Article:
Sun, Yan, Wan, Chuang, Zhang, Wenyang et al. (2024) A Multi-Kink Quantile Regression Model with Common Structure for Panel Data Analysis. Journal of Econometrics. 105304. ISSN: 0304-4076

https://doi.org/10.1016/j.jeconom.2022.04.012

A Multi-Kink Quantile Regression Model with Common Structure for Panel Data Analysis

Yan Sun[1], Chuang Wan[2], Wenyang Zhang[3] and Wei Zhong[2*]

Shanghai University of Finance and Economics[1],

Xiamen University[2] and The University of York[3]

April 30, 2022

Abstract

Stimulated by the analysis of a data set on financial portfolio returns, we propose a multi-kink quantile regression (MKQR) model with latent homogeneous structure for panel data analysis. The proposed model accounts for both homogeneity and heterogeneity among individuals and parameters in panel data analysis. From statistical modelling point of view, it well balances the risk of misspecification and the model parsimony. From practical point of view, it is able to reveal not only the impacts of covariates in the global sense, but also individual attributes. An estimation procedure is presented to estimate both the unknown parameters and the latent homogeneous structure in the proposed model. Computational issues with the implementation of the estimation procedure are also discussed. Asymptotic theory of the estimators is established. It shows the necessity of taking into account both homogeneity and heterogeneity in panel data analysis. Monte Carlo simulation studies are conducted to demonstrate the finite sample performance of the proposed estimation and the risk of ignoring the homogeneity or heterogeneity among individuals. Finally, we apply the

*Wei Zhong is the corresponding author, Email: wzhong@xmu.edu.cn.

proposed model and the estimation procedure to the data set which stimulates this work and reveal some interesting findings.

Keywords: Binary segmentation, common structure, homogeneity pursuit, multi-kink quantile regression, panel data analysis.

Running Head: MKQR with Common Structure

1 Introduction

In data analysis, the most important objective is to find a suitable model which can best accommodate to the intrinsic characteristics of the data concerned. Panel data analysis allows us to explore economic processes while accounting for both dynamic effects across times and heterogeneity across individuals (Hsiao, 2014; Greene, 2018). Although linear models have been commonly used in panel data analysis, there are two critical assumptions. The first one is the linearity assumption on the relationship between a response and regressors. It suffers from the risk of the model mis-specification problem which leads to both estimation biases and prediction inaccuracy. Some semiparametric and nonparametric models have been studied in the literature. For instance, Li et al. (2018) and Guo and Li (2022) proposed a nonparametric transformed Fama-French three factors model to better account for asset returns. Su et al. (2019) introduced a sieve-estimation-based procedure to estimate the functional coefficients for time-varying panel data models. Lian et al. (2021) studied the single index models embedded with homogeneity for panel data analysis where the conditional mean of the response is an unknown smooth function of the linear combination of regressors. Although these semiparametric and nonparametric models are able to capture the nonlinearity, they may lack the model interpretability and ignore some prior important information about the underlying relationship between the response and the covariates. On the other hand, the kink regression models provide a balance between linear models and nonparametric models. Hansen (2017) proposed a kink regression model which assumes that the relationship between the response and the threshold regressor is piecewise linear at two

different subregions but continuous at an unknown threshold. Zhang et al. (2017) extended the kink regression with an unknown threshold to the panel data framework.

The second one is the constant slope assumption that requires the slopes of regressors to be same over all individuals. The advantage of panel data analysis over cross-sectional data analysis is that it is flexible in modeling differences in behavior across individuals. Although the individual fixed effect can account for some heterogeneity across individuals, the constant slopes for all individuals ignore some individual attributes which are important in some applications such as personalized medicine, individualized recommendation and among others. On the other hand, if the complete slope heterogeneity that allows the slops to vary across individuals is assumed, too many unknown parameters are involved and the potential homogeneous structure may be ignored. As the result, the model is not parsimonious and the design matrix might be ill-posed. In practice, applied researchers first define different small homogeneous groups of individuals where individuals in each group share the same slope parameters and then conduct statistical modelling for each group. However, the estimation accuracy heavily depends on the quality of the prespecified subgroups. To detect the homogeneous groups in a data-driven way, some authors (Lin and Ng, 2012; Bonhomme and Manresa, 2015; Ando and Bai, 2016; Zhang et al., 2019) developed the K-means-clustering-based algorithms for panel data models. However, the number of subgroups is still needed to be prespecified. To this end, the regularization methods (Tibshirani et al., 2005; Fan et al., 2020) have been proposed to force similar parameters to be clustered into the same groups. For example, Su et al. (2016) proposed a classifier-LASSO based on an additive-multiplicative penalty to identify the subgroup memberships in panel data. Ke et al. (2015) proposed a clustering algorithm in regression via data-driven segmentation (CARDS) to explore the coefficients homogeneity. Wang et al. (2018) extended the CARDS approach to homogeneity pursuit in panel data models. Ke et al. (2016) introduced a computationally efficient binary segmentation approach to detect the slope homogeneity in panel data linear models.

These considerations motivate us to study a flexible modeling method with homogeneous structure for panel data analysis. In this paper, we consider a Multi-Kink Quantile Re-

gression (MKQR) model with unknown kink points for panel data. The key advantage of the kink regression is that it possesses both the interpretability of linear models and the flexibility of nonparametric regression models. To explore the individual heterogeneity, we first assign different parameters for each individual and obtain the preliminary estimates. Then, we sequentially detect homogeneous structures for different types of parameters using the binary segmentation approach. The final model can both account for necessary individual heterogeneity and detect the latent homogeneous structure to achieve better estimation efficiency and prediction accuracy. Furthermore, we establish the asymptotic properties of the estimators under three different scenarios: the overfitting case where the homogeneous structure is ignored, the correct fitting case which is the proposed method, and the underfitting case which considers all individuals as one homogeneous group. In the real data application, we extend the Fama-French three-factor model with latent homogeneous structure using the daily returns data of industry portfolios. We demonstrate that incorporating the homogeneity information can really improve the predictability.

Compared with existing methods in the literature, the proposed method has the following distinctive features. First, we extend kink regression models for independent data (Hansen, 2017; Zhong et al., 2021) to panel data and consider the latent homogeneous structure detection. To the best of our knowledge, it is the first work for homogeneity pursuit for kink regression models. Second, compared with the existing binary segmentation methods for homogeneity pursuit based on least squares estimation (Ke et al., 2016; Guo and Li, 2022), we consider the quantile regression for the multi-kink regression model and derive its associated theoretical properties. Quantile regression is able to depict the conditional distribution of the response given the covariates and is robust to analyze data with heterogeneous errors and outliers, especially financial data.

The rest of the paper is organized as follows. we first introduce the proposed MKQR model with homogeneous structure for panel data in Section 2. An estimation procedure as well as the computational algorithm to implement it are described in Section 3. Asymptotic properties are presented in Section 4. Simulation studies are conducted in Section 5. In Section 6, we apply the proposed model to a financial data set on industry portfolios.

Conditions, technical proofs and additional numerical studies are left in the Supplement.

2 Modelling Strategy

First, we introduce the Multi-Kink Quantile Regression (MKQR) model with latent homogeneous structure for panel data. Let x_{ij} be a threshold covariate which is piecewise linearly correlated with the response y_{ij} and $\mathbf{z}_{ij}$ be a p-dimensional vector of additional covariates at time j for the ith individual, where $i = 1, \cdots, m$ and $j = 1, \cdots, n_i$. For a given quantile level $\tau \in (0, 1)$ and any i, we assume the τth conditional quantile of y_{ij} given $(x_{ij}, \mathbf{z}_{ij}^{\mathrm{T}})$ is

$$Q_{i,y}(\tau|x_{ij},\ \mathbf{z}_{ij}) = \begin{cases} c_{i,1} + \alpha_{i,1}x_{ij} + \mathbf{z}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}_i, & \text{if } x_{ij} \le t_{i,1}, \\ c_{i,2} + \alpha_{i,2}x_{ij} + \mathbf{z}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}_i, & \text{if } t_{i,1} \le x_{ij} \le t_{i,2}, \\ \quad\vdots & \quad\vdots \\ c_{i,K_i+1} + \alpha_{i,K_i+1}x_{ij} + \mathbf{z}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}_i, & \text{if } t_{i,K_i} \le x_{ij}, \end{cases} \tag{2.1}$$

$$\text{subject to} \quad c_{i,k} + \alpha_{i,k}t_{i,k} = c_{i,k+1} + \alpha_{i,k+1}t_{i,k}, \quad k = 1,\ \cdots,\ K_i,\ i = 1,\ \cdots,\ m, \tag{2.2}$$

where K_i is the number of kink points $\{t_{i,1}, t_{i,2}, \cdots, t_{i,K_i}\}$ for individual i, $\{c_{i,1}, c_{i,2}, \cdots, c_{i,K_i+1}\}$ and $\{\alpha_{i,1}, \alpha_{i,2}, \cdots, \alpha_{i,K_i+1}\}$ are the intercepts that can be considered as the individual fixed effects and the regression coefficients of x_{ij} at different regions of x_{ij} for individual i, respectively, $\boldsymbol{\gamma}_i = (\gamma_{i,1}, \cdots, \gamma_{i,p})^{\mathrm{T}}$ is the p-dimensional vector of regression coefficients of $\mathbf{z}_{ij}$ which is invariant for different regions of x_{ij}. The continuity constraint (2.2) is imposed to ensure that the regression curve is continuous at kink points at the coordinate of x_{ij}. This model generates the kink regression models for independent data in Hansen (2017) and Zhong et al. (2021) to the panel data framework.

