A Class of Semiparametric Models with Homogeneous Structure for Panel Data Analysis

Tao Huang¹, Youquan Pei², Jinhong You¹, and Wenyang Zhang*³

¹School of Statistics and Management, Shanghai University of Finance and Economics, CN
²School of Economics, Shandong University, CN
³Department of Mathematics, University of York, UK

Abstract

Stimulated by the analysis of a dataset from China about Covid-19, we propose a class of semiparametric models for panel data analysis. The proposed models account for both homogeneity and heterogeneity among the individuals of a panel data. They strike a nice balance between parsimony and risk of misspecification. Although stimulated by the analysis of a particular dataset, the proposed models apply to very broad range of panel data analysis, they are powerful in exploring nonlinear dynamic patterns of impacts of covariates or transformed covariates. An estimation procedure is presented, and its asymptotic properties are established. Intensive simulation studies are also conducted to demonstrate how well the estimation procedure works and the risk of ignoring homogeneity or heterogeneity among individuals in panel data analysis. Finally, we apply the proposed models and estimation procedure to the Covid-19 data from China, and reveal some interesting dynamic patterns of the impacts of some important factors.

Keywords: Homogeneous structure, Semiparametric model, Panel data analysis

*The corresponding author, Email: wenyang.zhang@york.ac.uk
1 Introduction

The data, which stimulates this paper, come from 29 provinces (for the sake of convenience, municipalities are also called provinces in this paper) in China. It includes the daily number of Covid-19 infected cases, of people from Wuhan to a province, of cured cases, and maximum daily temperature in the province from 9 January 2020 to 25 March 2020, it also includes the population size of the province. This dataset is a typical panel dataset with each individual of length 77. We focus on the log ratio, $y_{i,t}$, of the cumulative number of infected cases in the $i$th province at time point $t$ to that at time point $t - 1$. We call $y_{i,t}$ the infection ratio at time point $t$ in the $i$th province. As cured cases may not have much effect on infection ratio, we exclude it from our analysis and focus on maximum daily temperature in a province, denoted by $x_{i,t,1}$, and the number of people from Wuhan to the province, denoted by $x_{i,t,2}$. What we are interested in are: how maximum daily temperature in a province and the number of people from Wuhan to the province affect the infection ratio in the province? Whether their impacts change over time? If they do, what are the dynamic patterns of the impacts? Whether the impacts vary over different provinces? Are there provinces sharing the same impacts? If there are, which provinces share the same impacts?

The varying coefficient models are a powerful tool to explore nonlinear dynamic patterns of impacts of covariates. There is much literature about the varying coefficient models for cross sectional data (Fan and Zhang, 1999, 2000; Zhang and Lee, 2000; Zhang et al., 2002, 2009; Cheng and Fan, 2016; Cheng et al., 2016), time series data (Cai et al., 2000; Lei et al., 2016) and panel data (Li et al., 2015, 2017; Feng et al., 2021, and references therein). We
start with the varying coefficient panel data model, which leads to

\[ y_{i,t} = b_i + a_{i,1}(u_{i,t})x_{i,t,1} + a_{i,2}(u_{i,t})x_{i,t,2} + \epsilon_{i,t}, \quad u_{i,t} = t/77. \]  (1.1)

Whilst the dynamic patterns of the impacts of \( x_{i,t,1} \) and \( x_{i,t,2} \) have been reflected and formulated by \( a_{i,1}(\cdot) \) and \( a_{i,2}(\cdot) \), this modelling assumes the contribution of covariates to response is linear, functional coefficients though. This may not be realistic, indeed, in Section 5, we will see this assumption is not valid, and the contribution of covariates is in fact through transformed covariates. Furthermore, this modelling has not taken into account the effects of previous infection ratios. To overcome these problems takes us to the following model

\[ y_{i,t} = b(u_{i,t}) + \sum_{j=1}^{p} a_{i,j}(u_{i,t})g_{i,j}(y_{i,t-j}) + a_{i,p+1}(u_{i,t})g_{i,p+1}(x_{i,t,1}) + a_{i,p+2}(u_{i,t})g_{i,p+2}(x_{i,t,2}) + \eta_i + \epsilon_{i,t}, \]  (1.2)

where \( b(\cdot) \), \( a_{i,k}(\cdot) \)s and \( g_{i,k}(\cdot) \)s are unknown functions, and \( \eta_i \)s are random effects. To answer the question that which provinces share the same impacts, we assume some of \( a_{i,k}(\cdot) \)s are the same, and some of \( g_{i,k}(\cdot) \)s are the same. We don’t assume which \( a_{i,k}(\cdot) \)s or \( g_{i,k}(\cdot) \)s are the same, and let data identify. Putting this modelling idea in a more formal and generic form leads to the following class of semiparametric models for panel data analysis.

Rather than confining us to the Covid-19 dataset, we assume \( y_{i,t} \) is the observation of the response variable of the \( i \)th individual at time point \( t \), \( (u_{i,t}, X_{i,t}) \) the vector of the corresponding covariates, where \( X_{i,t} = (x_{i,t,1}, \ldots, x_{i,t,q})^T \) is of dimension \( q \), \( u_{i,t} \) is a scalar.
We assume
\[ y_{i,t} = b(u_{i,t}) + \sum_{j=1}^{p} a_{i,j}(u_{i,t}) g_{i,j}(y_{i,t-j}) + \sum_{l=1}^{q} a_{i,p+l}(u_{i,t}) g_{i,p+l}(x_{i,t,l}) + \eta_i + \epsilon_{i,t} \]  \tag{1.3}
when \( t = p + 1, \cdots, T_i, \ i = 1, \cdots, n \). \( x_{i,t,l} \) is the \( l \)th component of \( X_{i,t} \), \( \eta_i \) a random effect with mean 0 and variance \( \sigma^2_\eta \). \( \epsilon_{i,t} \) is a random error with mean 0, variance \( \sigma^2 \). \( b(\cdot) \), \( a_{i,j}(\cdot) \)s, and \( g_{i,j}(\cdot) \)s are all unknown functions to be estimated.

Apparently, (1.3) is not identifiable. To make it identifiable, we assume
\[ E\{a_{i,j}(u_{i,t})\} = 1, \ E\{g_{i,j}(y_{i,t-j})\} = 0, \ E\{a_{i,p+l}(u_{i,t})\} = 1, \ E\{g_{i,p+l}(x_{i,t,l})\} = 0, \]  \tag{1.4}
for \( i = 1, \ldots, n, \ j = 1, \ldots, p \) and \( l = 1, \ldots, q \).

We also assume
\[ a_{i,j}(\cdot) = \begin{cases} \alpha_1(\cdot) & \text{if } (i,j) \in D_1 \\ \vdots & \vdots \\ \alpha_H(\cdot), & \text{if } (i,j) \in D_H \end{cases} \]
\[ g_{i,j}(\cdot) = \begin{cases} \beta_1(\cdot) & \text{if } (i,j) \in \Delta_1 \\ \vdots & \vdots \\ \beta_m(\cdot) & \text{if } (i,j) \in \Delta_m \end{cases} \]  \tag{1.5}
where \( \alpha_k(\cdot) \)'s and \( \beta_k(\cdot) \)'s are unknown functions to be estimated, and \( \{D_1, \ldots, D_H\} \) is an unknown partition of
\[ \{(i,j) : i = 1, \cdots, n; \ j = 1, \cdots, p+q\}, \]
and so is \( \{\Delta_1, \cdots, \Delta_m\} \). Except for the practical implication mentioned before, condition (1.5) also serves, in statistical modelling, to make model (1.3) more parsimonious.

\*We allow the time series length \( T_i \) to be heterogeneous across \( i \), which corresponding to unbalanced panel data.
Model (1.3) under conditions (1.4) and (1.5) is the model we are going to address in this paper.

We would like to stress that (1.3) is a large class of semiparametric models, which include nonparametric autoregressive models (Huang and Yang, 2004; Sun et al., 2014; Lei et al., 2016; Sun, 2016; Kalli and Griffin, 2018, and references therein), varying coefficient models and additive models (Wood et al., 2015; Chen et al., 2018; Sang et al., 2020, and references therein).

