Generalized Method of Moments with Grouped Heterogeneous Validity of Moment Conditions in Panel Data Models[*]

Zhan Gao[†]
Department of Economics
University of Southern California

November 12, 2024

Abstract

This paper provides a unified framework for the selection of valid moment conditions and detection of latent group structures based on the moment condition validity in general nonlinear generalized method of moments (GMM) panel data models. It accommodates a diverging number of moment conditions and group-specific heterogeneous validity of moment conditions across agents. The proposed method integrates the pairwise adaptive fused Lasso and the adaptive Lasso regularization into the GMM estimation. The estimator is shown to be consistent and simultaneously achieves classification and moment selection consistency. The asymptotic distribution of a post-regularization estimator is derived, and its oracle properties are established. The finite-sample performance of the proposed method is evaluated through a Monte Carlo simulation experiment. The method is applied to empirically investigate the impact of agricultural productivity shocks on rural-to-urban migration in China.

Key words: GMM, grouped panel, fused Lasso, adaptive Lasso, moment selection
JEL code: C13, C33, C36, C38, C52

[*] Job market paper. Please check here for the latest version.

[†] Correspondence: zhangao@usc.edu, Department of Economics, University of Southern California, Kaprielian Hall, 300, 3620 S. Vermont Ave, Los Angeles, CA 90089, USA. I am deeply indebted to my Ph.D. advisor Roger Moon for his continual guidance, encouragement and support. I am extremely grateful to Timothy Armstrong, Cheng Hsiao, Hashem Pesaran and Geert Ridder for their invaluable support and advice. I would also thank Yuehao Bai, Stephane Bonhommme, Hongqi Chen, Xiaohong Chen, Kei Hirano, Lingwei Kong, Jing Kong, Sokbae Simon Lee, Erica Xuenan Li, Yanbo Liu, Ziwei Mei, Marcelo Moreira, Zhentao Shi, Liangjun Su, Hayun Song, Yulong Wang, Yimeng Xie, Wan Zhang, Kang Zhou, as well as conference and seminar participants at IAAE, AMES, NASM-ES, EEA-ESEM, USC, SDU, PKU, for helpful comments and discussions. All remaining errors are my own.

1 Introduction

Panel data models are extensively employed in empirical economics research. These models utilize data that encompass various units—such as workers, firms, or countries—observed over time, and offer the distinct advantage of enabling the analysis of critical economic relationships while controlling for unobserved heterogeneity. However, this unobserved heterogeneity also presents statistical challenges for accurate estimation and inference, leading researchers to navigate trade-offs in the degree of unobserved heterogeneity accommodated by the model.

Recent years have seen a surge in interest towards developing panel data models that incorporate latent group structures, a concept pioneered by Hahn and Moon (2010); Bonhomme and Manresa (2015); Su et al. (2016). These models assume that units are categorized into a finite number of groups with homogeneous parameters within each group but heterogeneity across groups. This approach strikes a balance across model flexibility, statistical efficiency, and empirical interpretability.

Panel data models are often characterized by sets of moment conditions, such as endogenous regression models with instrumental variables and nonlinear structural models. In many cases, the available moment conditions, which are possibly misspecified, outnumber the parameters of interest. It is, therefore, crucial to selectively use valid moment conditions in the estimation process to prevent biased and misleading outcomes. For instance, approaches for shrinkage GMM estimation (Liao, 2013; Cheng and Liao, 2015) have been proposed to address these challenges by penalizing moment-specific slackness parameters that signify the validity of moment conditions.

Given the prevalence of unobserved heterogeneity in panel data, assuming universal validity of moment conditions across all units is overly restrictive. It is more pragmatic to acknowledge the heterogeneity in the validity of moment conditions across units. For example, in instrumental variable (IV) models, the exogeneity of instruments typically hinges on foundational model assumptions or empirical evidence. Given the inherent heterogeneity across units, it is unrealistic to assume a universal applicability of any single IV. Recognizing this, it becomes essential to accommodate the notion that varied units may necessitate distinct sets of IVs for accurately identifying causal relationships of interest. Similarly, in the context of moment conditions derived from economic theory, assuming a one-size-fits-all model that uniformly represents all facets of the data-generating process for every unit is overly optimistic. A more nuanced approach involves tailoring the model to accurately reflect specific subsets of features for different units, as defined by corresponding sets of moment conditions, thereby ensuring a more accurate and nuanced representation of the underlying economic realities.

This paper introduces a unified framework for estimating the structural parameters of interest, selecting valid moment conditions, and detecting latent group structures in nonlinear GMM panel data models. The framework accommodates an expanding number of moment conditions and group-specific heterogeneity in the validity of these conditions. By integrating the pairwise adaptive fused Lasso (Mehrabani, 2023) and the adaptive Lasso (Cheng and Liao, 2015) regularization for moment selection into a GMM framework, the method offers a comprehensive solution for simultaneously achieving consistent and efficient parameter estimation, accurate classification of latent groups, and the selection of valid moment conditions for each group. To the best of my knowledge, this is the inaugural work that integrates moment selection and group classification in nonlinear GMM panel data models. This integration bridges the gap between moment selection methodologies and panel data models featuring latent group structures and leverages the strengths of both approaches.

Under a large (N, T) asymptotic regime, the proposed method is shown to be consistent and achieves classification and moment selection consistency simultaneously. The asymptotic distribution of a post-regularization estimator is derived, allowing for statistical inference on the structural parameters of interest. The estimation procedure involves a large-scale optimization problem. A conic programming formulation is provided, which can be efficiently solved by off-the-shelf optimization solvers. The finite-sample performance of the proposed method is evaluated through Monte Carlo simulation experiments.

The proposed method is applied to empirically investigate the impact of agricultural productivity on rural-to-urban migration in China using unique datasets from the National Fixed-point Survey, a large-scale panel survey of rural households in China, and detailed weather data from the China Meteorological Data Service Center. Temperature and precipitation variables, as well as agricultural technology shocks, are used as instrumental variables to address potential endogeneity issues. We show that using all available instruments delivers opposite results compared to the estimates from the penalized GMM estimator. The results without accounting for the validity issue can be misleading since even the widely adopted precipitation variables are not universally exogenous as detected by our method. Meanwhile, the proposed method demonstrates efficiency gains over the IV estimation with only the temperature variables due to moment selection.

Literature Review. This work contributes to various strands of the literature. Firstly, it is related to the vast literature on the generalized method of moments (GMM) since Hansen (1982), with a particular emphasis on overidentifying moment conditions with possible misspecification. In early works, Andrews (1999); Andrews and Lu (2001) develop moment selection criteria based on J overidentification test statistic. Han and Phillips (2006) derive the asymptotic properties of the GMM estimator with many moment conditions. Liao (2013); Cheng and Liao (2015) propose shrinkage GMM estimation with adaptive Lasso regularization of the slackness parameters for moment selection. Another strand of the literature focuses on the empirical likelihood approach to moment models. For example, Moon and Schorfheide (2009) study the properties of the empirical likelihood estimator with overidentifying moment inequalities. Shi (2016); Chang et al. (2018); Ando and Sueishi (2019); Chang et al. (2022) investigate the penalized empirical likelihood estimation in high-dimensional environments. As an important class of moment condition models, linear IV models with high-dimensional instruments have attracted wide attention. Hahn et al. (2011) develop a Hausman test to test the validity of a set of strong but potentially invalid IVs in the presence of weak IVs. Selection of IVs via regularization has been extensively studied, see Okui (2011); Belloni et al. (2012); Fan and Liao (2014); Luo (2014); Windmeijer et al. (2019); Gautier and Rose (2021); Liang et al. (2022); Lin et al. (2022) and the references therein.

In this work, the wisdom from these works, particularly the shrinkage GMM estimation approach in Cheng and Liao (2015), is brought to the nonlinear panel data settings with latent group structures to account for heterogeneity in the validity of moment conditions.

Panel data models with latent group structure have become increasingly popular in the literature during the past decade. Several prominent approaches have been developed. The k-means algorithm has been introduced to the panel data literature for estimation of latent group structures and group-specific parameters by Lin and Ng (2012); Bonhomme and Manresa (2015) and this line of research has been flourishing, see Bonhomme et al. (2019, 2022); Liu et al. (2020, 2023); Miao et al. (2020); Okui and Wang (2021); Wang et al. (2023); Lumsdaine et al. (2023); Cytrynbaum (2020); Cheng et al. (2023). Some works borrow the sequential binary segmentation algorithm initally applied in the structural break detection literature, for examples Ke et al. (2016); Wang and Su (2021); Su et al. (2023), and recent works focus on spectral clustering algorithms, see Ma et al. (2022); Chetverikov and Manresa (2022); Yu et al. (2022). The Bayesian approach is proposed in Zhang (2023); Huang (2023). Last but not least, a major strand of the literature focuses on identifying group structures in panel data models via dedicated penalization schemes. Su et al. (2016) propose the classifier-Lasso (C-Lasso) penalty to identify group structure in a nonlinear profile likelihood framework and the linear IV models. Following their work, the method has been extended to a variety of settings, for example, Su and Lu (2017); Su and Ju (2018); Su et al. (2019); Huang et al. (2020, 2021), and the computational issues of C-Lasso is discussed in Gao and Shi (2021); Huang et al. (2023).

Many of the existing works on grouped panel models focus on linear or nonlinear regression models or linear IV models, with Cheng et al. (2023) as an exception in which the authors study a nonlinear GMM model with an emphasis on multi-dimensional grouped heterogeneity. This work contributes to the nonlinear GMM panel data models with latent group structure with a focus on heterogeneity in the validity of moment conditions. In addition, the model in this paper deals with parameters with increasing dimensions by allowing for a diverging number of moment conditions and facilitates variable selection in the estimation.

The current work adopts the pairwise adaptive fused Lasso (PAFL) penalization regularization to identify the grouped structure in the validity of moment conditions. The PAFL penalty originates from of idea of the adaptive Lasso (Zou, 2006), group Lasso (Yuan and Lin, 2006), fused Lasso (Tibshirani et al., 2005) and the group fused Lasso (Qian and Su, 2016; Okui and Wang, 2021; Lumsdaine et al., 2023) and is proposed for clustering problems as in Hocking et al. (2011); Radchenko and Mukherjee (2017). The method is applied for the detection of group structures in panel data models in Gu and Volgushev (2019) and Mehrabani (2023). This technique offers computational benefits over alternatives like the Classifier-Lasso, k-means clustering, and the sequential binary segmentation algorithm, thanks to its convex nature, and simplifies the training process with a single tuning parameter and provides a spectrum of tuning parameters to accurately determine the number of groups. Discussion and comparison of the PAFL penalty with other methods will be delineated in Remark 2.

Gu and Volgushev (2019) adopts a scaler version of the L_1-norm-baed PAFL penalization to study the quantile regression models with grouped fixed effects, while our model deals with increasing dimensional parameters. The PAFL penalty in our work has the same form as in Mehrabani (2023). However, the proof in Mehrabani (2023) was found to only support individual classification consistency with a few technical gaps in the development of the asymptotic theory. The present paper provides a novel proof to first rigorously establish uniform classification consistency with the pairwise adaptive fused Lasso, which can also serve as a remedy for Mehrabani (2023) in the linear panel data models with latent group structures.

The empirical application in this paper contributes to a large body of literature addressing a classic development issue: whether agricultural productivity shocks encourage rural-to-urban migration in the developing world (e.g., Matsuyama, 1992; Foster and Rosenzweig, 2007; Bustos et al., 2016; McArthur and McCord, 2017; Liu et al., 2023).

Notations. For any positive integer N, $\boldsymbol{I}_N \in \mathbb{R}^{N\times N}$, $\mathbf{0}_N \in \mathbb{R}^{N\times 1}$ and $\boldsymbol{\iota}_N \in \mathbb{R}^{N\times 1}$ denotes the $N \times N$ identity matrix, $N \times 1$ zero vector and $N \times 1$ vector of ones, respectively. Denote $[N] = \{1, 2, \cdots, N\}$. For generic vectors $\boldsymbol{a} \in \mathbb{R}^N$ and matrices $\boldsymbol{A} \in \mathbb{R}^{N\times K}$, $\|\cdot\|$ denotes Frobenius norm; $\boldsymbol{A}'$ is the transpose of $\boldsymbol{A}$; $\sigma_{\max}(\boldsymbol{A})$ and $\sigma_{\min}(\boldsymbol{A})$ denote the largest and smallest singular values of $\boldsymbol{A}$, respectively. Let a_i and a_{ik} denote the i-th element of $\boldsymbol{a}$ and the (i,k)-th element of $\boldsymbol{A}$, respectively. For a vector-valued function $f(\boldsymbol{a}): \mathbb{R}^N \to \mathbb{R}^K$, $\frac{\partial f(\boldsymbol{a})}{\partial \boldsymbol{a}'} \in \mathbb{R}^{K\times N}$ denotes the Jacobian matrix with (i,k)-th element as $\frac{\partial f_i(x)}{\partial a_k}$ where $f_i(\cdot)$ is the i-th component of $f(\cdot)$. Let $\mathcal{S} = \left\{i_{(1)}, i_{(2)}, \cdots, i_{(S)}\right\} \subset [N]$, denote $\boldsymbol{a}_{\mathcal{S}} = \left(a_{i_{(1)}}, a_{i_{(2)}}, \cdots, a_{i_{(S)}}\right)'$ and $f_{\mathcal{S}}(\cdot) = \left(f_{i_{(1)}}, f_{i_{(2)}}, \cdots, f_{i_{(S)}}\right)'$ as the subvectors of $\boldsymbol{a}$ and $f(\cdot)$, respectively. For a generic set $\mathcal{A}$, $|\mathcal{A}|$ is the cardinality of $\mathcal{A}$. We use C (c) to denote generic large (small) positive constants. Define $a \wedge b := \min(a,b)$ and $a \vee b := \max(a,b)$. $\Delta x_{it} := x_{it} - x_{i,t-1}$ denotes the first difference operator. "$\xrightarrow{p}$" and "$\xrightarrow{d}$" denote convergence in probability and convergence in distribution, respectively; "w.p.a.1" abbreviates "with probability approach 1". Denote $a_N = o_P(b_N)$ if for any $\varepsilon > 0$, $\Pr(|a_N/b_N| > \varepsilon) \to 0$ as $N \to \infty$; $a_N = \mathcal{O}_p(b_N)$ if for any $\varepsilon > 0$, there exist finite $C_\varepsilon > 0$ such that $\lim_{N\to\infty} \Pr(|a_N/b_N| \geq C_\varepsilon) < \varepsilon$. We use $(N,T) \to \infty$ to signify that N and T jointly go to infinity.

Layout. The rest of the paper is organized as follows. Section 2 presents the model setup and motivating examples and proposes the penalized GMM estimator. Section 3 investigates the asymptotic properties of the proposed method. The finite sample performance of the proposed method is evaluated through a Monte Carlo simulation experiment in Section 4. Section 5 applies the proposed method to examine the effects of agricultural productivity on rural-to-urban migration in China. Section 6 concludes. Proofs and computational details are relegated to the Appendix.

2 Framework

This section lays the foundation for our investigation of moment models with grouped heterogeneous validity of moment conditions with panel data. Firstly, we delineate the model's framework and examine illustrative econometric examples that motivate our study in Section 2.1 and 2.2. Subsequently, in Section 2.3 we introduce a penalized GMM estimator, specifically designed for parameter estimation, classification, and moment selection.

2.1 Model Setup

We examine an observed dataset wherein each observation $\boldsymbol{z}_{it} \in \mathbb{R}^{p_z}$, indexed by $i \in [N]$ and $t \in [T]$. This dataset may be conceptualized as either classical panel data or as clustered data. Within the framework of panel data, the index i represents the individual dimension, while t pertains to the time dimension. In the case of clustered data, on the other hand, i indicates the cluster dimension, with t identifying units within these clusters.

2.1.1 Moment Conditions

Consider $g(\boldsymbol{z}, \boldsymbol{\theta}) \in \mathbb{R}^L$, which represents a vector of moment functions associated with the parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{p_\theta}$, where p_θ is a predetermined positive integer. Let $\boldsymbol{\theta}^0$ denote the true parameter values of interest for $\boldsymbol{\theta}$.

There exists a subset $\mathcal{S} \subset [L]$ of the moment conditions, satisfying $p_\theta \leq L_{\mathcal{S}} = |\mathcal{S}| < L$, such that

$$\mathbb{E}\left[g_{\mathcal{S}}\left(\boldsymbol{z}_{it}, \boldsymbol{\theta}\right)\right] = 0 \tag{1}$$

is fulfilled if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}^0$, thus identifying the structural parameter $\boldsymbol{\theta}$ in accordance with (1). We assume that both p_θ and $L_{\mathcal{S}}$ are constant, while $L \equiv L_{NT}$ may increase with the sample sizes N and T.

The set $\mathcal{S}$ of *sure* moment conditions, which are correctly specified as per (1) and are sufficient for the identification of $\boldsymbol{\theta}$, is presumed to be known to the researchers based on economic theory or prior empirical evidence. Let $\mathcal{D} = [L] \setminus \mathcal{S}$, with $L_{\mathcal{D}} = L - L_{\mathcal{S}}$, denote the set of *doubtful* moment conditions, which may be potentially misspecified. Without loss of generality, let $\mathcal{D} = [L_{\mathcal{D}}]$. To model the pattern of misspecification, we introduce slackness parameters $\boldsymbol{\delta}_i \in \Theta_\delta \subset \mathbb{R}^{L_{\mathcal{D}}}$, $i \in [N]$, such that

$$\mathbb{E}\left[g_{\mathcal{D}}\left(\boldsymbol{z}_{it}, \boldsymbol{\theta}^0\right)\right] = \boldsymbol{\delta}_i, \tag{2}$$

thereby allowing for heterogeneous validity of moment conditions across i. Specifically, $\delta_{il} = 0$ indicates that moment condition $l \in \mathcal{D}$ is correctly specified for i, whereas $\delta_{il} \neq 0$ signifies that moment condition $l \in \mathcal{D}$ is misspecified for i. For each individual i, the set of doubtful moment conditions, $\mathcal{D}$, is partitioned into $\mathcal{V}_i$ and $\mathcal{I}_i$, where $\mathcal{V}_i = \{l \in [L_{\mathcal{D}}] : \delta_{il} = 0\}$ represents the subset of conditions correctly specified, and $\mathcal{I}_i = \{l \in [L_{\mathcal{D}}] : \delta_{il} \neq 0\}$ denotes those misspecified. The parameter

$$\zeta_{NT} = \min_{1 \leq i \leq N} \min_{l \in \mathcal{I}_i} |\delta_{il}^0|$$

quantifies the minimum degree of misspecification across individuals and conditions, which is permitted to approach zero.