Because model (2.1) admits different unknown parameters for different individuals, it is able to characterize the complete heterogeneity across all individuals. However, it is not parsimonious as it involves too many unknown parameters. On the other hand, it has not used up all information available and ignore some latent homogeneous structures where some individuals and covariates might share the same coefficients. To make (2.1)

more parsimonious and the final estimators more efficient, we impose the following common structure, also called homogeneous structure, on the unknown parameters in model (2.1),

$$\begin{aligned} &c_{i,1} = c_{(\ell)},\ \text{if } (i,1) \in \mathcal{C}_\ell,\ \ell = 1, \cdots, N_c; \quad \alpha_{i,k} = \alpha_{(\ell)},\ \text{if } (i,k) \in \mathcal{A}_\ell,\ \ell = 1, \cdots, N_\alpha; \\ &\gamma_{i,k} = \gamma_{(\ell)},\ \text{if } (i,k) \in \Gamma_\ell,\ \ell = 1, \cdots, N_\gamma; \quad t_{i,k} = t_{(\ell)},\ \text{if } (i,k) \in \mathcal{T}_\ell,\ \ell = 1, \cdots, N_t; \end{aligned} \tag{2.3}$$

where $\{\mathcal{C}_\ell :\ 1 \le \ell \le N_c\}$ is an unknown partition of $\mathcal{C} = \{(i,1) :\ 1 \le i \le m\}$, $\{\mathcal{A}_\ell :\ 1 \le \ell \le N_\alpha\}$ is an unknown partition of $\mathcal{S} = \bigcup_{i=1}^{m} S_i$ with $S_i = \{(i,k) :\ 1 \le k \le K_i + 1\}$, $\{\Gamma_\ell :\ 1 \le \ell \le N_\gamma\}$ is an unknown partition of $\Gamma = \{(i,k) :\ 1 \le i \le m;\ k = 1,\ \cdots,\ p\}$, and $\{\mathcal{T}_\ell :\ 1 \le \ell \le N_t\}$ is an unknown partition of $\mathcal{T} = \bigcup_{i=1}^{m} \{(i,k) :\ 1 \le k \le K_i\}$. Here, $c_{(\ell)}$'s, $\alpha_{(\ell)}$'s, $\gamma_{(\ell)}$'s, $t_{(\ell)}$'s and K_i's are all unknown group parameters to be estimated.

Model (2.1) together with the constraints (2.2) and (2.3) is the proposed MKQR model with latent homogeneous structure for panel data analysis, which is what this paper is going to address. From statistical modelling point of view, the proposed model strikes a nice balance between the risk of misspecification and the model parsimony by a data-driven approach. It accounts for both homogeneity and heterogeneity among individuals of panel data. Furthermore, we don't have the problem caused by cluster effects, because we don't use cluster effects to account for the difference between different individuals. From practical point of view, to find which individuals share the same impact of a covariate, we only need to estimate the unknown partitions in (2.3), namely identify the common structure of the unknown parameters, it is also called homogeneity pursuit.

We conclude this section by showing an equivalent representation of (2.1) with the continuity condition (2.2). By simple calculation, $Q_{i,y}(\tau|x_{ij}, \mathbf{z}_{ij})$ with (2.2) can be written as

$$Q_y(\tau;\ \boldsymbol{\eta}_i,\ \mathbf{t}_i|x_{ij}, \mathbf{z}_{ij}) = c_{i,1} + \beta_{i,0}x_{ij} + \sum_{k=1}^{K_i} \beta_{i,k}(x_{ij} - t_{i,k})I(x_{ij} > t_{i,k}) + \mathbf{z}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}_i, \tag{2.4}$$

where $\boldsymbol{\eta}_i = (c_{i,1},\ \beta_{i,0},\ \cdots,\ \beta_{i,K_i},\ \boldsymbol{\gamma}_i^{\mathrm{T}})^{\mathrm{T}}, \mathbf{t}_i = (t_{i,1},\ \cdots,\ t_{i,K_i})^{\mathrm{T}}, \beta_{i,0} = \alpha_{i,1}, \beta_{i,k} = \alpha_{i,k+1} - \alpha_{i,k}, c_{i,k+1} = c_{i,k} - \beta_{i,k}t_{i,k}, k = 1,\ \cdots,\ K_i$ and $I(x_{ij} > t_{i,k})$ denotes an indicator function on $x_{ij} > t_{i,k}$. We remark that the difference between the proposed kink model and a conventional threshold model is that the continuity condition (2.2) is required in the kink model. For

each i, when K_i is given, there are $3K_i + 2 + p$ unknown parameters in (2.1) without the continuity restriction (2.2). However, with the continuity condition (2.2) considered, there are only $2K_i + 2 + p$ unknown free parameters in (2.4). One can refer to Hidalgo et al. (2019) and Chen (2021) for more discussions on the difference between jump and kink at the change point.

3 Estimation Procedure

The proposed estimation procedure consists of three stages: preliminary estimation, homogeneity pursuit and final estimation.

Stage 1. Preliminary estimation.

We first obtain the preliminary estimates of all parameters for each i. Let $\rho_\tau(v) = v\{\tau - I(v < 0)\}$ be the quantile loss function. For each i, given an initial number of kink points K_i, we define the loss function as

$$L_0(\boldsymbol{\eta}_i,\ \mathbf{t}_i) = \sum_{j=1}^{n_i} \rho_\tau\Big(y_{ij} - Q_y(\tau;\ \boldsymbol{\eta}_i,\ \mathbf{t}_i | x_{ij}, \mathbf{z}_{ij})\Big). \tag{3.1}$$

We minimize the loss function $L_0(\boldsymbol{\eta}_i,\ \mathbf{t}_i)$ with respect to $\boldsymbol{\eta}_i$ and $\mathbf{t}_i$ and estimate the number of kink points K_i in the following three-step algorithm.

- Step 1. For each i, given an initial number of kink points $K_i = k$ where we begin with $k = 0$, we obtain the profile estimator of $\boldsymbol{\eta}_i$ conditional on $\mathbf{t}_i$ using the standard quantile linear regression by

$$\widehat{\boldsymbol{\eta}}_i^{(k)}(\mathbf{t}_i) = \arg\min_{\boldsymbol{\eta}_i \in \mathcal{B}} L_0(\boldsymbol{\eta}_i,\ \mathbf{t}_i), \tag{3.2}$$

  where $\mathcal{B} \subseteq \mathbb{R}^{2+k+p}$ is a compact set of $\boldsymbol{\eta}_i$. If $k = 0$, directly go to Step 3.

- Step 2. The estimator of $\mathbf{t}_i$ is obtained by minimizing the loss function given $\widehat{\boldsymbol{\eta}}_i^{(k)}(\mathbf{t}_i)$,

$$\widehat{\mathbf{t}}_i^{(k)} = \underset{\mathbf{t}\in\boldsymbol{\Lambda}_i}{\arg\min} L_0\left(\widehat{\boldsymbol{\eta}}_i^{(k)}(\mathbf{t}_i),\ \mathbf{t}_i\right), \tag{3.3}$$

where $\boldsymbol{\Lambda}_i = (L_x + \xi \le t_1 < t_2 < \cdots < t_k \le U_x - \xi)$, $[L_x, U_x]$ is the compact set of the threshold covariate x_{ij} and ξ is a sufficiently small positive number. The minimization of the linearly constrained optimization problem in (3.3) can be solved by the adaptive barrier algorithm using the R function "*constrOptim*".

- Step 3. Let $k = k + 1$ and repeat the first two steps until $k = K^*$, where K^* is a pre-specified maximum number of kink points. We determine the number of kink points for each i via minimizing the Schwarz-type quantile information criterion (SIC),

$$\mathrm{SIC}(k) = \log\left\{\frac{1}{n_i} L_0\left(\widehat{\boldsymbol{\eta}}_i^{(k)},\ \widehat{\mathbf{t}}_i^{(k)}\right)\right\} + D_{n_i}\frac{\log(n_i)}{2n_i}(2 + p + 2k), \tag{3.4}$$

where $k = 0, 1, \cdots, K^*$ and D_{n_i} is a positive constant that allows to approach the infinity slowly as n_i increases such as $D_{n_i} = \log(n_i)$. Then, the preliminary estimate for K_i is $\tilde{K}_i = \underset{k=0,1,\cdots,K^*}{\arg\min}\ \mathrm{SIC}(k)$. The resulting estimates for $c_{i,k}$'s, $\alpha_{i,k}$'s, $\gamma_{i,k}$'s and $t_{i,k}$'s corresponding to $k = \tilde{K}_i$ are their preliminary estimates for the later homogeneity pursuit, denoted by $\tilde{c}_{i,k}$'s, $\tilde{\alpha}_{i,k}$'s, $\tilde{\gamma}_{i,k}$'s and $\tilde{t}_{i,k}$'s, respectively.

