mathematics

MDPI

Article

Bicluster Analysis of Heterogeneous Panel Data via M-Estimation

Weijie Cui [1] and Yong Li [1,2,*]

[1] School of Management, University of Science and Technology of China, Hefei 230026, China; can@mail.ustc.edu.cn
[2] New Finance Research Center, International Institute of Finance, University of Science and Technology of China, Hefei 230026, China
* Correspondence: yonglee@ustc.edu.cn

Abstract: This paper investigates the latent block structure in the heterogeneous panel data model. It is assumed that the regression coefficients have group structures across individuals and structural breaks over time, where change points can cause changes to the group structures and structural breaks can vary between subgroups. To recover the latent block structure, we propose a robust biclustering approach that utilizes M-estimation and concave fused penalties. An algorithm based on local quadratic approximation is developed to optimize the objective function, which is more compact and efficient than the ADMM algorithm. Moreover, we establish the oracle property of the penalized M-estimators and prove that the proposed estimator recovers the latent block structure with a probability approaching one. Finally, simulation studies on multiple datasets demonstrate the good finite sample performance of the proposed estimators.

Keywords: heterogeneous panel data; block structure; bicluster; M-estimation; fused penalty

MSC: 62J07

Citation: Cui, W.; Li, Y. Bicluster Analysis of Heterogeneous Panel Data via M-Estimation. *Mathematics* **2023**, *11*, 2333. https://doi.org/10.3390/math11102333

Academic Editor: Alicia Nieto-Reyes

Received: 15 April 2023
Revised: 9 May 2023
Accepted: 12 May 2023
Published: 17 May 2023

1. Introduction

Panel data models can fully utilize both cross-sectional and time-series information, making them a popular tool in fields such as economics and finance. Traditional panel data models often assume that the regression coefficients are homogeneous across individuals and over periods, which is too rigid an assumption. In many real-world applications, heterogeneity in individual and/or time dimensions is often observed. For example, in precision medicine research, different subgroups of patients may respond differently to treatments, while in economics, events such as the 2009 European debt crisis led to varying debt-to-GDP ratios among European countries. Although these heterogeneous factors are unobserved and latent, modeling them will bring significant improvement to data analysis.

Numerous estimation methods have been developed for panel data models with heterogeneous coefficients, addressing two main sources of heterogeneity: individual and period. To account for heterogeneity across individuals, a commonly used assumption is that individuals can be classified into subgroups with identical coefficients within the same subgroup but different coefficients across subgroups. Penalty-based methods have been frequently used to cluster coefficients in the individual direction. Su et al. [1] propose C-Lasso, a modified version of Lasso, for subgroup identification and coefficient estimation. This method is based on a penalized objective function inspired by the fused Lasso method introduced by Tibshirani et al. [2]. Wang and Zhu [3] study high-dimensional panel data models using a concave fused penalty method for both subgroup identification and variable selection and proved the asymptotic properties of the proposed estimator under specific regularity conditions. To capture the heterogeneity that may exist over time, structural breaks are often assumed. Qian and Su [4] use a group fused Lasso method to estimate

both the number of breaks and model parameters simultaneously. Furthermore, Qian and Su [5] employ an adaptive group fused Lasso approach that applies the shrinkage method to PLS and PGMM estimations. Their method can consistently determine the number of breaks and estimate the break dates with a probability approaching one.

However, these studies have two primary limitations. Firstly, they only account for heterogeneity in one dimension, which may be insufficient for modeling complex data prevalent in the era of big data. Secondly, their objective functions employ the least squares loss which can result in substantial estimation bias when the data distribution contains a heavy tail or outliers. Consequently, further research has been conducted to address these issues. Some researchers have focused on the two-dimensional heterogeneous panel data model, where the coefficients exhibit both subgroup structure and structural breaks. Okui and Wang [6] allow the number, timing, and size of structural breaks to vary across different subgroups and employ the K-means method and adaptive group fused Lasso method to identify the individual group structure and structural breaks, respectively. Lumsdaine et al. [7] study cases where the group structure of coefficients changes after the unknown structural break and develop a novel iterative algorithm to estimate the coefficients and recover the unknown structure. Additionally, some researchers have focused on robust estimation. For example, Zhang et al. [8] set the objective function as the sum of L_1 loss and concave pairwise fused penalty when studying the panel data model with an individual group structure and provide an easy-to-implement algorithm based on the idea of local linear approximation [9] to find local minima. Cheng et al. [10] further generalize the L_1 loss to a general loss function under the framework of M-estimation.

In this paper, we generalize the coefficient structure studied by [6,7] to a more general block structure. The regression coefficients with a block structure exhibit both an individual-group structure and temporal–structural breaks, where the individual-group structure can change at change points, and the temporal–structural breaks can vary across different groups. Furthermore, the regression coefficients are identical within the same sub-block, while they exhibit heterogeneity across different sub-blocks. This block structure is highly flexible and more general compared to the structures studied previously. Additionally, the homogeneous panel data model, the panel data model with a group structure, and the panel data model with structural breaks can all be viewed as special cases of the model being investigated in this study.

We propose a robust biclustering method based on *M*-estimation and double concave fused penalties for simultaneously recovering the unknown block structure and estimating the regression coefficients. The *M*-estimator exhibits robustness to heavy-tailed distributions and outliers, while the double concave fused penalty can automatically identify potential block structure. We develop an effective algorithm utilizing local quadratic approximation to optimize the objective function, which is computationally more efficient than the Alternating Direction Method of Multipliers (ADMM) [11] algorithm. Moreover, we establish the asymptotic convergence property of the oracle estimator and prove that the proposed estimator can recover the latent block structure with a probability approaching one. Simulation experiments on multiple datasets demonstrate that the estimator proposed in this paper has an excellent performance in finite sample situations. Additionally, models based on L_1 loss and *Huber* loss functions can achieve more accurate results than those based on L_2 loss functions in the presence of heavy-tailed data distributions.

2. Materials and Methods

2.1. Model Setting

Given panel data observations $\{(\boldsymbol{x}_{it}, y_{it}), i = 1, \cdots, N; t = 1, \cdots, T\}$, this paper explores a linear panel data model that accounts for the heterogeneity of intercept and slope coefficients across both the individual and time dimensions. The model is represented as follows,

$$y_{it} = \mu_{it} + \boldsymbol{x}_{it}^{\top} \boldsymbol{\zeta}_{it} + \epsilon_{it}, \quad i = 1, \cdots, N; t = 1, \cdots T, \tag{1}$$

where $y_{it} \in \mathbb{R}$ is the response variable, and $\boldsymbol{x}_{it} \in \mathbb{R}^{P-1}$ is the explanatory variable. The intercept term μ_{it} and the slope coefficient $\boldsymbol{\zeta}_{it} \in \mathbb{R}^{P-1}$ can vary across both individual and time dimensions in the model. The random errors are assumed to be independent and identically distributed with a mean of 0 and a standard deviation of σ, and their distribution is denoted by f.

We assume that the observed data are collected from an unknown, L^0 different blocks, where the regression coefficients are homogeneous within the same block but heterogeneous across different blocks. It is worth noting that the group structure and structural breaks can be viewed as special cases of the block structure. Below, we provide the relevant notation to describe this block structure.

Let $\boldsymbol{\beta}_{it} = (\mu_{it}, \boldsymbol{\eta}_{it}^\top)^\top$ and $\boldsymbol{z}_{it} = (1, \boldsymbol{x}_{it}^\top)^\top$. Then, Equation (1) can be expressed as follows,

$$y_{it} = \boldsymbol{z}_{it}^\top \boldsymbol{\beta}_{it} + \epsilon_{it}, \quad i = 1, \cdots, N; t = 1, \cdots, T. \tag{2}$$

We denote the block structure as $\mathscr{B} = \{\mathscr{B}_1, \cdots, \mathscr{B}_{L^0}\}$, where $\mathscr{B}_k$ represents the index set of samples belonging to the kth sub-block. If the ith individual's observation at time t belongs to the kth sub-block, then $(i,t) \in \mathscr{B}_k$. Let $\boldsymbol{\alpha}_1^0, \cdots, \boldsymbol{\alpha}_{L^0}^0$ denote the real regression coefficients corresponding to the L^0 sub-blocks, and let $\boldsymbol{\beta}_{it}^0$ be the real value of $\boldsymbol{\beta}_{it}$. Then, we have

$$\boldsymbol{\beta}_{it}^0 = \begin{cases} \boldsymbol{\alpha}_1^0, & \text{if } (i,t) \in \mathscr{B}_1, \\ \boldsymbol{\alpha}_2^0, & \text{if } (i,t) \in \mathscr{B}_2, \\ \cdots & \\ \boldsymbol{\alpha}_{L^0}^0 & \text{if } (i,t) \in \mathscr{B}_{L^0}. \end{cases} \tag{3}$$

In practical scenarios, the real block structure is often unknown. To recover the block structure described above, it is necessary to estimate the number of sub-blocks, the index sets of each block, and the block-specific coefficients.