Condition (1.5) is also called a homogeneous structure. Homogeneity pursuit is a very useful statistical modelling idea, and recent literature about it includes Bonhomme and Manresa (2015), Bahadori et al. (2015), Ke et al. (2016), Ke et al. (2022), Su et al. (2016, 2019), Wang et al. (2018), Chen (2019), Li et al. (2019), Ren et al. (2019), Pei et al. (2020), Wu et al. (2020), Xu et al. (2020), Lian et al. (2021), Xiao et al. (2021), Guo and Li (2022) and references therein.

In this paper, we make contributions on three fronts. Firstly, we propose a general class of semiparametric panel data models with homogeneous structure, where dynamic nonparametric panel data models, varying coefficient models and additive models are special cases. Secondly, we develop a three step approach for identifying the latent group structure and estimating group-specific functions. Thirdly, we demonstrate the advantage of the proposed methodology by asymptotic theory and empirical analysis.

The rest of the paper is organised as follows. We describe our estimation procedure and the computational algorithm to implement it in Section 2. In Section 3, we present the asymptotic properties of the estimators resulted from either the proposed estimation, overfitting, or underfitting. Overfitting and underfitting will be defined at the beginning of Section 3. Intensive simulation studies are conducted in Section 4 to demonstrate how
well the proposed estimation procedure works and the risk of ignoring homogeneity or heterogeneity among the individuals in a panel dataset. In Section 5, we apply the proposed models and estimation procedure to the Covid-19 dataset, and explore how maximum daily temperature in a province and the number of people from Wuhan to the province affect the infection ratio in the province. We will identify which provinces in China share the same impacts, and find out the dynamic patterns of the impacts. All theoretical proofs and technical conditions are put in the supplementary material. Throughout this paper, a superscript $\tau$ indicates the transpose of a vector or a matrix. $\xrightarrow{D}$ indicates convergence in a distribution.

2 Estimation procedure

2.1 Estimation method

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

For the sake of convenience in the homogeneity pursuit in our estimation procedure, without loss of generality, we assume the range of each covariate involved in (1.3) and of $y_{i,t}$ is $[0, 1]$.

We apply B-spline decomposition, with knots being equally placed, to deal with the unknown functions in (1.3). The number of knots used for the final estimation is larger than that for the initial estimation due to the common structure identified being used, therefore, the basis functions for the final estimation are different to that for the initial estimation. Specifically, we use $B_\ell(\cdot)$, $\ell = 1, \cdots, K$, to denote the B-spline basis functions
used for the final estimation, $B^{(0)}(\cdot), \ell = 1, \cdots, K_0$, for the initial estimation. Let

$$
\bar{B}_{i,u,\ell} = \frac{1}{T_i} \sum_{t=1}^{T_i} B_\ell(u_{i,t}), \quad \bar{B}_{i,y,j,\ell} = \frac{1}{T_i} \sum_{t=1+j}^{T_i} B_\ell(y_{i,t-j}),
$$

$$
\bar{B}_{i,x,l,\ell} = \frac{1}{T_i} \sum_{t=1}^{T_i} B_\ell(x_{i,t,l}). \quad B_{i,0,t} = (B_1(u_{i,t}), \cdots, B_K(u_{i,t}))^\tau,
$$

$$
B_{i,u,t} = (B_1(u_{i,t}) - \bar{B}_{i,u,1}, \cdots, B_K(u_{i,t}) - \bar{B}_{i,u,K})^\tau,
$$

$$
B_{i,y,j,t} = (B_1(y_{i,t-j}) - \bar{B}_{i,y,j,1}, \cdots, B_K(y_{i,t-j}) - \bar{B}_{i,y,j,K})^\tau,
$$

$$
B_{i,x,l,t} = (B_1(x_{i,t,l}) - \bar{B}_{i,x,l,1}, \cdots, B_K(x_{i,t,l}) - \bar{B}_{i,x,l,K})^\tau.
$$

Under condition (1.4), in the final estimation,

$$
b(u_{i,t}) \approx B^\tau_{i,0,t} b, \quad a_{i,j}(u_{i,t}) \approx 1 + B^\tau_{i,u,t} a_{i,j}, \quad a_{i,p+l}(u_{i,t}) \approx 1 + B^\tau_{i,u,t} a_{i,p+l} \quad (2.1)
$$

$$
g_{i,j}(y_{i,t-j}) \approx B^\tau_{i,y,j,t} g_{i,j}, \quad g_{i,p+l}(x_{i,t,l}) \approx B^\tau_{i,x,l,t} g_{i,p+l},
$$

where $i = 1, \ldots, n$, $j = 1, \ldots, p$ and $l = 1, \ldots, q$. We use superscript (0) to denote the counterpart of this notation in the initial estimation, e.g. $B^{(0)}_{i,0,t}$ is the counterpart of $B_{i,0,t}$.

In the initial estimation,

$$
b(u_{i,t}) \approx B^{(0)\tau}_{i,0,t} b^{(0)}, \quad a_{i,j}(u_{i,t}) \approx 1 + B^{(0)\tau}_{i,u,t} a^{(0)}_{i,j}, \quad a_{i,p+l}(u_{i,t}) \approx 1 + B^{(0)\tau}_{i,u,t} a^{(0)}_{i,p+l} \quad (2.2)
$$

$$
g_{i,j}(y_{i,t-j}) \approx B^{(0)\tau}_{i,y,j,t} g^{(0)}_{i,j}, \quad g_{i,p+l}(x_{i,t,l}) \approx B^{(0)\tau}_{i,x,l,t} g^{(0)}_{i,p+l}.
$$

### 2.1.1 Initial estimation

We treat all unknown functions in (1.3) as different functions to estimate. Applying the least squares estimation and the approximations in (2.2), we have the following objective
function
\[
\sum_{t=p+1}^{T_i} \left\{ y_{i,t} - B_{i,0,t}^{(0)} b_i^{(0)} - \sum_{j=1}^{p} \left( 1 + B_{i,u,t}^{(0)} a_{i,j}^{(0)} \right) B_{i,y,j,t}^{(0)} g_{i,j}^{(0)} - \sum_{l=1}^{q} \left( 1 + B_{i,x,l,t}^{(0)} a_{i,p+l}^{(0)} \right) B_{i,x,l,t}^{(0)} g_{i,p+l}^{(0)} \right\}^2.
\] (2.3)

We minimise (2.3) with respect to \(b_i^{(0)}, a_{i,j}^{(0)}\), \(g_{i,j}^{(0)}\), \(a_{i,p+l}^{(0)}\), and \(g_{i,p+l}^{(0)}\), and denote the minimisers as \(b_i^{(0)}, a_{i,j}^{(0)}\), \(g_{i,j}^{(0)}\), \(a_{i,p+l}^{(0)}\) and \(g_{i,p+l}^{(0)}\). We use
\[
\tilde{b}_i(\cdot) = \left( B_1^{(0)}(\cdot), \ldots, B_{K_0}^{(0)}(\cdot) \right)^\tau b_i^{(0)},
\]
\[
\tilde{a}_{i,j}(\cdot) = 1 + \left( B_1^{(0)}(\cdot) - B_{i,u,1}^{(0)}, \ldots, B_{K_0}^{(0)}(\cdot) - B_{i,u,K_0}^{(0)} \right)^\tau a_{i,j}^{(0)},
\]
\[
\tilde{g}_{i,j}(\cdot) = \left( B_1^{(0)}(\cdot) - B_{i,y,j,1}^{(0)}, \ldots, B_{K_0}^{(0)}(\cdot) - B_{i,y,j,K_0}^{(0)} \right)^\tau g_{i,j}^{(0)},
\]
\[
\tilde{a}_{i,p+l}(\cdot) = 1 + \left( B_1^{(0)}(\cdot) - B_{i,u,1}^{(0)}, \ldots, B_{K_0}^{(0)}(\cdot) - B_{i,u,K_0}^{(0)} \right)^\tau a_{i,p+l}^{(0)},
\]
\[
\tilde{g}_{i,p+l} = \left( B_1^{(0)}(\cdot) - B_{i,x,l,1}^{(0)}, \ldots, B_{K_0}^{(0)}(\cdot) - B_{i,x,l,K_0}^{(0)} \right)^\tau g_{i,p+l}^{(0)},
\]
for \(i = 1, \ldots, n, j = 1, \ldots, p, l = 1, \ldots, q\), as the initial estimators of \(b(\cdot) + \eta_i\), \(a_{i,j}(\cdot)\), \(g_{i,j}(\cdot)\), \(a_{i,p+l}(\cdot)\), and \(g_{i,p+l}(\cdot)\), respectively.