2.1.2 Grouped Structure

Consider that $\mathcal{G} = \{\mathcal{G}_k\}_{k=1}^{K^0}$ constitutes a partition of the set $[N]$, such that $\bigcup_{k=1}^{K^0} \mathcal{G}_k = [N]$ and $\mathcal{G}_k \cap \mathcal{G}_j = \emptyset$ for all $j \neq k$. This partitioning organizes the sample into $K_0 \geq 1$ distinct groups. Define $N_k = \sum_{i=1}^N \mathbb{1}\left\{i \in \mathcal{G}_k\right\}$ as the number of observations within group

k, for $k \in [K^0]$. Furthermore, the group membership function $k(i) = \sum_{k=1}^{K^0} k \mathbb{1}\{i \in \mathcal{G}_k\}$ is introduced, alongside the indicator function $\lambda_{ik} = \mathbb{1}\{i \in \mathcal{G}_k\}$, for each $i \in [N]$ and $k \in [K^0]$.

Rather than assuming arbitrary patterns of heterogeneity in $\boldsymbol{\delta}_i$, we introduce a latent group structure by positing that

$$\boldsymbol{\delta}_i^0 = \sum_{k=1}^{K^0} \boldsymbol{\alpha}_k^0 \mathbb{1}\{i \in \mathcal{G}_k\} = \boldsymbol{\alpha}_{k(i)}^0, \tag{3}$$

where $\boldsymbol{\alpha}_k^0 \neq \boldsymbol{\alpha}_j^0$ for all $j \neq k$, signifying that the validity of moment conditions is homogeneous within groups but varies across different groups.[1] Accordingly, we define $\mathcal{V}_k = \{l \in \mathcal{D} : \alpha_{kl} = 0\}$ and $\mathcal{I}_k = \{l \in \mathcal{D} : \alpha_{kl} \neq 0\}$ for each group $k \in [K^0]$. Define

$$\begin{aligned} \mathcal{Z}_0 &= \{(i,j) \in [N] \times [N] : k(i) = k(j) \text{ and } i < j\}, \\ \mathcal{Z}_1 &= \{(i,j) \in [N] \times [N] : k(i) \neq k(j) \text{ and } i < j\}, \end{aligned} \tag{4}$$

to represent the sets of observation pairs within the same group and across different groups, according to their true latent group memberships, respectively. The notation of $\mathcal{Z}_0$ and $\mathcal{Z}_1$ is particularly helpful for rigorous analysis of the asymptotic properties of the estimator that will be proposed in Section 2.3 based on the adaptive fused Lasso penalty. Let

$$\rho_{NT} = \min_{1 \leq k < k' \leq K^0} \left\| \boldsymbol{\alpha}_k^0 - \boldsymbol{\alpha}_{k'}^0 \right\| = \min_{(i,j) \in \mathcal{Z}_1} \left\| \boldsymbol{\delta}_i^0 - \boldsymbol{\delta}_j^0 \right\|,$$

which delineates the minimum degree of separation between groups, a measure that is permitted to approach zero as the sample sizes N and T increase.

Remark 1. Define the matrix $\underset{N \times L_{\mathcal{D}}}{\boldsymbol{D}} = (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_N)'$, the group membership matrix $\underset{N \times K}{\boldsymbol{\Lambda}} = (\lambda_{ik})$, and the loading matrix $\underset{L_{\mathcal{D}} \times K}{\boldsymbol{A}} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_K)$. Consequently, in matrix notation, we have $\boldsymbol{D} = \boldsymbol{\Lambda}\boldsymbol{A}'$, which reveals a low-rank factor structure, as highlighted in the works of Ma et al. (2022); Chetverikov and Manresa (2022); Bonhomme et al. (2022). Within this framework, $\boldsymbol{\lambda}_i$ represents the factor loading, while $\boldsymbol{\alpha}_k$ denotes the latent factor. The interconnection between these two strands of literature and their implications for directions of extensions are briefly discussed in Remark 10.

[1] It is a direct extension to allow the structural parameters $\boldsymbol{\theta}_i$ to be heterogeneous and share the latent group structure with the slackness parameters $\boldsymbol{\delta}_i$. In this case, (1) and (2) are modified to $\mathbb{E}[g_{\mathcal{S}}(\boldsymbol{z}_{it}, \boldsymbol{\theta}_i)] = 0$, which is fulfilled if and only if $\boldsymbol{\theta}_i = \boldsymbol{\theta}_i^0$, and $\mathbb{E}\left[g_{\mathcal{D}}\left(\boldsymbol{z}_{it}, \boldsymbol{\theta}_i^0\right)\right] = \boldsymbol{\delta}_i$. Denote $\boldsymbol{\beta}_i = (\boldsymbol{\theta}_i', \boldsymbol{\delta}_i')'$. We impose the latent group structure on $\boldsymbol{\beta}_i$ by modifying (3), $\boldsymbol{\beta}_i^0 = \sum_{k=1}^{K^0} \boldsymbol{\alpha}_k^0 \mathbb{1}\{i \in \mathcal{G}_k\} = \boldsymbol{\alpha}_{k(i)}^0$, where $\boldsymbol{\alpha}_k^0 \neq \boldsymbol{\alpha}_j^0$ for all $j \neq k$. With minimal modification, we can propose the penalized GMM estimator for $\boldsymbol{\beta}$ and develop the same set of asymptotic properties as in Section 2.3 and 3.

The primary objective of this study is to facilitate the efficient estimation and inference of $\boldsymbol{\theta}^0$, alongside the detection of the group structure $\mathcal{G}$ and the validity parameters $\boldsymbol{\alpha}_k$, for $k \in [K^0]$.

2.2 Motivating Examples

Example 1 (Measurement Errors and Linear IV). Consider a first-differenced linear panel model, articulated as

$$\Delta y_{it} = \theta^0 \Delta x_{it} + \Delta \varepsilon_{it},$$

where i indexes individuals in the set $[N]$ and t denotes time periods within $[T]$. The researcher is concerned about the endogeneity of Δx_{it} caused by possible measurement errors, and identifies θ^0 by an exogenous IV, $z_{\mathcal{S},it}$, satisfying

$$\mathbb{E}\left[z_{\mathcal{S},it}\left(\Delta y_{it} - \theta^0 \Delta x_{it}\right)\right] = 0. \tag{5}$$

In addition to $z_{\mathcal{S},it}$, suppose there exists an alternative measure to x_{it}, $\widetilde{x}_{it}$, such that

$$\begin{aligned} x_{it} &= x_{it}^\star + u_{it} + \gamma_i v_{it},\ \gamma_i = \mathbb{1}\left\{i \in \mathcal{N}_1\right\}, \\ \widetilde{x}_{it} &= x_{it}^\star + \widetilde{u}_{it} + \widetilde{\gamma}_i \widetilde{v}_{it},\ \widetilde{\gamma}_i = \mathbb{1}\left\{i \in \mathcal{N}_2\right\}, \end{aligned}$$

where $x_{it}^\star$ is the latent explanatory variable, with $\mathbb{E}\left(\Delta \varepsilon_{it} \mid \Delta x_{it}^\star\right) = 0$, measured by observed proxy variables x_{it} and $\widetilde{x}_{it}$, u_{it} and $\widetilde{u}_{it}$ are classical measurement errors such that $\mathbb{E}\left(\Delta \varepsilon_{it} \mid \Delta u_{it}\right) = \mathbb{E}\left(\Delta \varepsilon_{it} \mid \Delta \widetilde{u}_{it}\right) = 0$ while v_{it} and $\widetilde{v}_{it}$ are nonclassical measurement errors that are correlated with the structural shocks $\Delta \varepsilon_{it}$ and consequently the source of endogeneity. $\mathcal{N}_1, \mathcal{N}_2 \subset [N]$ denote the set of individuals for which the proxy variables, x_{it} and $\widetilde{x}_{it}$, are endogenous due to the entering of v_{it} and $\widetilde{v}_{it}$, respectively. Let $\boldsymbol{z}_{\mathcal{D},it} = (\Delta x_{it}, \Delta \widetilde{x}_{it})$ and we then have the set of doubtful moment conditions,[2]

$$\mathbb{E}\left[\boldsymbol{z}_{\mathcal{D},it}\left(\Delta y_{it} - \theta^0 \Delta x_{it}\right)\right] = \boldsymbol{\delta}_i,$$

[2] If Δx_{it} is exogenous, the generalized least squares (GLS) estimator, which can be constructed as GMM estimator based on (5) together with $\mathbb{E}\left[\Delta x_{it}\left(\Delta y_{it} - \theta^0 \Delta x_{it}\right)\right] = 0$, is consistent and efficient. We can include Δx_{it} in $\boldsymbol{z}_{\mathcal{D},it}$ to check the exogeneity of Δx_{it}.

where

$$\boldsymbol{\delta}_i = \begin{cases} (\delta_1, \delta_2)' & \text{if } i \in \mathcal{N}_1 \cap \mathcal{N}_2, \\ (\delta_1, 0)' & \text{if } i \in \mathcal{N}_1 \setminus \mathcal{N}_2, \\ (0, \delta_2)' & \text{if } i \in \mathcal{N}_2 \setminus \mathcal{N}_1, \\ (0, 0)' & \text{otherwise,} \end{cases}$$

and $\delta_1 = \mathbb{E}\left(\Delta v_{it} \Delta \varepsilon_{it}\right) \neq 0$ and $\delta_2 = \mathbb{E}\left(\Delta \widetilde{v}_{it} \Delta \varepsilon_{it}\right) \neq 0$. $\boldsymbol{\delta}_i$ exhibits a group structure as in (3) capturing the heterogeneity of the validity of moment conditions across individuals. In a concrete empirical context, consider a simplified labor supply model, as in Liao (2013), that is specified as follows[3]

$$\Delta \log\left(y_{it}\right) = \theta^0 \Delta \log\left(x_{it}\right) + \Delta \varepsilon_{it},$$

where y_{it} denotes the annual hours worked, x_{it} represents the hourly wage rate and the parameter θ^0 captures the inter-temporal substitution elasticity of labor supply in response to evolutionary changes in wages. As discussed in MaCurdy (1981); Altonji (1986), researchers are concerned about measurement errors in $\Delta \log\left(x_{it}\right)$, which cause the OLS estimator to be inconsistent. MaCurdy (1981) suggest employing a set of family background variables (parents' education; parents' economic status when the individual is young; education, age and interaction between education and age) as IVs, whereas Altonji (1986) advocate alternative wage measures $\widetilde{x}_{i,t}$ to construct IVs.[4] The set $\mathcal{S}$ can be formed by including IVs for which there is a higher degree of confidence in their exogeneity, such as the economic status of the parents as in Liao (2013), and we investigate the potential group structure of the validity of moment conditions by including $\Delta \log\left(x_{it}\right)$, $\Delta \log\left(\widetilde{x}_{it}\right)$ and other IVs proposed in MaCurdy (1981) in $\boldsymbol{z}_{\mathcal{D},it}$.

Example 2 (Dynamic Panel). Consider the dynamic panel model

$$\Delta y_{it} = \theta^0 \Delta y_{i,t-1} + \Delta \varepsilon_{it},$$

$i \in [N]$ and $t \in [T]$. $\Delta y_{i,t-1}$ and $\Delta \varepsilon_{it}$ are naturally correlated, which leads OLS estimator to be inconsistent. Suppose ε_{it} is distributed independent across i and has no serial correlation, then $y_{i,t-2-j}$, $j = 0, 1, \cdots, t-2$ are valid instruments for $\Delta y_{i,t-1}$ (Arellano and Bond, 1991)

[3] We abstract the time-varying constant term from the model presented in Liao (2013) for illustration purposes.

[4] As also summarized in Liao (2013, Note 11), x_{it} is constructed by dividing the annual labor income of individual i by the annual labor supply and gross national product (GNP) price deflator in MaCurdy (1981); Altonji (1986). $\widetilde{x}_{it}$ is the hourly wage rate of individual i if the person is paid on hourly basis in Altonji (1986).

satisfying $\mathbb{E}\left(y_{i,t-2-j}\Delta y_{i,t-1}\right)\neq 0$ and

$$\mathbb{E}\left(y_{i,t-2-j}\Delta\varepsilon_{it}\right)=0.$$

The validity of the moment conditions relies on the *no serial correlation* assumption on the error term. In fact, the moment condition is still valid when the error term ε_{it} is serially correlated up to the order of j. If j is large, i.e. the lagged term is further from the current period, the validity is more robust. However, as noted in Arellano and Bover (1995) and Blundell and Bond (1998), the moment conditions in (6) may contain weak information about the structural parameter θ^0, particularly when θ^0 is close to 1 and j is large. There is a natural tension between the validity of moment conditions and the identification strength of the structural parameter. In practice, we can construct the $\mathcal{S}$ set, i.e.

$$g_{\mathcal{S}}\left(z_{it},\theta\right)=\left[y_{i,t-2-j}\left(\Delta y_{it}-\theta\Delta y_{i,t-1}\right)\right]_{j=\bar{j}\cdots,t-2}, \tag{6}$$

and use $\left[y_{i,t-2-j}\right]_{j=0,1,\cdots,\bar{j}}$ as instruments to form the $\mathcal{D}$ set.

We can derive additional moment conditions to form $\mathcal{D}$ by imposing more restrictions on the data-generating process. For example,

$$\begin{aligned}\mathbb{E}\left(y_{i,t}\Delta\varepsilon_{i,t+1}-y_{i,t+1}\Delta\varepsilon_{i,t+2}\right)=0,\ t=1,2,\ldots,T-2\\ \mathbb{E}\left(\bar{\varepsilon}_i\Delta\varepsilon_{i,t+1}\right)=0,\ t=1,2,\cdots,T-1,\end{aligned} \tag{7}$$

where $\bar{\varepsilon}_i=T^{-1}\sum_{t=1}^{T}\varepsilon_{it}$, hold under homoskedasticity across time $\mathbb{E}\left(\varepsilon_{it}^2\right)=\sigma_i^2$ (Ahn and Schmidt, 1995). However, the homoskedasticity assumption may not hold uniformly across unit i, say $\mathbb{E}\left(\varepsilon_{it}^2\right)=\sigma_i^2+\gamma_i\omega_{it}$, where $\gamma_i=\mathbb{1}\left(i\in\mathcal{N}\right)$ for some subset of individuals $\mathcal{N}\subset[N]$ and ω_{it} captures the heteroskedasticity across t. Additional linear or nonlinear moment conditions can be derived under homoskedasticity or initial condition restrictions (Arellano and Bover, 1995; Ahn and Schmidt, 1997; Blundell and Bond, 1998) and verifying the underlying assumptions may not be easy in practice.

In this scenario, our method accounts for the heterogeneity of the validity of moment conditions across individuals to robustly estimate the autoregressive parameter and also provides a procedure to test the validity of assumptions underlying the moment conditions.

Example 3 (Shift-share (Bartik) IV). The shift-share IV has become increasingly popular in empirical studies. Recent developments including Borusyak et al. (2022), Goldsmith-Pinkham et al. (2020) and Adao et al. (2019) provide theoretical justification of the shift-share IV. As shown in Goldsmith-Pinkham et al. (2020) that the Bartik IV estimator is equivalent to the GMM estimator with industry shares as IVs and a specific weight matrix.

Consider the model with N locations, T time periods, and J industries,

$$y_{it} = \theta^0 x_{it} + \varepsilon_{it},$$

where the endogenous variable employs the accounting identity $x_{it} = \sum_{j=1}^{J} s_{ijt} v_{ijt}$, and the location-industry-time shift can be decomposed as $v_{ijt} = v_{jt} + \widetilde{v}_{ijt}$, where v_{jt} is the industry-time shift and $\widetilde{v}_{ijt}$ is idiosyncratic shock. The shift-share IV is constructed by

$$z_{it} = \sum_{j=1}^{J} s_{ij0} v_{jt}, \tag{8}$$

which is the inner product of initial shares and aggregated level industry-time shifts. A notable example is provided by Autor et al. (2013), who investigate the causal impact of increased import penetration from China on local labor markets within the United States. In their analysis, the endogenous variable x_{it} quantifies the local exposure to the surge in imports from China, s_{ij0} represents the employment share of the manufacturing industry j within location i, measured a decade prior to each period t, and v_{jt} denotes the growth of imports of products in industry j from China into the eight comparable economies over the period t.

The validity of the shift-share instrumental variable (IV) hinges on the exogeneity of either v_{jt} or s_{ij0}, with the other variable being treated as fixed. However, the exogeneity assumption may not hold for each location-industry pair (i, j) since locations have different industrial structures and local amenities. In this case, the IV constructed in (8) by combining all industries for all i can be invalid. For instance, considering a scenario where the geographical location is Silicon Valley and the analysis incorporates the technology sector as part of (8), it is not plausible that the exogeneity condition holds. Consequently, it is crucial to select suitable subsets of industries for different groups of locations to construct valid shift-share IVs. In practice, we can construct the set of IVs as

$$z_{\mathcal{S},it} = \sum_{j \in \mathcal{J}_S} s_{ij0} v_{jt} \text{ and } z_{l,it} = \sum_{j \in \mathcal{J}_S \cup \widetilde{\mathcal{J}}_l} s_{ij0} v_{jt},$$

where $\mathcal{J}_S$ is a subset of industries for which the exogeneity condition holds based on prior knowledge or empirical evidence, and we add more industries in $\widetilde{\mathcal{J}}_l$, $l \in [L_{\mathcal{D}}]$, to construct additional IVs whose validity is subject to detection. Denote $\boldsymbol{z}_{\mathcal{D},it} = [z_{l,it}]_{l=1}^{L_{\mathcal{D}}}$. It is expected that the validity of constructed IVs is heterogeneous across locations.