2.2. Proposed Estimator

In this subsection, we propose a biclustering estimation method to automatically recover the block structure without specifying the number of blocks and provide robust estimates of regression coefficients via M-estimation and concave fused penalties. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_{11}^\top, \boldsymbol{\beta}_{12}^\top, \cdots, \boldsymbol{\beta}_{1T}^\top, \cdots, \boldsymbol{\beta}_{N1}^\top, \boldsymbol{\beta}_{N2}^\top, \cdots, \boldsymbol{\beta}_{NT}^\top)^\top$ denote the coefficients to be estimated. To recover the block structure under the fused sparse assumption, a natural idea is to shrink the coefficient differences $||\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{jt'}||$ between two samples (i,t) and $(j,t^{'})$ that belong to the same block $\mathscr{B}_l$ to zero. In the following, we present the objective function based on M-estimation and concave fused penalty as follows,

$$\begin{aligned} Q(\boldsymbol{\beta}; \lambda, \gamma) = &\sum_{i=1}^{N} \sum_{t=1}^{T} \rho(y_{it} - \boldsymbol{z}_{it}^\top \boldsymbol{\beta}_{it}) + \sum_{t=1}^{T} \sum_{i<j} P_\lambda(||\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{jt}||) \\ &+ \sum_{i=1}^{N} \sum_{t<t'} P_\gamma(||\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{it'}||), \end{aligned} \tag{4}$$

where the first term on the right-hand side is the regular loss function ρ in the M-estimation literature. It satisfies several conditions, including being a continuous convex function on $\mathbb{R}$, almost everywhere differentiable except for a finite set of points, having a unique global minimum at 0, and $\rho(0) = 0$. Commonly used loss functions such as least squares (L_2), absolute deviation (L_1), and Huber loss all satisfy these conditions. The second and third terms are two fused penalty terms designed to identify the individual-group structure and temporal–structure breaks.

Commonly used penalty terms include Lasso [12], Ridge [13], and Elastic net [14]. Ridge and Elastic net can shrink the fusion terms toward small values, but not exactly zero, which makes it challenging to cluster coefficients within the same block. Lasso, on the other hand, can shrink the fusion terms to exactly zero but has a tendency to introduce

bias in the estimated coefficients. This bias can impact the accuracy of the parameter estimates and potentially lead to suboptimal results in recovering the latent block structure. To achieve both automatic recovery of the parameter structure and unbiased or nearly unbiased estimation of coefficients, concave fused penalty functions have been proposed, such as the smoothly clipped absolute deviation (SCAD) [15] and the minimax concave penalty (MCP) [16]. In this paper, following the approach of Wang and Zhu [3], Ma and Huang [17], Wang et al. [18], we use the SCAD penalty function with a tuning parameter λ,

$$P_\lambda(k) = \lambda \int_0^k (1 - x/(\lambda a))_+ dx, \tag{5}$$

and the MCP penalty function with a tuning parameter γ,

$$P_\gamma(k) = \gamma \int_0^k \min\{1, (a - x/\gamma)/(a-1))\} dx, \tag{6}$$

where the fixed parameter a controls the concavity of the penalty function, and k represents pairwise differences in regression coefficients between individuals or periods. $P_\lambda(k)$ and $P_\gamma(k)$ can compress some pairwise difference values $||\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{jt}||$ and $||\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{it'}||$ to zero, thereby recovering the block structure of the regression coefficients.

For a given λ and γ, we define the proposed estimator as

$$\widehat{\boldsymbol{\beta}}(\lambda, \gamma) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{NTP}}{\arg\min}\, Q(\boldsymbol{\beta}; \lambda, \gamma). \tag{7}$$

In the following, we abbreviate $\widehat{\boldsymbol{\beta}}(\lambda, \gamma)$ as $\widehat{\boldsymbol{\beta}}$ when there is no ambiguity. Although the penalty term in Equation (4) is concave and global minimum points are difficult to obtain, local minimum points can be obtained through iterative algorithms. Ma and Huang [17] and Wang et al. [18] apply the ADMM algorithm to solve the penalized optimization problem by transforming it into a Lagrangian-constrained optimization problem. However, the algorithm is computationally intensive and involves cumbersome steps. In this paper, we develop a novel algorithm based on local quadratic approximation [15] to solve Equation (4). In the next section, we provide a detailed derivation.

2.3. Proposed Algorithm

We propose a local quadratic approximation-based algorithm to solve Equation (7). The local quadratic approximation algorithm was introduced by Fan and Li [15]. Specifically, given a non-zero value x_0 as the initial value for the penalty function $P_\lambda(|x|)$, we can apply a first-order Taylor expansion on x^2 as follows,

$$P_\lambda(|x|) = P_\lambda((x^2)^{\frac{1}{2}}) \approx P_\lambda(|x_0|) + \frac{P'_\lambda(|x_0|)}{2|x_0|}(x^2 - x_0^2). \tag{8}$$

Specifically, let $\boldsymbol{\beta}^{(k-1)}$ denote the estimate of $\boldsymbol{\beta}$ obtained after the $(k-1)$th iteration. At the kth iteration, we locally approximate $\rho(\cdot)$, $P_\lambda(\cdot)$, and $P_\gamma(\cdot)$ around $\boldsymbol{\beta}^{(k-1)}$, which yields

$$\begin{aligned} & Q^{(k)}(\boldsymbol{\beta}; \gamma, \lambda) \\ & = \sum_{i=1}^{N}\sum_{t=1}^{T} \frac{\phi(|y_{it} - z_{it}^T \boldsymbol{\beta}_{it}^{(k-1)}|)}{2|y_{it} - z_{it}^T \boldsymbol{\beta}_{it}^{(k-1)}|}(y_{it} - z_{it}^T \boldsymbol{\beta}_{it})^2 \\ & + \sum_{t=1}^{T}\sum_{i<j} \frac{P'_\lambda(||\boldsymbol{\beta}_{it}^{(k-1)} - \boldsymbol{\beta}_{jt}^{(k-1)}||)}{2||\boldsymbol{\beta}_{it}^{(k-1)} - \boldsymbol{\beta}_{jt}^{(k-1)}||}||\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{jt}||^2 \\ & + \sum_{i=1}^{N}\sum_{t<t'} \frac{P'_\gamma(||\boldsymbol{\beta}_{it}^{(k-1)} - \boldsymbol{\beta}_{it'}^{(k-1)}||)}{2||\boldsymbol{\beta}_{it}^{(k-1)} - \boldsymbol{\beta}_{it'}^{(k-1)}||}||\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{it'}||^2 \\ & + C, \end{aligned} \tag{9}$$

where $\phi(\cdot)$, $P_\lambda^{'}(\cdot)$, and $P_\gamma^{'}(\cdot)$ are the derivatives of $\rho(\cdot)$, $P_\lambda(\cdot)$, and $P_\gamma(\cdot)$, respectively. C depends on $\boldsymbol{\beta}^{(k-1)}$ and can be treated as a constant when solving for the kth estimator. By minimizing $Q^{(k)}(\boldsymbol{\beta};\gamma,\lambda)$, we obtain $\boldsymbol{\beta}^{(k)}$.

Equation (9) has an explicit solution. To simplify Equation (9), we first define some symbols as follows,

$$\boldsymbol{Y} = (y_{11},\cdots,y_{1T},\cdots,y_{N1},\cdots,y_{NT})^\top,$$
$$\boldsymbol{Z_i} = \mathrm{diag}(z_{i1}^\top,\cdots,z_{iT}^\top), \quad \boldsymbol{Z} = \mathrm{diag}(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_N),$$
$$\Delta_1 = \mathrm{diag}\left\{\sqrt{\phi(|\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\beta}^{(k-1)}|)/|\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\beta}^{(k-1)}|}\right\},$$

where the square root symbol $\sqrt{\cdot}$, the function $\phi(\cdot)$, the division symbol $\cdot/\cdot$, and the absolute value function $|\cdot|$ represent the corresponding operations performed on each element of the vector when they act on vectors.

Let $\boldsymbol{e}_i^{(c)}$ be an N-dimensional vector with 1 in the i-th dimension and 0 in the other dimensions, and let $\boldsymbol{e}_t^{(r)}$ be a T-dimensional vector with 1 in the tth dimension and 0 in the other dimensions. We define

$$\boldsymbol{\delta}_{i,j}^{(c)} = \mathrm{I}_{T\times T}\otimes\left[(\boldsymbol{e}_i^{(c)}-\boldsymbol{e}_j^{(c)})^\top\otimes\mathrm{I}_{P\times P}\right], \quad \boldsymbol{\delta}_{t,t'}^{(r)} = \mathrm{I}_{N\times N}\otimes\left[(\boldsymbol{e}_t^{(r)}-\boldsymbol{e}_{t'}^{(r)})^\top\otimes\mathrm{I}_{P\times P}\right],$$

where $\mathrm{I}_{T\times T}$, $\mathrm{I}_{P\times P}$, and $\mathrm{I}_{N\times N}$ are identity matrices, and $\otimes$ represents the Kronecker product. We concatenate $\boldsymbol{\delta}_{i,j}^{(c)}$ for $i<j$ and $\boldsymbol{\delta}_{t,t'}^{(r)}$ for $t<t^{'}$ to obtain

$$\boldsymbol{\delta}_c = (\boldsymbol{\delta}_{1,2}^{(c)\top},\cdots,\boldsymbol{\delta}_{N-1,N}^{(c)\top})^\top, \quad \boldsymbol{\delta}_r = (\boldsymbol{\delta}_{1,2}^{(r)\top},\cdots,\boldsymbol{\delta}_{T-1,T}^{(r)\top})^\top.$$

Let $\boldsymbol{U} = \boldsymbol{\delta}_c\boldsymbol{\beta}^{(k-1)}$ and $\boldsymbol{V} = \boldsymbol{\delta}_r\beta^{(k-1)}$. $\boldsymbol{U}$ and $\boldsymbol{V}$ are both NTP-dimensional vectors that can be expressed as $\boldsymbol{U} = (\boldsymbol{u}_{11}^\top,\cdots,\boldsymbol{u}_{NT}^\top)^\top$ and $\boldsymbol{V} = (\boldsymbol{v}_{11}^\top,\cdots,\boldsymbol{v}_{NT}^\top)^\top$, respectively, where $\boldsymbol{u}_{it}$ and $\boldsymbol{v}_{it}$ are P-dimensional vectors. Let $\overline{u}_{it}$ and $\overline{v}_{it}$ be the L_2 norms of $\boldsymbol{u}_{it}$ and $\boldsymbol{v}_{it}$, respectively. Let $\overline{\boldsymbol{U}} = (\overline{u}_{11},\cdots,\overline{u}_{NT})^\top$, $\overline{\boldsymbol{V}} = (\overline{v}_{11},\cdots,\overline{v}_{NT})^\top$.

