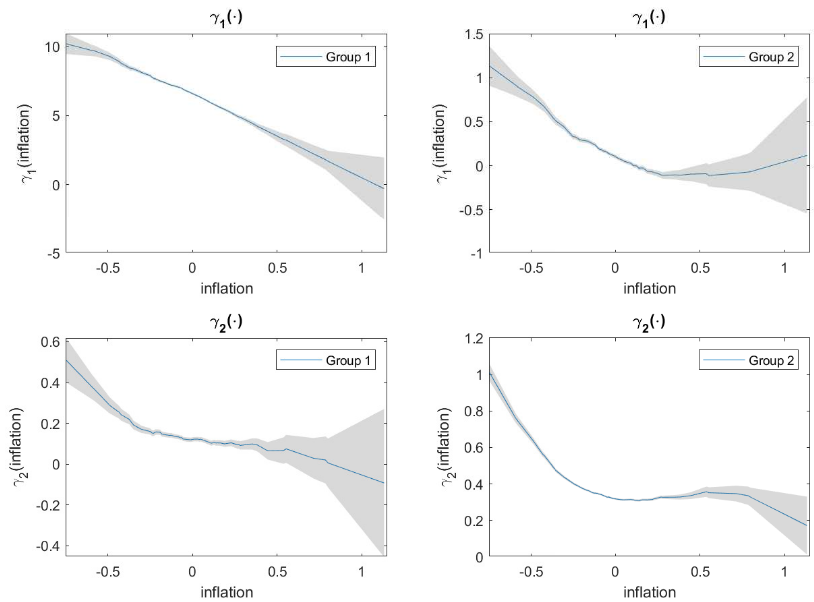

Figure 1: Post-grouping local linear estimates of the functional coefficients (with 95% confidence intervals) for population growth ($\gamma_1(\cdot)$, top row) and income growth ($\gamma_2(\cdot)$, bottom row) at $\tau = 0.25$: left - for Group 1; right - for Group 2.

results may be useful for policy makers in developing more targeted policies for specific groups of districts at specific quantiles.

6 Conclusion

In this paper, we propose a general functional-coefficient quantile regression model for large panel data and assume a latent group structure on the heterogenous functional coefficients. An estimation methodology which combines preliminary functional coefficient estimates (ignoring the latent group structure), an agglomerative clustering algorithm and a simple ratio criterion is introduced to consistently estimate the group number and membership. Furthermore, a post-grouping local linear quantile regression method is used to estimate the group-specific functional coefficients, aiming to achieve faster convergence rates than the preliminary local linear estimator. To asymptotically remove the influence of nuisance parameters and derive an asymptotic normal distribution theory comparable to that in the literature such as Kato et al. (2012), we impose a

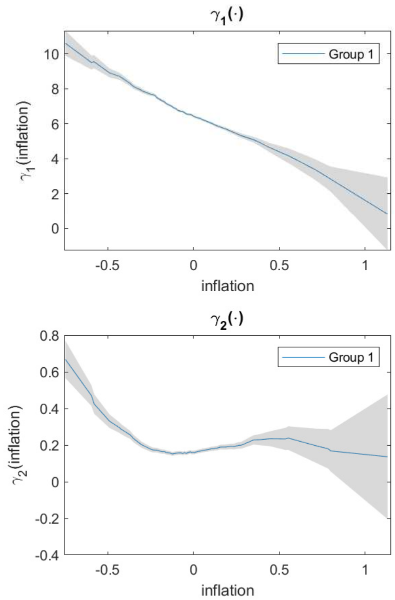

Figure 2: Post-grouping local linear estimates of the functional coefficients (with 95% confidence intervals) for population growth ($\gamma_1(\cdot)$, top row) and income growth ($\gamma_2(\cdot)$, bottom row) at $\tau = 0.50$: left - for Group 1; right - for Group 2.