2.3 Penalized GMM Estimation

To facilitate understanding, we introduce the following notations. Let

$$m\left(\boldsymbol{z},\boldsymbol{\theta}\right)=\begin{bmatrix} g_{\mathcal{S}}\left(\boldsymbol{z},\boldsymbol{\theta}\right)\\ g_{\mathcal{D}}\left(\boldsymbol{z},\boldsymbol{\theta}\right)\end{bmatrix},\ g\left(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\delta}\right)=\begin{bmatrix} g_{\mathcal{S}}\left(\boldsymbol{z},\boldsymbol{\theta}\right)\\ g_{\mathcal{D}}\left(\boldsymbol{z},\boldsymbol{\theta}\right)-\boldsymbol{\delta}\end{bmatrix}$$

represent the moment functions with and without slackness parameters, respectively. Correspondingly,

$$\begin{gathered}\overline{m}_i\left(\boldsymbol{\theta}\right)=\mathbb{E}\left[m\left(\boldsymbol{z}_{it},\boldsymbol{\theta}\right)\right],\overline{g}_i\left(\boldsymbol{\theta},\boldsymbol{\delta}\right)=\mathbb{E}\left[g\left(\boldsymbol{z}_{it},\boldsymbol{\theta},\boldsymbol{\delta}\right)\right],\\ \overline{m}_{\mathcal{S},i}\left(\theta\right)=\mathbb{E}\left[g_{\mathcal{S}}\left(\boldsymbol{z}_{it},\boldsymbol{\theta}\right)\right],\overline{m}_{\mathcal{D},i}\left(\theta\right)=\mathbb{E}\left[g_{\mathcal{D}}\left(\boldsymbol{z}_{it},\boldsymbol{\theta}\right)\right].\end{gathered}$$

The empirical counterparts are denoted as

$$\begin{gathered}\widehat{m}_{i,T}\left(\boldsymbol{\theta}\right)=\frac{1}{T}\sum_{t=1}^{T}m\left(\boldsymbol{z}_{it},\boldsymbol{\theta}\right),\widehat{g}_{i,T}\left(\boldsymbol{\theta},\boldsymbol{\delta}_i\right)=\frac{1}{T}\sum_{t=1}^{T}g\left(\boldsymbol{z}_{it},\boldsymbol{\theta},\boldsymbol{\delta}_i\right),\\ \widehat{m}_{\mathcal{S},i,T}\left(\boldsymbol{\theta}\right)=\frac{1}{T}\sum_{t=1}^{T}g_{\mathcal{S}}\left(\boldsymbol{z}_{it},\boldsymbol{\theta}\right),\widehat{m}_{\mathcal{D},i,T}\left(\boldsymbol{\theta}\right)=\frac{1}{T}\sum_{t=1}^{T}g_{\mathcal{D}}\left(\boldsymbol{z}_{it},\boldsymbol{\theta}\right).\end{gathered}$$

We also define the Jacobian matrices as

$$\Gamma_{\mathcal{S},i}\left(\boldsymbol{\theta}\right)=\frac{\partial\overline{m}_{\mathcal{S},i}\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}'},\ \Gamma_{\mathcal{D},i}\left(\boldsymbol{\theta}\right)=\frac{\partial\overline{m}_{\mathcal{D},i}\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}'},\ \Gamma_i\left(\boldsymbol{\theta}\right)=\begin{bmatrix}\Gamma_{\mathcal{S}}\left(\boldsymbol{\theta}\right) & \mathbf{0}_{L_{\mathcal{S}}\times L_{\mathcal{D}}}\\ \Gamma_{\mathcal{D}}\left(\boldsymbol{\theta}\right) & -\boldsymbol{I}_{L_{\mathcal{D}}\times L_{\mathcal{D}}}\end{bmatrix}.$$

We propose the penalized GMM estimator,

$$\left(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{D}}\right)=\underset{\boldsymbol{\theta}\in\Theta,\,\boldsymbol{D}\in{\Theta_\delta}^N}{\arg\min}\ \widehat{Q}_{NT}\left(\boldsymbol{\theta},\boldsymbol{D}\right)+P_{\psi_1,\psi_f}\left(\boldsymbol{D}\right),\tag{9}$$

where

$$\begin{gathered}\widehat{Q}_{NT}\left(\boldsymbol{\theta},\boldsymbol{D}\right)=\frac{1}{N}\sum_{i=1}^{N}\widehat{Q}_{i,T}\left(\boldsymbol{\theta},\boldsymbol{\delta}_i\right),\\ \widehat{Q}_{i,T}\left(\boldsymbol{\theta},\boldsymbol{\delta}_i\right)=\widehat{g}_{i,T}\left(\boldsymbol{\theta},\boldsymbol{\delta}_i\right)'\boldsymbol{W}_{i,T}\widehat{g}_{i,T}\left(\boldsymbol{\theta},\boldsymbol{\delta}_i\right),\end{gathered}$$

$$P_{\psi_1,\psi_c}\left(\boldsymbol{D}\right)=\frac{\psi_1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L_{\mathcal{D}}}\dot{w}_{il}\left|\delta_{il}\right|+\frac{\psi_f}{N^2}\sum_{1\le i<j\le N}\dot{\mu}_{ij}\left\|\boldsymbol{\delta}_i-\boldsymbol{\delta}_j\right\|,\tag{10}$$

and $\boldsymbol{W}_{i,T}$ is a positive definite weighting matrix.

The penalty scheme $P_{\psi_1,\psi_c}\left(\boldsymbol{D}\right)$ integrates a variant of the adaptive Lasso penalty (Zou, 2006; Cheng and Liao, 2015) and the pairwise adaptive fused Lasso (PAFL) penalty (Mehrabani, 2023) with the adaptive weights $\dot{w}_{il} = \left|\dot{\delta}_{il}\right|^{-\kappa_1}$ and $\dot{\mu}_{ij} = \left\|\dot{\boldsymbol{\delta}}_i - \dot{\boldsymbol{\delta}}_j\right\|^{-\kappa_f}$ based on a preliminary consistent estimator $\dot{\boldsymbol{D}}$, where $\kappa_1, \kappa_f \geq 2$.

The adaptive Lasso penalty is designed to differentiate the valid and invalid moment conditions by shrinking the slackness parameter δ_{il} associated with valid moment conditions, $l \in \mathcal{V}_i$, to zero while leaving δ_{il} stay away from zero for $l \in \mathcal{I}_i$. In the case of valid moment conditions, i.e. $l \in \mathcal{V}_i$, the adaptive weights $\dot{w}_{il}$ tends to be large since $\dot{\delta}_{il}$ is close to zero by its consistency, which imposes heavy penalties to ensure $\widehat{\delta}_{il} = 0$ for $l \in \mathcal{V}_i$. In contrast, invalid moment conditions are subject to small penalties so that they are asymptotically associated with nonzero estimated slackness parameters.

We employ the PAFL penalty to achieve the classification of individuals based on the validity of moment conditions. The penalty encourages $\boldsymbol{\delta}_i = \boldsymbol{\delta}_j$ if i and j belong to the same group, i.e. $\boldsymbol{\delta}_i^0 = \boldsymbol{\delta}_j^0$. The adaptive weights $\dot{\mu}_{ij}$ works similarly as $\dot{w}_{il}$ following the intuition of the adaptive Lasso (Zou, 2006).

Remark 2. It is also possible to utilize other methodologies to handle the latent group structure. For example, we can consider the classifier-Lasso (Su et al., 2016),

$$P_{\psi_c}(\boldsymbol{D}, \boldsymbol{A}) = \frac{\psi_c}{N} \sum_{i=1}^{N} \prod_{k=1}^{K} \left\|\boldsymbol{\delta}_i - \boldsymbol{\alpha}_k\right\|.$$

The estimation with the classifier-Lasso penalty, even associated with convex moment conditions, is a non-convex optimization problem. The numerical solution is approximated by solving a sequence of convex optimization problems (Gao and Shi, 2021), which is computationally expensive, and the convergence is not guaranteed. Furthermore, the number of groups K is a tuning parameter, in addition to the regularization parameter ψ_c, that needs to be selected in advance. On the contrary, the PAFL penalty is convex and relies on a single tuning parameter ψ_f to control the strength of the penalty.

2.3.1 Preliminary Estimator

Consider the GMM estimator using moment conditions in $\mathcal{S}$ for initial estimation of $\boldsymbol{\theta}^0$ and the plug-in estimator for $\boldsymbol{\delta}_i$,

$$\dot{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \widehat{m}_{\mathcal{S},NT}(\boldsymbol{\theta})' \boldsymbol{W}_{NT} \widehat{m}_{\mathcal{S},NT}(\boldsymbol{\theta}) \text{ and } \dot{\boldsymbol{\delta}}_i = \widehat{m}_{\mathcal{D},i,T}\left(\dot{\boldsymbol{\theta}}\right), \tag{11}$$

for $i \in [N]$, where $\widehat{m}_{\mathcal{S},NT}(\boldsymbol{\theta}) = N^{-1}\sum_{i=1}^{N} \widehat{m}_{\mathcal{S},i,T}(\boldsymbol{\theta})$ and $\boldsymbol{W}_{NT}$ is a positive definite weighting matrix. Denote $\dot{\boldsymbol{D}} = \left(\dot{\boldsymbol{\delta}}_1, \dot{\boldsymbol{\delta}}_2, \cdots, \dot{\boldsymbol{\delta}}_N\right)'$. The asymptotic properties of $\dot{\boldsymbol{D}}$ will be developed in Lemma A.1 in the appendix.

Remark 3. In the asymptotic analysis, only the convergence rate $\max_{1\leq i\leq N}\left\|\dot{\boldsymbol{\delta}}_i-\boldsymbol{\delta}_i^0\right\|=\mathcal{O}_p\left(\left(\frac{L_{\mathcal{D}}}{T}\right)^{\frac{1}{2}}(\log T)^3\right)$ is required. If the moment conditions in $\mathcal{S}$ strongly identifies $\boldsymbol{\theta}$, then $\dot{\boldsymbol{\delta}}_i$ defined in (11) is $\sqrt{NT}$-consistent, when $L_{\mathcal{D}}$ is fixed, for $\boldsymbol{\delta}_i^0$, which much faster than the required rate. Weakening the identification strength will slow down the convergence rate of $\dot{\boldsymbol{\delta}}_i$. In this sense, our method can allow moderately weak identification of moment conditions in $\mathcal{S}$ and still achieve regular convergence rate if we have strong identification of moment conditions in $\mathcal{D}$.

3 Asymptotic Properties

This section is devoted to investigating the asymptotic properties of the penalized GMM estimator introduced in Section 2.3. Initially, we outline the assumptions necessary for the estimator's consistency and proceed to derive the convergence rates of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\delta}}_i$. Subsequently, we establish the consistency of classification and moment selection, in the sense that the estimated group structure and the selected set of moment conditions for each group coincide with the true underlying sets with probability approaching 1. In the final part of this section, we study the asymptotic distribution of the penalized GMM estimator and its oracle properties. Detailed proofs are relegated in Appendix A.1.

3.1 Consistency and Preliminary Rate of Convergnce

To begin with, we introduce the following regularity conditions.

Assumption 1. *(i) $\boldsymbol{z}_i=\{\boldsymbol{z}_{it}:t\in[T]\}$ are independently distributed across i. For each i, $\boldsymbol{z}_i$ is stationary strong mixing with mixing coefficients $\boldsymbol{a}_i(\cdot)$, where*

$$\boldsymbol{a}(\cdot)=\sup_N\max_{1\leq i\leq N}\boldsymbol{a}_i(\cdot)$$

satisfies $\boldsymbol{a}(s)\leq c_a r^s$ for some $c_a>0$ and $r\in(0,1)$.

(ii) $\boldsymbol{\theta}^0$ lies in the interior of a compact set Θ.

(iii) There exists $f(\boldsymbol{z}_{it})$ s.t. $\sup_{\boldsymbol{\theta}\in\Theta}\|m(\boldsymbol{z}_{it},\boldsymbol{\theta})\|\leq f(\boldsymbol{z}_{it})$ and

$$\left\|m(\boldsymbol{z}_{it},\boldsymbol{\theta})-m\left(\boldsymbol{z}_{it},\overline{\boldsymbol{\theta}}\right)\right\|\leq f(\boldsymbol{z}_{it})\left\|\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}\right\|,$$

for all $\boldsymbol{\theta},\overline{\boldsymbol{\theta}}\in\Theta$. $\mathbb{E}\left|f(\boldsymbol{z}_{it})\right|^q<\infty$ for some $q\geq 6$.

(iv) $N = \mathcal{O}(T^{\frac{q}{2}-1})$ *where* $q \geq 6$ *is the constant in (iii).*

Remark 4. Assumption 1 is comparable to Assumption A1 in Su et al. (2016), which is essential to guarantee the convergence of the sample moments to the population moments uniformly over the parameter space and i at a desired rate, which is formally shown in Lemma A.1. The observations are assumed to be cross-sectionally independent and the dependence across t within each i is controlled by a mixing condition as in Assumption 1(i). Assumption 1(ii) regulates the parameter space. Assumption 1(iii) imposes a Lipschitz bound on the moment functions. Alternatively, one can follow Bonhomme and Manresa (2015); Liu et al. (2020) to impose tail conditions on $f(\boldsymbol{z}_{it})$ directly. Finally, Assumption 1(iv) specifies the relative growth rate of N and T depending on $q \geq 6$.

Assumption 2. *(i)* $\overline{m}_{\mathcal{S},i}(\boldsymbol{\theta})$ *is continuous in* $\boldsymbol{\theta}$ *for all* i*, and for any* $\epsilon > 0$*,*

$$\inf_{N} \min_{1\leq i\leq N} \inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|>\epsilon} \left\|\overline{m}_{\mathcal{S},i}(\boldsymbol{\theta})\right\| > 0.$$

(ii) $|\delta_{il}^0| \leq C$ *for* $i \in [N]$ *and* $l \in [L_{\mathcal{D}}]$*.*

(iii) There exists a sequence of constants $\tau_T \to 0$ *with* $\tau_T^{-1} = \mathcal{O}\left(T^{\frac{1}{2}}\right)$ *and a fixed constant* η *such that*

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|\leq\eta} \left\|\widehat{m}_{i,T}(\boldsymbol{\theta}) - \overline{m}_i(\boldsymbol{\theta})\right\| = \mathcal{O}_p(\tau_T),$$

for $i \in [N]$*.*

(iv) There exist nonrandom matrices $\boldsymbol{W}_{i,NT}$ *and some constant* $c_w, C_w > 0$ *such that*

$$\Pr\left(\max_{1\leq i\leq N} \left\|\boldsymbol{W}_{i,NT} - \boldsymbol{W}_i\right\| \geq \epsilon\right) = o(1)$$

for any $\epsilon > 0$*, and*

$$c_w < \inf_{N} \min_{1\leq i\leq N} \sigma_{\min}(\boldsymbol{W}_i) \leq \sup_{N} \max_{1\leq i\leq N} \sigma_{\max}(\boldsymbol{W}_i) < C_w.$$

(v) $\overline{m}_i(\boldsymbol{\theta})$ *is continuously differentiable for any* θ *in the local neighborhood of* $\boldsymbol{\theta}^0 \in \Theta$ *for all* i*, and there exists a constant* $c_\Gamma, C_\Gamma > 0$ *such that for some* $\eta > 0$

$$c_\Gamma < \inf_{N} \min_{1\leq i\leq N} \inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|\leq\eta} \sigma_{\min}\left(\Gamma_i(\boldsymbol{\theta})'\Gamma_i(\boldsymbol{\theta})\right) \leq \sup_{N} \max_{1\leq i\leq N} \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|\leq\eta} \sigma_{\max}\left(\Gamma_i(\boldsymbol{\theta})'\Gamma_i(\boldsymbol{\theta})\right) < C_\Gamma.$$

Remark 5. Assumption 2(i) ensures the structural parameter $\boldsymbol{\theta}$ is identifiable by the moment conditions in $\mathcal{S}$ as in (1). Assumption 2(ii) imposes a compactness condition on the true parameter $\boldsymbol{\delta}^0$, which is essentially assuming the existence of moments for moment conditions in $\mathcal{D}$. Assumption 2(iii) is a high-level condition on the convergence rate of the empirical process indexed by the moment function $m(\boldsymbol{z}_{it}, \boldsymbol{\theta})$. When the number of moment conditions, L is fixed, standard empirical process theory for dependent data implies $\tau_T = T^{-\frac{1}{2}}$ (Dehling and Philipp, 2002). The τ_T is introduced for an increasing number of moment conditions. Cheng and Liao (2015, Lemma D.1) provides sufficient conditions for $\tau_T = \sqrt{L/T}$ to hold. Assumption 2(iv) imposes regularity conditions on the weighting matrix as in Su et al. (2016, Assumption B1(iv)) and (v) regulates the first-order derivatives of the moment conditions.

Assumption 3. *Let* $\widetilde{\varkappa}_{NT} = \left(\frac{L_{\mathcal{D}}}{T}\right)^{\frac{1}{2}} (\log T)^3$ *and* $\varkappa_{NT} = \left(\frac{L_{\mathcal{D}}}{T}\right)^{\frac{1}{2}} (\log T)^{3+\upsilon}$ *for some* $\upsilon > 0$.

(i) $\left(\rho_{NT}^{-1} \vee \zeta_{NT}^{-1}\right) \varkappa_{NT} = o(1)$.

(ii) $\psi_f = \mathcal{O}\left(L_{\mathcal{D}}^{-\frac{1}{2}} \rho_{NT}^{\kappa_f} \tau_T\right)$ *and* $\psi_1 = \mathcal{O}\left(L_{\mathcal{D}}{}^{-\frac{1}{2}} \zeta_{NT}^{\kappa_1} \tau_T\right)$.

Remark 6. Assumption 3 is a set of rate conditions. As shown in Lemma A.1 and Theorem 2(i), the rates $\widetilde{\varkappa}_{NT}$ and $\varkappa_{NT}$ controls the uniform convergence of the preliminary estimator defined in (11) and the penalized GMM estimator, respectively. Assumption 3(i) ensures that the estimators converge to the true parameters faster than the minimal degree of group separation ρ_{NT} and the minimal degree of misspecification ζ_{NT} so that we can still correctly separate different groups and identify the invalid moment conditions based on $\widehat{\boldsymbol{\delta}}_i$ even when $\rho_{NT}, \zeta_{NT} \to 0$. Assumption 3(ii) specifies upper bounds on the tuning parameters ψ_1 and ψ_f to ensure that the penalty scheme cannot dominate the GMM objective so that the consistency of the penalized GMM estimator pertains.

With the assumptions outlined above, we can derive the rate of convergence of the penalized GMM estimator $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^{p_\theta}$ and $\widehat{\boldsymbol{D}} \in \mathbb{R}^{L_{\mathcal{D}} \times N}$ in the following Theorem.

Theorem 1. *Suppose Assumption 1 - 3 holds, then*

(i) $\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right\| = \mathcal{O}_p(\tau_T)$ *and* $\left\|\widehat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i^0\right\| = \mathcal{O}_p(\tau_T)$ *for* $i \in [N]$.

(ii) $N^{-1} \sum_{i=1}^N \left\|\widehat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}^0\right\|^2 = \mathcal{O}_p(\tau_T^2)$.

Remark 7. Theorem 1 establishes the pointwise and mean square convergence $\widehat{\boldsymbol{\delta}}_i$. As shown in (A.16) in the proof in Appendix A.1, the rate of convergence depends on $a_{NT} = \psi_f \left(\max_{(i,j)\in\mathcal{Z}_1} \dot{\mu}_{ij}\right)$ and $b_{NT} = \psi_1 \max_{1\le i\le N} \|\dot{\boldsymbol{w}}_{i,\mathcal{I}_i}\|$, where $\dot{\boldsymbol{w}}_{i,\mathcal{I}_i}$ is the subvector of $\dot{\boldsymbol{w}}_i$ with element w_{il}, for $l \in \mathcal{I}_i = \{l \in \mathcal{D} : \delta_{il}^0 \neq 0\}$, in addition to τ_T, which is comparable to the results in Cheng and Liao (2015, Theorem 3.2) and Su et al. (2016, Theorem 2.1). With Assumption 3(ii), we can simplify the results as in Theoreom 1 by showing $a_{NT} = \mathcal{O}_p(\tau_T)$ and $b_{NT} = \mathcal{O}_p(\tau_T)$.

3.2 Classification and Moment Selection Consistency

In addition to Assumptions in Section 3.1, we further introduce the following assumption.

Assumption 4. *(i)* $\psi_f = O\left(\rho_{NT}^{\kappa_f} L_{\mathcal{D}}^{-\frac{1}{2}} \widetilde{\varkappa}_{NT}^2\right)$ *and* $\psi_1 = O\left(\zeta_{NT}^{\kappa_1} L_{\mathcal{D}}^{-\frac{1}{2}} \widetilde{\varkappa}_{NT}^2\right)$.

(ii) $\psi_f^{-1} = o\left(\left(\tau_T \sqrt{L_{\mathcal{D}}} \widetilde{\varkappa}_{NT}^{\kappa_f}\right)^{-1}\right)$ *and* $\psi_1^{-1} = o\left(\left(\tau_T \widetilde{\varkappa}_{NT}^{\kappa_1}\right)^{-1}\right)$.