Thus, Equation (9) can be written as

$$Q^{(k)}(\boldsymbol{\beta};\gamma,\lambda) = \frac{1}{2}(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\beta})^\top\boldsymbol{\Delta_1}^\top\boldsymbol{\Delta_1}(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\beta}) + \frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\Delta_2}^\top\boldsymbol{\Delta_2}\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\Delta_3}^\top\boldsymbol{\Delta_3}\boldsymbol{\beta} + C, \tag{10}$$

where $\Delta_2 = \left[\mathrm{I}_{P\times P}\otimes\sqrt{P_\lambda^{'}(\overline{\boldsymbol{U}})/\overline{\boldsymbol{U}}}\right]\delta_c$, $\Delta_3 = \left[\mathrm{I}_{P\times P}\otimes\sqrt{P_\gamma^{'}(\overline{\boldsymbol{V}})/\overline{\boldsymbol{V}}}\right]\delta_r$.

By minimizing the above equation, we obtain the iteration formula for the kth step,

$$\boldsymbol{\beta}^{(k)} = (\boldsymbol{Z}^\top\boldsymbol{\Delta_1}^\top\boldsymbol{\Delta_1}\boldsymbol{Z} + \boldsymbol{\Delta_2}^\top\boldsymbol{\Delta_2} + \boldsymbol{\Delta_3}^\top\boldsymbol{\Delta_3})^{-1}\boldsymbol{Z}^\top\boldsymbol{\Delta_1}^\top\boldsymbol{\Delta_1}\boldsymbol{Y}. \tag{11}$$

We repeat this iteration process until the norm of the difference between $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\beta}^{(k-1)}$ is smaller than a given threshold δ (set to 1e-5 in our experiments), at which point the algorithm terminates. As noted by Hunter and Li [19], this algorithm belongs to the class of MM algorithms, and its convergence is guaranteed.

To perform the iterative process, it is necessary to specify the initial value of the regression coefficients. An appropriate initial value can reduce the number of iterations and computation time. Following the approach of Wang et al. [18], we use ridge regression to obtain the initial values. The specific formula is as follows,

$$\begin{aligned}\boldsymbol{\beta}^{(0)} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{NTP}}\{\sum_{i=1}^{N}\sum_{t=1}^{\top}(y_{it}-\boldsymbol{z}_{it}^{\top}\boldsymbol{\beta}_{it})^2+\lambda^*\sum_{t=1}^{\top}\sum_{i<j}||\boldsymbol{\beta}_{it}-\boldsymbol{\beta}_{jt}||^2 \\ &\qquad +\gamma^*\sum_{i=1}^{N}\sum_{t<t'}||\boldsymbol{\beta}_{it}-\boldsymbol{\beta}_{it'}||^2\} \\ &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{NTP}}\left\{||\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\beta}||^2+\lambda^*||\delta_c\boldsymbol{\beta}||^2+\gamma^*||\delta_r\boldsymbol{\beta}||^2\right\} \\ &= (\boldsymbol{Z}^{\top}\boldsymbol{Z}+\lambda^*{\delta_c}^{\top}\delta_c+\gamma^*{\delta_r}^{\top}\delta_r)^{-1}\boldsymbol{X}^{\top}\boldsymbol{Y}.\end{aligned} \tag{12}$$

Here, λ and γ are tuning parameters, which are set to 1e-3 in all subsequent experiments.

To select the optimal tuning parameters in the objective function, we use the modified Bayesian information criterion (mBIC) [20], which has been widely utilized for hyperparameter selection in the field of heterogeneous structure recovery [8,10,21,22]. Notably, Cheng et al. [10] have demonstrate the selection consistency of mBIC within the context of subgroup identification tasks under the framework of M-estimation. In this paper, we denote the mBIC as follows,

$$\mathrm{mBIC}(\lambda,\gamma)=\log\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\rho(y_{it}-{\boldsymbol{z}_{\boldsymbol{it}}}^{\top}\widehat{\boldsymbol{\beta}}_{\boldsymbol{it}})\right)+c\frac{\log\log(NT)}{NT}\log(NTP)\widehat{L(\lambda,\gamma)}P, \tag{13}$$

where $\widehat{L(\lambda,\gamma)}$ is an estimate of the number of sub-blocks, and c is a constant. In this paper, when ρ is the L_2 loss, we follow the setting of Ma and Huang [17] and set $c=10$; when ρ is the L_1 loss, we follow the setting of Zhang et al. [8] and set $c=5$; when ρ is Huber, we also set $c=5$. To search for the optimal values of the parameters λ, γ, and the controlling parameter a in the penalty functions, we use grid search to traverse the ranges of $[\lambda_{\min},\lambda_{\max}]$, $[\gamma_{\min},\gamma_{\max}]$, and $[a_{\min},a_{\max}]$, respectively, with a given step size. We calculate the mBIC value for each combination of λ, γ, and a. The combination that results in the minimum mBIC is selected as the optimal tuning parameters, which are then used to obtain the final estimation result.

2.4. Asymptotic Properties

We first investigate the property of oracle estimator. If the underlying block structure $\mathscr{B}=\{\mathscr{B}_l:l=1,...,L^0\}$ is known, the oracle estimator is defined by

$$\widetilde{\boldsymbol{\alpha}}=\arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^{PL^0}}\{\sum_{l=1}^{L}\sum_{(i,t)\in\mathscr{B}_l}\rho(y_{it}-z_{it}^{T}\alpha_l)\}, \tag{14}$$

where $\widetilde{\boldsymbol{\alpha}}=(\widetilde{\alpha}_1^T,...,\widetilde{\alpha}_L^T)^T$. The oracle estimator is unavailable in practice because it assumes the knowledge of the real block structure, but it plays a significant role in theoretical analysis.

First, we define some notations. Let $z_{it,p}$ denote the pth element of $\boldsymbol{z}_{it}$. $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the minimum and maximum eigenvalues of a matrix, respectively. Let ϕ denote the derivative of ρ, $\phi^{'}$ and $\phi^{''}$ denote the first and second derivatives of ϕ, respectively, and $f^{'}$ denotes the first derivative of the distribution function f of ϵ_{it}. $\mathrm{S}_P=\{d\in\mathscr{R}^P:||d||^2=1\}$ denotes the unit sphere in $\mathscr{R}^P$.

Since ρ may have nondifferentiable points, its derivative ϕ may have discontinuities. Therefore, we need to classify and discuss ϕ according to its properties. Following the method of He and Shao [23], we classify ϕ into two categories: smooth function and jump function. When ϕ is Lipschitz continuous on its domain, we call it a smooth function; when ϕ has a finite number of jump points but is Lipschitz continuous on the intervals between two adjacent jump points, we call it a jump function. It is clear that L_2 loss and Huber loss are smooth functions, while L_1 loss is a jump function.

We introduce the following assumptions.

(A1). There exists a constant M_1 such that

$$|x_{it,p}| \leq M_1, \quad \forall 1 \leq i \leq N, 1 \leq t \leq T, 1 \leq p \leq P,$$

and there exist two positive constants C_1 and C_2 such that

$$C_1 \leq \lambda_{\min}\left(\frac{1}{NT}\boldsymbol{Z}^\top \boldsymbol{Z}\right) \leq \lambda_{\max}\left(\frac{1}{NT}\boldsymbol{Z}^\top \boldsymbol{Z}\right) \leq C_2.$$

(A2). $L^0 P = O((NT)^{c_1})$, for some $0 < c_1 < \frac{1}{3}$

(A3). When the loss function is a smooth function, $\phi^{'}$ and $\phi^{''}$ are bounded by $c_0 = \mathrm{E}\phi^{'}(\epsilon_{it}) \in (0, \infty)$; when the loss function is a jump function, ϕ, f, and $f^{'}$ are bounded by $c_0 = -\int_{-\infty}^{\infty} \phi(r) f^{'}(r) dr \in (0, \infty)$.

(A4). $\sup_{\boldsymbol{d}_1, \boldsymbol{d}_2 \in \mathrm{S}_p} \sum_{i=1}^{N} \sum_{t=1}^{T} |\boldsymbol{z}_{it}^\top \boldsymbol{d}_1|^2 |\boldsymbol{z}_{it}^\top \boldsymbol{d}_2|^2 = O(NT)$.

Remark 1. *Assumption (A1) is a regularization assumption on the design matrix, where the minimum and maximum eigenvalues of $(NT)^{-1}\boldsymbol{Z}^\top \boldsymbol{Z}$ are bounded by constants, which is a common assumption in heterogeneity panel data analysis based on concave fused penalties; see Ma and Huang [17], Ma et al. [22], Wang and Zhu [3], etc. Assumption (A2) allows the real coefficients dimension $L^0 P$ to increase with the sample size NT but at a slower rate. Assumption (A3) provides bounds on the loss function and error term distribution for different types of ϕ. This assumption is used by He and Shao [23] to prove the asymptotic normality of the M-estimator in linear regression models. For commonly used loss functions in the M-estimation field (such as L_1, L_2, and Huber) and commonly used error term distributions (such as normal and t-distributions), it is easy to prove that assumption (A3) is satisfied. Assumption (A4) further imposes restrictions on the design matrix, where if z_{it} is a random sample from a P-variate distribution and for any $d \in S_P$, $E(|d^\top \boldsymbol{z}_{it}|^4)$ that is uniformly bounded, then assumption (A4) holds, obviously.*

Under these assumptions, we can obtain the consistency properties of the oracle estimator.