REMARK 1: The initial locations of kink points $\mathbf{t}_i$ play an important role in the performance of the above algorithm. In our simulations, for each k, we choose k initial kink locations which can split the range of the threshold covariate to $k + 1$ equal parts. For example, when $k = 2$, we choose $\{L_x + \frac{1}{3}(U_x - L_x), L_x + \frac{2}{3}(U_x - L_x)\}$ as the initial kink locations. To make the algorithm insensitive to the initial values, one can also apply the bootstrap restarting approach (Wood, 2001; Zhong et al., 2021) which iteratively updates the initial values using the bootstrap samples. The SIC in (3.4) is similar to the modified quantile Bayesian information criterion (BIC) which has been used by Lee et al. (2014) for consistent model selection of high dimensional linear quantile regression. When $D_{n_i} = 1$, the SIC in (3.4) becomes the standard quantile BIC.

Stage 2. Homogeneity pursuit.

Stage 2.1. Refinement of $\tilde{K}_i$'s. We remark that the estimation of K_i's is affected by the homogeneous structure of $\alpha_{i,j}$'s. Thus, we first apply the binary segmentation algorithm to conduct homogeneity pursuit for $\alpha_{i,j}$'s and then refine the estimates of K_i's. To this end, we sort $\tilde{\alpha}_{i,j}$, $(i,j) \in \bigcup_{i=1}^{m} \{(i,k) : \ 1 \le k \le \tilde{K}_i + 1\}$ in the ascending order and denote them by

$$a_{(1)} \le a_{(2)} \le \cdots \le a_{(M)}, \quad M = \sum_{i=1}^{m} (\tilde{K}_i + 1).$$

We use R_{ij} to denote the rank of $\tilde{\alpha}_{i,j}$. Identifying the homogeneity among $\alpha_{i,j}$'s is equivalent to detecting the change points among $a_{(l)}$'s, $l = 1, \cdots, M$. To this end, we apply the binary segmentation algorithm as follows. For any $1 \le i < j \le M$, let

$$\Delta_{ij}(\kappa) = \sqrt{\frac{(j-\kappa)(\kappa - i + 1)}{j - i + 1}} \left| \frac{\sum_{l=\kappa+1}^{j} a_{(l)}}{j - \kappa} - \frac{\sum_{l=i}^{\kappa} a_{(l)}}{\kappa - i + 1} \right|, \tag{3.5}$$

where κ is a point to separate the preliminary estimates $\{a_{(i)}, a_{(i+1)}, \cdots, a_{(j)}\}$ into two disjoint parts: $\{a_{(i)}, \cdots, a_{(\kappa)}\}$ and $\{a_{(\kappa+1)}, \cdots, a_{(j)}\}$. Given a threshold δ, the binary segmentation algorithm to detect the change points works as follows.

(1) Find $\widehat{k}_1$ such that

$$\Delta_{1,M}(\widehat{k}_1) = \max_{1 \le \kappa < M} \Delta_{1,M}(\kappa).$$

If $\Delta_{1,M}(\widehat{k}_1) \le \delta$, there is no change point among $a_{(l)}$'s, $l = 1, \cdots, M$, and the process of detection ends. Otherwise, add $\widehat{k}_1$ to the set of change points and divide the region $\{\kappa : \ 1 \le \kappa \le M\}$ into two subregions: $\{\kappa : \ 1 \le \kappa \le \widehat{k}_1\}$ and $\{\kappa : \ \widehat{k}_1 + 1 \le \kappa \le M\}$.

(2) Detect the change points in the two subregions obtained in (1), respectively. Let us deal with the subregion $\{\kappa : \ 1 \le \kappa \le \widehat{k}_1\}$ first. Find $\widehat{k}_2$ such that

$$\Delta_{1,\widehat{k}_1}(\widehat{k}_2) = \max_{1 \le \kappa < \widehat{k}_1} \Delta_{1,\widehat{k}_1}(\kappa).$$

If $\Delta_{1,\widehat{k}_1}(\widehat{k}_2) \le \delta$, there is no change point in the region $\{\kappa : \ 1 \le \kappa \le \widehat{k}_1\}$. Otherwise,

add $\widehat{k}_2$ to the set of change points and divide the region $\{\kappa:\ 1 \le \kappa \le \widehat{k}_1\}$ into two subregions: $\{\kappa:\ 1 \le \kappa \le \widehat{k}_2\}$ and $\{\kappa:\ \widehat{k}_2 + 1 \le \kappa \le \widehat{k}_1\}$. For the subregion $\{\kappa:\ \widehat{k}_1 + 1 \le \kappa \le M\}$, we find $\widehat{k}_3$ such that

$$\Delta_{\widehat{k}_1+1,M}(\widehat{k}_3) = \max_{\widehat{k}_1+1 \le \kappa < M} \Delta_{\widehat{k}_1+1,M}(\kappa).$$

If $\Delta_{\widehat{k}_1+1,M}(\widehat{k}_3) \le \delta$, there is no change point in the region $\{\kappa:\ \widehat{k}_1 + 1 \le \kappa \le M\}$. Otherwise, add $\widehat{k}_3$ to the set of change points and divide the region $\{\kappa:\ \widehat{k}_1 + 1 \le \kappa \le M\}$ into two subregions: $\{\kappa:\ \widehat{k}_1 + 1 \le \kappa \le \widehat{k}_3\}$ and $\{\kappa:\ \widehat{k}_3 + 1 \le \kappa \le M\}$.

(3) For each subregion obtained in (2), we do exactly the same as that for the subregion $\{\kappa:\ 1 \le \kappa \le \widehat{k}_1\}$ or $\{\kappa:\ \widehat{k}_1 + 1 \le \kappa \le M\}$ in (2), and keep doing so until there is no subregion containing any change point.

We sort the estimated change point locations in the ascending order and denote them by

$$\widehat{k}_{(1)} < \widehat{k}_{(2)} < \cdots < \widehat{k}_{(\widehat{N}_{-1})},$$

where $\widehat{N}_{-1}$ is the number of change points detected. In addition, we denote $\widehat{k}_{(0)} = 0$, $\tilde{N}_\alpha = \widehat{N}_{-1} + 1$, and $\widehat{k}_{(\tilde{N}_\alpha)} = M$. We use $\tilde{N}_\alpha$ to estimate N_α. Let $\tilde{\mathcal{A}}_\ell = \{(i,j):\ \widehat{k}_{(\ell-1)} < R_{ij} \le \widehat{k}_{(\ell)}\}, 1 \le \ell \le \tilde{N}_\alpha,$. We use $\left\{\tilde{\mathcal{A}}_\ell:\ 1 \le \ell \le \tilde{N}_\alpha\right\}$ to estimate the partition $\{\mathcal{A}_\ell:\ 1 \le \ell \le N_\alpha\}$. We consider all the $\alpha_{i,j}$'s with the subscript (i,j) in the same group of the estimated partition having the same value.

Remark 2: The finite sample performances of both the CARDS algorithms (Ke et al., 2015; Wang et al., 2018) and the binary segmentation approaches (Ke et al., 2016; Guo and Li, 2022; Lian et al., 2021) depend on the quality of the preliminary estimates. Thus, the number of observations for each individual should be relatively large to ensure that the ranking of preliminary estimates is reasonable to represent the order of the true parameters.

Remark 3: The tuning parameter δ plays a crucial role in the homogeneity pursuit to balance model complexity and estimation accuracy. Selecting the tuning parameter is actually equivalent to selecting the number of elements in the partition. If δ is too small, the

redundant homogeneous subgroups will be identified and the estimation efficiency may loss. If δ is too large, some latent homogeneous structure will be missed and the final estimator could be biased. In our simulations, we select an appropriate value of δ via minimizing the Bayesian information criterion (BIC),

$$\mathrm{BIC}(\delta) = \log\left\{ m^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \rho_\tau \left(y_{ij} - Q_y(\tau; \widehat{\boldsymbol{\eta}}_i(\delta), \widehat{\mathbf{t}}_i(\delta) | x_{ij}, \mathbf{z}_{ij}) \right) \right\} + \mathrm{df}(\delta) \cdot \frac{\log m}{2m},$$

where $\widehat{\boldsymbol{\eta}}_i(\delta), \widehat{\mathbf{t}}_i(\delta)$ are the estimators associated with δ and $\mathrm{df}(\delta)$ is the effective degree of freedom associated with δ. If the prediction accuracy is the primary concern, the multifold cross-validation prediction error can be also considered as the criterion (Zhang, 1993).

Then, for each i, the estimate of K_i is further refined by

$$\widehat{K}_i = \tilde{K}_i - \sum_{k=1}^{\tilde{K}_i} \sum_{\ell=1}^{\tilde{N}_\alpha} I\{(i,k) \in \tilde{\mathcal{A}}_\ell\} I\{(i,k+1) \in \tilde{\mathcal{A}}_\ell\}, \tag{3.6}$$

where the second part means that if two successive subregions are clustered into one group by the binary segmentation algorithm, then there is no kink point between these two successive subregions and thus refine the estimate of K_i by subtracting one from $\tilde{K}_i$.

Stage 2.2. Estimation of the partitions.