### 2.1.2 Homogeneity pursuit

Throughout this paper, for any function \(f(\cdot)\) on \([0,1]\), we define
\[
\|f\|_2 = \int_0^1 f^2(v) dv.
\]
We first estimate the partition \( \{ \Delta_1, \cdots, \Delta_m \} \) by using the idea in Vogt and Linton (2017). We start with \( \tilde{g}_{1,1}(\cdot) \), and compute

\[
\delta_{ij} = \frac{1}{\| \tilde{g}_{1,1} \|^2} \int_0^1 \left\{ \tilde{g}_{1,1}(v) - \tilde{g}_{i,j}(v) \right\}^2 dv, \quad (i, j) \in S,
\]

where \( S = \{(i, j) : i = 1, \ldots, n; j = 1, \ldots, p+q\} \). Let \( \hat{\Delta}_1 \) be the set of all \((i, j)\)s that satisfy \( \delta_{ij} < C \), where \( C \) is a given threshold. In practice, it can be selected by cross-validation.

We select an element, say \((i_0, j_0)\), from \( S - \hat{\Delta}_1 \), and compute

\[
\delta_{ij} = \frac{1}{\| \tilde{g}_{i_0,j_0} \|^2} \int_0^1 \left\{ \tilde{g}_{i_0,j_0}(v) - \tilde{g}_{i,j}(v) \right\}^2 dv, \quad (i, j) \in S - \hat{\Delta}_1.
\]

Let \( \hat{\Delta}_2 \) be the set of all \((i, j)\)s in \( S - \hat{\Delta}_1 \) that satisfy \( \delta_{ij} < C \). Continuously doing so, we get \( \{ \hat{\Delta}_1, \ldots, \hat{\Delta}_m \} \) and use it to estimate \( \{ \Delta_1, \ldots, \Delta_m \} \).

Using exactly the same approach, we can get the estimator \( \{ \hat{D}_1, \ldots, \hat{D}_H \} \) of \( \{ D_1, \ldots, D_H \} \).

### 2.1.3 Final estimation

Let \( M_i = (F_{i,p+1}, \ldots, F_{i,T_i})^\top \) and \( F_{i,t} \) be

\[
\mathbf{B}_{i,0,t}^\top \mathbf{b} + \sum_{j=1}^p (1 + \mathbf{B}_{i,u,t} \mathbf{a}_{i,j}) \mathbf{B}_{i,y,j,t}^\top \mathbf{g}_{i,j} + \sum_{l=1}^q (1 + \mathbf{B}_{i,u,t} \mathbf{a}_{i,p+l}) \mathbf{B}_{i,x,l,t}^\top \mathbf{g}_{i,p+l}
\]

with \( \mathbf{a}_{i,\ell}, \ell = 1, \cdots, n, \ell = 1, \cdots, p+q \), being replaced by \( \mathbf{\alpha}_k \) if \((\ell, \ell) \in \hat{\Delta}_k\), \( \mathbf{g}_{i,\ell} \), \( \ell = 1, \cdots, n, \ell = 1, \cdots, p+q \), being replaced by \( \mathbf{\beta}_k \) if \((\ell, \ell) \in \hat{\Delta}_k\).

To take the within cluster correlation into account in the final estimation, we have to get the estimators of \( \sigma_\eta^2 \) and \( \sigma^2 \) first. We estimate \( \sigma_\eta^2 \) and \( \sigma^2 \) based on the residuals of
working independence fitting. Specifically, we minimise

$$\sum_{i=1}^{n} (Y_i - M_i)^\tau (Y_i - M_i)$$

with respect to \((b, \alpha_1, \ldots, \alpha_H, \beta_1, \ldots, \beta_m)\), and denote the minimiser by

\((\hat{b}, \hat{\alpha}_1, \ldots, \hat{\alpha}_H, \hat{\beta}_1, \ldots, \hat{\beta}_m)\).

Let \(\tilde{M}_i\) be \(M_i\) with \(b, \alpha_k\)s and \(\beta_k\)s being replaced by \(\tilde{b}, \tilde{\alpha}_k\)s and \(\tilde{\beta}_k\)s. We have the following objective function:

$$\sum_{i=1}^{n} \left\| (Y_i - \tilde{M}_i)(Y_i - \tilde{M}_i)^\tau - \sigma^2 I_{T_i} - \sigma^2_\eta 1_{T_i}1_{T_i}^\tau \right\|_2$$

(2.4)

Applying the weighted least squares estimation, by simple calculation, we have the objective function:

$$L(b, \alpha_1, \ldots, \alpha_H, \beta_1, \ldots, \beta_m) = \sum_{i=1}^{n} (Y_i - M_i)^\tau \left( I_{T_i} - \frac{\sigma^2_\eta}{T_i\sigma^2_\eta + \hat{\sigma}^2} 1_{T_i}1_{T_i}^\tau \right) (Y_i - M_i)$$

(2.5)

where \(Y_i = (y_{i,p+1}, \ldots, y_{i,T_i})^\tau\).

We minimise (2.5) with respect to \((b, \alpha_1, \ldots, \alpha_H, \beta_1, \ldots, \beta_m)\), and denote the minimiser by

\((\hat{b}, \hat{\alpha}_1, \ldots, \hat{\alpha}_H, \hat{\beta}_1, \ldots, \hat{\beta}_m)\).
The final estimators of $b(\cdot)$, $a_{i,j}(\cdot)$, $g_{i,j}(\cdot)$, $a_{i,p+l}(\cdot)$ and $g_{i,p+l}(\cdot)$ are

$$\hat{b}(\cdot) = (B_1(\cdot), \ldots, B_K(\cdot))^\tau \hat{b},$$

$$\hat{a}_{i,j}(\cdot) = 1 + (B_1(\cdot) - \bar{B}_{i,u,1}, \ldots, B_K(\cdot) - \bar{B}_{i,u,K})^\tau \hat{\alpha}_k, \quad \text{if } (i, j) \in \hat{D}_k,$n

$$\hat{g}_{i,j}(\cdot) = (B_1(\cdot) - \bar{B}_{i,y,j,1}, \ldots, B_K(\cdot) - \bar{B}_{i,y,j,K})^\tau \hat{\beta}_k, \quad \text{if } (i, j) \in \hat{\Delta}_k,$n

$$\hat{a}_{i,p+l}(\cdot) = 1 + (B_1(\cdot) - \bar{B}_{i,u,1}, \ldots, B_K(\cdot) - \bar{B}_{i,u,K})^\tau \hat{\alpha}_k, \quad \text{if } (i, p + l) \in \hat{D}_k,$n

$$\hat{g}_{i,p+l}(\cdot) = (B_1(\cdot) - \bar{B}_{i,x,l,1}, \ldots, B_K(\cdot) - \bar{B}_{i,x,l,K})^\tau \hat{\beta}_k, \quad \text{if } (i, p + l) \in \hat{\Delta}_k.$n

### 2.2 Computational algorithm

The main hurdle in the implementation of the proposed estimation method is the minimisation of (2.3) and (2.5). Because neither of them has a minimiser with closed form, we are going to minimise them by an iterative approach.