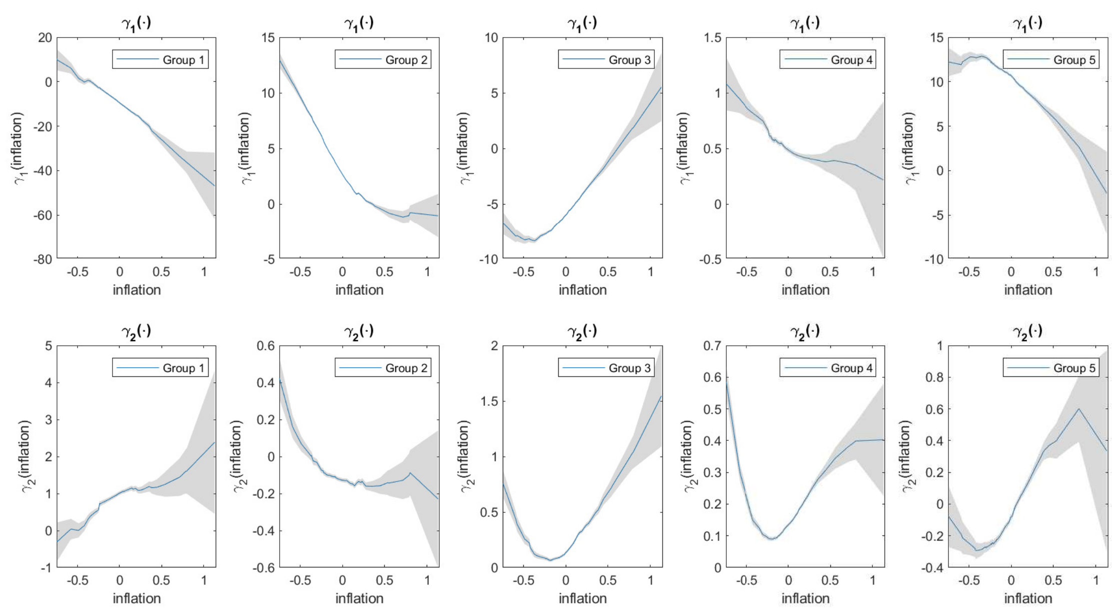

Figure 3: Post-grouping local linear estimates of the functional coefficients (with 95% confidence intervals) for population growth ($\gamma_1(\cdot)$, top row) and income growth ($\gamma_2(\cdot)$, bottom row) at $\tau = 0.75$: from left to right - Group 1, Group 2, Group 3, Group 4, Group 5.

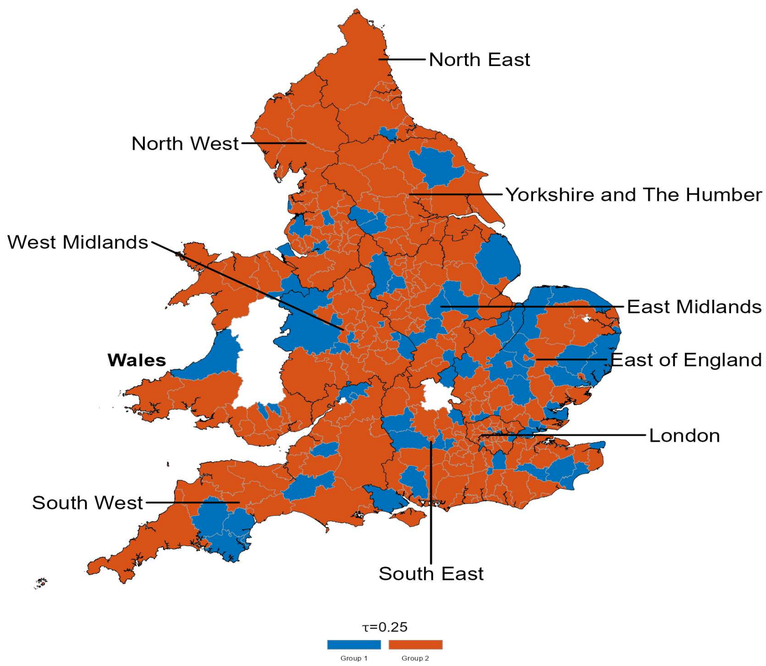

Figure 4: Spatial pattern of estimated groups for $\tau = 0.25$: blue LADs - Group 1; red LADs - Group 2.