Theorem 1. *Under assumptions (A1)–(A4), we have*

$$||\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0|| = O_p\left(\sqrt{\frac{L^0 P}{NT}}\right), ||\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0|| = O_p\left(\sqrt{\frac{L^0 P |\mathscr{B}_{\max}|}{NT}}\right),$$

$$\sup_{i,t} ||\widetilde{\boldsymbol{\beta}}_{it} - \boldsymbol{\beta}_{it}^0|| = O_p\left(\sqrt{\frac{L^0 P}{NT}}\right).$$

We can also provide the asymptotic normal theory for the oracle estimator (Proof of Theorem 1 in Appendix A.1).

Theorem 2. *Under assumptions (A1)–(A4), we have*

$$NT\boldsymbol{d}^T(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0)/\sigma(\boldsymbol{d}) \to N(0,1),$$

where $\sigma^2(\boldsymbol{d}) = (c_0 E\phi^2(\epsilon_{it}))^{-1}\boldsymbol{d}^\top(\boldsymbol{Z}^\top \boldsymbol{Z})\boldsymbol{d}$, and d is a $PL \times 1$ vector such that $||\boldsymbol{d}|| = 1$ (Proof of Theorem 2 in Appendix A.2).

Let b denote the minimum difference between coefficients of any two sub-blocks, i.e., $b = \min_{l \neq l'} ||\boldsymbol{\alpha}_l^0 - \boldsymbol{\alpha}_{l'}^0||$. Let $|\mathscr{B}_{\min}|$ denote the sample size of the smallest sub-block. Let $p_\lambda(s) = \lambda^{-1} P_\lambda(s)$ and $p_\gamma(s) = \gamma^{-1} P_\gamma(s)$ denote the standardized penalty functions, where

$p_\lambda^{'}(s)$ and $p_\gamma^{'}(s)$ are their derivatives. To derive the asymptotic properties of our proposed estimator, additional assumptions are required.

(A5). For any $c_1 < c_2 \leq 1$, there exists $M_2 > 0$ such that

$$(NT)^{(1-c_2)/2} b \geq M_2,$$

where c_1 is defined in Assumption (A2).

(A6). There exist two positive constants c_3 and c_4 such that for any ϵ_{it} and $c \in [-c_3, c_3]$, we have

$$\mathbb{P}(|\phi(\epsilon_{it} + c)| > x) \leq 2\exp(-c_4 x^2).$$

(A7). The normalized penalty functions $p_\lambda(s)$ and $p_\gamma(s)$ are symmetric non-decreasing functions that are convex on $[0, \infty)$. We have $p_\lambda(0) = p_\gamma(0) = 0$. There exist positive numbers $a > 0$ and $a^{'} > 0$ such that $p_\lambda(s)$ and $p_\gamma(s^{'})$ are constant when $s \geq a\lambda$ and $s^{'} \geq a^{'}\gamma$, respectively. The derivatives $p_\lambda^{'}(s)$ and $p_\gamma^{'}(s)$ are continuous except at a finite number of points, and $p_\lambda^{'}(0+) = p_\gamma^{'}(0+) = 1$.

Remark 2. *Assumption (A5) provides the minimum difference in regression coefficients between different sub-blocks, which is essential for the separability of the coefficients. Assumption (A6) further restricts the error term, which is relatively mild for M-estimators. Specifically, when ρ is the L_2 loss, $\phi(\epsilon_{it}) = 2\epsilon_{it}$, and assumption (A6) is equivalent to requiring that the error term ϵ_{it} has sub-Gaussian tails, which is a common assumption in the field of high-dimensional statistics. When ρ is the L_1 or Huber loss, assumption (A6) obviously holds because ϕ is bounded by a constant. Assumption (A7) restricts concave penalty functions, which can be easily verified to be satisfied by SCAD and MCP, where positive a and a' control the concavity of the penalty function. It should be noted that when $s \geq a\lambda$, $p_\lambda^{'}(s)$ is constant. This means that when (i,t) and (j,t') belong to different sub-blocks, the fused penalty term $p_\lambda(||\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{jt'}||)$ tends to be constant; i.e., it does not compress the coefficient differences of different sub-blocks.*

Theorem 3. *Under assumptions (A1)–(A7) and $max(\lambda, \gamma) = o((NT)^{-(1-c_2)/2})$, $\frac{\sqrt{P}\sqrt{\log(NT)}}{\min(\lambda,\gamma)|\mathscr{B}_{\min}|} = o(1)$, the oracle estimator is a local minimizer of the objective function with probability tending to one, i.e., as both N and $T \to \infty$, we have*

$$\lim_{N,T\to\infty} \mathbb{P}(\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}) \to 1.$$

Under the conditions of Theorems 2 and 3, we can obtain the following corollary (Proof of Theorem 3 in Appendix A.3).

Corollary 1.

$$NT\boldsymbol{d}^T(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0)/\sigma(\boldsymbol{d}) \to N(0,1),$$

where $\sigma^2(\boldsymbol{d}) = \left(c_0 E\phi^2(\epsilon_{it})\right)^{-1} \boldsymbol{d}^\top(\boldsymbol{Z}^\top \boldsymbol{Z})\boldsymbol{d}$, and d is a $PL \times 1$ vector such that $||\boldsymbol{d}|| = 1$.

In practice, the distribution of ϵ_{it} is unknown, and the estimate of $\mathrm{E}\phi^2(\epsilon_{it})$ can be taken as

$$\widehat{\mathrm{E}\phi^2(\epsilon_{it})} = (NT - \widehat{L}P)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{\top} \left[\phi(y_{it} - \boldsymbol{z}_{it}^\top \widehat{\boldsymbol{\beta}}_{it})\right]^2.$$

When $\phi(\cdot)$ is a smooth function, the estimate of c_0 is denoted as

$$\widehat{c}_0 = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{\top} \phi^{'}(y_{it} - \boldsymbol{z}_{it}^\top \widehat{\boldsymbol{\beta}}_{it}).$$

3. Simulation

3.1. Simulation Setting

This section presents two artificially constructed examples to investigate the finite-sample performance of the proposed estimator. The simulated data are independently generated from the following model,

$$y_{it} = \mu_{it} + \boldsymbol{x}_{it}^T \boldsymbol{\eta}_{it} + \epsilon_{it}, \quad i = 1, ..., N; \quad t = 1, ..., T,$$

We consider different combinations of the number of individuals N, the time range T, and the dimension of the regression coefficients P. For each combination of (N, T), we change the distribution of ϵ_{it} from normal to heavy-tailed and compared the estimation results under different loss functions, including L1, Huber, and L2 loss.

We use the MCP penalty function in the experiment, as SCAD yields similar results and is therefore not presented. The maximum number of iterations is set to 50, and the threshold δ is set to 10^{-5}. The algorithm terminates when the number of iterations exceeds 50 or the update range of coefficients is less than 10^{-5}. Each group of experiments is repeated $\mathcal{R} = 100$ times. In each experiment, we perform a grid search to select the optimal tuning parameters λ, γ, and a by comparing the mBIC values. The range of λ and γ is $[0.1, 1.5]$ with a step size of 0.2, and the range of a is $[2, 10]$ with a step size of 2. We utilize the following metrics to assess the systematic error of the regression coefficient estimation and the precision of the block structure recovery.

1. RMSE: root mean square error between the estimated parameter $\widehat{\boldsymbol{\beta}}$ and the real parameter $\boldsymbol{\beta}^0$.
$$\frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \sqrt{\frac{1}{NTP} ||\widehat{\boldsymbol{\beta}}^r - \boldsymbol{\beta}^0||^2}$$
2. Bias: bias between the estimated parameter $\widehat{\boldsymbol{\beta}}$ and the real parameter $\boldsymbol{\beta}^0$.
$$\frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \left[\frac{1}{NTP} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{p=1}^{P} |\widehat{\beta}_{it,p}^r - \beta_{it,p}^0| \right]$$
3. Per: the percentage that the estimated number of blocks and the real number of blocks are equal.
$$\frac{1}{\mathcal{R}} \sum_{r=1}^{\mathcal{R}} \mathrm{I}(\widehat{L}^r = L^0)$$
4. ERI: The Rand Index (RI) is used to evaluate the accuracy of clustering, which ranges between 0 and 1, with higher values indicating better performance. Motivated by the formation of RI, we can calculate individual or period-specific RIs, denoted as RI_t or RI_i, respectively. We define ERI as the averages over all periods and individuals, as follows,
$$ERI = \frac{1}{2} \left[\frac{1}{T} \sum_{t=1}^{T} RI_t + \frac{1}{N} \sum_{i=1}^{N} RI_i \right]$$

3.2. Simulation Examples

Example 1. *In this example, we generate simulated data from the following model,*

$$y_{it} = \mu_{it} + x_{it}\eta_{it} + \epsilon_{it}, \quad i = 1, \cdots, N; \quad t = 1, \cdots, T,$$

where $x_{it} = 2 \times e_{it}$ and e_{it} are independent and identically distributed standard normal random variables. The intercept term μ_{it} and slope coefficients η_{it} have dimension 1 and exhibit the same block structure. As shown in Figure 1, we construct three sets of panel data by varying the combinations of N and T. Each set of data can be partitioned into two sub-blocks $\mathscr{B}_1$ and $\mathscr{B}_2$, corresponding to coefficients $\boldsymbol{\beta}_1 = (2, 3)$ and $\boldsymbol{\beta}_2 = (2, 5)$, respectively.

To verify the robustness of the proposed method, we consider three scenarios for generating ϵ_{it}. Scenario 1 (normal distribution): $\epsilon_{it} \sim \mathcal{N}(0,1)$. Scenario 2 (heavy-tailed distribution): $\epsilon_{it} \sim 0.5 \times t(3)$, where $t(3)$ denotes the Student's t-distribution with 3 degrees of freedom. Scenario 3 (mixture distribution): $\epsilon_{it} \sim 0.3 \times \mathcal{N}(0, 0.5^2) + 0.2 \times \mathcal{N}(0, 5^2)$.