(iii) $\lim_{N\to\infty} \min_{1\le k\le K^0} N_k/N \to \pi_{\min} \in (0,1)$.

Remark 8. As we will see in Theorem 2, the classification and moment selection consistency relies on the uniform consistency of the penalized GMM estimator across i, which is a stronger result than Theorem 1(i). Consequently, it necessitates a more restrictive upper bound on the tuning parameters ψ_f and ψ_1 in Assumption 4(i) than that in Assumption 3(ii). Assumption 4(ii) delineates lower bounds for ψ_f and ψ_1. This ensures that for any $(i,j) \in \mathcal{Z}_0$ and $l \in \mathcal{V}_i$, ψ_f and ψ_1, in conjunction with the adaptive weights $\dot{\mu}_{ij}$ and $\dot{w}_{il}$, levy sufficiently heavy penalties on $\|\boldsymbol{\delta}_i - \boldsymbol{\delta}_j\|$ and $|\delta_{il}|$, respectively, which is imperative for the consistency of the classification and moment selection. The upper and lower bounds for the tuning parameters also hinge on the range we can allow for ζ_{NT}, ρ_{NT} and the number of moments $L_{\mathcal{D}}$, and guide the choice of κ_1 and κ_f. Consider the case where $\tau_T = \sqrt{\frac{L_{\mathcal{D}}}{N}}$. Let $\zeta_{NT} = T^{-\phi_\zeta}$, $\rho_{NT} = T^{-\phi_\rho}$ and $L_{\mathcal{D}} = T^{\phi_L}$ take polynomial rates of T for some $\phi_\zeta, \phi_\rho, \phi_L > 0$. It is required that $\frac{\tau_T \widetilde{\varkappa}_{NT}^{\kappa_1}}{\zeta_{NT}^{\kappa_1} L_{\mathcal{D}}^{-\frac{1}{2}} \widetilde{\varkappa}_{NT}^2} = o_p(1)$, which leads to $\frac{1}{2} - \frac{1}{2}(1 - 2\phi_\zeta - \phi_L)\kappa_1 < 0$. With $\kappa_1 = 2$, we can allow

$$2\phi_\zeta + \phi_L < \frac{1}{2}. \tag{12}$$

Similarly, we can derive conditions for κ_f, ϕ_ρ and ϕ_L that $\left(\frac{1}{2} + \frac{\phi_L}{2}\right) - \frac{1}{2}(1 - 2\phi_\rho - \phi_L)\kappa_f < 0$, which call for a large choice of κ_f. With $\kappa_f = 3$, we can allow

$$3\phi_\rho + 2\phi_L < 1. \tag{13}$$

With (12) and (13), Assumption 3(i), which requires $\phi_L + 2(\phi_\zeta \vee \phi_\rho) < 1$, is automatically satisfied under the current setting.

Remark 9. In instances where the degree of misspecification of moment conditions is minimal, as indicated by a large value of ϕ_ζ, or when groups are not well separated, i.e. ϕ_ρ takes a large value, our methodology may not consistently identify invalid moment conditions or accurately discern the underlying group structure. Moreover, as we shall show in Theorem 2, the penalized GMM estimator can detect invalid moment conditions up to the rate of convergence $\varkappa_{NT}$ while it fails to achieve consistent moment selection when $\zeta_{NT} \asymp \frac{1}{\sqrt{T}}$, which is the case when the moment conditions are locally misspecified for at least one group. These intricacies prompt the exploration of robust bias-aware inference techniques in future research. Armstrong and Kolesár (2021) propose bias-aware confidence intervals, in the presence of local misspecification at $\sqrt{T}$-rate, that can be constructed by taking the GMM estimator and adding and subtracting the standard error times a critical value that takes into account the potential bias from misspecification of the moment conditions. It is a fruitful direction to consider the post-regularization estimation as in (18) and constructing bias-aware confidence intervals for the structural parameters following Armstrong and Kolesár (2021).

Remark 10. As noticed in Remark 1, the group structure admits a factor structure $\boldsymbol{D} = \boldsymbol{\Lambda} \boldsymbol{A}'$. The approximate moment conditions with $\alpha_{kl} = \frac{c}{\sqrt{T}}$ for a constant $c \neq 0$ for some $l \in [L_{\mathcal{D}}]$ corresponds to weak factor issues in the interactive fixed effects models, for which the robust inference method is investigated in the recent work by Armstrong et al. (2023).

Remark 11. Assumption 4(iii) imposes the condition that no group has an asymptotically trivial size, which is a convenient handle for the technical derivation. At the cost of cumbersomeness in notations and derivations, we can relax this condition to allow for $\pi_{\min} = 0$.

Now we are readily to establish the following theorem, which directly implies classification and moment selection consistency which will be elaborated in Remark 14.

Theorem 2. *Suppose Assumption 1 - 4 hold, then as* $(N, T) \to \infty$*,*

(i) $\max_{1 \le i \le N} \left\| \widehat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i^0 \right\| = O_p\left(\varkappa_{NT}\right).$

(ii) $\Pr\left(\max_{(i,j) \in \mathcal{Z}_0} \left\| \widehat{\boldsymbol{\delta}}_i - \widehat{\boldsymbol{\delta}}_j \right\| > 0\right) \to 0$ *and* $\Pr\left(\min_{(i,j) \in \mathcal{Z}_1} \left\| \widehat{\boldsymbol{\delta}}_i - \widehat{\boldsymbol{\delta}}_j \right\| > 0\right) \to 1.$

(iii) $\Pr\left(\max_{1 \le i \le N} \max_{l \in \mathcal{V}_i} \left| \widehat{\delta}_{il} \right| > 0\right) \to 0$ *and* $\Pr\left(\min_{1 \le i \le N} \min_{l \in \mathcal{I}_i} \left| \widehat{\delta}_{il} \right| > 0\right) \to 1.$

Remark 12. As in Appendix A.1, we show the first statement of Theorem 2(ii) in the first place by investigating the Karush-Kuhn-Tucker (KKT) conditions of (9) and making use of Assumption 4(ii). With this result, for any $i \in [N]$ such that $\left\| \widehat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i^0 \right\| > \varkappa_{NT}$, we have $\left\| \widehat{\boldsymbol{\delta}}_j - \boldsymbol{\delta}_i^0 \right\| > \varkappa_{NT}$ for all $j \in \mathcal{G}_{k(i)}$, which is essential to prove the uniform consistency in Theorem 2(i). It is then guaranteed that we can separate different groups and detect invalid moment conditions up to the rate $\varkappa_{NT}$. Together with Assumption 3(i), we can show the second part of both Theorem 2(ii) and (iii).

Remark 13. Mehrabani (2023, Theorem 3.2) claim a similar results to the first statement in Theorem 2(ii). However, the proof in Mehrabani (2023) actually attempts to show

$$\Pr\left(\min_{j \in \mathcal{G}_{k(i)}} \left\| \widehat{\boldsymbol{\delta}}_i - \widehat{\boldsymbol{\delta}}_j \right\| > 0\right) \to 0$$

as $(N,T)\to\infty$ for each individual $i\in[N]$, which is neither a proper notion of classification consistency nor a property uniform over i. Since the uniform consistency is not established in Mehrabani (2023), the second statement in Theorem 2(ii) is not guaranteed and hence the classification consistency is not established as claimed therein. The proof in A.1 can serve as a remedy in their setting.

Remark 14. Denote

$$\widehat{\mathcal{Z}}_0=\left\{(i,j)\in[N]\times[N]:\widehat{\boldsymbol{\delta}}_i=\widehat{\boldsymbol{\delta}}_j, i<j\right\} \text{ and } \widehat{\mathcal{Z}}_1=\left\{(i,j)\in[N]\times[N]:\widehat{\boldsymbol{\delta}}_i\neq\widehat{\boldsymbol{\delta}}_j, i<j\right\}.$$

Theorem 2(ii) shows that

$$\Pr\left(\mathcal{Z}_0\subset\widehat{\mathcal{Z}}_0\right)\to 1,\ \Pr\left(\mathcal{Z}_1\subset\widehat{\mathcal{Z}}_1\right)\to 1, \tag{14}$$

as $(N,T)\to\infty$. By triangle inequality, Theorem 2(i) and the rate condition Assumption 3(ii), we have

$$\begin{aligned}\Pr\left(\max_{(i,j)\in\widehat{\mathcal{Z}}_0}\|\boldsymbol{\delta}_i^0-\boldsymbol{\delta}_j^0\|\geq\rho_{NT}\right)&\leq\Pr\left(\max_{(i,j)\in\widehat{\mathcal{Z}}_0}\left\|\widehat{\boldsymbol{\delta}}_i-\widehat{\boldsymbol{\delta}}_j\right\|+\left\|\widehat{\boldsymbol{\delta}}_i-\boldsymbol{\delta}_i^0\right\|+\left\|\widehat{\boldsymbol{\delta}}_j-\boldsymbol{\delta}_j^0\right\|\geq\rho_{NT}\right)\\&\leq\Pr\left(\max_{1\leq i\leq N}\left\|\widehat{\boldsymbol{\delta}}_i-\boldsymbol{\delta}_i^0\right\|\geq\frac{\rho_{NT}}{2}\right)\to 0,\end{aligned}$$

as $(N,T)\to\infty$, which implies that $\Pr\left(\widehat{\mathcal{Z}}_0\subset\mathcal{Z}_0\right)\to 1$ and together with (14), we have

$$\Pr\left(\widehat{\mathcal{Z}}_0=\mathcal{Z}_0,\ \widehat{\mathcal{Z}}_1=\mathcal{Z}_1\right)\to 1 \text{ as } (N,T)\to\infty. \tag{15}$$

Similarly, Theorem 2(iii) implies

$$\Pr\left(\bigcap_{i=1}^{N}\left\{\widehat{\mathcal{V}}_i=\mathcal{V}_i\right\}\right)\to 1 \text{ as } (N,T)\to\infty, \tag{16}$$

where $\widehat{\mathcal{V}}_i=\{l\in[L_{\mathcal{D}}]:\delta_{il}=0\}$, for $i\in[N]$.

Let $\{\widehat{\boldsymbol{\alpha}}_1,\widehat{\boldsymbol{\alpha}}_2,\cdots,\widehat{\boldsymbol{\alpha}}_{\widehat{K}}\}$ be the distinct values of $\left\{\widehat{\boldsymbol{\delta}}_1,\widehat{\boldsymbol{\delta}}_2,\cdots,\widehat{\boldsymbol{\delta}}_N\right\}$. For $k=1,2,\cdots,\widehat{K}$, let $\widehat{\mathcal{G}}_k=\left\{i\in[N]:\widehat{\boldsymbol{\delta}}_i=\widehat{\boldsymbol{\alpha}}_k\right\}$ and $\widehat{\mathcal{V}}_k=\{l\in\mathcal{D}:\widehat{\boldsymbol{\alpha}}_{kl}=0\}$ denote the estimated group membership and the set of selected moment conditions for group k, respectively. We formalized the classification and moment selection consistency in the following corollary.

Corollary 3. *Suppose Assumption 1 - 4 hold, then*

$$\Pr\left(\left\{\widehat{K}=K^0\right\}\right)\to 1,\ \Pr\left(\bigcap_{k=1}^{K^0}\left\{\widehat{\mathcal{G}}_k=\mathcal{G}_{(k)}\right\}\right)\to 1 \textit{ and } \Pr\left(\bigcap_{k=1}^{K^0}\left\{\widehat{\mathcal{V}}_k=\mathcal{V}_{(k)}\right\}\right)\to 1, \quad (17)$$

as $(N,T)\to\infty$, where $\{(1),(2),\cdots,(K^0)\}$ is a suitable permutation of $[K^0]$.

Corollary 3 can be directly implied from Theorem 2 and Remark 14. Note that $\widehat{\mathcal{Z}}_0\subset\mathcal{Z}_0$ implies $K^0\leq\widehat{K}$ and $\mathcal{Z}_0\subset\widehat{\mathcal{Z}}_0$ implies $\widehat{K}\leq K^0$. By (15), we have the consistency of the estimated number of groups. With $\widehat{K}=K^0$, $\widehat{\mathcal{Z}}_0\subset\mathcal{Z}_0$ implies $\widehat{\mathcal{G}}_k\subset\mathcal{G}_{(k)}$, for $k\in[K^0]$ and $\{(1),(2),\cdots,(K^0)\}$ is a permutation of $[K^0]$; conversely $\mathcal{Z}_0\subset\widehat{\mathcal{Z}}_0$ implies $\mathcal{G}_{(k)}\subset\widehat{\mathcal{G}}_k$, for $k\in[K^0]$, which implies classification consistency. Together with (16), we have the moment selection consistency.

3.3 Asymptotic Distribution

Under Theorem 2 and Corollary 3, $\widehat{\boldsymbol{\alpha}}_{k,\mathcal{V}_k}=\mathbf{0}$ w.p.a.1 for $k\in[K^0]$. It remains to develop the asymptotic distribution for $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}_{k,\mathcal{I}_k}$. Denote $\boldsymbol{\beta}_k^0=\left(\boldsymbol{\theta}^{0\prime},\boldsymbol{\alpha}_{k,\mathcal{I}_k}^{0}{}'\right)'$. For $\boldsymbol{\alpha}_{k,\widehat{\mathcal{I}}_k}\in\mathbb{R}^{|\widehat{\mathcal{I}}_k|}$, denote $\widetilde{\boldsymbol{\alpha}}_k\in\mathbb{R}^{L_\mathcal{D}}$ with $\widetilde{\boldsymbol{\alpha}}_{k,\mathcal{I}_k}=\boldsymbol{\alpha}_{k,\mathcal{I}_k}$ and $\widehat{\boldsymbol{\alpha}}_{k,\mathcal{V}_k}=\mathbf{0}$. Let

$$\widetilde{g}_{i,T}^{(k)}\left(\boldsymbol{\theta},\boldsymbol{\alpha}_{k,\widehat{\mathcal{I}}_k}\right)=\frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix}g_\mathcal{S}\left(\boldsymbol{z}_{it},\boldsymbol{\theta}\right)\\ g_\mathcal{D}\left(\boldsymbol{z}_{it},\boldsymbol{\theta}\right)-\widetilde{\boldsymbol{\alpha}}_k\end{bmatrix}.$$

We define the post-regularization estimator $\widehat{\boldsymbol{\beta}}_k^{\text{post}}$, $k\in\left[\widehat{K}\right]$, as

$$\left(\widehat{\boldsymbol{\theta}}^{\text{post}\prime},\widehat{\boldsymbol{\alpha}}_{k,\widehat{\mathcal{I}}_k}^{\text{post}\prime}\right)'=\underset{\boldsymbol{\theta}\in\Theta,\boldsymbol{\alpha}_{k,\widehat{\mathcal{I}}_k}\in\Theta_\delta^{|\widehat{\mathcal{I}}_k|}}{\arg\min}\left(\frac{1}{\widehat{N}_k}\sum_{i\in\widehat{\mathcal{G}}_k}\widetilde{g}_{i,T}^{(k)}\left(\boldsymbol{\theta},\boldsymbol{\alpha}_{k,\widehat{\mathcal{I}}_k}\right)\right)'\boldsymbol{W}_{k,NT}\left(\frac{1}{\widehat{N}_k}\sum_{i\in\widehat{\mathcal{G}}_k}\widetilde{g}_{i,T}^{(k)}\left(\boldsymbol{\theta},\boldsymbol{\alpha}_{k,\widehat{\mathcal{I}}_k}\right)\right), \quad (18)$$

and we let $\widehat{\boldsymbol{\beta}}_k^{\text{post}}=\left(\widehat{\boldsymbol{\theta}}^{\text{post}\prime},\widehat{\boldsymbol{\alpha}}_{k,\widehat{\mathcal{I}}_k}^{\text{post}\prime}\right)'$

Assumption 5. *(i) Let $\widehat{\nu}_{i,T}=\widehat{m}_{i,T}\left(\boldsymbol{\theta}\right)-\overline{m}_{i,T}\left(\boldsymbol{\theta}\right)$. There exists a sequence of constants $\varsigma_T\to 0$ such that*

$$\sup_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\{\boldsymbol{\theta}\in\Theta:\|\boldsymbol{\theta}-\boldsymbol{\theta}^0\|\leq\eta_T\}}\frac{\|\widehat{\nu}_{i,T}(\boldsymbol{\theta}_1)-\widehat{\nu}_{i,T}(\boldsymbol{\theta}_2)\|}{T^{-\frac{1}{2}}+\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|}=\mathcal{O}_p(\varsigma_T), \quad (19)$$

for some sequence $\eta_T\to 0$ with $\eta_T^{-1}\tau_T=o\left(1\right)$, $\forall i\in[N]$.

(ii) For $k \in [K^0]$, define the variance of the moment conditions as

$$\boldsymbol{\Omega}_{i,T} = \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}\left[g\left(\boldsymbol{z}_{it}, \boldsymbol{\theta}^0, \boldsymbol{\delta}_i^0\right) g\left(\boldsymbol{z}_{is}, \boldsymbol{\theta}^0, \boldsymbol{\delta}_i^0\right)'\right] \text{ and } \boldsymbol{\Omega}_k = \lim_{(N,T)\to\infty} N_k^{-1}\sum_{i\in\mathcal{G}_k}\boldsymbol{\Omega}_{i,T},$$

and the Jacobian matrix is

$$\boldsymbol{\Gamma}_k = \lim_{N\to\infty}\sum_{i=1}^{N}\begin{bmatrix}\Gamma_{\mathcal{S}}\left(\boldsymbol{\theta}^0\right) & \mathbf{0}_{L_{\mathcal{S}}\times L_{\mathcal{D}}} \\ \Gamma_{\mathcal{D}}\left(\boldsymbol{\theta}^0\right) & -\Upsilon_k\end{bmatrix}$$

where $\Upsilon = \operatorname{diag}(\boldsymbol{\upsilon})$ in which $\boldsymbol{\upsilon}_{\mathcal{I}_k} = -\boldsymbol{\iota}_{|\mathcal{I}_k|}$ and $\upsilon_{\mathcal{V}_k} = \mathbf{0}_{|\mathcal{V}_k|}$. Assume that $\boldsymbol{\Omega}_k$ and Γ_k exists and

$$c < \sigma_{\min}\left(\boldsymbol{\Omega}_k\right) \le \sigma_{\max}\left(\boldsymbol{\Omega}_k\right) < C,$$
$$c < \sigma_{\min}\left(\boldsymbol{\Gamma}_k'\boldsymbol{\Gamma}_k\right) \le \sigma_{\max}\left(\boldsymbol{\Gamma}_k'\boldsymbol{\Gamma}_k\right) < C,$$

for constant $c, C > 0$, for $k \in K^0$.

(iii) Assumption 2(iv) holds with $\boldsymbol{W}_{i,NT}$ and $\boldsymbol{W}_i$ replaced by $\boldsymbol{W}_{k,NT}$ and $\boldsymbol{W}_k$, respectively, for $k \in [K^0]$.