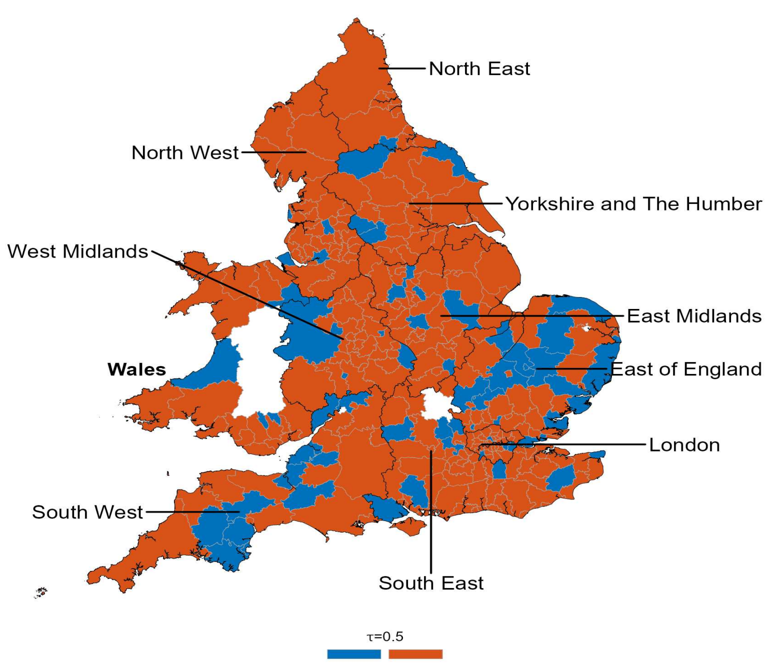

Figure 5: Spatial pattern of estimated groups for $\tau = 0.50$: blue LADs - Group 1; red LADs - Group 2.

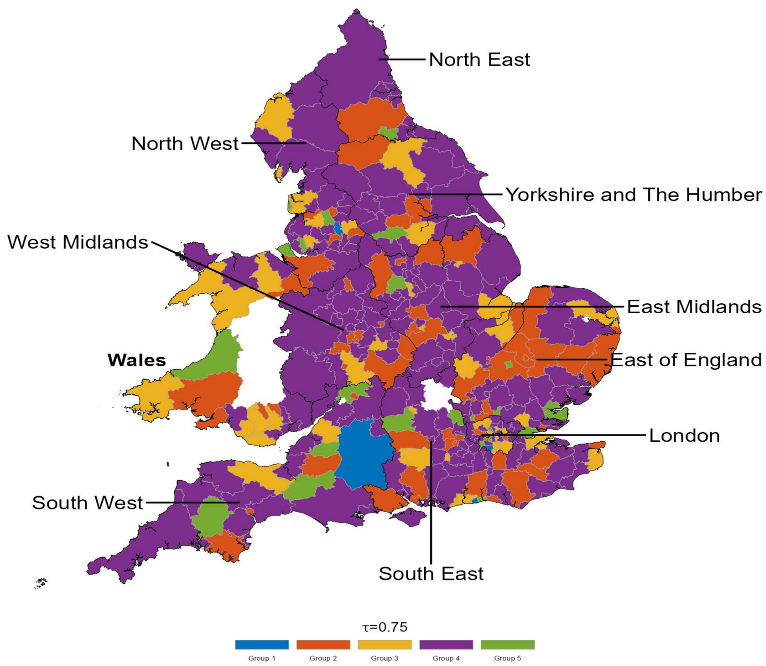

Figure 6: Spatial pattern of estimated groups for $\tau = 0.75$: blue LADs - Group 1; red LADs - Group 2; yellow LADs - Group 3; purple LADs - Group 4; green LADs - Group 5.

relatively restrictive condition on the divergence rate of the group size, but allow weak cross-sectional and temporal dependence for large panel observations. The simulation studies show that the proposed methods have reliable finite-sample performance. The empirical application to the UK house price data reveals that the latent structures vary over different quantile levels with more heterogeneity observed at the upper quartile. In addition, the methodology is modified to estimate the latent group structure in linear panel quantile regression uniformly over quantile levels, and the main methodology and theory are applicable to the time-varying coefficient panel quantile regression with the latent structure.

Acknowledgements

The authors would like to thank an Associate Editor and two reviewers for the constructive comments, which helped to substantially improve the paper. The authors also thank Dr Chaowen Zheng for collecting the empirical data and preparing Figures 4-6. The first author is partially supported by Natural Science Foundation of Zhejiang Province (No. LY22A010006), National Social Science Foundation of China (No. 17BTJ027) and the Fundamental Research Funds for Provincial Universities of Zhejiang. The second author is partially supported by the Economic and Social Research Council in the UK (No. ES/T01573X/1). The third author is partially supported by the National Natural Science Foundation of China (No. 72033002). The usual disclaimer applies.