We obtain nine groups of panel data by varying the combinations of (N, T) and the distribution of ϵ_{it}. For each group of data, we compare the performance of the oracle estimator, the L_1-loss-based estimator, the L_2-loss-based estimator, and the Huber-loss-based estimator. The oracle estimator is the estimator when the block structure is known, and for convenience, we use the L_2 loss as its loss function, so the oracle estimator has an explicit solution. The estimators under an unknown block structure are influenced by the tuning parameters λ, γ, and a. We use a grid search method to obtain the optimal hyperparameter combination following the setting in Section 2.3. Additionally, there is an extra parameter δ in the Huber loss function to control the shape of the loss function, and we use the default value of 1.345.

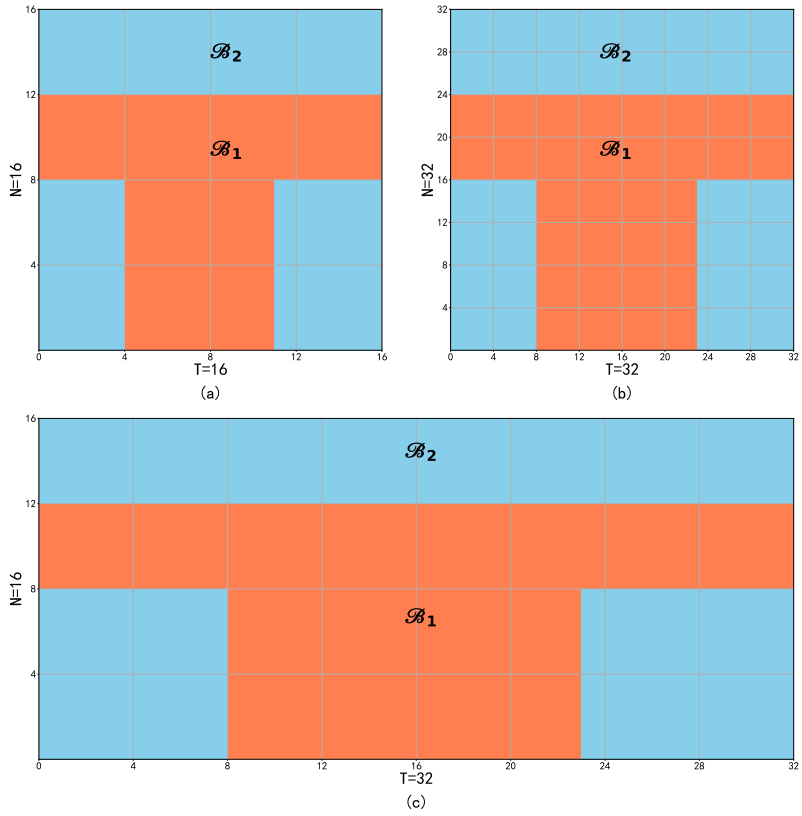

Figure 1. Block structures corresponding to different combinations of N and T in Example 1. The orange regions correspond to block 1, and the blue regions correspond to block 2. (**a**) shows the block structure for $N = T = 16$. (**b**) shows the block structure for $N = T = 32$. (**c**) shows the block structure for $N = 16$ and $T = 32$.

Example 2. *In this example, we adopt the block partitioning method under different combinations of* (N, T) *as in Example 1 with the difference being an increase in the dimension of the regression coefficients from 2 to 4. The model is specified as follows,*

$$y_{it} = \mu_{it} + x_{i1}\eta_{it,1} + x_{i2}\eta_{it,2} + x_{i3}\eta_{it,3} + \epsilon_{it}, i = 1, \cdots, N; t = 1, \cdots, T,$$

where the real coefficients of the first block are $\boldsymbol{\beta}_1 = (2, 3, -2, 1)$*, and those of the second block are* $\boldsymbol{\beta}_2 = (-2, 1, 3, -1)$*. The explanatory variables* (x_{i1}, x_{i2}, x_{i3}) *are generated from a multivariate normal distribution with mean* $(0, 0, 0)$ *and covariance matrix*

$$\begin{pmatrix} 1 & 0.5 & 0.25 \\ 0.5 & 1 & 0.5 \\ 0.25 & 0.5 & 1 \end{pmatrix}$$

The error term ϵ_{it} *is generated according to the same mixture distribution as in Example 1. We obtain three sets of panel data for different combinations of N and T.*

For each set of data, we compute the oracle estimator and the estimator for the unknown block structure following the same procedure as in Example 1. Then, we analyze various result indicators to investigate the performance of the proposed method as the dimension of the regression coefficients increases.

Example 3. *To evaluate the effectiveness of a dual-penalty in handling two-dimensional heterogeneous panel data, we conduct an ablation experiment by removing either the individual dimension penalty or the time dimension penalty from the objective function (4) and comparing the results with the full objective function that includes both penalties.*

We use a set of simulated data from Example 1, where the data have a dimension of $N = T = 32$, the error term follows a standard normal distribution, and the loss term in the objective function is the L_2 loss.

3.3. Simulation Results

Tables 1–3 display the simulation results for Example 1, corresponding to three different distributions of the error term. In each table, we consider three combinations of (N, T), reporting the results of the oracle estimator when the block structure is known, and the results based on three different loss functions when the block structure is unknown. The objective function for obtaining the oracle estimator is set to use the L_2 loss. Since obtaining the oracle estimator is difficult in practice, we use it only for comparison. We report the mean and standard deviation of 100 repeated experiments, with the standard deviation displayed in parentheses.

Table 1. Experimental Indicators of Each Combination When the Error Term is Normally Distributed in Example 1.

(N, T)	Model	Rmse	Bias	Per	ERI
(16, 16)	Oracle	0.034 (0.025)	0.056 (0.021)		
	L_1	0.040 (0.031)	0.087 (0.041)	0.99	0.998 (0.002)
	L_2	0.039 (0.026)	0.073 (0.030)	1	0.999 (0.001)
	Huber	0.039 (0.026)	0.075 (0.031)	1	0.998 (0.001)
(16, 32)	Oracle	0.020 (0.015)	0.039 (0.014)		
	L_1	0.024 (0.020)	0.068 (0.034)	1	0.999 (0.001)
	L_2	0.021 (0.016)	0.040 (0.020)	1	0.999 (0.001)
	Huber	0.022 (0.016)	0.041 (0.018)	1	0.998 (0.001)
(32, 32)	Oracle	0.015 (0.012)	0.028 (0.011)		
	L_1	0.016 (0.014)	0.035 (0.021)	1	0.998 (0.001)
	L_2	0.016 (0.011)	0.028 (0.012)	1	0.999 (0.001)
	Huber	0.016 (0.010)	0.028 (0.012)	1	0.999 (0.001)

In Table 1, the error terms follow the standard normal distribution. Simulations based on L_1 loss and Huber loss perform similarly to those based on L_2 loss. In terms of coefficient estimation, when $N = T = 16$, the L_2 loss slightly outperforms L_1 and Huber losses, but as

N and T increase, the differences among the three simulation results diminish, and they all approach the oracle estimator. In terms of structural recovery, the Per metric equals 1 in the second and third combinations of (N, T), indicating that the estimated number of blocks equals the actual number of blocks. The ERI metric approaches 1, indicating that the method accurately classifies the samples into the correct sub-blocks.

Table 2. Experimental Indicators of Each Combination When the Error Term is t-Distributed in Example 1.

(N, T)	Model	Rmse	Bias	Per	ERI
(16, 16)	Oracle	0.024 (0.020)	0.046 (0.022)		
	L_1	0.023 (0.018)	0.058 (0.032)	0.98	0.998 (0.001)
	L_2	0.026 (0.021)	0.063 (0.039)	0.89	0.983 (0.011)
	Huber	0.024 (0.019)	0.060 (0.035)	0.99	0.998 (0.001)
(16, 32)	Oracle	0.016 (0.013)	0.030 (0.011)		
	L_1	0.017 (0.013)	0.041 (0.019)	1	0.999 (0.001)
	L_2	0.023 (0.018)	0.054 (0.033)	0.91	0.987 (0.006)
	Huber	0.018 (0.014)	0.043 (0.028)	1	0.998 (0.002)
(32, 32)	Oracle	0.012 (0.009)	0.021 (0.009)		
	L_1	0.012 (0.009)	0.032 (0.012)	1	0.999 (0.001)
	L_2	0.015 (0.012)	0.037 (0.023)	0.94	0.993 (0.004)
	Huber	0.012 (0.010)	0.033 (0.022)	1	0.999 (0.001)

In Table 2, the error terms follow a heavy-tailed t distribution. It is evident that the results based on L_1 loss and Huber loss outperform those based on L_2 loss. When $N = T = 16$, we even observe that the RMSE metric based on L_1 loss is slightly lower than the oracle estimator based on L_2 loss. This is because the heavy-tailed distribution increases the probability of outliers in the simulated data, and L_2 loss is more sensitive to outliers than L_1 loss. As N and T increase, the results based on L_1 and Huber losses approach the oracle estimator, and the RMSE and Bias metrics for coefficient estimation become increasingly closer to those of the oracle estimator. The Per and ERI metrics reach or approach 1, indicating that the method accurately recovers the block structure.

In Table 3, the error terms follow a mixture of normal distributions with a stronger heavy-tailed effect. The simulation results are worse than those of the previous two cases, but those based on L_1 loss and Huber loss are still better than those based on L_2 loss. As N and T increase, the Per and ERI metrics gradually approach 1. This simulation experiment also confirms the robustness and block structure recovery ability of the proposed estimator.

Table 3. Experimental Indicators of Each Combination When the Error Term Follows a Mixture Distribution in Example 1.