(iv) For any $\boldsymbol{\gamma} \in \mathbb{R}^{L_{\mathcal{D}}}$ with $\|\gamma\| = 1$, $\boldsymbol{\gamma}'\left(\frac{1}{\sqrt{NT}}\sum_{i\in\mathcal{G}_k}\sum_{t=1}^{T} g\left(\boldsymbol{z}_{it}, \theta^0, \alpha_k^0\right)\right) \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{\gamma}'\boldsymbol{\Omega}_k\boldsymbol{\gamma}\right)$.

(v) $\sqrt{L_{\mathcal{D}}}\tau_T^2 = o\left(T^{-\frac{1}{2}}\right)$ and $\varsigma_T\tau_T = o\left(T^{-\frac{1}{2}}\right)$.

Remark 15. Assumption 5(i) restates the stochastic equicontintuity condition in Cheng and Liao (2015, Assumption 3.5). The rate of convergence ς_T is introduced to accommodate a diverging number of moments $L_{\mathcal{D}}$. If $L_{\mathcal{D}}$ is fixed, we can replace the right-hand side of (19) with $o_p(1)$. Cheng and Liao (2015, Lemma D.2) provides primitive low-level conditions under which (19) holds with $\varsigma_T = \sqrt{L_{\mathcal{D}}/T}$. Assumption 5(ii) and (iii) regulate the covariance of the moment conditions, the Jacobian matrix and the weight matrix. Assumption 5(iv) assumes the Lindeberg-Feller central limit theorem holds for the moment conditions, for which Su et al. (2016, Lemma S1.12) provides verification details based on the same set of Assumptions.

Theorem 4. *Let $\boldsymbol{\Sigma}_k = \left(\boldsymbol{\Gamma}_k'\boldsymbol{W}_k\boldsymbol{\Gamma}_k\right)^{-1}\left(\boldsymbol{\Gamma}_k'\boldsymbol{W}_k\boldsymbol{\Omega}_k\boldsymbol{W}\boldsymbol{\Gamma}_k\right)\left(\boldsymbol{\Gamma}_k'\boldsymbol{W}_k\boldsymbol{\Gamma}_k\right)^{-1}$. For any $\boldsymbol{\gamma} \in \mathbb{R}^{p_\theta+|\mathcal{I}_k|}$ with $\|\boldsymbol{\gamma}\| = 1$,*

$$\sqrt{N_k T}\boldsymbol{\gamma}'\left(\widehat{\boldsymbol{\beta}}_k^{\text{post}} - \boldsymbol{\beta}_{(k)}^0\right) \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{\gamma}'\boldsymbol{\Sigma}_k\boldsymbol{\gamma}\right),$$

as $(N, T) \to \infty$, for all $k \in [K^0]$, where $\{(1), (2), \cdots, (K^0)\}$ is a suitable permutation of $[K^0]$.

Remark 16. The post-regularization estimator has the same asymptotic distribution as the infeasible estimator based on the true group membership and the set of invalid moment conditions being known, i.e. it possesses the oracle property.

4 Monte Carlo Simulation

In this section, we evaluate the performance of the proposed penalized GMM estimator with simulation experiments.

4.1 Simulation Design

DGP 1 (Linear IV Model). The structural equation is given by:

$$y_{it} = \theta x_{it} + \varepsilon_{it}^{(0)},$$

for $i \in [N]$ and $t \in [T]$, where $\varepsilon_{it}^{(0)}$ is the structural error and x_{it} is an endogenous regressor. The reduced-form equation for the endogenous variable is:

$$x_{it} = \boldsymbol{z}_{it}'\boldsymbol{\gamma} + \varepsilon_{it}^{(1)},$$

where $\varepsilon_{it}^{(1)}$ is the reduced-form error. $\boldsymbol{z}_{it} \in \mathbb{R}^L$ is a vector of instruments. Let

$$\begin{pmatrix} \varepsilon_{it}^{(0)} \\ \varepsilon_{it}^{(1)} \end{pmatrix} \sim N\left(\mathbf{0}_{2\times 1}, \begin{pmatrix} 1 & \rho_\varepsilon \\ \rho_\varepsilon & 1 \end{pmatrix}\right),$$

and

$$z_{it,l} = \delta_{il}\varepsilon_{it}^{(0)} + \sqrt{1-\delta_{il}^2}\tilde{z}_{it,l},$$

where $\tilde{\boldsymbol{z}}_{it} \sim N\left(\mathbf{0}_{L\times 1}, \boldsymbol{\Sigma}_z\right)$ and $\boldsymbol{\Sigma}_z = \begin{pmatrix} 1 & \rho_z & \rho_z & \dots & \rho_z \\ \rho_z & 1 & \rho_z & \dots & \rho_z \\ \rho_z & \rho_z & 1 & \dots & \rho_z \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_z & \rho_z & \rho_z & \dots & 1 \end{pmatrix}$. The non-zero correlation between $\varepsilon_{it}^{(0)}$ and $\varepsilon_{it}^{(1)}$, ρ_ε, is the source of endogeneity. $|\rho_z| > 0$ ensures the relevance conditions hold for each instrument z_{il}. There are 3 groups, i.e. $K^0 = 3$. Denote N_k as the number of observations in each group and let $N_1 : N_2 : N_3 = 0.3 : 0.3 : 0.4$. The validity of instruments is characterized by $\boldsymbol{D} = (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \cdots, \boldsymbol{\delta}_N)' \in \mathbb{R}^{N\times L}$, where $\boldsymbol{D} = \boldsymbol{\Lambda}\boldsymbol{A}'$, where $\lambda_{ik} = 1$ if $i \in \mathcal{G}_k$ and $\lambda_{ik} = 0$ otherwise, for $i \in [N]$ and $k \in [K^0]$. $\boldsymbol{A} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3)$, where $\alpha_{k1} = 0$ for all

$k = 1, 2, 3$, $\alpha_{12} = \alpha_{13} = \alpha_{23} = \alpha_{24} = \alpha_{34} = \alpha_{45} = 0$, and we generate $\alpha_{kl} \sim \text{unif}\,(0.5, 0.9)$ elsewhere and fix it across replications. This means that the sure set $\mathcal{S} = \{1\}$ and $\mathcal{V}_1 = \{2, 3\}$, $\mathcal{V}_2 = \{3, 4\}$ and $\mathcal{V}_3 = \{4, 5\}$. In the experiment, $\theta = 1$, $\boldsymbol{\gamma} = \left(\frac{1}{\sqrt{T}}, 1, 1, \cdots, 1\right)$, $\rho_\varepsilon = 0.5$, $\rho_z = 0.5$ and we fix $L = 8$. The $\sqrt{T}$-rate in $\boldsymbol{\gamma}_1$ weakens the identification strength of the sure set of instruments.

DGP 2 (Dynamic Panel). Consider the following dynamic panel data model:

$$y_{it} = \mu_i + \theta y_{i,t-1} + \varepsilon_{it},$$

for $i \in [N]$ and $t \in [T]$, where μ_i is the individual fixed effect, $y_{i,t-1}$ is the lagged dependent variable, and ε_{it} is the error term. Let $\mu_i \sim \text{unif}\,(-1, 1)$ and be fixed across replications. Similar to DGP 1, we have three groups with $N_1 : N_2 : N_3 = 0.3 : 0.3 : 0.4$. For $i \in \mathcal{G}_1$, the error term $\varepsilon_{it} \sim \text{i.i.d.}N\left(0, \sigma_\varepsilon^2\right)$; for $i \in \mathcal{G}_2$, the error term ε_{it} follows a moving average process of order 1 (MA(1)), i.e., $\varepsilon_{it} = u_{it} + \rho u_{i,t-1}$, where $u_{it} \sim \text{i.i.d.}N\,(0, 1)$; for $i \in \mathcal{G}_3$, the error term ε_{it} follows an MA(2) process, i.e., $\varepsilon_{it} = v_{it} + \gamma_1 v_{i,t-1} + \gamma_2 v_{i,t-2}$, where $v_{it} \sim \text{i.i.d.}N\,(0, 1)$. Note that groups 2 and 3 have serially correlated errors.

Consider the moment conditions corresponding to the Arellano and Bond (1991) estimator. We construct the $\mathcal{S}$ set containing the moment condition that uses the lagged dependent variable $z_{\mathcal{S},it} = y_{i,t-5}$ as an instrument:

$$\mathbb{E}\left(z_{\mathcal{S},it}\left(\Delta y_{it} - \theta \Delta y_{i,t-1}\right)\right) = 0,$$

and the doubtful set consists of the moment conditions that use $\boldsymbol{z}_{\mathcal{D},it} = \left(y_{i,t-2}, y_{i,t-3}, y_{t-4}\right)'$ as instruments,

$$\mathbb{E}\left(\boldsymbol{z}_{\mathcal{D},it}\left(\Delta y_{it} - \theta \Delta y_{i,t-1}\right)\right) = \boldsymbol{\delta}_i,$$

where

$$\boldsymbol{\delta}_i = \begin{cases} \mathbf{0}_{3\times 1}, & i \in \mathcal{G}_1 \\ \left(-\rho, 0, 0\right)', & i \in \mathcal{G}_2 \\ \left(\left(1-\theta\right)\gamma_2 - \gamma_1 - \gamma_1\gamma_2, -\gamma_2, 0\right)', & i \in \mathcal{G}_3 \end{cases}.$$

In the experiment, we set $\theta = 0.5$, $\rho = 0.7$, $\boldsymbol{\theta} = (0.8, 0.6)'$.

4.2 Implementation

In practice, we need to determine κ_1 and κ_f and choose the tuning parameters ψ_1 and ψ_f. Throughout the simulation and application sections, we fix $\kappa_1 = 2$ and $\kappa_f = 3$. Cheng and Liao (2015) provide guidance on choosing the tuning parameters based on the asymptotic regime for moment selection. Following existing practice and the discussion in Remark 8, the tuning parameters are set as $\psi_1 = c_1 \operatorname{Var}(\boldsymbol{y}) T^{-\frac{6}{5}}$ and $\psi_f = c_f \operatorname{Var}(\boldsymbol{y}) T^{-\frac{3}{2}}$ based on $\frac{\log L}{\log T} \approx \frac{1}{3}$ according to the simulation design for the sake of computation. In Section 5, we use the same guidance to choose ψ_1 and choose ψ_f based on the information criterion

$$IC(\psi_f) = \widehat{Q}_{NT}\left(\widehat{\theta}^{\psi_f}, \widehat{\boldsymbol{D}}^{\psi_f}\right) + \varphi_{NT}\left(p_\theta + \widehat{K}^{\psi_f} L_{\mathcal{D}}\right), \tag{20}$$

where $\varphi_{NT} = c_\varphi \frac{\log NT}{NT}$ with a constant c_φ adapted from Mehrabani (2023); Cheng et al. (2023).

For a linear IV model, the optimization problem (9) is nicely convex. Leveraging the modeling techniques summarized in Gao and Shi (2021), (9) can be formulated as a canonical conic programming (QCQP) problem, which is detailed in Appendix A.2. Modern convex optimization solvers, such as `MOSEK` (MOSEK ApS, 2024), can efficiently handle the heavy lifting of solving the optimization problem. For the linear IV case, we can also extend the alternating direction method of multipliers (ADMM) algorithm to the multi-block case, given that we have one more penalty term. It is easy to verify the sufficient condition for the convergence of multi-block ADMM provided in Chen et al. (2016). If (9) is nonconvex, we can use the Gauss-Newton algorithm, which has been justified for nonconvex GMM problems by Forneron and Zhong (2023).

As noted by Mehrabani (2023) and Wang et al. (2018), when T is small or the groups are not sufficiently distinct, the PAFL penalty may create trivial groups with only a few individuals. To address this, we follow the approach of Park et al. (2007), Wang et al. (2018), and Mehrabani (2023) by setting a minimum group proportion of 5% and performing post-estimation hierarchical clustering to eliminate these trivial groups. Additionally, the criterion of classifying i and j into the same group if $\widehat{\boldsymbol{\delta}}_i = \widehat{\boldsymbol{\delta}}_j$ is too strict in practice and often results in many trivial groups. Therefore, we classify i and j into the same group if $\left\|\widehat{\boldsymbol{\delta}}_i - \widehat{\boldsymbol{\delta}}_j\right\| \leq \epsilon_{\text{tol}}$ for a small tolerance ϵ_{tol}.

4.3 Results

In the experiment, we consider sample sizes $N = 50, 100, 200$ and $T = 25, 50, 100$ and run 1,000 replications for each setup. Table 1 reports the bias and root mean squared error (RMSE) of the estimates for the structural parameter of interest θ, as well as the finite sample performance of classification and moment selection of the penalized GMM estimator. This performance is measured in terms of the percentage of correct classification, the percentage of invalid IV, and the percentage of correct number of groups. As the sample size T increases, the precision of the estimator improves and the bias becomes negligible, which is consistent with the asymptotic properties. In addition, the penalized GMM estimator achieves a high correct classification ratio and moment selection accuracy, which supports the classification and moment selection consistency as proven in Theorem 2.

Table 1: Finite Sample Performance of the Penalized GMM Estimator

				Classification	Moment Selection	
N	T	Bias	RMSE	% $\widehat{k(i)} = k(i)$	% Invalid IV	$\%\widehat{K} = K^0$
		DGP 1 Linear IV Model ($\theta = 1$)				
50	25	0.023	0.193	79.4	6.8	97.5
50	50	0.005	0.065	94.8	1.2	99.2
50	100	0.003	0.027	99.2	0.6	100
100	25	0.016	0.154	78.2	4.2	98.2
100	50	0.002	0.063	94.4	0.9	100
100	100	0.003	0.022	99.6	0.2	100
200	25	0.008	0.152	77.3	5.3	95.7
200	50	0.001	0.056	95.1	1.4	99.6
200	100	0.002	0.015	99.3	0.8	100
		DGP 2 Dynamic Panel Model ($\theta = 0.5$)				
50	25	-0.019	0.097	82.4	5.2	97.6
50	50	-0.007	0.037	98.8	1.6	99.2
50	100	0.002	0.023	99.5	0.3	100.0
100	25	-0.011	0.086	91.6	4.7	96.3
100	50	-0.004	0.033	99.1	1.1	99.7
100	100	-0.001	0.019	99.8	0.1	100
200	25	-0.01	0.065	90.3	4.2	96.5
200	50	0.003	0.024	98.9	1.2	99.4
200	100	0.001	0.012	99.7	0.0	100

Notes: Generically, bias and RMSE are calculated by $R^{-1}\sum_{r=1}^{R}\left(\hat{\theta}^{(r)} - \theta_0\right)$ and $\sqrt{R^{-1}\sum_{r=1}^{R}\left(\hat{\theta}^{(r)} - \theta_0\right)^2}$, respectively, for true parameter θ_0 and its estimate $\hat{\theta}^{(r)}$, across $R = 1,000$ replications. We relabel the estimated group with $\widehat{\boldsymbol{\alpha}}_k$ to the closest $\boldsymbol{\alpha}_k^0$ and compute the percentage of correct classification by $R^{-1}\sum_{r=1}^{R} N^{-1}\sum_{i=1}^{N} \mathbb{1}\left(i \in \widehat{\mathcal{G}}_{k(i)}\right)$. The percentage of invalid IV is calculated by $R^{-1}\sum_{r=1}^{R}\sum_{k=1}^{K} |\mathcal{I}_k|^{-1} \sum_{l \in \mathcal{I}_k} \mathbb{1}\left(\widehat{\alpha}_{kl} \neq 0\right)$. The percentage of correct number of group is calculated by $R^{-1}\sum_{r=1}^{R} \mathbb{1}\left(\widehat{K} = K^0\right)$.

5 Empirical Application

China has experienced significant rural-to-urban migration over the past few decades, driven by rapid economic growth and urbanization. Agricultural productivity plays a crucial role in this migration process. This study aims to investigate the impact of agricultural productivity shocks, arising from extreme weather and other factors, on rural-to-urban migration in China. Understanding the relationship is critically important for rural households, as encouraging migration out of less productive rural areas can yield substantial productivity improvements and thus produce large welfare gains, as quantified in the recent literature by Bryan and Morten (2019) and Lagakos, Mobarak, and Waugh (2023).

Specifically, we examine the following linear model,

$$\texttt{migr}_{it} = \mu_i + \theta \texttt{prod}_{it} + \beta x_{it} + u_{it}, \tag{21}$$

where α_i represents the village fixed effect, θ is the parameter of interest that denotes the effect of agricultural productivity, $\texttt{prod}_{it}$, on rural-to-urban migration rate, $\texttt{migr}_{it}$, and x_{it} includes control variables such as average household income (log transformed), $\texttt{income}_{it}$, and the average household head's education level, $\texttt{edu}_{it}$, within the village i at year t.

To address the potential endogeneity issues, we use weather conditions and agricultural technology adoption as instrumental variables. The penalized GMM estimator is applied to estimate the effects and examine the validity of available instruments.

5.1 Empirical Strategy and Data

We utilize a unique dataset, the National Fixed-point Survey, maintained by the Research Center of Rural Economy (RCRE) at the Ministry of Agriculture of China, which annually surveys 27,385 rural households in 293 counties across China and tracks agricultural production and rural household migration behavior over more than two decades, from 1990 to 2013. It asks the sample households to keep a diary on their income, expenditure, as well as detailed information on their economic activities including the agricultural inputs and outputs at the household level. This dataset provides a comprehensive longitudinal view of rural China.

The survey was intended to track the same households through time when it was first conceptualized, but severe attrition is inevitable for such long panel survey covering the most rapid changing periods in China (Zhang et al., 2014). In this study, we aggregate the household-level data to the village level, focusing on the rice cultivation area, which results in a sample of $N = 79$ villages across $T = 19$ years (1995 - 2013). The data are further merged with granular rainfall and temperature records from the weather stations closest to the surveyed villages. These rainfall and temperature data were obtained from the China Meteorological Data Service Center, which is an affiliate of the National Meteorological Information Center of China. The data include daily records of maximum, minimum, and average temperatures, as well as precipitation, from 2,423 weather stations across China.

The summary statistics are reported in Table 2.

The outcome variable $\texttt{migr}_{it}$, rural-to-urban migration rate, is measured by the ratio of the total number of rural-to-urban migrant workers over the size of the labor force in the village i and year t. This variable captures temporal migration as the migrant workers keep their Hukou (household registration system) registration in the original county. We calculate the agricultural productivity, $\texttt{prod}_{it}$, as the rice output divided by the total acreage of farmland for each village-year observation. The cultivation of crops exhibits strong regional characteristics. For instance, wheat is primarily grown in the northern regions, while rice is predominantly cultivated in the southeastern areas. In this study, we focus on the villages that adopt rice as the main crop to avoid the comparison among different crops.

Agricultural productivity is potentially endogenous in (21) due to omitted variable bias. Several factors may not be fully accounted for even after controlling for the specified control variables. These factors include public administration, demographic characteristics, transportation infrastructure, price of input factors, and cultural aspects (White and Lindstrom, 2005). These can be broadly categorized into population, capital, and technology. For instance, machinery adoption in agriculture can lead to automation and reduced labor demand, which may not be captured adequately in the model.

To identify the causal effect, we employ two categories of productivity shocks as instrumental variables. The first category corresponds to extreme weather conditions including temperature extremes and precipitation extremes, as in Minale (2018).