Applying the three-step spline estimation method proposed in Hu et al. (2019), we can easily get the initial values for $g_{i,j}(0)$’s and $g_{i,p+l}(0)$’s in (2.3). Replacing $g_{i,j}(0)$’s and $g_{i,p+l}(0)$’s in (2.3) by their initial values and minimising (2.3) with respect to $b_i(0)$, $a_{i,j}(0)$’s and $a_{i,p+l}(0)$’s, we take the resulting minimisers as the initial values for $b_i(0)$, $a_{i,j}(0)$’s and $a_{i,p+l}(0)$’s. Replacing $b_i(0)$, $a_{i,j}(0)$’s and $a_{i,p+l}(0)$’s in (2.3) by their initial values and minimising (2.3) with respect to $g_{i,j}(0)$’s and $g_{i,p+l}(0)$’s, we take the resulting minimisers as the updated values for $g_{i,j}(0)$’s and $g_{i,p+l}(0)$’s. Continuing this iterative process until convergence, we get the minimiser of (2.3).

The minimisation of (2.5) is similar to that of (2.3). We start by choosing the initial values for $\beta_k$’s in (2.5) based on $\tilde{g}_{i,j}$’s and $\tilde{g}_{i,p+l}$’s obtained in Section 2.1.1. Specifically, the
initial value for $\beta_k$ is taken to be

$$\frac{1}{|\Delta_k|} \sum_{(i,j) \in \Delta_k} \tilde{g}_{i,j}. $$

We replace $\beta_k$s in (2.5) by their initial values, then minimise (2.5) with respect to $b$ and $\alpha_k$s. We take the resulting minimisers as the initial values for $b$ and $\alpha_k$s and substitute them for $b$ and $\alpha_k$s in (2.5), then minimise (2.5) with respect to $\beta_k$’s. We take the resulting minimisers as the updated values for $\beta_k$s, and continue this iterative process until convergence to get the minimiser of (2.5).

2.3 A couple of remarks

Remark 1. The estimation of the partition in Section 2.1.2 can be further improved. Taking $\{\hat{\Delta}_1, \ldots, \hat{\Delta}_m\}$ as an example, we can apply the following iterative process to improve the estimation:

1. Compute

$$\bar{g}_k(\cdot) = \frac{1}{|\hat{\Delta}_k|} \sum_{(i,j) \in \hat{\Delta}_k} \tilde{g}_{i,j}(\cdot), \quad k = 1, \ldots, \hat{m},$$

where $|\hat{\Delta}_k|$ is the cardinality of $\hat{\Delta}_k$.

2. For each $(i, j) \in S$, we compute

$$c_k = \int_0^1 \{\bar{g}_k(v) - \tilde{g}_{i,j}(v)\}^2 dv, \quad k = 1, \ldots, \hat{m}.$$ 

If $c_{k_0}$ is the smallest $c_k$, then $(i, j)$ belongs to set $\hat{\Delta}_{k_0}^{(1)}$. This leads to a partition $\{\hat{\Delta}_1^{(1)}, \ldots, \hat{\Delta}_m^{(1)}\}$ of $S$. 

12
We treat \( \{\hat{\Delta}_1^{(1)}, \ldots, \hat{\Delta}_m^{(1)}\} \) as \( \{\hat{\Delta}_1, \ldots, \hat{\Delta}_m\} \), and repeat (1) and (2). Continuously doing so until convergence, we get an improved estimator of the partition \( \{\Delta_1, \ldots, \Delta_m\} \).

**Remark 2.** In the final estimation, an iterative process can be used to further improve the final estimators. Specifically, we substitute the minimiser of (2.5) for \( \tilde{b}, \tilde{\alpha}_k, \tilde{\beta}_k \) in (2.4), and minimise (2.4). We treat the resulting minimiser as updated \( \hat{\sigma}_2^\eta \) and \( \hat{\sigma}^2 \), and substitute them for the \( \hat{\sigma}_2^\eta \) and \( \hat{\sigma}^2 \) in (2.5), then minimise (2.5) and substitute the resulting minimiser for \( \tilde{b}, \tilde{\alpha}_k, \tilde{\beta}_k \) in (2.4), and minimise (2.4) to get updated \( \hat{\sigma}_2^\eta \) and \( \hat{\sigma}^2 \). Continue this iterative process until convergence, and substitute the converged \( \hat{\sigma}_2^\eta \) and \( \hat{\sigma}^2 \) for the \( \hat{\sigma}_2^\eta \) and \( \hat{\sigma}^2 \) in (2.5), then minimise (2.5) to get the final \( \tilde{b}, \tilde{\alpha}_k, \tilde{\beta}_k \) based on which improved final estimators of \( b(\cdot), a_{i,j}(\cdot), g_{i,j}(\cdot), a_{i,p+l}(\cdot), \) and \( g_{i,p+l}(\cdot) \) are obtained.

### 2.4 Selecting the values of the tuning parameters

In this section, we address several practical problems regarding the selection of the values of the tuning parameters in the proposed methods.

#### 2.4.1 Selecting the optimal number of knots in the spline decomposition

Following the lead of Hu et al. (2019), we select the number of knots \( K_0 \) via a Bayesian information criterion (BIC) approach:

\[
\text{BIC}(K_0) = \log(\text{RSS}) + N \frac{\log T}{T}.
\]

The residual sum of squares (RSS) is defined in (2.3), which reflects the goodness of fit. Here, \( N = (2p + 2q + 1)K_0 \) controls the complexity of the model, and we assume \( T_i = T \) for the sake of clarification. The optimal number of knots \( K_0 \) can be estimated by minimising the above BIC. According to Theorem 1, it is reasonable to choose the optimal number of
knots $K_0$ on the interval $[0.5T^{1/5}, 2T^{1/5}]$, where $[a]$ denotes the largest integer not larger than $a$. The number of knots $K$ in the final estimation can be selected using the same approach.

### 2.4.2 Selecting the threshold $C$ in the homogeneity pursuit

The homogeneity pursuit method relies on prior information about the threshold $C$. However, it is usually unknown in practical applications and needs to be determined via a data-driven method. Actually, selecting $C$ is equivalent to selecting $\hat{H}$ and $\hat{m}$. In this paper, following the lead of Lian et al. (2021), we also apply the cross-validation method to select the two tuning parameters, $\hat{H}$ and $\hat{m}$.

In particular, we implement a $V$-fold cross-validation approach. For a given pair $\{H, m\}$, we remove the $1/V$th member of the observed time points for

$$\{(y_{i,t}, u_{i,t}, X_{i,t}), i = 1, \ldots, n, t = 1, \ldots, T_i\}$$

as a validation set. We estimate the semiparametric model (1.3) with identified homogeneity structure on the remaining data, compute the squared error between $y_{i,t}$ and the fitted values on the validation set, and repeat this procedure $V$ times to calculate the cross-validated mean squared error. The optimal $\hat{H}$ and $\hat{m}$ can be estimated by minimising the cross-validation mean squared error.

### 3 Asymptotic theory

In this section, we are going to demonstrate the advantage of the proposed methodology by asymptotic theory. We will consider three different approaches:
• **Over-fitting:** Treat all unknown functions in model (1.3) as different functions to estimate. Namely, directly apply the final estimation described in Section 2.1.3 under the assumption that either of the two partitions involved is \( \{(i,j)\} : i = 1, \ldots, n; j = 1, \ldots, p+q \} \).

• **Under-fitting:** Directly apply the final estimation to estimate the unknown functions under the assumption

\[
a_{1,j}(\cdot) = \cdots = a_{n,j}(\cdot), \quad g_{1,j}(\cdot) = \cdots = g_{n,j}(\cdot),
\]

for \( j = 1, \ldots, p+q \).

• **Correct fitting:** The proposed estimation is used to estimate the unknown functions.

In this section, we assume \( T_i \to \infty, n \) possibly diverges to infinity, but \( H \) and \( m \) are fixed. This agrees with many real applications where \( H \) and \( m \) are expected to be small, and thus, there is a significant reduction of the unknown functions by clustering similar functions.