For each i, given the estimated number of kink points $\widehat{K}_i$, we obtain the refined preliminary estimates of $\alpha_{i,k}$'s, $c_{i,k}$'s, $\gamma_{i,k}$'s and $t_{i,k}$'s using the method in Step 1. With a slight abuse of notation, we still use $\tilde{\alpha}_{i,k}$s, $\tilde{c}_{i,k}$s, $\tilde{\gamma}_{i,k}$s and $\tilde{t}_{i,k}$s to denote the resulting preliminary estimates. Then, applying the binary segmentation algorithm in Stage 2.1 sequentially to $\tilde{t}_{i,k}$'s, $\tilde{\alpha}_{i,k}$'s, $\tilde{\gamma}_{i,k}$'s and $\tilde{c}_{i,k}$'s, respectively, we get the estimators of homogeneous structures of $t_{i,k}$'s, $\alpha_{i,k}$'s, $\gamma_{i,k}$'s and $c_{i,k}$'s, denoted by $\{\widehat{\mathcal{T}}_\ell : \ 1 \le \ell \le \widehat{N}_t\}$, $\{\widehat{\mathcal{A}}_\ell : \ 1 \le \ell \le \widehat{N}_\alpha\}$, $\{\widehat{\Gamma}_\ell : \ 1 \le \ell \le \widehat{N}_\gamma\}$, and $\{\widehat{\mathcal{C}}_\ell : \ 1 \le \ell \le \widehat{N}_c\}$. They are the final estimators of the partitions, $\{\mathcal{T}_\ell : \ 1 \le \ell \le N_t\}$, $\{\mathcal{A}_\ell : \ 1 \le \ell \le N_\alpha\}$, $\{\Gamma_\ell : \ 1 \le \ell \le N_\gamma\}$, and $\{\mathcal{C}_\ell : \ 1 \le \ell \le N_c\}$ in (2.3).

Stage 3. Final estimation.

Given $K_i = \widehat{K}_i$, we let $\boldsymbol{\theta}_i = (c_{i,1}, \ \alpha_{i,1}, \ \cdots, \ \alpha_{i,K_i+1}, \ \boldsymbol{\gamma}_i^{\mathrm{T}})^{\mathrm{T}}$, $\mathcal{F}_y(\tau; \ \boldsymbol{\theta}_i, \ \mathbf{t}_i | x_{ij}, \ \mathbf{z}_{ij})$ be

$Q_y(\tau;\ \boldsymbol{\eta}_i,\ \mathbf{t}_i|x_{ij},\ \mathbf{z}_{ij})$ with $\beta_{i,k}$'s being expressed in terms of $\alpha_{i,k}$'s, and

$$\boldsymbol{\theta} = (c_{(1)},\ \cdots,\ c_{(\widehat{N}_c)},\ \alpha_{(1)},\ \cdots,\ \alpha_{(\widehat{N}_\alpha)},\ \gamma_{(1)},\ \cdots,\ \gamma_{(\widehat{N}_\gamma)},\ t_{(1)},\ \cdots,\ t_{(\widehat{N}_t)})^{\mathrm{T}}.$$

Let $L(\boldsymbol{\theta})$ be

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} \rho_\tau \left(y_{ij} - \mathcal{F}_y(\tau;\ \boldsymbol{\theta}_i,\ \mathbf{t}_i|x_{ij},\ \mathbf{z}_{ij})\right) \tag{3.7}$$

with $c_{i,1}$ being replaced by $c_{(k)}$ if $(i,1) \in \widehat{\mathcal{C}}_k$, $\alpha_{i,j}$ by $\alpha_{(k)}$ if $(i,j) \in \widehat{\mathcal{A}}_k$, $\gamma_{i,j}$ by $\gamma_{(k)}$ if $(i,j) \in \widehat{\Gamma}_k$, and $t_{i,j}$ by $t_{(k)}$ if $(i,j) \in \widehat{\mathcal{T}}_k$. To minimize the objective function (3.7) in practice, we apply the Nelder-Mead algorithm in the function "*optim*" in the R package *stats*. This minimization is computationally fast when good initial values are used. To this end, we choose the averages of preliminary estimates within the same group as the initial values for the group-specific parameters. We denote the minimizer of $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ as

$$\widehat{\boldsymbol{\theta}} = (\widehat{c}_{(1)},\ \cdots,\ \widehat{c}_{(\widehat{N}_c)},\ \widehat{\alpha}_{(1)},\ \cdots,\ \widehat{\alpha}_{(\widehat{N}_\alpha)},\ \widehat{\gamma}_{(1)},\ \cdots,\ \widehat{\gamma}_{(\widehat{N}_\gamma)},\ \widehat{t}_{(1)},\ \cdots,\ \widehat{t}_{(\widehat{N}_t)})^{\mathrm{T}}. \tag{3.8}$$

The final estimators of $c_{i,1}$, $\alpha_{i,j}$, $\gamma_{i,j}$ and $t_{i,j}$ are $\widehat{c}_{(k)}$ if $(i,1) \in \widehat{\mathcal{C}}_k$, $\widehat{\alpha}_{(k)}$ if $(i,j) \in \widehat{\mathcal{A}}_k$, $\widehat{\gamma}_{(k)}$ if $(i,j) \in \widehat{\Gamma}_k$, and $\widehat{t}_{(k)}$ if $(i,j) \in \widehat{\mathcal{T}}_k$, respectively. We denote them by $\widehat{c}_{i,1}$, $\widehat{\alpha}_{i,j}$, $\widehat{\gamma}_{i,j}$ and $\widehat{t}_{i,j}$. The estimator of $c_{i,k+1}$ can be obtained by $\widehat{c}_{i,k+1} = \widehat{c}_{i,1} - \sum_{\ell=1}^{k}(\widehat{\alpha}_{i,\ell+1} - \widehat{\alpha}_{i,\ell})\widehat{t}_{i,\ell}, k = 1, \cdots, \widehat{K}_i$.

4 Asymptotic Properties

In this section, we study the asymptotic properties of the estimators for different parameters in the MKQR model under the following three different cases.

- *Overfitting case*: we assign different parameters of the MKQR model to different individual i without considering any homogeneous structure of individuals and parameters. The estimators are essentially the preliminary estimators obtained in Stage 1.
- *Correct-fitting case*: we consider homogeneous structures of individuals and parameters and use the proposed estimation procedure to obtain the estimators.

- *Underfitting case*: the estimation based on the global homogeneity assumption for all individuals. That is, $c_{1,1} = \cdots = c_{m,1}, \boldsymbol{\gamma}_1 = \cdots = \boldsymbol{\gamma}_m, K_1 = \cdots = K_m = K$, and $\alpha_{1,k} = \cdots = \alpha_{m,k}, t_{1,k} = \cdots = t_{m,k} for k = 1, \cdots, K$, where K is the unknown number of kink points which is same for all individuals.

We present notations and the necessary conditions in Section A of the Supplement. Next, we first derive the asymptotic properties of estimators under the overfitting case.

Theorem 4.1. *(Overfitting case). Under conditions (C1)-(C5), for each individual i, when $n_i \to \infty$ and $D_{n_i} \log n_i / n_i \to 0$, the estimators under the overfitting case enjoy the following asymptotic properties.*

(1) Given the true number K_i of kink points, we have

$$\begin{aligned}
n_i^{1/2}(\tilde{c}_{i,1} - c_{i,1}) &\xrightarrow{d} N\Big(0,\ \boldsymbol{\iota}_1^T \Sigma_i^{-1} \Omega_i \Sigma_i^{-1} \boldsymbol{\iota}_1\Big), \\
n_i^{1/2}(\tilde{\alpha}_{i,k} - \alpha_{i,k}) &\xrightarrow{d} N\Big(0,\ (\sum_{j=2}^{k+1} \boldsymbol{\iota}_j)^T \Sigma_i^{-1} \Omega_i \Sigma_i^{-1} (\sum_{j=2}^{k+1} \boldsymbol{\iota}_j)\Big),\ k = 1, \cdots, K_i + 1, \\
n_i^{1/2}(\tilde{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i) &\xrightarrow{d} N\Big(0,\ \boldsymbol{\nu}_\gamma^T \Sigma_i^{-1} \Omega_i \Sigma_i^{-1} \boldsymbol{\nu}_\gamma\Big), \\
n_i^{1/2}(\tilde{t}_{i,k} - t_{i,k}) &\xrightarrow{d} N\Big(0,\ \boldsymbol{\iota}_{K_i+2+p+k}^T \Sigma_i^{-1} \Omega_i \Sigma_i^{-1} \boldsymbol{\iota}_{K_i+2+p+k}\Big),\ k = 1, \cdots, K_i,
\end{aligned}$$

where $\xrightarrow{d}$ denotes the convergence in distribution, $\boldsymbol{\iota}_j$ $(j = 1, \cdots, 2+p+2K_i)$ is a $(2+p+2K_i)$-dimensional unit vector with the jth element being one and all other elements being zero, $\boldsymbol{\nu}_\gamma = (\boldsymbol{\iota}_{K_i+2+1}, \cdots, \boldsymbol{\iota}_{K_i+2+p})$, Ω_i and Σ_i are defined in the Supplement.

(2) As $n_i \to \infty$, for individual i, we have that $P(\tilde{K}_i = K_i) \longrightarrow 1$.