We use the cold temperature variable, $\texttt{cold}_{it}$, defined as the number of days with temperatures below the lower bound of the optimal growth range for rice based on (Huang and Zhang, 2024), to form the sure set of instruments. The choice is consistent with a substantial body of literature examining the consequences of climate change (e.g., Colmer, 2021; Liu et al., 2023; Cui and Zhong, 2024; Cui and Tang, 2024; Wang et al., 2024) where temperature is considered exogenous.

We use the Standardized Precipitation Index (SPI), denoted by $\texttt{SPI}_{it}$, to measure precipitation extremes (Branco and Féres, 2021). An advantage of SPI relative to other indices is that it can be compared across regions with markedly different climates. This is a crucial consideration given the substantial climatic variations among the counties in our study. The SPI has been successfully utilized by many researchers to study extreme climate events, including droughts in Northeast China (Yu and Ma, 2022) as well as other regions globally (Angelidis et al., 2012; Alsumaiei, 2020). The SPI is calculated using a long-term precipitation record to measure the gamma distribution function, which is then transformed into a normal distribution. SPI values are expressed in standard deviations, with negative values below zero signifying precipitation levels lower than the median, indicating drought conditions, and positive values above zero denoting precipitation exceeding the median, characterizing wet conditions (Edwards et al., 1997). As discussed by Mellon (2021), the exclusion condition of precipitation can be violated in various scenarios. Rainfall impacts not only agricultural productivity but also local infrastructure, environmental, and health-related variables, creating pathways for endogeneity. Specifically, Kleemans and Magruder (2018) and Pugatch and Yang (2011) provide evidence that people may migrate in response to extreme rainfall. Therefore, we classify $\texttt{SPI}_{it}$ as part of the doubtful set.

The second category corresponds to agricultural technology adoption, which is measured by the quantity of *pesticides* and *fertilizers* applied per acre of farmland. The variables are labeled as $\texttt{pest}_{it}$ and $\texttt{fert}_{it}$. While technological advancements are often used as instruments for production activities due to the exogeneity of their accessibility (e.g., Foster and Rosenzweig, 1995; Bustos, Caprettini, and Ponticelli, 2016; Hjort and Poulsen, 2019; Gollin, Hansen, and Wingender, 2021; Hjort and Poulsen, 2019), productivity shocks resulting from technology adoption may be endogenous due to factors such as self-selection and unobserved local heterogeneity. For instance, farmers who invest more in agricultural technology may systematically differ from those who do not, often in ways that are correlated with productivity, such as risk tolerance, access to credit, or education level (Foster and Rosenzweig, 2010; Bryan, Chowdhury, and Mobarak, 2014; Meriggi, Bulte, and Mobarak, 2021). Together with $\texttt{SPI}_{it}$, $\texttt{pest}_{it}$, and $\texttt{fert}_{it}$ form the doubtful set. This approach allows us to leverage the exogeneity of extreme temperature events while accounting for potential endogeneity in precipitation extremes and agricultural technology adoption measures.

Table 2: Summary Statistics

Variable	Mean	S.D.	Median
migr	0.23	0.18	0.21
prod	447.65	118.00	446.36
cold	198.57	39.60	197.00
SPI	-0.09	1.04	-0.14
fert	4.82	0.75	4.81
pest	1.15	0.76	1.00
income	9.97	0.73	9.85
edu	6.73	1.11	6.72

Notes: Variable labels are defined in Section 5.1. The sample size is $N = 79$ villages and $T = 19$ years. "S.D." stands for the sample standard deviation.

5.2 Results

We estimate the model (21) using three different estimation methods: 2SLS with only the sure set of instruments, 2SLS with both the sure and doubtful set of instruments, and penalized GMM with both the sure and doubtful set of instruments. The results presented in Table 3 show the estimated effects associated with the standard errors. Firstly, including instruments `SPI` and per acre usage of pesticides and fertilizers alters the results of the 2SLS estimation with only the sure set of instruments, the extreme temperature variable $\texttt{cold}_{it}$. It delivers a positive effect of agricultural productivity on rural-to-urban migration that is statistically significant at the 10% confidence level. However, as we discussed in 5.1, and also based on the moment selection from the penalized GMM estimator, the doubtful set of instruments are potentially invalid, so the results can be misleading.

The estimates from the penalized GMM estimator are qualitatively consistent with the 2SLS with only the sure set of instruments, while the standard errors is substantially smaller, which demonstrates the efficiency gain from including overidentifying moment conditions.

Table 3: Estimated Effects of Agricultural Productivity on Rural-to-Urban Migration

	2SLS ($\boldsymbol{z}_{\mathcal{S}}$)		2SLS ($\boldsymbol{z}_{\mathcal{S}}$ and $\boldsymbol{z}_{\mathcal{D}}$)		P-GMM	
	Est.	S.E.	Est.	S.E.	Est.	S.E.
`prod` ($\theta \times 100$)	-0.0062	0.0192	0.0057	0.0035	0.0021	0.0044
`income` (β_1)	0.1355	0.0285	0.1466	0.0081	0.1735	0.0090
`edu` (β_2)	-0.0912	0.0286	-0.0804	0.0087	-0.0863	0.0089

Notes: The left panel corresponds to the 2SLS estimation with only the sure set of instruments, middle panel corresponds to the 2SLS estimation with both the sure and doubtful set of instruments, and the right panel corresponds to the penalized GMM estimation with both the sure and doubtful set of instruments. "S.E." stands for the standard error for the estimates.

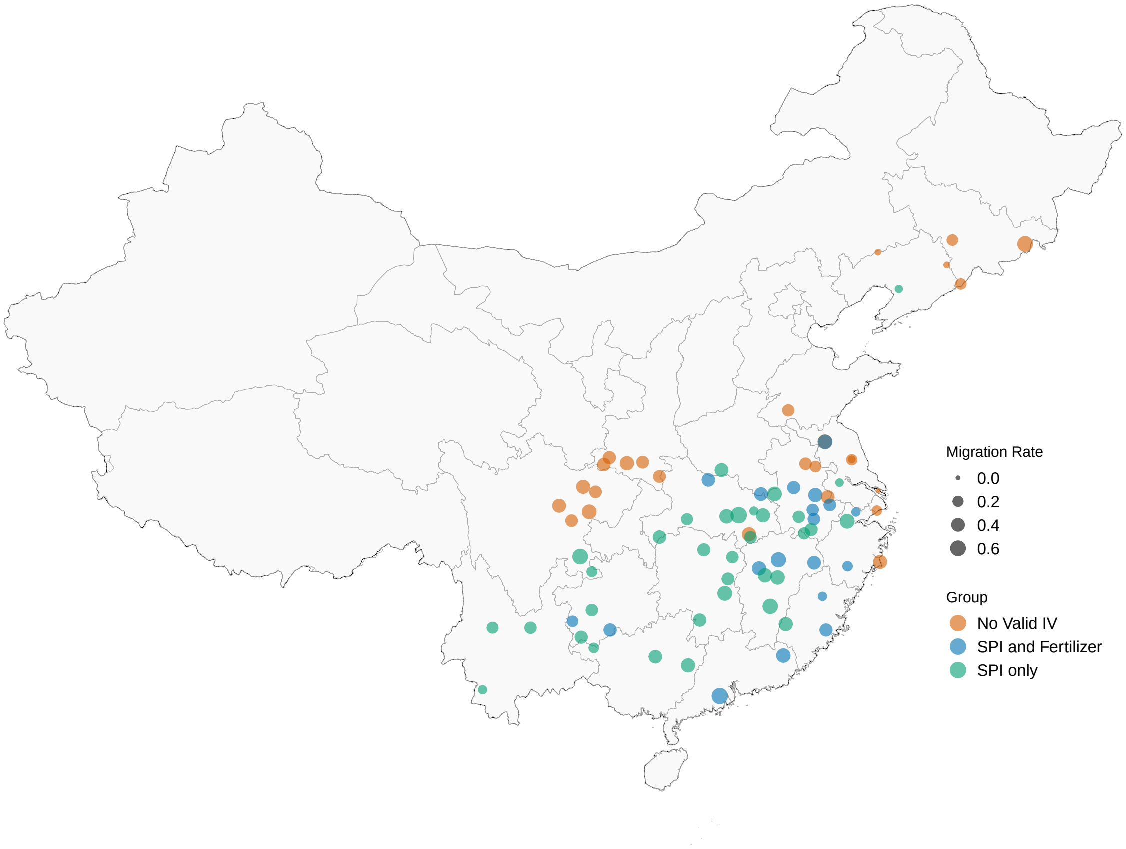
Table 4 provides insights into the validity of the moment conditions by groups. The table reports the group structure detected based on the estimated slackness parameters. According to the results, we can identify at least one more valid moment conditions for 67% of the villages, which suggests imperfect instruments can still be informative when we properly account for the heterogeneity in validity. In Figure 1, we visualize the group membership of villages detected by the penalized GMM estimator. The groups demonstrate geographic clustering while no clear pattern is observed between the outcome variable migration rate and the group structure.

Table 4: Estimated Moment Condition Validity by Groups

Slackness Parameter $\boldsymbol{\alpha}_k$	Group (%)	$\texttt{SPI}_{it}$ ($\boldsymbol{\alpha}_{k,1}$)	$\texttt{fert}_{it}$ ($\boldsymbol{\alpha}_{k,2}$)	$\texttt{pest}_{it}$ ($\boldsymbol{\alpha}_{k,3}$)
Group 1 (No Valid IV)	32.9%	0.0066	0.0059	0.0054
Group 2 ($\texttt{SPI}_{it}$ and $\texttt{fert}_{it}$)	24.1%	0	0	0.0043
Group 3 ($\texttt{SPI}_{it}$ Only)	43.0%	0	0.0125	0.0155

Notes: The reports the group structure detected based on the estimated slackness parameters. The second column reports the percentage of the villages in each group. The third, fourth, and fifth columns report the estimates of the slackness parameters corresponding to the instruments in the doubtful set, respectively.

Figure 1: Group Memebership of Villages Detected by the Penalized GMM Estimator



Notes: The villages are located in the map based on their latitude and longitude coordinates. The color of the village indicates the group membership detected by the penalized GMM estimator and the size of the scatter is proportional to the rural-to-urban migration rate, $\texttt{migr}_{i,2000}$ at year 2000.

6 Conclusion

In this paper, we provide a unified framework for the selection of valid moment conditions and detection of latent group structure based on the moment condition validity in general nonlinear generalized method of moments (GMM) panel data models, which can accommodate a diverging number of moment conditions and group-specific heterogeneous validity of moment conditions across agents. The proposed penalized GMM estimator is shown to be consistent and achieve classification and moment selection consistency simultaneously. The asymptotic distribution of a post-regularization estimator is derived, and its oracle properties are established. The finite-sample performance of the proposed method is evaluated through a Monte Carlo simulation experiment. The empirical application demonstrates the effectiveness of the proposed method in identifying the impact of agricultural productivity on rural-to-urban migration in China.

References

Adao, R., M. Kolesár, and E. Morales (2019). Shift-share designs: Theory and inference. *The Quarterly Journal of Economics 134*(4), 1949–2010.

Ahn, S. C. and P. Schmidt (1995). Efficient estimation of models for dynamic panel data. *Journal of Econometrics 68*(1), 5–27.

Ahn, S. C. and P. Schmidt (1997). Efficient estimation of dynamic panel data models: Alternative assumptions and simplified estimation. *Journal of Econometrics 76*(1-2), 309–321.

Alsumaiei, A. A. (2020). Monitoring hydrometeorological droughts using a simplified precipitation index. *Climate 8*(2), 19.

Altonji, J. G. (1986). Intertemporal substitution in labor supply: Evidence from micro data. *Journal of Political Economy 94*(3, Part 2), S176–S215.

Ando, T. and N. Sueishi (2019). On the convergence rate of the scad-penalized empirical likelihood estimator. *Econometrics 7*(1), 15.

Andrews, D. W. (1999). Consistent moment selection procedures for generalized method of moments estimation. *Econometrica 67*(3), 543–563.

Andrews, D. W. and B. Lu (2001). Consistent model and moment selection procedures for gmm estimation with application to dynamic panel data models. *Journal of Econometrics 101*(1), 123–164.

Angelidis, P., F. Maris, N. Kotsovinos, and V. Hrissanthou (2012). Computation of drought index spi with alternative distribution functions. *Water resources management 26*, 2453–2473.

Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The Review of Economic Studies 58*(2), 277–297.

Arellano, M. and O. Bover (1995). Another look at the instrumental variable estimation of error-components models. *Journal of econometrics 68*(1), 29–51.

Armstrong, T. B. and M. Kolesár (2021). Sensitivity analysis using approximate moment condition models. *Quantitative Economics 12*(1), 77–108.

Armstrong, T. B., M. Weidner, and A. Zeleneev (2023). Robust estimation and inference in panels with interactive fixed effects. *arXiv preprint arXiv:2210.06639*.

Autor, D. H., D. Dorn, and G. H. Hanson (2013). The china syndrome: Local labor market effects of import competition in the united states. *American Economic Review 103*(6), 2121–2168.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Blundell, R. and S. Bond (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics 87*(1), 115–143.

Bonhomme, S., T. Lamadon, and E. Manresa (2019). A distributional framework for matched employer employee data. *Econometrica 87*(3), 699–739.

Bonhomme, S., T. Lamadon, and E. Manresa (2022). Discretizing unobserved heterogeneity. *Econometrica 90*(2), 625–643.

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

Borusyak, K., P. Hull, and X. Jaravel (2022). Quasi-experimental shift-share research designs. *The Review of Economic Studies 89*(1), 181–213.

Branco, D. and J. Féres (2021). Weather shocks and labor allocation: Evidence from rural brazil. *American Journal of Agricultural Economics 103*(4), 1359–1377.

Bryan, G., S. Chowdhury, and A. M. Mobarak (2014). Underinvestment in a profitable technology: The case of seasonal migration in bangladesh. *Econometrica 82*(5), 1671–1748.

Bryan, G. and M. Morten (2019). The aggregate productivity effects of internal migration: Evidence from indonesia. *Journal of Political Economy 127*(5), 2229–2268.

Bustos, P., B. Caprettini, and J. Ponticelli (2016). Agricultural productivity and structural transformation: Evidence from brazil. *American Economic Review 106*(6), 1320–1365.

Chang, J., Z. Shi, and J. Zhang (2022). Culling the herd of moments with penalized empirical likelihood. *Journal of Business & Economic Statistics*, 1–15.

Chang, J., C. Y. Tang, and T. T. Wu (2018). A new scope of penalized empirical likelihood with high-dimensional estimating equations. *The Annals of Statistics 46*(6B), 3185–3216.

Chen, C., B. He, Y. Ye, and X. Yuan (2016). The direct extension of admm for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming 155*(1), 57–79.

Cheng, X. and Z. Liao (2015). Select the valid and relevant moments: An information-based lasso for gmm with many moments. *Journal of Econometrics 186*(2), 443–464.

Cheng, X., F. Schorfheide, and P. Shao (2023). Clustering for multi-dimensional heterogeneity with an application to production function estimation. *Working paper*.

Chetverikov, D. and E. Manresa (2022). Spectral and post-spectral estimators for grouped panel data models. *arXiv preprint arXiv:2212.13324*.

Colmer, J. (2021). Temperature, labor reallocation, and industrial production: Evidence from india. *American Economic Journal: Applied Economics 13*(4), 101–124.

Cui, X. and Q. Tang (2024). Extreme heat and rural household adaptation: Evidence from northeast china. *Journal of Development Economics 167*, 103243.

Cui, X. and Z. Zhong (2024). Climate change, cropland adjustments, and food security: Evidence from china. *Journal of Development Economics 167*, 103245.

Cytrynbaum, M. (2020). Blocked clusterwise regression. *arXiv preprint arXiv:2001.11130*.

Dehling, H. and W. Philipp (2002). Empirical process techniques for dependent data. In *Empirical process techniques for dependent data*, pp. 3–113. Springer.

Edwards, D. C., T. B. McKee, et al. (1997). Characteristics of 20th century drought in the united states at multiple time scales.

Fan, J. and Y. Liao (2014). Endogeneity in high dimensions. *Annals of Statistics 42*(3), 872.

Forneron, J.-J. and L. Zhong (2023). Convexity not required: Estimation of smooth moment condition models. *arXiv preprint arXiv:2304.14386*.

Foster, A. D. and M. R. Rosenzweig (1995). Learning by doing and learning from others: Human capital and technical change in agriculture. *Journal of political Economy 103*(6), 1176–1209.

Foster, A. D. and M. R. Rosenzweig (2007). Economic development and the decline of agricultural employment. *Handbook of development economics 4*, 3051–3083.

Foster, A. D. and M. R. Rosenzweig (2010). Microeconomics of technology adoption. *Annu. Rev. Econ. 2*(1), 395–424.

Gao, Z. and Z. Shi (2021). Implementing convex optimization in R: Two econometric examples. *Computational Economics 58*, 1127–1135.

Gautier, E. and C. Rose (2021). High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454*.

Goldsmith-Pinkham, P., I. Sorkin, and H. Swift (2020). Bartik instruments: What, when, why, and how. *American Economic Review 110*(8), 2586–2624.

Gollin, D., C. W. Hansen, and A. M. Wingender (2021). Two blades of grass: The impact of the green revolution. *Journal of Political Economy 129*(8), 2344–2384.

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

Hahn, J., J. C. Ham, and H. R. Moon (2011). The hausman test and weak instruments. *Journal of Econometrics 160*(2), 289–299.

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

Han, C. and P. C. Phillips (2006). Gmm with many moment conditions. *Econometrica 74*(1), 147–192.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*(4), 1029–1054.

Hjort, J. and J. Poulsen (2019). The arrival of fast internet and employment in africa. *American Economic Review 109*(3), 1032–1079.

Hocking, T. D., A. Joulin, F. Bach, and J.-P. Vert (2011). Clusterpath an algorithm for clustering using convex fusion penalties. In *28th international conference on machine learning*, pp. 1.

Huang, J. (2023). Quasi-bayesian inference for grouped panels.

Huang, K. and P. Zhang (2024). Aggregate temperature measures and overestimation of the impact of global warming on crop yield. *CUHK Working paper*.

Huang, W., S. Jin, P. C. Phillips, and L. Su (2021). Nonstationary panel models with latent group structures and cross-section dependence. *Journal of Econometrics 221*(1), 198–222.

Huang, W., S. Jin, and L. Su (2020). Identifying latent grouped patterns in cointegrated panels. *Econometric Theory 36*(3), 410–456.

Huang, W., Y. Wang, and L. Zhou (2023). Classifylasso: Stata module to identify latent group structures via classifier-lasso.

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

Kleemans, M. and J. Magruder (2018). Labour market responses to immigration: Evidence from internal migration driven by weather shocks. *The Economic Journal 128*(613), 2032–2065.

Lagakos, D., A. M. Mobarak, and M. E. Waugh (2023). The welfare effects of encouraging rural–urban migration. *Econometrica 91*(3), 803–837.

Liang, X., E. Sanderson, and F. Windmeijer (2022). Selecting valid instrumental variables in linear models with multiple exposure variables: adaptive lasso and the median-of-medians estimator. *arXiv preprint arXiv:2208.05278*.