To keep the presentation concise, we state the asymptotic theorems in this section and leave all technical proofs in the supplementary material. Let

\[
B^{(0)}(u) = \left( B_{1}^{(0)}(u), \ldots, B_{K_0}^{(0)}(u) \right)^\tau,
\]

\[
B_i^{(0)}(u) = \left( B_{1}^{(0)}(u) - \tilde{B}_{i,u,1}^{(0)}, \ldots, B_{K_0}^{(0)}(u) - \tilde{B}_{i,u,K_0}^{(0)} \right)^\tau,
\]

\[
B_{i,j}^{(0)}(x) = \left( B_{1}^{(0)}(x) - \tilde{B}_{i,x,j,1}^{(0)}, \ldots, B_{K_0}^{(0)}(x) - \tilde{B}_{i,x,j,K_0}^{(0)} \right)^\tau,
\]

\[
\tilde{A}_{i,j}(u) = (0^\tau, \ldots, B_i^{(0)r}(u), \ldots, 0^\tau)^\tau, \quad \tilde{C}_{i,j}(x) = (0^\tau, \ldots, B_{i,j}^{(0)r}(x), \ldots, 0^\tau)^\tau.
\]
\(\tilde{A}_{i,j}(u)\) is a \((p + q + 1)K_0\)-dimensional vector consisting of \((p + q + 1)\) blocks of length \(K_0\), with the \((j + 1)\)th block being \(B_i^{(0)\tau}(u)\), others all being 0. \(\tilde{G}_{i,j}(x)\) is a \((p + q)K_0\)-dimensional vector consisting of \((p + q)\) blocks of length \(K_0\), with the \(j\)th block being \(B_{i,j}^{(0)\tau}(x)\), others all being 0.

**Theorem 1** (Over-fitting case). For any \(i, i = 1, \ldots, n\), and \(1 \leq j \leq p + q\), under the technical conditions (C1)–(C3) and (C5) in the supplementary material, if \(K_0 = O(T_i^{1/5})\), we have

\[
T_i^{2/5} \left( \hat{b}(u) - b(u) - r(u) \right) \xrightarrow{D} N \left( 0, \tilde{\Sigma}_1^{-1} \tilde{\Xi}_1 \tilde{\Sigma}_1^{-1} \tilde{A}_{i,0}(u) \right),
\]

\[
T_i^{2/5} \left( \hat{a}_{i,j}(u) - a_{i,j}(u) - r_{i,j}(u) \right) \xrightarrow{D} N(0, \tilde{\Sigma}_1^{\tau}(u) \tilde{\Xi}_1 \tilde{\Sigma}_1^{\tau} \tilde{A}_{i,j}(u)),
\]

\[
T_i^{2/5} \left( \hat{g}_{i,j}(x) - g_{i,j}(x) - d_{i,j}(x) \right) \xrightarrow{D} N(0, \tilde{\Sigma}_2^{\tau}(x) \tilde{\Xi}_2 \tilde{\Sigma}_2^{\tau} \tilde{G}_{i,j}(x)),
\]

where \(\hat{b}(u), \hat{a}_{i,j}(u)\) and \(\hat{g}_{i,j}(x)\) are obtained by overfitting, the bias terms \(r(u) = b(u) - B^{(0)\tau}(u)B^{(0)}(u), r_{i,j}(u) = a_{i,j}(u) - B^{(0)\tau}(u)a_{i,j}(0), d_{i,j}(x) = g_{i,j}(x) - B^{(0)\tau}(x)g_{i,j}(0)\) are of order \(O(K_0^{-2})\). \(\tilde{\Xi}_1\) and \(\tilde{\Xi}_2\) are defined in condition (C5). \(\tilde{\Sigma}_1 = (\sigma^2 + \sigma_\eta^2) \int_0^1 E \left[ \tilde{D}_i(u)\tilde{D}_i(u) \right] du, \tilde{\Sigma}_2 = (\sigma^2 + \sigma_\eta^2) \int_0^1 E \left[ \tilde{Z}_i(x)\tilde{Z}_i(x) \right] dx,\) and

\[
\tilde{D}_i(u) = \left( B^{(0)\tau}(u), g_{i,1}(x_{i,1})B_i^{(0)\tau}(u), \ldots, g_{i,p+q}(x_{i,p+q})B_i^{(0)\tau}(u) \right)^\tau,
\]

\[
\tilde{Z}_i(x) = \left( a_{i,1}(u_{i,1})B_{i,1}^{(0)\tau}(x), \ldots, a_{i,p+q}(u_{i,p+q})B_{i,p+q}(x) \right)^\tau.
\]

From Theorem 1, it is easy to see that the convergence rate of the estimators \(\hat{b}(u), \hat{a}_{i,j}(u)\) and \(\hat{g}_{i,j}(x)\) are of order \(T_i^{-2/5}\), which is as expected, as we assume that the functions are twice differentiable.

**Theorem 2** (Under-fitting case). Suppose the functions \(a_{i,j}(\cdot)\)s are sufficiently separated,
\[ \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{1} \{a_{i,j}(u) - \bar{a}_1(u)\}^2 du > C > 0, \quad \bar{a}_1(u) = \frac{1}{n} \sum_{i=1}^{n} a_{i,j}(u), \]

then we have

\[ \|\hat{a}_{i,j} - a_{i,j}\|_2 > C. \]

Similarly, we have

\[ \|\hat{b} - b\|_2 > C, \quad \|\hat{g}_{i,j} - g_{i,j}\|_2 > C, \]

where \(\hat{b}(u), \hat{a}_{i,j}(u)\) and \(\hat{g}_{i,j}(x)\) are obtained by underfitting.

Theorem 2 shows that the estimators obtained by underfitting are not consistent.

Before presenting Theorem 3, we introduce some notations: let

\[
\begin{align*}
B(u) &= (B_1(u), \ldots, B_K(u))^\tau, & B_i(u) &= (B_1(u) - \bar{B}_{i,u,1}, \ldots, B_K(u) - \bar{B}_{i,u,K})^\tau, \\
B_{i,j}(x) &= (B_1(x) - \bar{B}_{i,x,j,1}, \ldots, B_K(x) - \bar{B}_{i,x,j,K})^\tau, \\
A_{i,j}(u) &= (0^\tau, \ldots, B_{i,j}^\tau(u), \ldots, 0^\tau)^\tau, & G_{i,j}(x) &= (0^\tau, \ldots, B_{i,j}^\tau(x), \ldots, 0^\tau)^\tau.
\end{align*}
\]

\(A_{i,j}(u)\) is a \((p + q + 1)K\)-vector consisting of \((p + q + 1)\) blocks of length \(K\), with the \((j + 1)th\) block being \(B_{i,j}^\tau(u)\), others all being \(0\). \(G_{i,j}(x)\) is a \((p + q)K\)-vector consisting of \((p + q)\) blocks of length \(K\), with the \(j\)th block being \(B_{i,j}^\tau(x)\), others all being \(0\).