Theorem 4.1 establishes the consistency of the preliminary estimators which ensures that the ranking of preliminary estimates in the binary segmentation algorithm is reasonable to retain the underlying relative magnitudes of the true parameters if the sample size is large enough.

Next, we present the asymptotic properties of the estimators obtained by using the proposed method for the MKQR model with homogeneous structure.

Theorem 4.2. *(Correct-fitting case). Under conditions (C1)-(C7), given the true number K_i of kink points, we have that*

$$\begin{aligned}
N^{1/2}\sigma_{c,i1}^{-1}(\widehat{c}_{i,1} - c_{i,1}) &\xrightarrow{d} N(0,\ 1), \\
N^{1/2}\sigma_{\alpha,ik}^{-1}(\widehat{\alpha}_{i,k} - \alpha_{i,k}) &\xrightarrow{d} N(0,\ 1),\ k = 1, \cdots, K_i + 1, \\
N^{1/2}\Sigma_{\gamma,i}^{-1/2}(\widehat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i) &\xrightarrow{d} N(0,\ I_p), \\
N^{1/2}\sigma_{t,ik}^{-1}(\widehat{t}_{i,k} - t_{i,k}) &\xrightarrow{d} N(0,\ 1),\ k = 1, \cdots, K_i,
\end{aligned}$$

where $\sigma_{c,i1}^2, \sigma_{\alpha,ik}^2, \Sigma_{\gamma,i}, \sigma_{t,ik}^2$ are defined in the proof of Theorem 4.2 in the Supplement.

Theorem 4.2 shows that the final estimators of parameters with the homogeneous structure conditions have the convergence rate of order $N^{-1/2} = (\sum_{i=1}^m n_i)^{-1/2}$, which is much faster than $n_i^{-1/2}$ under the overfitting case. It suggests that it is necessary to consider the homogeneous structure among individuals and parameters.

Next, we present the asymptotic result of the estimators under the global homogeneity condition, i.e. the underfitting case. We denote these estimators as $\check{c}_{i,1}, \check{\alpha}_{i,k}, \check{\boldsymbol{\gamma}}_i$ and $\check{t}_{i,k}$ for $c_{i,1}, \alpha_{i,k}, \boldsymbol{\gamma}_i$ and $t_{i,k}$, respectively.

Theorem 4.3. *(Underfitting case). Suppose the $c_{i,1}$'s are sufficiently separated in the sense that $m^{-1}\sum_{i=1}^m (c_{i,1} - \bar{c}_1)^2 > C, \bar{c}_1 = m^{-1}\sum_{i=1}^m c_{i,1}$, then the estimators $\check{c}_{i,1}$ obtained by the underfitting case satisfies the following properties $m^{-1}\sum_{i=1}^m (\check{c}_{i,1} - c_{i,1})^2 > C$. Similarly, if the other parameters are all sufficiently separated as $c_{i,1}$, then $m^{-1}\sum_{i=1}^m (\check{\alpha}_{i,k} - \alpha_{i,k})^2 > C$, for $k = 1, \cdots, K_i + 1$, $m^{-1}\sum_{i=1}^m \|\check{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i\|^2 > C$, and $m^{-1}\sum_{i=1}^m (\check{t}_{i,k} - t_{i,k})^2 > C$, for $k = 1, \cdots, K_i$.*

Theorem 4.3 shows that the estimators under the underfitting case are inconsistent. Thus, it is of very importance to consider the heterogeneity among individuals. The serious inconsistency problem of the estimators in the underfitting case will be also shown in the simulations.

5 Monte Carlo Simulations

In this section, we evaluate the finite sample performances of our proposed method by using Monte Carlo simulations. We consider the following data generation processes:

Model 1 ($K = 0$)**:** $y_{ij} = c_{i,1} + \gamma_i z_{ij} + (1 + \rho z_{ij})e_{it}$,

Model 2 ($K > 0$)**:** $y_{ij} = c_{i,1} + \beta_{i,0} x_{ij} + \sum_{k=1}^{K} \beta_{i,k}(x_{ij} - t_{i,k})_+ + \gamma_i z_{ij} + (1 + \rho z_{ij})e_{ij}$,

where $i = 1, \cdots, m = 60$, $j = 1, \cdots, n$, $x_{ij} \overset{i.i.d}{\sim} Uniform(-5, 5)$, $z_{ij} = 0.5c_{i,1} + u_{ij}$ and $u_{ij} \overset{i.i.d}{\sim} Unifrom(0, 1)$. Here ρ controls the error heterogeneity, i.e. $\rho = 0$ corresponding to a homogeneous model and $\rho = 0.5$ corresponding to a heterogeneous model. For all models, we considers two distributions for the error term e_{ij}: $N(0, 0.5^2)$ and $t(3)/3$. The intercepts $c_{i,1}$'s are randomly assigned to be $-\sqrt{3}$ or $\sqrt{3}$ with equal probability and γ_i is randomly set to be -1.5 or 1.5 with equal probability. For Model 2, we consider two different settings with $K = 1$ and 2. For $K = 1$, $\boldsymbol{\beta}_i = (\beta_{i,0}, \beta_{i,1})$ is generated randomly from $(-2, 2)$ and $(2, -2)$ with equal probability and $t_{i,1}$ randomly takes values -2 or 2 with equal probability. For $K = 2$, $\boldsymbol{\beta}_i = (\beta_{i,0}, \beta_{i,1}, \beta_{i,2})$ is randomly set to be $(1, -3, 4)$ or $(1, -4, 4)$ with equal probability, $\mathbf{t}_i = (t_{i,1}, t_{i,2})$ randomly takes $(-2, 3)$ or $(-2, 1)$ with equal probability. For saving the space, we only present the simulation results for the error from $t(3)/3$ and relegate the results for the normal error to the Supplement.

We first evaluate the finite sample performance of detecting the number of kink points across different quantile levels $\tau = 0.25, 0.5$, and 0.75. We consider the mean of correct classification rates (MCCR) over $R = 500$ replications as the evaluation criterion which is defined as MCCR $= R^{-1} \sum_{r=1}^{R} \left\{ m^{-1} \sum_{i=1}^{m} I(\hat{K}_i^{(r)} = K_i) \right\}$, where $\hat{K}_i^{(r)}$ is the estimated number of kink points for individual i at the rth replication, $r = 1, \cdots, R$, and K_i is the true number of kink points for individual i. The simulation results are reported in Table 1 for $n = 50$, 100 and 200. We observe that as the number of observations n increases, the selection rates gradually approach to 100% for all model settings, which illustrates the selection consistency of using the SIC criterion in (3.4). In addition, the detection results at the median quantile ($\tau = 0.5$) perform generally better than that at the other two quantiles,

Table 1: MCCR for the estimated number of kink points for Model 1 ($K = 0$) and Model 2 ($K = 1$ and $K = 2$) when the error $\sim t(3)/3$ for $\tau = 0.25, 0.5, 0.75$ and $n = 50, 100, 200$.

Model	ρ	$n = 50$			$n = 100$			$n = 200$		
		0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
Model 1($K = 0$)	0	0.902	0.945	0.901	0.942	0.973	0.941	0.962	0.985	0.962
	0.5	0.877	0.928	0.879	0.919	0.956	0.920	0.943	0.969	0.944
Model 2($K = 1$)	0	0.890	0.949	0.892	0.956	0.981	0.955	0.978	0.993	0.979
	0.5	0.878	0.941	0.880	0.943	0.975	0.924	0.967	0.987	0.966
Model 2($K = 2$)	0	0.862	0.927	0.871	0.951	0.979	0.951	0.975	0.991	0.975
	0.5	0.847	0.922	0.861	0.940	0.974	0.938	0.956	0.966	0.957

which is quite common in quantile based clustering such as Zhang et al. (2019), because the effective sample size at the median quantile is larger than that at the tail quantiles.

Next, we evaluate the estimation accuracy and the necessity of homogeneous structure detection of the proposed method for the MKQR model. For the purpose of comparison, we take three different methods into consideration: the overfitting case ('Overfit') that fits each individual separately, the oracle case ('Oracle') based on the true homogeneous structure as a prior, and the misspecified underfitting case ('Underfit') that treats all individuals as one homogeneous group. To measure the estimation accuracy for an estimator such as $\widehat{\boldsymbol{\beta}}_i$, we consider the absolute estimation bias ('AbsBias') averaged over all m individuals, which is defined as AbsBias $= m^{-1}\sum_{i=1}^{m}\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0}\|$, where $\widehat{\boldsymbol{\beta}}_i$ and $\boldsymbol{\beta}_{i0}$ are the estimated and true parameters for $\boldsymbol{\beta}_i$. To evaluate the accuracy of the identified homogeneous structure detected by the proposed method, we apply the commonly-used normalized mutual information (NMI) (Ke et al., 2015) to measure the similarity between the estimated and true homogeneous clusters. It belongs to $[0, 1]$ and a larger NMI value indicates a better performance in the homogeneous structure identification. Specially, we denote $\mathbb{C} = \{C_1, C_2, \cdots\}$ and $\mathbb{D} = \{D_1, D_2, \cdots\}$ as the estimated and true homogeneous structures, respectively, and define their NMI as

$$\mathrm{NMI}(\mathbb{C}, \mathbb{D}) = \frac{2I(\mathbb{C}, \mathbb{D})}{H(\mathbb{C}) + H(\mathbb{D})}, \tag{5.1}$$

where $I(\mathbb{C}, \mathbb{D}) = \sum_{k,j}\left(|C_k \cap D_j|/n\right)\log\left(n|C_k \cap D_j|/|C_k||D_j|\right)$, and $|C_k|$ denotes the cardinality of the set C_k, and $H(\mathbb{C}) = -\sum_k\left(|C_k|/n\right)\log(|C_k|/n)$ is the entropy of $\mathbb{C}$.