(N, T)	Model	Rmse	Bias	Per	ERI
(16, 16)	Oracle	0.044 (0.037)	0.084 (0.038)		
	L_1	0.054 (0.038)	0.122 (0.056)	0.79	0.986 (0.012)
	L_2	0.081 (0.066)	0.162 (0.098)	0.42	0.943 (0.037)
	Huber	0.057 (0.065)	0.151 (0.117)	0.74	0.980 (0.011)
(16, 32)	Oracle	0.034 (0.025)	0.062 (0.025)		
	L_1	0.049 (0.034)	0.112 (0.530)	0.81	0.986 (0.010)
	L_2	0.088 (0.075)	0.207 (0.129)	0.49	0.953 (0.028)
	Huber	0.051 (0.040)	0.118 (0.073)	0.76	0.984 (0.010)
(32, 32)	Oracle	0.023 (0.017)	0.044 (0.017)		
	L_1	0.032 (0.024)	0.086 (0.066)	0.89	0.989 (0.005)
	L_2	0.061 (0.073)	0.149 (0.106)	0.62	0.971 (0.021)
	Huber	0.044 (0.046)	0.093 (0.068)	0.83	0.987 (0.007)

Below, we analyze the performance of the proposed estimator as the dimension of the coefficient P increases. In Table 4, the error term follows a heavy-tailed mixture normal distribution. Increasing P makes coefficient estimation and block structure recovery more challenging, but the performance metrics corresponding to L_1 loss and Huber loss remain superior to those of L_2 loss. As N and T increase, all performance metrics of the estimator improve significantly, and the *Per* metric corresponding to L_1 loss approaches 0.9 when $N = T = 32$.

Table 4. Experimental Indicators of Each Combination When the Error Term Follows a Mixture Distribution and P = 4 in Example 2.

(N, T)	Model	Rmse	Bias	Per	ERI
(16, 16)	Oracle	0.044 (0.037)	0.084 (0.038)		
	L1	0.054 (0.038)	0.122 (0.056)	0.79	0.986 (0.012)
	L2	0.073 (0.062)	0.157 (0.093)	0.42	0.943 (0.037)
	Huber	0.057 (0.045)	0.151 (0.087)	0.74	0.980 (0.011)
(16, 32)	Oracle	0.034 (0.025)	0.062 (0.025)		
	L1	0.049 (0.034)	0.112 (0.053)	0.81	0.986 (0.010)
	L2	0.070 (0.063)	0.154 (0.091)	0.49	0.953 (0.028)
	Huber	0.051 (0.040)	0.118 (0.073)	0.76	0.984 (0.010)
(32, 32)	Oracle	0.023 (0.017)	0.044 (0.017)		
	L1	0.032 (0.024)	0.086 (0.066)	0.89	0.989 (0.005)
	L2	0.061 (0.073)	0.149 (0.106)	0.62	0.971 (0.002)
	Huber	0.044 (0.036)	0.093 (0.068)	0.83	0.987 (0.007)

Finally, we present the results of the ablation experiment with fused penalty terms. Since the simulated data have a block structure, setting penalty terms only in one dimension cannot effectively compress the coefficient differences in the other dimension and thus cannot recover the block structure. Hence, we report only the results of the RMSE and Bias metrics in Table 5. The performance of the double-penalty method exceed those of any single-dimensional penalty method by a large margin, indicating the necessity of biclustering analysis for block panel data. The double-penalty method not only recovers unknown structures but also greatly reduces the estimation error of the coefficients.

Table 5. Ablation Experiment of Penalty Terms in Example 3.

Model	RMSE	Bias
Oracle	0.015 (0.012)	0.028 (0.011)
Double Penalties	0.016 (0.011)	0.028 (0.012)
Individual Penalty Only	0.155 (0.119)	0.799 (0.342)
Temporal Penalty Only	0.226 (0.130)	0.440 (0.154)

4. Discussion

Panel data models with heterogeneous coefficients have gained a lot of attention in various fields due to their ability to capture complex data patterns. In this paper, we extend the existing literature by proposing a more general block structure that captures heterogeneity in individual and time dimensions in a flexible manner. Our proposed model exhibits both an individual-group structure that can change at change points and temporal–structural breaks that can vary across different groups. A robust biclustering method based on *M*-estimation and double concave fused penalties is developed to estimate the coefficients, which can handle heavy-tailed data and outliers. Under certain regularity conditions, we established the asymptotic normality of the oracle estimator and the proposed estimator. Numerical simulations have validated the excellent finite-sample performance of our proposed method by evaluating the recovery of unknown structures as well as the estimated bias of regression coefficients. Furthermore, in our numerical simula-

tions, we specifically investigate the performance of our proposed model in the presence of heavy-tailed distributions, which highlights the superior performance of our proposed method in handling outliers. We believe that our method has potential applications in various fields where data exhibit complex heterogeneity.

Despite the progress made, there are still some limitations and further research topics that warrant further exploration. First among these is the lack of convergence proof for the Bayesian information criterion. To achieve this, more regularization assumptions would be required for the distribution of covariates and error terms. For relevant work in this area, researchers can consider the methodology proposed by Cheng et al. [10] for clustering individual group structures. A second challenge arises from the high-dimensional matrix calculations involved in the algorithm for solving the objective function. The computational requirements increase exponentially with N and T. To address this issue, the divide-and-conquer method can be used for parallel computation to improve the algorithm's efficiency. Finally, variable selection through L1 penalty on covariates is another promising area for further research, particularly in situations with high-dimensional covariate dimensions P. These areas of investigation will be the topic of future studies.

Author Contributions: Methodology, W.C.; Software, W.C.; Formal analysis, W.C.; Investigation, W.C.; Resources, Y.L.; Data curation, W.C.; Writing—original draft, W.C.; Writing—review & editing, Y.L.; Supervision, Y.L.; Project administration, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All data generated or analyzed during this study are included in this article.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

In the appendix, we provide the proofs of Theorems 1–3.

Appendix A.1. Proof of Theorem 1

Here, we present the proof of Theorem 1. The first conclusion of Theorem 1 follows directly from Example 1 in He and Shao [23].

Since $||\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0|| = O_p\left(\sqrt{\frac{L^0 P}{NT}}\right)$, there exists a positive constant C such that $||\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{\mathbf{0}}|| \leq C\sqrt{\frac{L^0 P}{NT}}$. Using the relationship between $\widetilde{\boldsymbol{\alpha}}$ and $\widetilde{\boldsymbol{\beta}}$, we have

$$\begin{aligned} ||\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0||^2 &= \sum_{l=1}^{L^0} \sum_{(i,t)\in\mathscr{B}_l} ||\widetilde{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_{l^0}||^2 \\ &\leq |\mathscr{B}_{\max}| \sum_{l=1}^{L^0} ||\widetilde{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}_{l^0}||^2 \\ &= |\mathscr{B}_{\max}| ||\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0||^2 \\ &\leq \frac{C^2 L^0 P |\mathscr{B}_{\max}|}{NT}. \end{aligned} \tag{A1}$$

Thus, $||\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}}|| \leq C\sqrt{\frac{L^0 P|\mathscr{B}_{\max}|}{NT}}$.
Finally, applying a simple inequality yields

$$\sup_{i,t} ||\widetilde{\boldsymbol{\beta}}_{it} - \boldsymbol{\beta}^0_{it}|| = \sup_{l} ||\widetilde{\boldsymbol{\alpha}}_l - \boldsymbol{\alpha}^0_l|| \leq ||\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0|| \leq C\sqrt{\frac{L^0 P}{NT}}. \tag{A2}$$

This completes the proof of Theorem 1.

Appendix A.2. Proof of Theorem 2

By assumption (A2), there exists $0 < c_1 < \frac{1}{3}$ such that $L^0 P = O((NT)^{c_1})$. Clearly,

$$(L^0 P)^3 (\log(L^0 P))^2 = o(NT). \tag{A3}$$

Therefore, by Example 1 in He and Shao [23], the result of Theorem 2 follows directly.

Appendix A.3. Proof of Theorem 3

We partition the block structure of regression coefficients as follows: we first divide the samples into C groups according to the group structure in the dimension of individuals; then, we partition them into K time periods according to the structural breaks in the time dimension (if there exist structural breaks for any individuals at given time period, we perform segmentation on all individuals instead of splitting individuals under given groups). In this way, we obtain KC sub-blocks, which obviously exceeds the number of true sub-blocks. Let $\mathscr{B}_{lc}$ be the set of sample indexes that belong to both the cth individual group and the l-th true sub-block, and let $\mathscr{B}_{lk}$ be the set of sample indexes that belong to both the kth time group and the lth true sub-block. Denote

$$L(\beta) = \sum_{i=1}^{N} \sum_{t=1}^{T} \rho(y_{it} - \boldsymbol{z}_{it}^T \boldsymbol{\beta}_{it}), \tag{A4}$$

$$P(\beta) = \sum_{t=1}^{T} \sum_{i<j} P_\lambda(||\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{jt}||) + \sum_{i=1}^{N} \sum_{t<t'} P_\gamma(||\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{it'}||), \tag{A5}$$

$$L^{\mathscr{B}}(\alpha) = \sum_{l=1}^{L} \sum_{(i,t)\in\mathscr{B}_l} \rho(y_{it} - \boldsymbol{z}_{it}^\top \boldsymbol{\alpha}_l), \tag{A6}$$

$$\begin{aligned} P^{\mathscr{B}}(\boldsymbol{\alpha}) = \lambda \sum_{l<l'} \sum_{c=1}^{C} (|\mathscr{B}_{lc}||\mathscr{B}_{l'c}|)\rho_\lambda(||\boldsymbol{\alpha}_l - \boldsymbol{\alpha}_{l'}||) \\ + \gamma \sum_{l<l'} \sum_{k=1}^{K} (|\mathscr{B}_{lk}||\mathscr{B}_{l'k}|))\rho_\gamma(||\boldsymbol{\alpha}_l - \boldsymbol{\alpha}_{l'}||), \end{aligned} \tag{A7}$$

where the variable $|\mathscr{B}_{lc}|$ denotes the number of samples belonging to both the cth individual group and the lth true sub-block, while $|\mathscr{B}_{lk}|$ represents the number of samples belonging to both the kth time group and the lth true sub-block. To simplify the notation without creating confusion, we use $\boldsymbol{Q}(\boldsymbol{\beta})$ to refer to the objective function in Equation (4), which is defined as

$$Q(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) + P(\boldsymbol{\beta}), \tag{A8}$$

and let

$$Q^{\mathscr{B}}(\boldsymbol{\alpha}) = L^{\mathscr{B}}(\boldsymbol{\alpha}) + P^{\mathscr{B}}(\boldsymbol{\alpha}). \tag{A9}$$