**Theorem 3** (Correct fitting case). Under the technical conditions (C1)–(C5) in the supplementary material, let \(N = \sum_{i=1}^{n} T_i\). When \(K_0 = O(T_i^{1/5})\) and \(K = O(N^{1/5})\), we have

\[
N^{2/5} \left( \hat{b}(u) - b(u) - r(u) \right) \xrightarrow{D} N \left( 0, \ A_{i,0}^\tau(u) \Sigma_1^{-1} \Sigma_1^{-1} A_{i,0}(u) \right),
\]

\[
N^{2/5} \left( \hat{a}_{i,j}(u) - a_{i,j}(u) - r_{i,j}(u) \right) \xrightarrow{D} N \left( 0, \ A_{i,j}^\tau(u) \Sigma_1^{-1} \Sigma_1^{-1} A_{i,j}(u) \right),
\]
\[
N^{2/5}\left(\hat{g}_{i,j}(x) - g_{i,j}(x) - d_{i,j}(x)\right) \overset{D}{\to} N\left(0, \mathbf{G}_{i,j}^*(x) \Xi_2^{-1} \Sigma_2 \Xi_2^{-1} \mathbf{G}_{i,j}(x)\right),
\]
where \(\hat{b}(u), \hat{a}_{i,j}(u)\) and \(\hat{g}_{i,j}(x)\) are obtained by correct fitting, and the bias terms \(r(u), r_{i,j}(u), d_{i,j}(x)\) are defined as Theorem 1 with order of \(K^{-2}\). \(\Xi_1\) and \(\Xi_2\) are defined in condition (C5), and

\[
\Sigma_1 = (\sigma^2 + \sigma^2) \int_0^1 E[D_k(u)D_k(u)^\tau] du, \quad \Sigma_2 = (\sigma^2 + \sigma^2) \int_0^1 E[Z_k(x)Z_k(x)^\tau] dx,
\]

\[
D_k(u) = (B^\tau(u), \beta_k(x_{i,t,1})B^\tau_i(u), \ldots, \beta_k(x_{i,t,p+q})B^\tau_i(u))^\tau, \text{ if } (i, j) \in \mathcal{D}_k,
\]

\[
Z_k(x) = (\alpha_k(u_{i,t})B^\tau_{i,1}(x), \ldots, \alpha_k(u_{i,t})B^\tau_{i,p+q}(x))^\tau, \text{ if } (i, j) \in \Delta_k.
\]

Theorem 3 shows that the convergence rates of the estimators \(\hat{b}(u), \hat{a}_{i,j}(u)\) and \(\hat{g}_{i,j}(x)\) are of order \(N^{-2/5}\). This together with Theorem 1 shows that the estimators obtained by correct fitting have convergence rates with a higher order than those obtained by overfitting. Therefore, they are more accurate.

Because the asymptotic variance of the above estimators has a very complicated form and it is not clear how to estimate it consistently, constructing a statistical inference based on the asymptotic normality established in Theorem 3 can be very challenging. In this paper, we do not consider the inference problem and leave it as an open question.

4 Simulations

In this section, we use a simulated example to demonstrate how well the proposed estimation procedure works and the risk of ignoring the homogeneity or heterogeneity among individuals.

In particular, the data are generated from the following semiparametric panel data
model:
\[ y_{i,t} = b(u_{i,t}) + a_{i,1}(u_{i,t})g_{i,1}(y_{i,t-1}) + a_{i,2}(u_{i,t})g_{i,2}(x_{i,t}) + \eta_i + \epsilon_{i,t}, \]

where \( u_{i,t} \) are generated from a uniform distribution \( U(0, 1) \). The exogenous covariates are generated via a first-order autoregressive (AR) process with parameters \( \rho \) and \( \sigma \), denoted by AR(1; \( \rho, \sigma \)), which is given by
\[ x_{i,t} = \rho x_{i,t-1} + \sigma \zeta_{i,t}, \]
where \( \zeta_{i,t} \) are generated from i.i.d. standard normal random variables. In particular, \( x_{i,t} \sim AR(1; 0.6, 0.5) \), and \( \epsilon_{i,t} \) and \( \eta_i \) are independently generated from \( N(0, 0.1^2) \).

The homogeneous coefficient functions and additive functions are generated as follows:

\[ b(u) = 1.5 \cos(2\pi u), \]

and for each \( i = 1, 2, \ldots, n \),

\[ a_{i,1}(u) = \begin{cases} 
1.3u \sin(2\pi u) + 1, & \text{when } i = 1, 2, \ldots, n/2, \\
1.3u \cos(2\pi u) + 1, & \text{when } i = n/2 + 1, \ldots, n. 
\end{cases} \]

\[ a_{i,2}(u) = 2 \sin(1.5\pi u) - 1.2(u - 0.5)(1 - u) + 1, \]

and

\[ g_{i,1}(y_{i,t-1}) = -0.4(3 - y_{i,t-1}^2)/(1 + y_{i,t-1}^2), \]

\[ g_{i,2}(x_{i,t,2}) = \begin{cases} 
2 \cos(\pi x_{i,t,2}/2) + 1.8 \sin(\pi x_{i,t,2}/3), & \text{when } i = 1, 2, \ldots, n/2, \\
1.5 \sin(\pi x_{i,t,2}/4) - 1.2 \cos(\pi x_{i,t,2}/3), & \text{when } i = n/2 + 1, \ldots, n. 
\end{cases} \]

For the latent structure functions, Figure 1 describes a toy example of \( a_{i,j}(\cdot) \) and \( g_{i,j}(\cdot) \) \((i = 1, \ldots, n \text{ and } j = 1, 2)\), where different colours denote different functions.

We run the simulated example 100 times with various \( n \) and \( T \) and compare our proposed approach to its potential competitors based on the following performance metrics:
Figure 1: Toy example of the latent structures imposed on (a) $a_{i,j}(\cdot)$ and (b) $g_{i,j}(\cdot)$.

**Estimation accuracy** For an estimator $\hat{a}_{i,j}(\cdot)$ of $a_{i,j}(\cdot)$, its estimation accuracy can be evaluated based on the mean integrated squared error:

$$
\text{MISE}(\hat{a}_{i,j}) = E\left\{ \int (\hat{a}_{i,j}(u) - a_{i,j}(u))^2 du \right\}.
$$

To prevent the performance from being dominated by the poor boundary behaviour, we let the integral domain be a non-boundary region, which is between the 1st and 99th quantiles of $\{u_{it}\}$.

**Homogeneity pursuit consistency** To evaluate the distance between the detected homogeneous structure and the true one, we use the normalised mutual information (NMI) (Ke et al., 2016), which measures the similarity between two partitions. Suppose that $C = \{C_1, C_2, \ldots\}$ and $D = \{D_1, D_2, \ldots\}$ are two partitions of $\{1, \ldots, n\}$, the NMI is defined as

$$
\text{NMI}(C, D) = \frac{I(C, D)}{\left[H(C) + H(D)\right]/2},
$$

where $I(C, D)$ is the mutual information between $C$ and $D$.
where
\[
I(\mathcal{C}, \mathcal{D}) = \sum_{k,j} \left( \frac{|C_k \cap D_j|}{n} \log \left( \frac{n |C_k \cap D_j|}{|C_k||D_j|} \right) \right),
\]
and
\[
H(\mathcal{C}) = -\sum_k \left( \frac{|C_k|}{n} \log \left( \frac{|C_k|}{n} \right) \right).
\]
The NMI takes values in $[0,1]$ with larger values indicating a higher level of similarity between the two partitions. For an estimated partition $\hat{\mathcal{D}} = \{\mathcal{D}_1, \ldots, \mathcal{D}_{\hat{H}}\}$ of $\{(i,j) : 1, \ldots, n, j = 1, \ldots, p + q\}$, obtained in stage 2 of the proposed estimation procedure in Section 2.1, we calculate $\text{NMI}(\hat{\mathcal{D}}, \mathcal{D})$ to assess how close the true homogeneous structure in $a_{i,j}(\cdot)$ is to the estimated one. Similarly, we can calculate $\text{NMI}(\hat{\Delta}, \Delta)$ to assess how close the true homogeneous structure in $g_{i,j}(\cdot)$ is to the estimated one.

The sample size $T_i = 100, 200$ or $400$ and $n = 20$ or $40$. We estimate the unknown functions by overfitting, correct fitting and underfitting, respectively. The mean and standard deviation of NMI for the resulting estimators are presented in Table 1 and the mean squared errors of the resulting estimators are presented in Table 2.

### 4.1 Initial estimation

The initial estimates for subject-specific varying-coefficient functions $a_{i,j}(\cdot)$ and additive functions $g_{i,j}(\cdot)$ are presented in Figures 2 and 3. From these figures, we can see that the initial estimates for $a_{i,1}(\cdot)$ and $g_{i,2}(\cdot)$ have an obvious group structure, which is a preliminary validation of our simulation setup.
Figure 2: Average estimated varying-coefficient functions $a_{i,j}(\cdot)$ based on 100 simulations when $n = 20$ and $T = 200$. Left: Estimates of $a_{i,1}(\cdot)$. Right: Estimates of $a_{i,2}(\cdot)$.