Liao, Z. (2013). Adaptive gmm shrinkage estimation with consistent moment selection. *Econometric Theory 29*(5), 857–904.

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

Lin, Y., F. Windmeijer, X. Song, and Q. Fan (2022). On the instrumental variable estimation with many weak and invalid instruments. *arXiv preprint arXiv:2207.03035*.

Liu, M., Y. Shamdasani, and V. Taraz (2023). Climate change and labor reallocation: Evidence from six decades of the indian census. *American Economic Journal: Economic Policy 15*(2), 395–423.

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

Liu, Y., P. C. Phillips, and J. Yu (2023). A panel clustering approach to analyzing bubble behavior. *International Economic Review 64*(4), 1347–1395.

Lumsdaine, R. L., R. Okui, and W. Wang (2023). Estimation of panel group structure models with structural breaks in group memberships and coefficients. *Journal of Econometrics 233*(1), 45–65.

Luo, Y. (2014). Selecting informative moments via lasso.

Ma, S., L. Su, and Y. Zhang (2022). Detecting latent communities in network formation models. *The Journal of Machine Learning Research 23*(1), 13971–14031.

MaCurdy, T. E. (1981). An empirical model of labor supply in a life-cycle setting. *Journal of Political Economy 89*(6), 1059–1085.

Matsuyama, K. (1992). Agricultural productivity, comparative advantage, and economic growth. *Journal of economic theory 58*(2), 317–334.

McArthur, J. W. and G. C. McCord (2017). Fertilizing growth: Agricultural inputs and their effects in economic development. *Journal of development economics 127*, 133–152.

Mehrabani, A. (2023). Estimation and identification of latent group structures in panel data. *Journal of Econometrics 235*(2), 1464–1482.

Mellon, J. (2021). Rain, rain, go away: 194 potential exclusion-restriction violations for studies using weather as an instrumental variable. *American Journal of Political Science*.

Meriggi, N. F., E. Bulte, and A. M. Mobarak (2021). Subsidies for technology adoption: Experimental evidence from rural cameroon. *Journal of Development Economics 153*, 102710.

Miao, K., L. Su, and W. Wang (2020). Panel threshold regressions with latent group structures. *Journal of Econometrics 214*(2), 451–481.

Minale, L. (2018). Agricultural productivity shocks, labour reallocation and rural–urban migration in china. *Journal of Economic Geography 18*(4), 795–821.

Moon, H. R. and F. Schorfheide (2009). Estimation with overidentifying inequality moment conditions. *Journal of Econometrics 153*(2), 136–154.

MOSEK ApS (2024). *The MOSEK optimization toolbox for Rmosek package manual. Version 10.1.25.*

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics 4*, 2111–2245.

Okui, R. (2011). Instrumental variable estimation in the presence of many moment conditions. *Journal of Econometrics 165*(1), 70–86.

Okui, R. and W. Wang (2021). Heterogeneous structural breaks in panel data models. *Journal of Econometrics 220*(2), 447–473.

Park, M. Y., T. Hastie, and R. Tibshirani (2007). Averaged gene expressions for regression. *Biostatistics 8*(2), 212–227.

Pugatch, T. and D. Yang (2011). The impact of mexican immigration on us labor markets: Evidence from migrant flows driven by rainfall shocks. *University of Michigan, Ann Arbor, US*.

Qian, J. and L. Su (2016). Shrinkage estimation of common breaks in panel data models via adaptive group fused lasso. *Journal of Econometrics 191*(1), 86–109.

Radchenko, P. and G. Mukherjee (2017). Convex clustering via l 1 fusion penalization. *Journal of the Royal Statistical Society Series B: Statistical Methodology 79*(5), 1527–1546.

Shi, Z. (2016). Econometric estimation with high-dimensional moment equalities. *Journal of Econometrics 195*(1), 104–119.

Su, L. and G. Ju (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics 206*(2), 554–573.

Su, L. and X. Lu (2017). Determining the number of groups in latent panel structures with an application to income and democracy. *Quantitative Economics 8*(3), 729–760.

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

Su, L., W. Wang, and X. Xu (2023). Identifying latent group structures in spatial dynamic panels. *Journal of Econometrics 235*(2), 1955–1980.

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

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology 67*(1), 91–108.

Wang, D., P. Zhang, S. Chen, and N. Zhang (2024). Adaptation to temperature extremes in chinese agriculture, 1981 to 2010. *Journal of Development Economics 166*, 103196.

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

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

Wang, Y., P. C. Phillips, and L. Su (2023). Panel data models with time-varying latent group structures. *arXiv preprint arXiv:2307.15863*.

White, M. J. and D. P. Lindstrom (2005). Internal migration. In *Handbook of population*, pp. 311–346. Springer.

Windmeijer, F., H. Farbmacher, N. Davies, and G. Davey Smith (2019). On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association 114*(527), 1339–1350.

Yu, L., J. Gu, and S. Volgushev (2022). Group structure estimation for panel data–a general approach. *arXiv preprint arXiv:2201.01793*.

Yu, X. and Y. Ma (2022). Spatial and temporal analysis of extreme climate events over northeast china. *Atmosphere 13*(8), 1197.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology 68*(1), 49–67.

Zhang, B. (2023). Incorporating prior knowledge of latent group structure in panel data models. *arXiv preprint arXiv:2211.16714*.

Zhang, J., L. Gan, L. C. Xu, and Y. Yao (2014). Health shocks, village elections, and household income: Evidence from rural china. *China Economic Review 30*, 155–168.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*(476), 1418–1429.

Appendix

A.1 Proofs of Main Results

We first introduce the following notations for simplicity of exposition in the proofs. Denote $\widehat{\nu}_{i,T}\left(\boldsymbol{\theta}\right) \coloneqq \widehat{m}_{i,T}\left(\boldsymbol{\theta}\right) - \overline{m}_{i}\left(\boldsymbol{\theta}\right)$ as in Assumption 5(i) and $\widehat{R}_{i,T}\left(\boldsymbol{\theta}\right) \coloneqq \widehat{\nu}_{i,T}\left(\boldsymbol{\theta}\right)' \boldsymbol{W}_i \widehat{\nu}_{i,T}\left(\boldsymbol{\theta}\right).$ In addition, denote $\overline{Q}_i\left(\boldsymbol{\theta}, \boldsymbol{\delta}\right) = \overline{g}_i\left(\boldsymbol{\theta}, \boldsymbol{\delta}\right) \boldsymbol{W}_i \overline{g}_i\left(\boldsymbol{\theta}, \boldsymbol{\delta}\right)$. For any two sequences a_N and b_N, let $a_N \lesssim b_N$ denote $Ca_N \leq b_N$ where $C > 0$ is some fixed finite constant, and $a_N \gtrsim b_N$ denote $b_N \lesssim a_N$. With abuse of notation, we reuse the notation Ξ to denote algebraic terms for convenience of exposition.

Lemma A.1. *Let* $\widetilde{\varkappa}_{NT} = \sqrt{\frac{L_{\mathcal{D}}}{T}} \left(\log T\right)^3$. $\dot{\boldsymbol{\delta}}_i$ *is the preliminary estimator defined in* (11). $\dot{w}_{il} = \left|\dot{\delta}_{il}\right|^{-\kappa_1}$ *and* $\dot{\mu}_{ij} = \left\|\dot{\boldsymbol{\delta}}_i - \dot{\boldsymbol{\delta}}_j\right\|^{-\kappa_f}$ *are the adaptive weights introduced in Section* 2.3. *Under Assumption* 1 - 3*,*

(i) $\max_{1\leq i\leq N} \left\|\widehat{m}_{i,T}\left(\boldsymbol{\theta}\right) - \overline{m}_i\left(\boldsymbol{\theta}\right)\right\| = \mathcal{O}_p\left(\widetilde{\varkappa}_{NT}\right)$

(ii) $\max_{1\leq i\leq N} \left\|\dot{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i^0\right\| = \mathcal{O}_p\left(\widetilde{\varkappa}_{NT}\right)$

(iii) $\max_{(i,j)\in\mathcal{Z}_1} \dot{\mu}_{ij} = \mathcal{O}_p\left(\rho_{NT}^{-\kappa_f}\right)$.

(iv) $\max_{1\leq i\leq N} \max_{l\in\mathcal{I}_i} \dot{w}_{il} = \mathcal{O}_p\left(\zeta_{NT}^{-\kappa_1}\right)$ *and* $\max_{1\leq i\leq N} \left\|\dot{\boldsymbol{w}}_{i,\mathcal{I}_i}\right\| = \mathcal{O}_p\left(\sqrt{L_{\mathcal{D}}}\zeta_{NT}^{-\kappa_1}\right)$ *where* $\dot{\boldsymbol{w}}_{i,\mathcal{I}_i}$ *is the subvector of* $\dot{\boldsymbol{w}}_i$ *with element* w_{il}*, for* $l \in \mathcal{I}_i = \{l \in [L_{\mathcal{D}}] : \delta_{il}^0 \neq 0\}$.

Proof of Lemma A.1. Part (i) restates Lemma S1.2(i) in Su et al. (2016) with the dimension of the moment function, $L_{\mathcal{D}}$, diverging. By applying the proof arguments element-by-element to $\left|\widehat{m}_{il,T}\left(\boldsymbol{\theta}\right) - \overline{m}_{il}\left(\boldsymbol{\theta}\right)\right|^2$, we have the modified convergence rate $\varkappa_{NT}$ with the new term $\sqrt{L_{\mathcal{D}}}$ showing up compared to the original rate in Su et al. (2016).

Part (ii). Note that we leverage on the identification of $\boldsymbol{\theta}$ with fixed dimensional moment conditions $\overline{m}_{\mathcal{S},i}\left(\boldsymbol{\theta}\right) = 0$ to construct the initial GMM estimator for $\boldsymbol{\theta}^0$, standard asymptotic theory as in Newey and McFadden (1994) yield the $\sqrt{T}$ rate of convergence, i.e. $\left\|\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right\| = \mathcal{O}_p\left(T^{-\frac{1}{2}}\right)$. By mean value theorem,

$$\begin{aligned}
\dot{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i^0 &= \widehat{m}_{\mathcal{D},i,T}\left(\dot{\boldsymbol{\theta}}\right) - \overline{m}_{\mathcal{D},i}\left(\dot{\boldsymbol{\theta}}\right) + \overline{m}_{\mathcal{D},i}\left(\dot{\boldsymbol{\theta}}\right) - \overline{m}_{\mathcal{D},i}\left(\boldsymbol{\theta}^0\right) \\
&= \widehat{m}_{\mathcal{D},i,T}\left(\dot{\boldsymbol{\theta}}\right) - \overline{m}_{\mathcal{D},i}\left(\dot{\boldsymbol{\theta}}\right) + \Gamma_{\mathcal{D},i}\left(\widetilde{\boldsymbol{\theta}}\right)\left(\dot{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right),
\end{aligned}$$

where $\widetilde{\boldsymbol{\delta}}$ is between $\boldsymbol{\theta}^0$ and $\dot{\boldsymbol{\theta}}$. Then

$$\begin{aligned}\max_{1\leq i\leq N}\left\|\dot{\boldsymbol{\delta}}_i-\boldsymbol{\delta}_i^0\right\| &\leq \max_{1\leq i\leq N}\left\|\widehat{m}_{\mathcal{D},i,T}\left(\dot{\boldsymbol{\theta}}\right)-\overline{m}_{\mathcal{D},i}\left(\dot{\boldsymbol{\theta}}\right)\right\|+\left(\max_{1\leq i\leq N}\left\|\Gamma_{\mathcal{D},i}\left(\widetilde{\boldsymbol{\theta}}\right)\right\|\right)\left\|\dot{\boldsymbol{\theta}}-\boldsymbol{\theta}^0\right\| \\ &\leq \mathcal{O}_p\left(\widetilde{\varkappa}_{NT}\right)+\mathcal{O}_p\left(T^{-\frac{1}{2}}\right)=\mathcal{O}_p\left(\widetilde{\varkappa}_{NT}\right),\end{aligned}$$

where the first line follows from triangle inequality and Cauchy-Schwartz inequality, and the second inequality holds under part (i) and Assumption 2(v).

Part (iii). Note that $\max_{(i,j)\in\mathcal{Z}_1}\dot{\mu}_{ij}=\left(\min_{(i,j)\in\mathcal{Z}_1}\left\|\dot{\boldsymbol{\delta}}_i-\dot{\boldsymbol{\delta}}_j\right\|\right)^{-\kappa_f}$. For $(i,j)\in\mathcal{Z}_1$ and sufficiently large (N,T),

$$\min_{(i,j)\in\mathcal{Z}_1}\left\|\dot{\boldsymbol{\delta}}_i-\dot{\boldsymbol{\delta}}_j\right\|\geq\min_{(i,j)\in\mathcal{Z}_1}\left|\left\|\boldsymbol{\delta}_i^0-\boldsymbol{\delta}_j^0\right\|-\left\|\left(\dot{\boldsymbol{\delta}}_i-\boldsymbol{\delta}^0\right)-\left(\dot{\boldsymbol{\delta}}_j-\boldsymbol{\delta}^0\right)\right\|\right|\geq\rho_{NT}-2\max_{1\leq i\leq N}\left\|\dot{\boldsymbol{\delta}}_i-\boldsymbol{\delta}_i^0\right\|. \tag{A.1}$$

by triangle inequality. By the uniform convergence in part (ii) and the rate condition in Assumption 3(i), $\max_{(i,j)\in\mathcal{Z}_1}\dot{\mu}_{ij}=O_p\left(\rho_{NT}^{-\kappa_f}\right)$.

Part (iv). By parallel arguments as in part (iii), we have $\max_{1\leq i\leq N}\max_{l\in\mathcal{I}_i}\dot{w}_{il}=\mathcal{O}_p\left(\zeta_{NT}^{-\kappa_1}\right)$,

$$\max_{1\leq i\leq N}\left\|\dot{\boldsymbol{w}}_{i,\mathcal{I}_i}\right\|=\mathcal{O}_p\left(\sqrt{L_{\mathcal{D}}}\zeta_{NT}^{-\kappa_1}\right)$$

holds by noting that $\left\|\dot{\boldsymbol{w}}_{i,\mathcal{I}_i}\right\|\leq\sqrt{L_{\mathcal{D}}}\max_{l\in\mathcal{I}_i}\dot{w}_{il}$.

□

Lemma A.2. *Under Assumption 1 - 3,* $\widehat{\boldsymbol{\theta}}\xrightarrow{p}\boldsymbol{\theta}^0$ *as* $(N,T)\to\infty$.

Proof of Lemma A.2. By the optimality in (9),

$$\begin{aligned}\widehat{Q}_{NT}\left(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{D}}\right)\leq\widehat{Q}_{NT}\left(\boldsymbol{\theta}^0,\boldsymbol{D}^0\right)&+\frac{\psi_f}{N^2}\sum_{1\leq i<j\leq N}\dot{\mu}_{ij}\left(\left\|\boldsymbol{\delta}_i^0-\boldsymbol{\delta}_j^0\right\|-\left\|\widehat{\boldsymbol{\delta}}_i-\widehat{\boldsymbol{\delta}}_j\right\|\right) \\ &+\frac{\psi_1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L_{\mathcal{D}}}\dot{w}_{il}\left(\left|\delta_{il}^0\right|-\left|\widehat{\delta}_{il}\right|\right)\end{aligned} \tag{A.2}$$

Note that, $\widehat{g}_{i,T}\left(\boldsymbol{\theta}^0,\boldsymbol{\delta}_i^0\right)=\widehat{\nu}_{i,T}\left(\boldsymbol{\theta}^0\right)=\widehat{m}_{i,T}\left(\boldsymbol{\theta}^0\right)-\overline{m}_i\left(\boldsymbol{\theta}^0\right)$, then by Assumption 2(iv) and Lemma A.1(i),

$$\begin{aligned}\widehat{Q}_{NT}\left(\boldsymbol{\theta}^0,\boldsymbol{D}^0\right)=&\frac{1}{N}\sum_{i=1}^{N}\widehat{\nu}_{i,T}\left(\boldsymbol{\theta}^0\right)'\boldsymbol{W}_i\widehat{\nu}_{i,T}\left(\boldsymbol{\theta}^0\right)+\widehat{\nu}_{i,T}\left(\boldsymbol{\theta}^0\right)'\left(\boldsymbol{W}_{i,NT}-\boldsymbol{W}_i\right)\widehat{\nu}_{i,T}\left(\boldsymbol{\theta}^0\right) \\ \leq&\frac{1}{N}\sum_{i=1}^{N}\left(\sigma_{\max}\left(\boldsymbol{W}_i\right)+\left\|\boldsymbol{W}_{i,NT}-\boldsymbol{W}_i\right\|\right)\left\|\widehat{\nu}_{i,T}\left(\boldsymbol{\theta}^0\right)\right\|^2\end{aligned}$$

$$
\begin{aligned}
&\le \left(C_w + \max_{1\le i\le N} \left\|\boldsymbol{W}_{i,NT} - \boldsymbol{W}_i\right\|\right)\left(\max_{1\le i\le N}\left\|\widehat{\nu}_{i,T}\left(\boldsymbol{\theta}^0\right)\right\|\right)^2 \\
&= \left(C_w + o_p\left(1\right)\right) o_p\left(1\right) = o_p\left(1\right). \qquad \text{(A.3)}
\end{aligned}
$$

The second term on R.H.S. of (A.2) due to the fused Lasso penalty is bounded by

$$
\begin{aligned}
&\frac{\psi_f}{N^2}\sum_{1\le i<j\le N}\dot{\mu}_{ij}\left(\left\|\boldsymbol{\delta}_i^0 - \boldsymbol{\delta}_j^0\right\| - \left\|\widehat{\boldsymbol{\delta}}_i - \widehat{\boldsymbol{\delta}}_j\right\|\right) \\
=&\frac{\psi_f}{N^2}\left[-\sum_{(i,j)\in\mathcal{Z}_0}\dot{\mu}_{ij}\left\|\widehat{\boldsymbol{\delta}}_i - \widehat{\boldsymbol{\delta}}_j\right\| + \sum_{(i,j)\in\mathcal{Z}_1}\dot{\mu}_{ij}\left(\left\|\boldsymbol{\delta}_i^0 - \boldsymbol{\delta}_j^0\right\| - \left\|\widehat{\boldsymbol{\delta}}_i - \widehat{\boldsymbol{\delta}}_j\right\|\right)\right] \\
\le&\frac{\psi_f}{N^2}\sum_{(i,j)\in\mathcal{Z}_1}\dot{\mu}_{ij}\left(\left\|\widehat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i^0\right\| + \left\|\widehat{\boldsymbol{\delta}}_j - \boldsymbol{\delta}_j^0\right\|\right) \\
\le&\psi_f\left(\max_{(i,j)\in\mathcal{Z}_1}\dot{\mu}_{ij}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i^0\right\|\right) \qquad \text{(A.4)} \\
\le&\mathcal{O}_p\left(\psi_f\rho_{NT}^{-\kappa_f}\sqrt{L_{\mathcal{D}}}\right) = o_p(1)
\end{aligned}
$$

where the third line follows from (reverse) triangle inequality, and the last line is due to Lemma A.1(iii) and Assumption 2(ii) and Assumption 3(ii).