SUPPLEMENT

The supplemental document contains proofs of the main asymptotic theorems, technical lemmas with proofs, some extensions of the developed method and theory, and additional simulation and empirical results.

References

Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*(3), 1203–1227.

Ando, T. and J. Bai (2017). Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures. *Journal of the American Statistical Association 112*(519), 1182–1198.

Belloni, A., V. Chernozhukov, D. Chetverikov, and I. Fernández-Val (2019). Conditional quantile processes based on series or many regressors. *Journal of Econometrics 213*(1), 4–29.

Bonhomme, S. and E. Manresa (2015). Grouped patterns of heterogeneity in panel data. *Econometrica 83*(3), 1147–1184.

Cai, Z. (2002). Regression quantiles for time series. *Econometric Theory 18*(1), 169–192.

Cai, Z., L. Chen, and Y. Fang (2018). A semiparametric quantile panel data model with an application to estimating the growth effect of fdi. *Journal of Econometrics 206*(2), 531–553.

Cai, Z. and X. Xu (2008). Nonparametric quantile estimations for dynamic smooth coefficient models. *Journal of the American Statistical Association 103*(484), 1595–1608.

Canay, I. A. (2011). A simple approach to quantile regression for panel data. *The Econometrics Journal 14*(3), 368–386.

Chen, J. (2019). Estimating latent group structure in time-varying coefficient panel data models. *The Econometrics Journal 22*(3), 223–240.

Chen, J., D. Li, L. Wei, and W. Zhang (2021). Nonparametric homogeneity pursuit in functional-coefficient models. *Journal of Nonparametric Statistics 33*(3-4), 387–416.

Chen, J., Y. Shin, and C. Zheng (2022). Estimation and inference in heterogeneous spatial panels with a multifactor error structure. *Journal of Econometrics 229*(1), 55–79.

Chen, L. (2021). Nonparametric quantile regressions for panel data models with large T. *Working paper available at https://english.phbs.pku.edu.cn/uploadfile/2021/0128/20210 128073111708.pdf*.

Chernozhukov, V. and C. Hansen (2006). Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics 132*(2), 491–525.

Chetverikov, D., B. Larsen, and C. Palmer (2016). Iv quantile regression for group-level treatments, with an application to the distributional effects of trade. *Econometrica 84*(2), 809—833.

Dagum, E. B. and P. A. Cholette (2006). *Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series*, Volume 186 of *Lecture Notes in Statistics*. Springer.

Denton, F. T. (1971). Adjustment of monthly or quarterly series to annual totals: an approach based on quadratic minimization. *Journal of the American Statistical Association 66*(333), 99–102.

Everitt, B. S., S. Landau, M. Leese, and D. Stahl (2011). *Cluster Analysis (5th Edition)*. Wiley Series in Probability and Statistics, Wiley.

Galvao, A. F., J. Gu, and S. Volgushev (2020). On the unbiased asymptotic normality of quantile regression with fixed effects. *Journal of Econometrics 218*(1), 178–215.

Galvao, A. F. and K. Kato (2016). Smoothed quantile regression for panel data. *Journal of Econometrics 193*(1), 92–112.

Galvao, A. F., C. Lamarche, and L. R. Lima (2013). Estimation of censored quantile regression for panel data with fixed effects. *Journal of the American Statistical Association 108*(503), 1075–1089.

Gu, J. and S. Volgushev (2019). Panel data quantile regression with grouped fixed effects. *Journal of Econometrics 213*(1), 68–91.

Hahn, J. and H. R. Moon (2010). Panel data models with finite number of multiple equilibria. *Econometric Theory 26*(3), 863–881.

Honda, T. (2004). Quantile regression in varying coefficient models. *Journal of Statistical Planning and Inference 121*(1), 113–125.

Kai, B., R. Li, and H. Zou (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *The Annals of Statistics 39*(1), 305–332.

Kato, K. (2012). Estimation in functional linear quantile regression. *The Annals of Statistics 40*(6), 3108–3136.

Kato, K., A. F. Galvao Jr, and G. V. Montes-Rojas (2012). Asymptotics for panel quantile regression models with individual effects. *Journal of Econometrics 170*(1), 76–91.

Ke, Y., J. Li, and W. Zhang (2016). Structure identification in panel data analysis. *The Annals of Statistics 44*(3), 1193–1233.