When there is no kink point in Model 1, it becomes a linear model for panel data. In

Table 2: Averages and standard deviations in parentheses of the AbsBias and the NMI over 500 replications for Model 1 with $K = 0$ when $n = 200$ and the error $\sim t(3)/3$.

ρ		LS		$\tau = 0.25$		$\tau = 0.5$		$\tau = 0.75$	
		$\mathbf{c}$	$\boldsymbol{\gamma}$	$\mathbf{c}$	$\boldsymbol{\gamma}$	$\mathbf{c}$	$\boldsymbol{\gamma}$	$\mathbf{c}$	$\boldsymbol{\gamma}$
0	Overfit	17.301	16.639	17.160	16.714	13.955	13.392	17.274	16.636
		(1.957)	(1.832)	(1.651)	(1.526)	(1.378)	(1.251)	(1.623)	(1.592)
	Undefit	173.205	151.430	173.205	153.605	173.205	175.375	173.205	152.867
		(8.148)	(11.783)	(7.029)	(11.748)	(5.412)	(8.570)	(7.965)	(12.238)
	Oracle	2.637	2.544	2.486	2.443	2.065	1.984	2.478	2.397
		(1.640)	(1.627)	(1.522)	(1.540)	(1.222)	(1.105)	(1.469)	(1.383)
	Proposed	2.637	2.544	2.486	2.443	2.065	1.984	2.478	2.397
		(1.640)	(1.627)	(1.522)	(1.540)	(1.222)	(1.105)	(1.469)	(1.383)
	NMI	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.5	Overfit	17.798	17.947	16.236	15.891	13.107	12.920	16.265	15.875
		(1.734)	(1.920)	(1.654)	(1.719)	(1.339)	(1.350)	(1.610)	(1.740)
	Undefit	173.205	151.839	173.205	179.325	173.205	183.778	173.205	148.589
		(7.927)	(12.402)	(6.816)	(11.151)	(11.934)	(16.098)	(9.164)	(12.178)
	Oracle	2.819	2.905	2.354	1.589	1.890	1.255	2.433	1.569
		(1.652)	(1.766)	(1.321)	(1.123)	(1.049)	(0.859)	(1.347)	(1.063)
	Proposed	2.886	3.010	2.354	1.589	1.890	1.255	2.433	1.569
		(2.042)	(2.136)	(1.321)	(1.123)	(1.049)	(0.859)	(1.347)	(1.063)
	NMI	0.999	0.999	1.000	1.000	1.000	1.000	1.000	1.000

this case, we consider the mean-based regression method ('LS') similar to Ke et al. (2016) as an alternative method. Table 2 summarizes the simulation results of different methods at various error settings. For quantile-based regression methods, we consider three different quantile levels $\tau = 0.25, 0.5$ and 0.75. According to Table 2, the proposed method is able to correctly identify the latent homogeneous structure. When the error is generated from the heavy-tailed distribution $t(3)/3$ especially with error heteroscedasticity, the quantile-based methods outperform the mean-based method. It demonstrates the robustness of the quantile-based homogeneity pursuit method. When there is no kink point, the proposed method can be considered as an extension of the mean-based homogeneity pursuit method in Ke et al. (2016) to a robust framework.

Tables 3 reports the simulation results of regression coefficients of Model 2 with $K = 1$ and $K = 2$ for $\tau = 0.5$ when the error is from $t(3)/3$. We can draw the following conclusions. First, the overfitting estimators, the oracle estimators and the proposed estimators have relatively small estimation biases. However, the underfitting estimators are obviously inconsistent with quite large biases since some necessary heterogeneity across individuals is

Table 3: Averages and standard deviations in parentheses of the AbsBias and the NMI over 500 replications for Model 2 with $K = 1$ and 2 when $n = 200$ and the error $\sim t(3)/3$.

K	τ	ρ		Overfit	Underfit	Oracle	Proposed	NMI
1	0.25	0	$\mathbf{c}$	16.610 (2.348)	173.598 (3.255)	1.062 (0.705)	1.079 (0.799)	1.000
			$\boldsymbol{\beta}$	5.975 (0.622)	208.401 (28.922)	0.250 (0.167)	0.251 (0.172)	1.000
			$\boldsymbol{\gamma}$	5.604 (0.547)	181.164 (26.727)	1.034 (0.718)	1.046 (0.775)	1.000
			$\mathbf{t}$	7.484 (0.747)	197.855 (15.208)	0.574 (0.325)	0.574 (0.325)	1.000
		0.5	$\mathbf{c}$	15.155 (2.466)	173.852 (6.733)	1.140 (0.774)	1.276 (1.237)	0.997
			$\boldsymbol{\beta}$	5.328 (0.646)	210.365 (28.022)	0.161 (0.110)	0.164 (0.115)	1.000
			$\boldsymbol{\gamma}$	5.409 (0.615)	187.549 (32.591)	0.846 (0.671)	0.904 (0.799)	1.000
			$\mathbf{t}$	6.681 (0.812)	198.556 (13.294)	0.337 (0.216)	0.345 (0.223)	1.000
	0.5	0	$\mathbf{c}$	13.225 (1.941)	289.847 (380.103)	0.850 (0.556)	0.850 (0.556)	1.000
			$\boldsymbol{\beta}$	4.760 (0.484)	196.285 (79.729)	0.197 (0.126)	0.197 (0.126)	1.000
			$\boldsymbol{\gamma}$	4.524 (0.440)	158.350 (22.406)	0.786 (0.538)	0.786 (0.538)	1.000
			$\mathbf{t}$	6.002 (0.611)	206.475 (32.106)	0.462 (0.295)	0.462 (0.295)	1.000
		0.5	$\mathbf{c}$	12.152 (1.820)	293.030 (295.637)	0.839 (0.598)	0.870 (0.721)	0.999
			$\boldsymbol{\beta}$	4.274 (0.470)	197.480 (62.500)	0.134 (0.107)	0.139 (0.130)	1.000
			$\boldsymbol{\gamma}$	4.398 (0.477)	157.514 (22.070)	0.618 (0.538)	0.634 (0.573)	1.000
			$\mathbf{t}$	5.288 (0.617)	205.642 (32.137)	0.285 (0.180)	0.295 (0.244)	1.000
	0.75	0	$\mathbf{c}$	16.573 (2.341)	174.071 (7.039)	1.058 (0.767)	1.077 (0.859)	0.999
			$\boldsymbol{\beta}$	5.920 (0.604)	209.986 (26.982)	0.246 (0.150)	0.248 (0.151)	1.000
			$\boldsymbol{\gamma}$	5.606 (0.606)	179.493 (30.000)	1.036 (0.746)	1.048 (0.764)	1.000
			$\mathbf{t}$	7.463 (0.798)	199.170 (15.005)	0.550 (0.333)	0.554 (0.340)	1.000
		0.5	$\mathbf{c}$	15.132 (2.348)	173.974 (6.107)	1.079 (0.799)	1.141 (1.067)	0.998
			$\boldsymbol{\beta}$	5.341 (0.626)	211.293 (28.672)	0.162 (0.131)	0.169 (0.181)	1.000
			$\boldsymbol{\gamma}$	5.385 (0.618)	170.798 (25.789)	0.750 (0.625)	0.792 (0.761)	1.000
			$\mathbf{t}$	6.699 (0.795)	197.413 (15.355)	0.352 (0.230)	0.373 (0.444)	1.000
2	0.25	0	$\mathbf{c}$	10.996 (2.275)	513.213 (51.033)	0.286 (0.416)	0.296 (0.430)	0.999
			$\boldsymbol{\beta}$	1.383 (0.217)	35.049 (28.143)	0.025 (0.027)	0.028 (0.049)	1.000
			$\boldsymbol{\gamma}$	0.499 (0.091)	418.437 (43.190)	0.017 (0.023)	0.019 (0.029)	1.000
			$\mathbf{t}$	0.361 (0.059)	83.403 (17.391)	0.009 (0.011)	0.010 (0.014)	1.000
		0.5	$\mathbf{c}$	11.855 (3.221)	520.458 (51.460)	0.348 (0.548)	2.247 (4.776)	0.944
			$\boldsymbol{\beta}$	1.501 (0.313)	33.850 (26.683)	0.026 (0.033)	0.138 (0.256)	0.999
			$\boldsymbol{\gamma}$	0.561 (0.128)	458.706 (44.688)	0.021 (0.036)	0.078 (0.192)	1.000
			$\mathbf{t}$	0.391 (0.091)	87.504 (17.176)	0.009 (0.013)	0.043 (0.088)	1.000
	0.5	0	$\mathbf{c}$	6.950 (1.419)	324.585 (33.133)	0.223 (0.298)	0.223 (0.298)	1.000
			$\boldsymbol{\beta}$	0.898 (0.138)	33.233 (10.525)	0.017 (0.017)	0.017 (0.017)	1.000
			$\boldsymbol{\gamma}$	0.334 (0.065)	526.984 (93.729)	0.013 (0.019)	0.013 (0.019)	1.000
			$\mathbf{t}$	0.225 (0.035)	58.877 (10.448)	0.007 (0.009)	0.007 (0.009)	1.000
		0.5	$\mathbf{c}$	7.460 (2.040)	328.233 (36.859)	0.202 (0.315)	0.242 (0.541)	0.997
			$\boldsymbol{\beta}$	0.940 (0.188)	35.208 (11.128)	0.015 (0.018)	0.025 (0.070)	1.000
			$\boldsymbol{\gamma}$	0.382 (0.092)	528.264 (86.124)	0.010 (0.017)	0.011 (0.020)	1.000
			$\mathbf{t}$	0.242 (0.056)	59.220 (11.779)	0.006 (0.008)	0.007 (0.015)	1.000
	0.75	0	$\mathbf{c}$	11.226 (2.440)	510.040 (49.299)	0.398 (0.547)	0.540 (1.620)	0.998
			$\boldsymbol{\beta}$	1.380 (0.202)	16.987 (9.574)	0.028 (0.030)	0.036 (0.071)	1.000
			$\boldsymbol{\gamma}$	0.502 (0.091)	422.548 (41.402)	0.020 (0.031)	0.022 (0.035)	1.000
			$\mathbf{t}$	0.355 (0.061)	73.342 (14.241)	0.009 (0.009)	0.011 (0.026)	1.000
		0.5	$\mathbf{c}$	11.512 (3.287)	504.319 (53.230)	0.286 (0.502)	1.197 (3.429)	0.956
			$\boldsymbol{\beta}$	1.479 (0.317)	16.032 (8.866)	0.021 (0.029)	0.089 (0.176)	0.999
			$\boldsymbol{\gamma}$	0.569 (0.141)	382.581 (40.175)	0.013 (0.022)	0.033 (0.105)	1.000
			$\mathbf{t}$	0.387 (0.123)	71.871 (13.531)	0.006 (0.007)	0.024 (0.151)	1.000