Let $\mathcal{M}_{\mathscr{B}}$ denote the set of $\mathbb{R}^{NTP}$ coefficients with a block structure. We define a mapping $T : \mathcal{M}_{\mathscr{B}} \to \mathbb{R}^{LP}$, which maps a block-structured NTP-dimensional vector β to an LP-dimensional vector α. Here, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top, \cdots, \boldsymbol{\alpha}_L^\top)^\top$ is the concatenation of L P-dimensional vectors, with the lth vector $\boldsymbol{\alpha}_l$ representing the coefficients for the lth sub-block. Additionally, we define a mapping $T^{\mathscr{B}} : \mathbb{R}^{NTP} \to \mathbb{R}^{LP}$, which maps any NTP-dimensional vector $\boldsymbol{\beta}$ to an LP-dimensional vector $\boldsymbol{\alpha}$ based on the block structure of $\mathscr{B}$ as follows. For any $\boldsymbol{\beta} = \{\boldsymbol{\beta}_{11}^\top, \cdots, \boldsymbol{\beta}_{1T}^\top, \cdots \boldsymbol{\beta}_{N1}^\top, \cdots, \boldsymbol{\beta}_{NT}^\top\}$ and $\mathscr{B} = \{\mathscr{B}_1, \cdots \mathscr{B}_L\}$, the mapping process is given by

$$T^{\mathscr{B}}(\boldsymbol{\beta}) = \left\{ |\mathscr{B}_1|^{-1} \sum_{(i,t)\in\mathscr{B}_1} \boldsymbol{\beta}_{it}^\top, \cdots, |\mathscr{B}_L|^{-1} \sum_{(i,t)\in\mathscr{B}_L} \boldsymbol{\beta}_{it}^\top \right\}. \tag{A10}$$

From the properties of the mappings T and $T^{\mathscr{B}}$, it follows that for any $\boldsymbol{\beta} \in \mathcal{M}_{\mathscr{B}}$, we have $T(\boldsymbol{\beta}) = T^{\mathscr{B}}(\boldsymbol{\beta})$, $P(\boldsymbol{\beta}) = P^{\mathscr{B}}(T(\boldsymbol{\beta}))$. Let $\boldsymbol{\alpha} = T(\boldsymbol{\beta})$, then $P(T^{-1}(\boldsymbol{\alpha})) = P^{\mathscr{B}}(\boldsymbol{\alpha})$. Therefore, we conclude that,

$$Q(\boldsymbol{\beta}) = Q^{\mathscr{B}}(T(\boldsymbol{\beta})), \quad Q^{\mathscr{B}}(\boldsymbol{\alpha}) = Q(T^{-1}(\boldsymbol{\alpha})). \tag{A11}$$

We denote the true regression coefficients as $\boldsymbol{\beta}^0$ and $\boldsymbol{\alpha}^0$, and the oracle estimators as $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\alpha}}$. Then, we define the set

$$\Theta_1 = \left\{ \boldsymbol{\beta} \in \mathbb{R}^{NTP} : \sup_{it} ||\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{it}^0|| \leq \sqrt{\frac{L^0 P}{NT}} \right\}. \tag{A12}$$

We now prove Theorem 2 in two steps.

(i) Let $\boldsymbol{\beta} \in \Theta_1$, and denote $\boldsymbol{\beta}^* = T^{-1}(T^{\mathscr{B}}(\boldsymbol{\beta}))$. For all $\boldsymbol{\beta}^* \neq \widetilde{\boldsymbol{\beta}}$, we have $Q(\boldsymbol{\beta}^*) > Q(\widetilde{\boldsymbol{\beta}})$ with a probability tending to 1.

(ii) Define the set $\Theta_2 = \{\boldsymbol{\beta}_{it} : \sup_{it} ||\boldsymbol{\beta}_{it} - \widehat{\boldsymbol{\beta}}_{it}|| \leq s\}$, where s is a positive sequence. When s is small enough, we have $Q(\boldsymbol{\beta}) \geq Q(\boldsymbol{\beta}^*)$ with a probability tending to 1 for all $\boldsymbol{\beta} \in \Theta_1 \cap \Theta_2$.

Clearly, if (i) and (ii) are proved, then for any $\boldsymbol{\beta} \in \Theta_1 \cap \Theta_2$, we have $Q(\boldsymbol{\beta}) > Q(\widetilde{\boldsymbol{\beta}})$, which means that the oracle estimator $\widetilde{\boldsymbol{\beta}}$ is a local minimum of $Q(\boldsymbol{\beta})$, and this conclusion holds with probability tending to 1.

The proof for (i) is as follows. Let $\boldsymbol{\alpha} = T^{\mathscr{B}}(\boldsymbol{\beta})$; then,

$$\begin{aligned}
\sup_l \left\| \boldsymbol{\alpha}_l - \boldsymbol{\alpha}_l^0 \right\|^2 &= \sup_l \left\| |\mathscr{B}_l|^{-1} \sum_{(i,t)\in\mathscr{B}_l} \boldsymbol{\beta}_{it} - \boldsymbol{\alpha}_l^0 \right\|^2 \\
&= \sup_l \left\| |\mathscr{B}_l|^{-1} \sum_{(i,t)\in\mathscr{B}_l} \left(\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{it}^0 \right) \right\|^2 \\
&\leq \sup_l |\mathscr{B}_l|^{-1} \sum_{(i,t)\in\mathscr{B}_l} \left\| \boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{it}^0 \right\|^2 \\
&\leq \sup_{i,t} \left\| \boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{it}^0 \right\|^2 \\
&\leq \frac{L^0 P}{NT},
\end{aligned} \tag{A13}$$

therefore, all l and l', the following inequation holds,

$$||\boldsymbol{\alpha}_l - \boldsymbol{\alpha}_{l'}|| \geq ||\boldsymbol{\alpha}_l^0 - \boldsymbol{\alpha}_{l'}^0|| - 2 \sup_l ||\boldsymbol{\alpha}_l - \boldsymbol{\alpha}_l^0|| \geq b - 2\sqrt{\frac{L^0 P}{NT}}. \tag{A14}$$

Using the inequality from assumption (A5), $(NT)^{(1-c_2)/2} b \geq M_2$, and the condition $\max(\lambda, \gamma) = o((NT)^{-(1-c_2)/2})$, we can obtain

$$b - 2\sqrt{\frac{L^0 P}{NT}} > \max(a\lambda, a'\gamma), \tag{A15}$$

Therefore, assumption (A7) implies $P^{\mathscr{B}}(\boldsymbol{\alpha}) = C$, where C is a constant, and $Q^{\mathscr{B}}(\boldsymbol{\alpha}) = L^{\mathscr{B}}(\boldsymbol{\alpha}) + C$. Since $\widetilde{\boldsymbol{\alpha}}$ is the global minimum of $L^{\mathscr{B}}(\boldsymbol{\alpha})$, it follows that $Q^{\mathscr{B}}(\boldsymbol{\alpha}) > Q^{\mathscr{B}}(\widetilde{\boldsymbol{\alpha}})$ for any $\boldsymbol{\alpha} \neq \widetilde{\boldsymbol{\alpha}}$. Finally, using (A11), we have $Q^{\mathscr{B}}(\boldsymbol{\alpha}) = Q(T^{-1}(\boldsymbol{\alpha})) = Q(\boldsymbol{\beta})$ and $Q^{\mathscr{B}}(\widetilde{\boldsymbol{\alpha}}) = Q(\widetilde{\boldsymbol{\beta}})$, so $Q(\boldsymbol{\beta}^*) > Q(\widetilde{\boldsymbol{\beta}})$ for any $\boldsymbol{\beta} \neq \widetilde{\boldsymbol{\beta}}$. This completes the proof of conclusion (i).

Continuing with the proof for result (ii), we define two functions

$$P_1(\boldsymbol{\beta}) = \sum_{t=1}^{\top} \sum_{i<j} P_\lambda(||\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{jt}||), \quad P_2(\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{t<t'} P_\gamma(||\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{it'}||). \tag{A16}$$

By Taylor expanding around $\boldsymbol{\beta}_{it}$, we can decompose the difference of the objective function into three parts,

$$Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}^*) = \Gamma_1 + \Gamma_2 + \Gamma_3 \tag{A17}$$

where each part takes the following form:

$$\Gamma_1 = L(\boldsymbol{\beta}) - L(\boldsymbol{\beta}^*), \tag{A18}$$

$$\Gamma_2 = \sum_{t=1}^{\top} \sum_{i=1}^{N} \frac{\partial P_1(\boldsymbol{\beta}^m)}{\partial \boldsymbol{\beta}_{it}} (\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{it}^*), \tag{A19}$$

$$\Gamma_3 = \sum_{i=1}^{N} \sum_{t=1}^{\top} \frac{\partial P_2(\boldsymbol{\beta}^m)}{\partial \boldsymbol{\beta}_{it}} (\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{it}^*). \tag{A20}$$

Here, $\boldsymbol{\beta}^m = \theta\boldsymbol{\beta} + (1-\theta)\boldsymbol{\beta}^*$, where θ is a scalar between 0 and 1.