Kim, M.-O. (2007). Quantile regression with varying coefficients. *The Annals of Statistics 35*(1), 92–108.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis 91*(1), 74–89.

Koenker, R. (2005). *Quantile regression*. Cambridge University Press.

Koenker, R. and G. Bassett (1978). Quantile regression. *Econometrica 46*(1), 33–50.

Koenker, R., V. Chernozhukov, X. He, and L. Peng (2017). *Handbook of Quantile Regression*. Chapman and Hall/CRC.

Koenker, R. and Z. Xiao (2006). Quantile autoregression. *Journal of the American Statistical Association 101*(475), 980–990.

Lam, C. and Q. Yao (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics 40*(2), 694–726.

Li, D., J. Chen, and J. Gao (2011). Non-parametric time-varying coefficient panel data models with fixed effects. *The Econometrics Journal 14*(3), 387–408.

Li, D., Q. Li, and Z. Li (2021). Nonparametric quantile regression estimation with mixed discrete and continuous data. *Journal of Business & Economic Statistics 39*(3), 741–756.

Li, D., P. M. Robinson, and H. L. Shang (2020). Long-range dependent curve time series. *Journal of the American Statistical Association 115*(530), 957–971.

Li, Q., J. Lin, and J. S. Racine (2013). Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. *Journal of Business & Economic Statistics 31*(1), 57–65.

Lian, H., X. Qiao, and W. Zhang (2021). Homogeneity pursuit in single index models based panel data analysis. *Journal of Business & Economic Statistics 39*(2), 386–401.

Liu, R., Z. Shang, Y. Zhang, and Q. Zhou (2020). Identification and estimation in panel models with overspecified number of groups. *Journal of Econometrics 215*(2), 574–590.

Phillips, P. C. and D. Sul (2007). Transition modeling and econometric convergence tests. *Econometrica 75*(6), 1771–1855.

Phillips, P. C. and Y. Wang (2022). Functional coefficient panel modeling with communal smoothing covariates. *Journal of Econometrics 227*(2), 371–407.

Rencher, A. C. and W. F. Christensen (2012). *Methods of Multivariate Analysis (3rd Edition)*. Wiley Series in Probability and Statistics, Wiley.

Sá, F. (2015). Immigration and house prices in the uk. *The Economic Journal 125*(587), 1393–1424.

Su, L. and T. Hoshino (2016). Sieve instrumental variable quantile regression estimation of functional coefficient models. *Journal of Econometrics 191*(1), 231–254.

Su, L., Z. Shi, and P. C. Phillips (2016). Identifying latent structures in panel data. *Econometrica 84*(6), 2215–2264.

Su, L., X. Wang, and S. Jin (2019). Sieve estimation of time-varying panel data models with latent structures. *Journal of Business & Economic Statistics 37*(2), 334–349.

Tang, Y., X. Song, H. J. Wang, and Z. Zhu (2013). Variable selection in high-dimensional quantile varying coefficient models. *Journal of Multivariate Analysis 122*, 115–132.

Vogt, M. and O. Linton (2017). Classification of non-parametric regression functions in longitudinal data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(1), 5–27.

Vogt, M. and O. Linton (2020). Multiscale clustering of nonparametric regression curves. *Journal of Econometrics 216*(1), 305–325.

Wang, H. J., Z. Zhu, and J. Zhou (2009). Quantile regression in partially linear varying coefficient models. *The Annals of Statistics 37*(6B), 3841–3866.

Wang, W. and L. Su (2021). Identifying latent group structures in nonlinear panels. *Journal of Econometrics 220*(2), 272–295.

Yan, K. X. and Q. Li (2018). Nonparametric estimation of a conditional quantile function in a fixed effects panel data model. *Journal of Risk and Financial Management 11*(3), 44.

Yu, K. and M. Jones (1998). Local linear quantile regression. *Journal of the American statistical Association 93*(441), 228–237.

Yu, K. and Z. Lu (2004). Local linear additive quantile regression. *Scandinavian Journal of Statistics 31*(3), 333–346.

Zhang, Y., H. J. Wang, and Z. Zhu (2019a). Quantile-regression-based clustering for panel data. *Journal of Econometrics 213*(1), 54–67.

Zhang, Y., H. J. Wang, and Z. Zhu (2019b). Robust subgroup identification. *Statistica Sinica 29*(4), 1873–1889.