ignored and the model is misspecified. Second, the estimation biases of estimators in the overfitting case are still larger than those of the proposed method since the latent homogeneous structure is ignored in the overfitting case. It indicates that it is necessary to consider the homogeneous structure among individuals and parameters to improve the estimation accuracy. Third, our proposed estimators perform closely to the oracle estimators with the true homogeneous structure. The NMI values for different parameters across various error settings are also close to 1. It demonstrates that our proposed method can correctly identify the true homogeneous subgroup pattern for the MKQR model. It provides a nice balance between the heterogeneity and the homogeneous structures across individuals and parameters in a data-driven manner.

6 Real Data Analysis

In financial studies, the asset pricing models using firm characteristics as risk factors are commonly adopted to account for the stock returns of firms. In this section, we consider the panel MKQR model based on the Fama-French three factors (Fama and French, 1993) for a data set about $m = 49$ industry portfolios daily returns during the period of July 1, 2019 to June 30, 2020 ($n = 253$). The selected period may contain some potential heterogeneous kink effects in some portfolios which are triggered by the COVID-19 pandemic. This dataset can be freely downloaded from Prof. Kenneth French's website `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`. We consider the Fama-French three factors, which are market-to-book ratio (Rm-Rf), the log of the firm's asset size (SMB) and the book value of debt over market value (HML). Some certain degrees of homogeneity across portfolios and regression coefficients may exist because the portfolios from the close industry sectors may share the same homogeneous effects of some factors. Thus, we apply the proposed method to detect the latent homogeneous structure.

Here, we take $\tau = 0.5$ as an illustration and the analysis at $\tau = 0.05$ and 0.95 are reported in the Supplement. Our first step is to determine the threshold covariate among the three factors. We apply the quantile score-based test in Zhong et al. (2021) to test the existence of

kink effect for each portfolio and find that the factor SMB possesses the kink effects in more portfolios than other two factors. For a simple illustration, we consider the factor SMB as the threshold covariate denoted by x_{ij} and denote the other two factors as $\mathbf{z}_{ij} = (Z_{ij,1}, Z_{ij,2})^{\mathrm{T}}$. The MKQR model is represented as

$$y_{ij} = c_{i,1} + \beta_{i,0} x_{ij} + \sum_{k=1}^{K_i} \beta_{i,k} (x_{ij} - t_{i,k})_+ + \mathbf{z}_{ij}^{\mathrm{T}} \boldsymbol{\gamma}_i + \epsilon_{ij}, \tag{6.1}$$

where y_{ij} is the daily return for the ith portfolio at the jth time. The homogeneous structure is also assumed to exist in the individual fixed effects $c_{i,1}$ similar to Gu and Volgushev (2019). We remark that we just focus on one threshold covariate in the model, which may ignores some potential kink effects in other covariates, but it simplifies the latent structure identification to achieve the purpose of the model parsimony. One can of course conduct homogeneity pursuit for multiple threshold covariates and the estimation procedure is similar.

Next, we explore the homogeneous structure of the MKQR model by using the proposed binary segmentation method. We find that 36 portfolios have no kink effect in the factor SMB, 9 portfolios experience one kink point in the factor SMB and the remaining 4 portfolios have two kink points in the factor SMB. Next, we detect the homogeneous structures for three clusters of portfolios with the same estimated kink points. To choose the tuning parameter δ in the binary segmentation method, we use the cross-validation prediction errors (CVPE) as the evaluation criterion. That is, we divide the observations for each portfolio into L equal-sized subsamples, denoted by $\mathcal{J}_l$ for $l = 1, \cdots, L$. Then, the CVPE for each cluster of portfolios with $K = 0, 1, 2$ kink points is defined as

$$\mathrm{CVPE}_K = \frac{1}{m_K L} \sum_{l=1}^{L} \sum_{i \in \mathcal{I}_K} \sum_{j \in \mathcal{J}_l} \rho_\tau \left(Y_{ij} - \widehat{Y}_{ij}^{(-l)} \right), \tag{6.2}$$

where $\mathcal{I}_K$ is the subset of portfolios corresponding to K estimated kink points with $K = 0, 1, 2$, $m_K = |\mathcal{I}_K|$ is the cardinality of $\mathcal{I}_K$, and $\widehat{Y}_{ij}^{(-l)}$ is the predicted values obtained by the data without $\mathcal{J}_l$ using the proposed MKQR model. In our analysis, we choose $L = 10$, i.e. the 10-fold cross-validation.

For the group of portfolios with no kink effect, the latent homogeneous structures are identified for the regression coefficients of the Fama-French three factors and the intercepts, which can be regarded as an extension of Ke et al. (2016) into quantile regression. The optimal threshold parameters selected by CVPE is $(\delta_c, \delta_{\boldsymbol{\gamma}}) = (0.001, 0.001)$. The estimated coefficients and the grouping results for the 36 industry portfolios are reported in the Supplement, respectively. We observe that the portfolio industries that are closely related tend to be clustered together or share the same regression coefficients of some factors. For example, Food (food products), Soda (candy, soft drinks, snacks, etc.), Beer (beer, beverages, wine and liquor) and Util (utilities for daily life) * industries which concern human's daily consumption share the most common loading coefficients for the intercept and three factors. Similar phenomena are also presented in the Steel (steel works), Mach (machinery), Autos (automobiles and trucks), Ships (shipbuilding and railroad equipment) and Trans (transportation) industries. The estimated coefficients reflect that the three factors show a positive impact on the most industry portfolios. However, the HML factor for Drugs (pharmaceutical products) and Gold (precious metals) appears a negative impact on the portfolio returns. The potential reason may be that the precious metal industry and the pharmaceutical products are easily affected by the COVID-19 pandemic in this period. For better visualization, we also draw the scatter plot of estimated coefficients before and after performing homogeneity pursuit in Figure 1. We can see that the preliminary coefficients that are close enough tend to be grouped together via homogeneity pursuit.