First, let us handle the first part,

$$\begin{aligned} \Gamma_1 &= \sum_{i=1}^{N} \sum_{t=1}^{\top} (\rho(y_{it} - \boldsymbol{z}_{it}^\top \boldsymbol{\beta}_{it}) - \rho(y_{it} - \boldsymbol{z}_{it}^\top \boldsymbol{\beta}_{it}^*)) \\ &= -\sum_{i=1}^{N} \sum_{t=1}^{\top} \phi(y_{it} - \boldsymbol{z}_{it}^\top \boldsymbol{\beta}_{it}^m) \boldsymbol{z}_{it}^\top (\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{it}^*) \\ &= -\sum_{i=1}^{N} \sum_{t=1}^{\top} \phi(\epsilon_{it} + \boldsymbol{z}_{it}^\top (\boldsymbol{\beta}_{it}^0 - \boldsymbol{\beta}_{it}^m)) \boldsymbol{z}_{it}^\top (\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{it}^*). \end{aligned} \tag{A21}$$

By Assumption (A6) and $\boldsymbol{z}_{it}^\top (\boldsymbol{\beta}_{it}^0 - \boldsymbol{\beta}_{it}^m) = o_p(1)$, we can deduce that,

$$P(\phi(\epsilon_{it} + \boldsymbol{z}_{it}^\top (\boldsymbol{\beta}_{it}^0 - \boldsymbol{\beta}_{it}^m)) > \sqrt{\log(NT)}) \leq 2(NT)^{-c_4}, \tag{A22}$$

which implies that as N and T approach infinity, $\phi(\epsilon_{it} + \boldsymbol{z}_{it}^\top (\boldsymbol{\beta}_{it}^0 - \boldsymbol{\beta}_{it}^m)) \leq \sqrt{\log(NT)}$ holds with probability tending to one. Thus, we can bound Γ_1 as follows,

$$\begin{aligned} \Gamma_1 &\geq -\sum_{i=1}^{N} \sum_{t=1}^{\top} \sqrt{c_4^{-1}} \sqrt{\log(NT)} \boldsymbol{z}_{it}^\top (\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{it}^*) \\ &= -\sum_{l=1}^{L} \sum_{(i,t) \in \mathscr{B}_l} \sqrt{c_4^{-1}} \sqrt{\log(NT)} \boldsymbol{z}_{it}^\top (\boldsymbol{\beta}_{it} - |\mathscr{B}_l|^{-1} \sum_{(j,t') \in \mathscr{B}_l} \boldsymbol{\beta}_{jt'}) \\ &= -\sum_{l=1}^{L} \sum_{(i,t),(j,t') \in \mathscr{B}_l} \sqrt{c_4^{-1}} \sqrt{\log(NT)} |\mathscr{B}_l|^{-1} \boldsymbol{z}_{it}^\top (\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{jt'}) \\ &\geq -\sum_{l=1}^{L} \sum_{(i,t),(j,t') \in \mathscr{B}_l} \sqrt{c_4^{-1}} \sqrt{\log(NT)} |\mathscr{B}_{min}|^{-1} \sqrt{P} M_1 ||\boldsymbol{\beta}_{it} - \boldsymbol{\beta}_{jt'}||. \end{aligned} \tag{A23}$$

For Γ_2 and Γ_3, using the results from Wang et al. [18] and Fang et al. [24], we have

$$\begin{aligned}\Gamma_2 &\geq \lambda \sum_{t=1}^{\top}\sum_{l=1}^{L}\sum_{(i,t),(j,t)\in\mathscr{B}_l,i<j}\rho^{'}(4s)||\boldsymbol{\beta}_{it}-\boldsymbol{\beta}_{jt}||,\\ \Gamma_3 &\geq \gamma \sum_{i=1}^{N}\sum_{l=1}^{L}\sum_{(i,t),(i,t')\in\mathscr{B}_l,t<t'}\rho^{'}(4s)||\boldsymbol{\beta}_{it}-\boldsymbol{\beta}_{it'}||.\end{aligned} \tag{A24}$$

Finally, by combining the above results, we can obtain

$$\begin{aligned}Q(\boldsymbol{\beta})-Q(\boldsymbol{\beta}^*) &\geq -4\sum_{l=1}^{L}\sum_{(i,t),(j,t')\in\mathscr{B}_l,i<j,t<t'}\sqrt{c_4^{-1}}\sqrt{\log(NT)}|\mathscr{B}_{min}|^{-1}\sqrt{P}M_1||\boldsymbol{\beta}_{it}-\boldsymbol{\beta}_{jt'}||\\ &\quad+\sum_{l=1}^{L}[\sum_{t=1}^{\top}\sum_{(i,t),(j,t)\in\mathscr{B}_l,i<j}\lambda\rho^{'}(4s)+\sum_{i=1}^{N}\sum_{(i,t),(i,t')\in\mathscr{B}_l,t<t'}\gamma\rho^{'}(4s)]||\boldsymbol{\beta}_{it}-\boldsymbol{\beta}_{jt'}||\\ &\geq -4\sum_{l=1}^{L}\sum_{(i,t),(j,t')\in\mathscr{B}_l,i<j,t<t'}\sqrt{c_4^{-1}}\sqrt{\log(NT)}|\mathscr{B}_{min}|^{-1}\sqrt{P}M_1||\boldsymbol{\beta}_{it}-\boldsymbol{\beta}_{jt'}||\\ &\quad+\sum_{l=1}^{L}\sum_{(i,t),(j,t')\in\mathscr{B}_l,i<j,t<t'}\min(\lambda,\gamma)\rho^{'}(4s)||\boldsymbol{\beta}_{it}-\boldsymbol{\beta}_{jt'}||.\end{aligned} \tag{A25}$$

As $s\to 0$, $\rho^{'}(4s)\to 1$. Moreover, since $\frac{\sqrt{P}\sqrt{\log(NT)}}{\min(\lambda,\gamma)|\mathscr{B}_{\min}|}=o(1)$, we have $Q(\boldsymbol{\beta})-Q(\boldsymbol{\beta}^*)\geq 0$. This completes the proof.

References

1. Su, L.; Shi, Z.; Phillips, P.C. Identifying latent structures in panel data. *Econometrica* **2016**, *84*, 2215–2264. [CrossRef]
2. Tibshirani, R.; Saunders, M.; Rosset, S.; Zhu, J.; Knight, K. Sparsity and smoothness via the fused LASSO. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 91–108. [CrossRef]
3. Wang, W.; Zhu, Z. Group structure detection for a high-dimensional panel data model. *Can. J. Stat.* **2022**, *50*, 852–866. [CrossRef]
4. Qian, J.; Su, L. Shrinkage estimation of regression models with multiple structural changes. *Econom. Theory* **2016**, *32*, 1376–1433. [CrossRef]
5. Qian, J.; Su, L. Shrinkage estimation of common breaks in panel data models via adaptive group fused Lasso. *J. Econom.* **2016**, *191*, 86–109. [CrossRef]
6. Okui, R.; Wang, W. Heterogeneous structural breaks in panel data models. *J. Econom.* **2021**, *220*, 447–473. [CrossRef]
7. Lumsdaine, R.L.; Okui, R.; Wang, W. Estimation of panel group structure models with structural breaks in group memberships and coefficients. *J. Econom.* **2023**, *233*, 45–65. [CrossRef]
8. Zhang, Y.; Wang, H.J.; Zhu, Z. Robust subgroup identification. *Stat. Sin.* **2019**, *29*, 1873–1889. [CrossRef]
9. Zou, H.; Li, R. One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Stat.* **2008**, *36*, 1509–1533.
10. Cheng, C.; Feng, X.; Li, X.; Wu, M. Robust analysis of cancer heterogeneity for high-dimensional data. *Stat. Med.* **2022**, *41*, 5448–5462. [CrossRef]
11. Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends® Mach. Learn.* **2011**, *3*, 1–122.
12. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [CrossRef]
13. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
14. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2005**, *67*, 301–320. [CrossRef]
15. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
16. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [CrossRef]
17. Ma, S.; Huang, J. A concave pairwise fusion approach to subgroup analysis. *J. Am. Stat. Assoc.* **2017**, *112*, 410–423. [CrossRef]
18. Wang, W.; Yan, X.; Ren, Y.; Xiao, Z. Bi-Integrative Analysis of Two-Dimensional Heterogeneous Panel Data Model. *arXiv* **2021**, arXiv:econ.EM/2110.10480. Available online: http://xxx.lanl.gov/abs/2110.10480 (accessed on 10 October 2021).
19. Hunter, D.R.; Li, R. Variable selection using MM algorithms. *Ann. Stat.* **2005**, *33*, 1617. [CrossRef]

20. Wang, H.; Li, R.; Tsai, C.L. Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method. *Biometrika* **2007**, *94*, 553–568. [CrossRef]
21. Wang, H.; Li, B.; Leng, C. Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2009**, *71*, 671–683. [CrossRef]
22. Ma, S.; Huang, J.; Zhang, Z.; Liu, M. Exploration of heterogeneous treatment effects via concave fusion. *Int. J. Biostat.* **2019**, *16*, 20180026. [CrossRef] [PubMed]
23. He, X.; Shao, Q.M. On parameters of increasing dimensions. *J. Multivar. Anal.* **2000**, *73*, 120–135. [CrossRef]
24. Fang, K.; Chen, Y.; Ma, S.; Zhang, Q. Biclustering analysis of functionals via penalized fusion. *J. Multivar. Anal.* **2022**, *189*, 104874. [CrossRef] [PubMed]