Figure 3: Average estimated additive functions $g_{i,j}(\cdot)$ based on 100 simulations when $n = 20$ and $T = 200$. Left: Estimates of $g_{i,1}(\cdot)$. Right: Estimates of $g_{i,2}(\cdot)$.

4.2 Homogeneity pursuit

Based on the initial estimation, we can identify the latent group structure using the procedure in Section 2.1.2. Note that in our simulation, the true partition for both $a_{i,j}(\cdot)$ and
$g_{i,j}(\cdot)$ has three groups. Figures 4 and 5 present one simulation result of this step, which is aligned with our simulation setting. For example, in Figure 4, the left-hand side shows the initial estimates for $a_{i,j}(\cdot)$, the middle is the $L_2$ distance matrix calculated based on $a_{i,j}(\cdot)$ and the right-hand side is the result of the homogeneity pursuit, which has three groups. The group membership is denoted by the colours.

![Figure 4: (a) The estimates, $\hat{a}_{i,j}(\cdot)$, of specific coefficient functions for each subject $i$. (b) The distance matrix constructed by the $L_2$ distance. (c) The identified group structure based on homogeneity pursuit.](image)

The mean and standard deviation of the NMIs for $a_{i,j}(\cdot)$ and $g_{i,j}(\cdot)$ are shown in Table 1. We can observe that as $T$ gets larger, the performance becomes better. This makes sense since the initial estimates improve as the sample size $T$ increases.

### 4.3 Final estimation

We have illustrated the performance of homogeneity pursuit. We now assume that the true group structure is known in this step and adopt the procedure in Section 2.1.3 to estimate the group-specific functions. The final estimates are shown in Figures 6 and 7. In particular, we compare the estimation performance by plotting the final estimates and the
true functions in the same figure. Obviously, the final estimates are all close to the true functions.

Finally, in Table 2, we compare the performance of correct fitting with overfitting and underfitting. We can observe that among these three methods, correct fitting provides the best performance with the lowest MISE. For correct fitting and overfitting, the overall estimation accuracy improves as $T$ increases, which is somewhat expected, since as $T$ increases, the initial estimates are better and the subsequent procedures rely more on the initial estimates. Note that the functions estimated by underfitting are not consistent,
Figure 6: Average estimated varying-coefficient functions $\alpha_k(\cdot)$ based on 100 simulations when $n = 20, T = 200$.

Figure 7: Average estimated additive functions $\beta_k(\cdot)$ based on 100 simulations when $n = 20, T = 200$.

which agrees with Theorem 2.

Note that overfitting and underfitting, which either ignore or mistakenly specify the homogeneous structure, perform worse than correct fitting, which illustrates the importance of incorporating the homogeneous structure in panel data analysis.
Table 2: The mean and standard deviation of MISE for the estimated functions based on 100 simulations when \( n = 20 \).

| Function | Method          | \( T = 100 \)          | \( T = 200 \)          | \( T = 400 \)          |
|----------|-----------------|------------------------|------------------------|------------------------|
| \( b(\cdot) \) | Correct fitting | 0.0445(0.0140)         | 0.0340(0.0141)         | 0.0268(0.0116)         |
|          | Overfitting     | 0.1881(0.0217)         | 0.1370(0.0196)         | 0.1336(0.0189)         |
|          | Underfitting    | 0.1258(0.0177)         | 0.1363(0.0146)         | 0.1521(0.0192)         |
| \( a_{i,1}(\cdot) \) | Correct fitting | 0.0972(0.0284)         | 0.0497(0.0102)         | 0.0406(0.0080)         |
|          | Overfitting     | 0.2148(0.0493)         | 0.0964(0.0099)         | 0.0723(0.0134)         |
|          | Underfitting    | 0.3915(0.0097)         | 0.3970(0.0062)         | 0.4027(0.0099)         |
| \( a_{i,2}(\cdot) \) | Correct fitting | 0.0487(0.0083)         | 0.0106(0.0024)         | 0.0061(0.0016)         |
|          | Overfitting     | 0.1290(0.0190)         | 0.0417(0.0034)         | 0.0262(0.0023)         |
|          | Underfitting    | 0.1885(0.0071)         | 0.0902(0.0041)         | 0.0713(0.0108)         |
| \( g_{i,1}(\cdot) \) | Correct fitting | 0.0400(0.0100)         | 0.0164(0.0046)         | 0.0156(0.0039)         |
|          | Overfitting     | 0.1216(0.0148)         | 0.0619(0.0058)         | 0.0518(0.0055)         |
|          | Underfitting    | 0.1584(0.0136)         | 0.2242(0.0153)         | 0.3124(0.0197)         |
| \( g_{i,2}(\cdot) \) | Correct fitting | 0.1323(0.0092)         | 0.0572(0.0047)         | 0.0548(0.0019)         |
|          | Overfitting     | 0.1806(0.0235)         | 0.0579(0.0057)         | 0.0403(0.0050)         |
|          | Underfitting    | 0.5974(0.0008)         | 0.6119(0.0006)         | 0.6163(0.0008)         |

5 Analysis of Covid-19 data from China

With the severity of the new coronavirus disease (Covid-19) outbreak, much literature focused on the prediction of growth trajectories, such as (Tang et al., 2021; Liu et al., 2021; Li and Linton, 2021, and references therein). Now we apply our method to analyse the Covid-19 data from a different perspective, mentioned in Section 1. For each day, we collected the cumulative number of confirmed cases \( (Z_{i,t}) \), the cumulative number of cured cases (https://github.com/CSSEGISandData/COVID-19), the number of people travelling from Wuhan to other provinces (https://qianxi.baidu.com/) and the maximum daily temperature (http://www.weather.com.cn).

The growth rate of cumulative confirmed cases (GRCC) is the response variable. It is
denoted by $y_{i,t}$ and is measured by $\log (Z_{i,t}) - \log (Z_{i,t-1})$.\footnote{The time series plot of $y_{i,t}$ for 29 provinces in China from 23 January to 8 April is presented in the supplementary material.} To explore how the influential factors contribute to the GRCC, we focus on two covariates: (1) the proportion of the population of Wuhan travelling to the $i$th province on day $t - 14$, denoted by $x_{i,t,1}$; (2) the maximum daily temperature in the $i$th province on day $t$, denoted by $x_{i,t,2}$. We normalise the covariates $x_{i,t,1}$ and $x_{i,t,2}$ to mean 0 and variance 1 first, then apply the following time-varying-coefficient additive model to fit the data

$$y_{i,t} = b(t/T) + a_{i,1}(t/T)g_{i,1}(x_{i,t,1}) + a_{i,2}(t/T)g_{i,2}(x_{i,t,2}) + \eta_i + \varepsilon_{i,t},$$

where $i = 1, \ldots, 29$ and $t = 1, \ldots, 77$. The proposed estimation procedure is implemented to estimate the unknown functions in the model, and explore how the maximum daily temperature in a province and the number of people who travelled from Wuhan to the province affect the infection ratio in the province. We also identify the provinces which share the same impacts of the two influential factors, and the dynamic patterns of the impacts.

Specifically, we first apply the proposed method to estimate the subject-specific coefficients, the obtained initial estimates of time-varying coefficients and additive functions are presented in Figure 8. We can see that both $\hat{a}_{i,j}(\cdot)$ and $\hat{g}_{i,j}(\cdot)$ have group structures, which verifies the necessity of considering a latent group structure model to characterise this homogeneity. A question naturally arises: How many groups are there and which provinces share similar impacts?