To save the space, the results for the group with one estimated kink point are summarized in the Supplement. Table 4 reports the estimated parameters before and after conducting homogeneity pursuit for the group with two estimated kink points. The optimal tuning parameters $(\delta_c, \delta_{\boldsymbol{\gamma}}, \delta_{\boldsymbol{\beta}}, \delta_{\mathbf{t}})$ are $(0, 0, 0, 0.01)$ selected by minimizing CVPE. When the tuning parameter δ is 0, there is no homogeneous group of the corresponding parameter. From Table 4, we observe that the BusSv (business services) and Meals(restaurants, hotels, motels) industries experience the two same kink points. From these tables, we can conclude that the homogeneity pursuit can help to reduce the number of the unknown parameters and achieve

*The detailed definitions of portfolio names are provided in Prof. Kenneth French's website `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_49_ind_port.html`

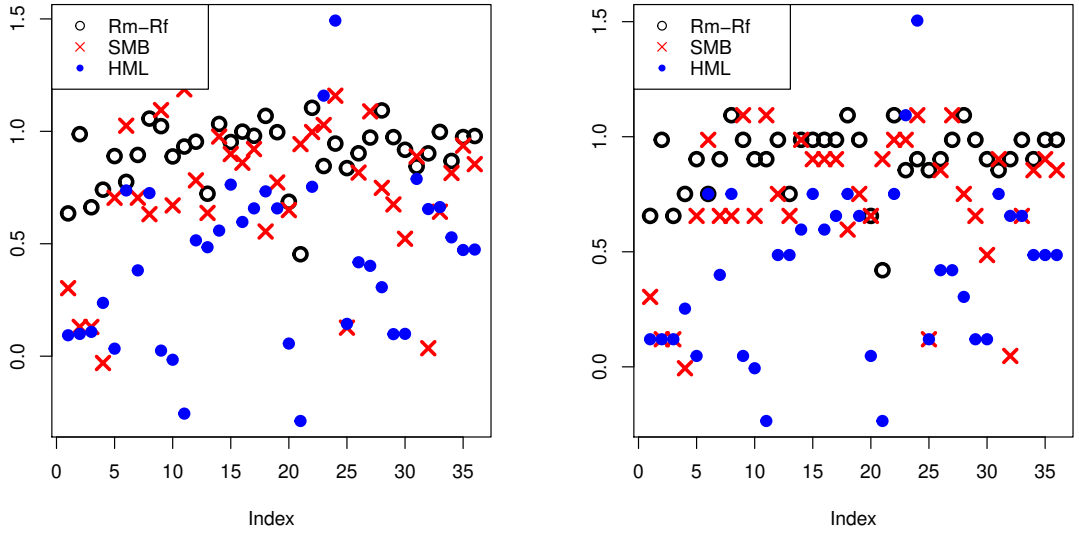

Figure 1: The estimated coefficients of three factors before (left) and after (right) homogeneity pursuit for the 36 industrial portolios without kink effect at $\tau = 0.5$.

a parsimonious model for panel data analysis. Meanwhile, it can provide more insights on the latent grouping structure of the portfolios which may be useful for portfolio investment.

Table 4: The estimated coefficients for portfolios with two kink points before and after conducting homogeneity pursuit.

		$\widehat{c}$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{t}_1$	$\widehat{t}_2$
Before	BusSv	0.889	1.378	-0.940	0.790	0.949	0.432	-0.960	0.935
	Meals	1.514	1.730	-1.373	2.580	0.905	0.580	-1.100	1.160
	Banks	0.498	1.131	-1.089	0.818	0.858	0.827	-0.445	0.228
	Other	8.464	4.705	-4.425	1.362	0.736	0.193	-1.929	1.080
After	BusSv	0.471	1.212	-0.813	0.861	0.951	0.430	**-0.581**	**1.080**
	Meals	0.954	1.591	-1.513	2.866	0.912	0.558	**-0.581**	**1.080**
	Banks	0.634	1.223	-1.054	0.704	0.860	0.826	**-0.581**	0.390
	Other	8.465	4.706	-4.425	1.362	0.736	0.194	-1.929	**1.080**

We further show the necessity of homogeneity pursuit for the predictability of the MKQR models under three different cases like the simulations: the overfitting model, the proposed model with homogeneity pursuit and the underfitting model which assumes all portfolios share the same homogeneous model. We compare their 10-fold cross-validation prediction errors (CVPE) for each group of portfolios with the different numbers of kink points. We

define the relative prediction error (RPE) of each case relative to the underfitting case as

$$\text{RPE} = \frac{\text{CVPE(the case of interest)}}{\text{CVPE(the underfitting case)}}. \tag{6.3}$$

Figure 2 displays the relative prediction errors across 10 repetitions for each group with different kink points. From the figure, our proposed method (the green dotted line) with homogeneity pursuit has the smallest relative prediction errors for all different portfolio groups. It further demonstrates that incorporating the homogeneity information can also improve the predictability of the MKQR model for panel data analysis.

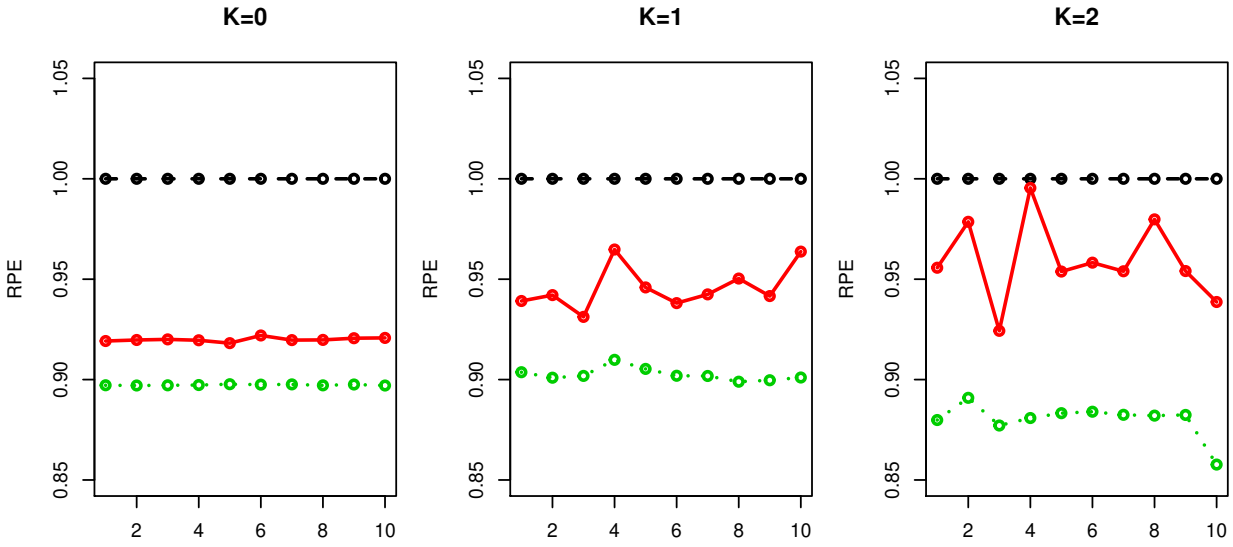

Figure 2: The relative prediction errors across 10 repetitions for $K = 0, 1$ and 2 at $\tau = 0.5$. The black dashed line represents the underfitting case, the red solid line represents the overfitting case and the green dotted line represents the proposed method.

Supplementary Material. The notations, the regularity conditions, the proofs of Theorems, some additional simulation results and real data analysis as well as discussions are included in a separate online supplemental file.

Acknowledgement. We thank the Editor, the Associate Editor and two referees for their encouragements and insightful comments which have substantially improved the paper. All authors equally contributed to the paper and are listed in the alphabetical order. Sun's research was supported by National Natural Science Foundation of China grants 71873085

and 71833004. Zhang's research was supported by NNSFC grant 11931014. Zhong's research was supported by NNSFC grants 11922117 and 71988101, Fujian Provincial Science Fund 2019J06004 and the 111 Project B13028.

References

Ando, T. and Bai, J. (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, 31:163–191.

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

Chen, Y. (2021). Jump or kink: on super-efficiency in segmented linear regression breakpoint estimation. *Biometrika*, 108(1):215–222.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fan, J., Li, R., Zhang, C., and Zou, H. (2020). *Statistical Foundations of Data Science*. Chapman and Hall/CRC.

Greene, W. H. (2018). *Econometric Analysis, 8th Edition*. Pearson.

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

Guo, C. and Li, J. (2022). Homogeneity and structure identification in semiparametric factor models. *Journal of Business & Economic Statsitics*, 40:408–422.

Hansen, B. E. (2017). Regression kink with an unknown threshold. *Journal of Business & Economic Statistics*, 35(2):228–240.

Hidalgo, J., Lee, J., and Seo, M. H. (2019). Robust inference for threshold regression models. *Journal of Econometrics*, 210(2):291–309.

Hsiao, C. (2014). *Analysis of Panel Data*. Cambridge University Press.

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

Ke, Z. T., Fan, J., and Wu, Y. (2015). Homogeneity pursuit. *Journal of the American Statistical Association*, 110(509):175–194.

Lee, E. R., Noh, H., and Park, B. U. (2014). Model selection via bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, 109(505):216–229.

Li, J., Zhang, W., and Kong, E. (2018). Factor models for asset returns based on transformed factors. *Journal of Econometrics*, 207:432–448.

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

Lin, C.-C. and Ng, S. (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods*, 1:42–55.

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

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 67:91–108.

Wang, W., Phillips, P., and Su, L. (2018). Homogeneity pursuit in panel data models: Theory and application. *Journal of Applied Econometrics*, 33:797–815.

Wood, S. N. (2001). Minimizing model fitting objectives that contain spurious local minima by bootstrap restarting. *Biometrics*, 57(1):240–244.

Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics*, pages 299–313.

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

Zhang, Y., Zhou, Q., and Jiang, L. (2017). Panel kink regression with an unknown threshold. *Economics Letters*, 157:116–121.

Zhong, W., Wan, C., and Zhang, W. (2021). Estimation and inference for multi-kink quantile regression. *Journal of Business & Economic Statistics*, forthcoming:https://doi.org/10.1080/07350015.2021.1901720.