We calculate the $L_2$ distance between different functions, and identify the latent group structure based on the proposed method in Section 2.1.2. The obtained results are presented in the supplementary material. For each identified group, we plot the initial estimates of
Figure 8: (a) The estimated time-varying-coefficient functions $\hat{a}_{i,j}(\cdot)$ for each subject $i$. (b) The estimated additive functions $\hat{g}_{i,j}(\cdot)$ for each subject $i$.

time-varying coefficients or additive functions in that group, and present them in Figures 9 and 10. From Figures 9 and 10, we can roughly see that the coefficients $\hat{a}_{i,j}(\cdot)$ and $\hat{g}_{i,j}(\cdot)$ belonging to the same group have similar patterns, which illustrates that our proposed homogeneity pursuit approach performs well.

Finally, based on the identified group structure, we estimate $b(\cdot)$, $\alpha_k(\cdot)$ and $\beta_k(\cdot)$ according to the procedure proposed in Section 2.1.3. The results are presented in Figure 11. We summarise the group-specific functions and the corresponding functional changes in Table 3.
Figure 9: Identified group structure for varying coefficients $\hat{a}_{i,j}$, where $i = 1, \ldots, 29$, $j = 1, 2$.

Table 3: Final estimates for the functions and its corresponding descriptions.

| Function | Group | Function description |
|----------|-------|----------------------|
| $b(\cdot)$ | Group 1 | The coefficients decreases from positive to zero gradually. |
| $\alpha_k(\cdot)$ | Group 1 | The coefficients first increases, then decreases and finally increases to zero. |
| | Group 2 | The coefficients first decreases, then increases and finally decreases to zero. |
| | Group 3 | The coefficients first increases and then decreases to zero. |
| | Group 4 | The coefficients first decreases and then increases to zero. |
| | Group 5 | The coefficients increases from negative to zero gradually. |
| | Group 6 | The coefficients decreases from positive to zero gradually. |
| $\beta_k(\cdot)$ | Group 1 | The function decreases from positive to negative. |
| | Group 2 | The function increases gradually and then remain the same. |
| | Group 3 | The function first decreases, then increases and finally remain the same. |
| | Group 4 | The function first decreases, then increases. |
Figure 10: Identified group structure for additive functions $\hat{g}_{i,j}(\cdot)$, where $i = 1, \ldots, 29, j = 1, 2$.

6 Concluding remarks

This paper has presented a mechanism for identifying and estimating latent group structures in a class of semiparametric panel data models in which the varying coefficients and additive functions are heterogeneous across groups but homogeneous within a group and
the group membership is unknown. In particular, we have considered identifying the group membership by a distance-based approach and estimating the group-specific functions with a three-stage method. We also established the asymptotic properties for the estimators obtained by overfitting, underfitting and correct fitting. The results of the simulations and an analysis of real data were provided to illustrate the finite-sample performance of the proposed method.

References

Bahadori, M. T., Kale, D., Fan, Y., and Liu, Y. (2015). Functional subspace clustering with application to time series. *ICML’15: Proceedings of the 32nd International Conference on Machine Learning*, 37:228–237.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.
Cai, Z., Fan, J., and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, 95(451):941–956.

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

Chen, Z., Fan, J., and Li, R. (2018). Error variance estimation in ultrahigh-dimensional additive models. *Journal of the American Statistical Association*, 113(521):315–327.

Cheng, M.-Y. and Fan, J. (2016). Peter hall’s contributions to nonparametric function estimation and modeling. *The Annals of Statistics*, 44(5):1837–1853.

Cheng, M.-Y., Honda, T., and Zhang, J.-T. (2016). Forward variable selection for sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association*, 111(515):1209–1221.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics*, 27(5):1491–1518.

Fan, J. and Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying coefficient models. *Scandinavian Journal of Statistics*, 27(4):715–731.

Feng, S., Li, G., Peng, H., and Tong, T. (2021). Varying coefficient panel data model with interactive fixed effects. *Statistica Sinica*, 31(2):935–957.

Guo, C. and Li, J. (2022). Homogeneity and structure identification in semiparametric factor models. *Journal of Business and Economic Statistics*, 40(1):408–422.

Hu, L., Huang, T., and You, J. (2019). Estimation and identification of a varying-coefficient additive model for locally stationary processes. *Journal of the American Statistical Association*, 114(527):1191–1204.
Huang, J. Z. and Yang, L. (2004). Identification of non-linear additive autoregressive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):463–477.

Kalli, M. and Griffin, J. E. (2018). Bayesian nonparametric vector autoregressive models. *Journal of Econometrics*, 203(2):267–282.

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

Ke, Y., Lian, H., and Zhang, W. (2022). High dimensional dynamic covariance matrices with homogeneous structure. *Journal of Business and Economic Statistics*, 40(1):96–110.

Lei, H., Xia, Y., and Qin, X. (2016). Estimation of semivarying coefficient time series models with arma errors. *Annals of Statistics*, 44(4):1618–1660.

Li, D., Ke, Y., and Zhang, W. (2015). Model selection and structure specification in ultra-high dimensional generalised semi-varying coefficient models. *Annals of Statistics*, 43(6):2676–2705.

Li, J., Huang, C., and Zhu, H. (2017). A functional varying-coefficient single-index model for functional response data. *Journal of the American Statistical Association*, 112(519):1169–1181.

Li, J., Yue, M., and Zhang, W. (2019). Subgroup identification via homogeneity pursuit for dense longitudinal/spatial data. *Statistics in Medicine*, 38(17):3256–3271.

Li, S. and Linton, O. (2021). When will the covid-19 pandemic peak? *Journal of Econometrics*, 220(1):130–157.
Lian, H., Qiao, X., and Zhang, W. (2021). Homogeneity pursuit in single index models based panel data analysis. *Journal of Business and Economic Statistics*, 39(2):386–401.

Liu, L., Moon, H. R., and Schorfheide, F. (2021). Panel forecasts of country-level covid-19 infections. *Journal of Econometrics*, 220(1):2–22.

Pei, Y., Huang, T., Peng, H., and You, J. (2020). Network-based clustering for varying coefficient panel data models. *Journal of Business and Economic Statistics*, page to appear.

Ren, Z., Kang, Y., Fan, Y., and Lv, J. (2019). Tuning-free heterogeneous inference in massive networks. *Journal of the American Statistical Association*, 114(528):1908–1925.

Sang, P., Wang, L., and Cao, J. (2020). Estimation of sparse functional additive models with adaptive group lasso. *Statistica Sinica*, 30(3):1191–1211.

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

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

Sun, Y. (2016). Functional-coefficient spatial autoregressive models with nonparametric spatial weights. *Journal of Econometrics*, 195(1):134–153.

Sun, Y., Yan, H., Zhang, W., and Lu, Z. (2014). A semiparametric spatial dynamic model. *Annals of Statistics*, 42(2):700–727.

Tang, F., Feng, Y., Chiheb, H., and Fan, J. (2021). The interplay of demographic variables and social distancing scores in deep prediction of us covid-19 cases. *Journal of the American Statistical Association*, 116(534):492–506.
Vogt, M. and Linton, O. (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.

Wang, W., Phillips, P. C., and Su, L. (2018). Homogeneity pursuit in panel data models: Theory and application. *Journal of Applied Econometrics*, 33(6):797–815.

Wood, S. N., Goude, Y., and Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1):139–155.

Wu, H., Fan, Y., and Lv, J. (2020). Statistical insights into deep neural network learning in subspace classification. *Stat*, 9(1):e273.

Xiao, D., Ke, Y., and Li, R. (2021). Homogeneity structure learning in large-scale panel data with heavy-tailed errors. *Journal of Machine Learning Research*, 22(13):1–42.

Xu, J., Yue, M., and Zhang, W. (2020). A new multilevel modelling approach for clustered survival data. *Econometric Theory*, 36(4):707–750.

Zhang, W., Fan, J., and Sun, Y. (2009). A semiparametric model for cluster data. *Annals of Statistics*, 37(5A):2377–2408.

Zhang, W. and Lee, S.-Y. (2000). Variable bandwidth selection in varying-coefficient models. *Journal of Multivariate Analysis*, 74(1):116–134.

Zhang, W., Lee, S.-Y., and Song, X. (2002). Local polynomial fitting in semivarying coefficient model. *Journal of Multivariate Analysis*, 82(1):166–188.

35