It follows from triangle inequality, Assumption 2(ii) and 3(ii) and $\delta_{il}^0 = 0$ for $l\in\mathcal{I}_i$ that

$$
\begin{aligned}
\psi_1\max_i\sum_{l=1}^{L_{\mathcal{D}}}\dot{w}_{il}\left(\left|\delta_{il}^0\right| - \left|\widehat{\delta}_{il}\right|\right) \le&\psi_1\max_i\left\{\sum_{l\in\mathcal{I}_i}\dot{w}_{il}\left(\left|\delta_{il}^0\right| - \left|\widehat{\delta}_{il}\right|\right) - \psi_1\sum_{l\in\mathcal{V}_i}\dot{w}_{il}\left|\widehat{\delta}_{il}\right|\right\} \\
\le&\psi_1\max_i\sum_{l\in\mathcal{I}_i}\dot{w}_{il}\left|\delta_{il}^0 - \widehat{\delta}_{il}\right| \lesssim \psi_1\max_i\sum_{l\in\mathcal{I}_i}\dot{w}_{il} \\
\le&\psi_1\max_i\left\|\dot{w}_{i,\mathcal{I}_i}\right\| = o_p(1), \qquad \text{(A.5)}
\end{aligned}
$$

by Lemma A.1(iv) and Assumption 3(ii), which implies the third term of R.H.S. of (A.2) is of order $o_p(1)$.

Therefore,

$$
\widehat{Q}_{NT}\left(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{D}}\right) = \frac{1}{N}\sum_{i=1}^{N}\widehat{g}_{i,T}\left(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\delta}}_i\right)'\boldsymbol{W}_{i,T}\widehat{g}_{i,T}\left(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\delta}}_i\right) = o_p(1).
$$

On the other side,

$$\begin{aligned}\widehat{Q}_{NT}\left(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{D}}\right) \geq & \left(\min_i \sigma_{\min}\left(\boldsymbol{W}_i\right)\right)\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{g}_{i,T}\left(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\delta}}_i\right)\right\|^2\right)+o_p(1)\\ \geq & c_w\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{g}_{i,T}\left(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\delta}}_i\right)\right\|^2\right)+o_p(1),\end{aligned}$$

which implies

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{g}_{i,T}\left(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\delta}}_i\right)\right\|^2=o_p(1). \tag{A.6}$$

Note that

$$\left\|\widehat{g}_{i,T}\left(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\delta}}_i\right)\right\|^2=\left\|\widehat{m}_{\mathcal{S},i,T}\left(\widehat{\boldsymbol{\theta}}\right)\right\|^2+\left\|\widehat{m}_{\mathcal{D},i,T}\left(\widehat{\boldsymbol{\theta}}\right)-\widehat{\boldsymbol{\delta}}_i\right\|^2\geq\left\|\widehat{m}_{\mathcal{S},i,T}\left(\widehat{\boldsymbol{\theta}}\right)\right\|^2, \tag{A.7}$$

and by triangle inequality,

$$\left\|\widehat{m}_{\mathcal{S},i,T}\left(\widehat{\boldsymbol{\theta}}\right)\right\|\geq\left|\left\|\overline{m}_{\mathcal{S},i}\left(\widehat{\boldsymbol{\theta}}\right)\right\|-\left\|\widehat{m}_{\mathcal{S},i,T}\left(\widehat{\boldsymbol{\theta}}\right)-\overline{m}_{\mathcal{S},i}\left(\widehat{\boldsymbol{\theta}}\right)\right\|\right|. \tag{A.8}$$

Combine (A.6), (A.7) and (A.8), we have

$$\begin{aligned}o_p(1)=&\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{g}_{i,T}\left(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\delta}}_i\right)\right\|^2\\ \geq&\min_i\left\|\overline{m}_{\mathcal{S},i}\left(\widehat{\boldsymbol{\theta}}\right)\right\|^2-2\left(\max_i\left\|\overline{m}_{\mathcal{S},i}\left(\widehat{\boldsymbol{\theta}}\right)\right\|\right)\left(\max_i\left\|\widehat{m}_{\mathcal{S},i,T}\left(\widehat{\boldsymbol{\theta}}\right)-\overline{m}_{\mathcal{S},i}\left(\widehat{\boldsymbol{\theta}}\right)\right\|\right)\\ =&\min_i\left\|\overline{m}_{\mathcal{S},i}\left(\widehat{\boldsymbol{\theta}}\right)\right\|^2-2O_p\left(1\right)o_p(1)\\ =&\min_i\left\|\overline{m}_{\mathcal{S},i}\left(\widehat{\boldsymbol{\theta}}\right)\right\|^2+o_p(1),\end{aligned}$$

which follows from Assumption 1(ii) and Lemma A.1(i). Then, $\min_i\left\|\overline{m}_{\mathcal{S},i}\left(\widehat{\boldsymbol{\theta}}\right)\right\|=o_p(1)$. By identification of $\boldsymbol{\theta}^0$ imposed by Assumption 2(i), we reach the desired result $\widehat{\boldsymbol{\theta}}\xrightarrow{p}\boldsymbol{\theta}^0$. □

Lemma A.3 (Lemma B.1 in Su et al. (2016))**.** *Suppose Assumption 2(iv) holds, then*

$$\Pr\left(c_Q\left(\frac{1}{2}\overline{Q}_i\left(\boldsymbol{\theta},\boldsymbol{\delta}_i\right)-\widehat{R}_{i,T}\left(\boldsymbol{\theta}\right)\right)\leq\widehat{Q}_{i,T}\left(\boldsymbol{\theta},\boldsymbol{\delta}_i\right)\leq C_Q\left(2\overline{Q}_i\left(\boldsymbol{\theta},\boldsymbol{\delta}_i\right)+2\widehat{R}_{i,T}\left(\boldsymbol{\theta}\right)\right)\right)=1-o\left(1\right)$$

for $\boldsymbol{\theta}\in\Theta$, $\boldsymbol{\theta}_i\in\Theta_\delta$, where c_Q and C_Q are positive constants with $0<c_Q<1<C_Q<\infty$.

Proof of Theorem 1. For simplicity, we denote

$$a_{NT} = \psi_f \left(\max_{(i,j)\in\mathcal{Z}_1} \dot{\mu}_{ij} \right) \text{ and } b_{NT} = \psi_1 \max_{1\leq i\leq N} \|\dot{\boldsymbol{w}}_{i,\mathcal{I}_i}\|,$$

where $\dot{\boldsymbol{w}}_{i,\mathcal{I}_i}$ is the subvector of $\dot{\boldsymbol{w}}_i$ with element w_{il}, for $l \in \mathcal{I}_i = \{l \in [L_{\mathcal{D}}] : \delta_{il}^0 \neq 0\}$. By Lemma A.1(iii) - (iv) and the rate condition Assumption 3(ii), we have $a_{NT} = \mathcal{O}_p(\tau_T)$ and $b_{NT} = \mathcal{O}_p(\tau_T)$.

Part (i). The proof is again starting with (A.2). From (A.5) and Cauchy Schwartz inequality, we have

$$\psi_1 \sum_{l=1}^{L_{\mathcal{D}}} \dot{w}_{il} \left(|\delta_{il}^0| - \left|\widehat{\delta}_{il}\right| \right) \leq \psi_1 \|\dot{\boldsymbol{w}}_{i,\mathcal{I}_i}\| \left\|\widehat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i^0\right\| \leq b_{NT} \left\|\widehat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i^0\right\|. \tag{A.9}$$

From (A.2), (A.4) and (A.9), we have

$$\frac{1}{N}\sum_{i=1}^{N} \left\{ \widehat{Q}_{i,NT}\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\delta}}_i\right) - \widehat{Q}_{i,NT}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\delta}_i^0\right) - (a_{NT} + b_{NT}) \left\|\widehat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i^0\right\| \right\} \leq 0 \tag{A.10}$$

By Lemma A.3, w.p.a.1,

$$\begin{aligned} &\widehat{Q}_{i,NT}\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\delta}}_i\right) - \widehat{Q}_{i,NT}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\delta}_i^0\right) \\ &\qquad \geq \frac{c_Q}{2}\overline{Q}_i\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\delta}}_i\right) - 2C_Q\overline{Q}_i\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\delta}_i^0\right) - (c_Q + 2C_Q)\widehat{R}_{i,T}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\delta}_i^0\right). \end{aligned} \tag{A.11}$$

With sufficiently large (N,T),

$$\widehat{R}_{i,T}\left(\widehat{\boldsymbol{\theta}}\right) = \widehat{\nu}_{i,T}\left(\widehat{\boldsymbol{\theta}}\right)' W_i \widehat{\nu}_{i,T}\left(\widehat{\boldsymbol{\theta}}\right) \leq C_w \left\|\widehat{\nu}_{i,T}\left(\widehat{\boldsymbol{\theta}}\right)\right\|^2 = \mathcal{O}_p\left(\tau_T^2\right), \tag{A.12}$$

where the first inequality follows from Assumption 2(iv) and the last equality is due to Assumption 2(iii) and consistency of $\widehat{\boldsymbol{\theta}}$ shown in Lemma A.2.

Similarly, with sufficiently large (N,T), by first-order Taylor expansion, Assumption 2(iv) and (v) and Lemma A.2,

$$\overline{Q}_i\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\delta}_i^0\right) \leq C_w \left\|\overline{g}_i\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\delta}_i^0\right)\right\|^2 = C_w \left\|\Gamma_i\left(\widetilde{\boldsymbol{\theta}}\right) \begin{bmatrix} \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 \\ \mathbf{0}_{L_{\mathcal{D}}} \end{bmatrix}\right\|^2 \leq C_w C_\Gamma \left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right\|^2, \tag{A.13}$$

$$\overline{Q}_i\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\delta}}_i\right) \geq c_w \left\|\overline{g}_i\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\delta}}_i\right)\right\|^2 = c_w \left\|\Gamma_i\left(\overline{\boldsymbol{\theta}}\right) \begin{bmatrix} \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 \\ \widehat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i^0 \end{bmatrix}\right\|^2 \geq c_w c_\Gamma \left(\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right\|^2 + \left\|\widehat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i^0\right\|^2\right), \tag{A.14}$$

where $\widetilde{\boldsymbol{\theta}}$, $\overline{\boldsymbol{\theta}}$ are between $\boldsymbol{\theta}^0$ and $\widehat{\boldsymbol{\theta}}$.

Combine (A.10), (A.11), (A.4), (A.9), (A.12), (A.13) and (A.14), we have

$$\frac{1}{N}\sum_{i=1}^{N}\left\{\left\|\widehat{\boldsymbol{\delta}}_i-\boldsymbol{\delta}_i^0\right\|^2-(a_{NT}+b_{NT})\left\|\widehat{\boldsymbol{\delta}}_i-\boldsymbol{\delta}_i^0\right\|-\mathcal{O}_p\left(\tau_T^2\right)-\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0\right\|^2\right\}\lesssim 0, \tag{A.15}$$

which, together with Lemma A.1(ii) and Assumption 3(ii), implies

$$\left\|\widehat{\boldsymbol{\delta}}_i-\boldsymbol{\delta}_i^0\right\|\lesssim\mathcal{O}_p\left(\tau_T+a_{NT}+b_{NT}\right)=\mathcal{O}_p\left(\tau_T\right). \tag{A.16}$$

Note that with sufficiently large (N,T) and consistency of $\widehat{\boldsymbol{\theta}}$, (A.13) implies

$$\overline{Q}_i\left(\widehat{\boldsymbol{\theta}},\boldsymbol{\delta}^0\right)\lesssim\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0\right\|,$$

which further implies a twist of (A.15),

$$\frac{1}{N}\sum_{i=1}^{N}\left\{\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0\right\|^2-\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0\right\|+\left\|\widehat{\boldsymbol{\delta}}_i-\boldsymbol{\delta}_i^0\right\|^2-(a_{NT}+b_{NT})\left\|\widehat{\boldsymbol{\delta}}_i-\boldsymbol{\delta}_i^0\right\|-\mathcal{O}_p\left(\tau_T^2\right)\right\}\lesssim 0.$$

Plug in the rate of $\widehat{\boldsymbol{\delta}}_i$ in (A.16), we have

$$\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0\right\|=\mathcal{O}_p\left(\tau_T\right) \tag{A.17}$$

which completes the proof of part (i).

Part (ii). Let $\widehat{\boldsymbol{D}}=\boldsymbol{D}^0+\tau_T\widehat{\boldsymbol{V}}$ where $\widehat{\boldsymbol{V}}=(\widehat{\boldsymbol{v}}_1,\ldots,\widehat{\boldsymbol{v}}_N)\in\mathbb{R}^{L_{\mathcal{D}}\times N}$. Note that

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\boldsymbol{\delta}}_i-\boldsymbol{\delta}^0\right\|^2=\tau_T^2\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\boldsymbol{v}}_i\right\|^2\right),$$

and we want to show that

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\boldsymbol{v}}_i\right\|^2=\mathcal{O}_p\left(1\right). \tag{A.18}$$

For any $\widetilde{\boldsymbol{D}}=\boldsymbol{D}^0+\tau_T\boldsymbol{V}$ with $\left(N^{-1}\sum_{i=1}^{N}\left\|\boldsymbol{v}_i\right\|^2\right)^{-1}=o_p\left(1\right)$,

$$\begin{aligned}&\tau_T^{-2}\left[\left(\widehat{Q}_{NT}\left(\widehat{\boldsymbol{\theta}},\widetilde{\boldsymbol{D}}\right)+P_{\psi_1,\psi_f}\left(\widetilde{\boldsymbol{D}}\right)\right)-\left(\widehat{Q}_{NT}\left(\boldsymbol{\theta}^0,\boldsymbol{D}^0\right)+P_{\psi_1,\psi_f}\left(\boldsymbol{D}^0\right)\right)\right]\\ =&\tau_T^{-2}\left(\widehat{Q}_{NT}\left(\widehat{\boldsymbol{\theta}},\widetilde{\boldsymbol{D}}\right)-\widehat{Q}_{NT}\left(\boldsymbol{\theta}^0,\boldsymbol{D}^0\right)\right)+\tau_T^{-2}\left[\frac{\psi_f}{N^2}\sum_{1\le i<j\le N}\dot{\mu}_{ij}\left(\left\|\boldsymbol{\delta}_i^0+\tau_T\boldsymbol{v}_i-\boldsymbol{\delta}_j^0-\tau_T\boldsymbol{v}_j\right\|-\left\|\boldsymbol{\delta}_i^0-\boldsymbol{\delta}_j^0\right\|\right)\right]\end{aligned}$$

$$
\begin{aligned}
&+\tau_T^{-2}\left[\frac{\psi_1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L_{\mathcal{D}}}\dot{w}_{il}\left(\left|\delta_{il}^0+v_{il}\right|-\left|\delta_{il}^0\right|\right)\right]\\
&:=\Xi_{1,NT}+\Xi_{2,NT}+\Xi_{3,NT}. \qquad \text{(A.19)}
\end{aligned}
$$

The first term in (A.19) can be bounded w.p.a.1

$$
\begin{aligned}
\Xi_{1,NT} \geq&\tau_T^{-2}\left(\frac{c_Q}{2N}\sum_{i=1}^{N}\overline{Q}_i\left(\widehat{\boldsymbol{\theta}},\boldsymbol{\delta}_i^0+\tau_T\boldsymbol{v}_i\right)-\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{R}_{i,T}\left(\widehat{\boldsymbol{\theta}}\right)+\widehat{R}_{i,T}\left(\boldsymbol{\theta}^0\right)\right)\right)\\
\geq&\frac{c_Q c_w c_\Gamma}{2N}\sum_{i=1}^{N}\left(\tau_T^{-2}\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0\right\|^2+\left\|\boldsymbol{v}_i\right\|^2-\mathcal{O}_p(1)\right)\\
\gtrsim&\frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{v}_i\right\|^2+\mathcal{O}_p(1), \qquad \text{(A.20)}
\end{aligned}
$$

where the first inequality follows from Lemma A.3 and $\widehat{Q}_{NT}\left(\boldsymbol{\theta}^0,\boldsymbol{D}^0\right)=N^{-1}\sum_{i=1}^{N}\widehat{R}_{i,T}\left(\boldsymbol{\theta}^0\right)$, the second inequality is due to (A.12) and (A.14) and the last line holds with the result in part (i).

By (A.4), Assumption 3(ii) and Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
\Xi_{2,NT}\geq&-\tau_T^{-2}\left[\frac{\psi_f}{N^2}\sum_{1\leq i<j\leq N}\dot{\mu}_{ij}\left(\left\|\boldsymbol{\delta}_i^0-\boldsymbol{\delta}_j^0\right\|-\left\|\boldsymbol{\delta}_i^0+\tau_T\boldsymbol{v}_i-\boldsymbol{\delta}_j^0-\tau_T\boldsymbol{v}_j\right\|\right)\right]\\
\gtrsim&-\tau_T^{-1}\psi_f\left(\max_{(i,j)\in\mathcal{Z}_1}\dot{\mu}_{ij}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{v}_i\right\|\right)\gtrsim-\mathcal{O}\left(1\right)\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{v}_i\right\|^2\right)^{\frac{1}{2}}. \qquad \text{(A.21)}
\end{aligned}
$$

By (A.9), Assumption 3(ii) and Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
\Xi_{3,NT}=&-\tau_T^{-2}\frac{\psi_1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L_{\mathcal{D}}}\dot{w}_{il}\left(\left|\delta_{il}^0\right|-\left|\widetilde{\delta}_{il}\right|\right)\\
\gtrsim&-\tau_T^{-1}b_{NT}\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{v}_i\right\|\right)\gtrsim-\mathcal{O}\left(1\right)\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{v}_i\right\|^2\right)^{\frac{1}{2}}. \qquad \text{(A.22)}
\end{aligned}
$$

Combine (A.19), (A.20) and (A.21), for sufficiently large (N,T),

$$
\begin{aligned}
&\tau_T^{-2}\left[\left(\widehat{Q}_{NT}\left(\widehat{\boldsymbol{\theta}},\widetilde{\boldsymbol{D}}\right)+P_{\psi_1,\psi_f}\left(\widetilde{\boldsymbol{D}}\right)\right)-\left(\widehat{Q}_{NT}\left(\boldsymbol{\theta}^0,\boldsymbol{D}^0\right)+P_{\psi_1,\psi_f}\left(\boldsymbol{D}^0\right)\right)\right]\\
\gtrsim&\frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{v}_i\right\|^2-\mathcal{O}\left(1\right)\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{v}_i\right\|^2\right)^{\frac{1}{2}}+\mathcal{O}_p(1)
\end{aligned}
$$

$$>0 \tag{A.23}$$

w.p.a.1 since $N^{-1}\sum_{i=1}^{N}\|\boldsymbol{v}_i\|^2$ is diverging, which implies $\left\{\widehat{\boldsymbol{\theta}},\widetilde{\boldsymbol{D}}\right\}$ does not minimize (9) and hence (A.18) holds, which completes the proof of part (ii). □

Lemma A.4. *Suppose the conditions in Lemma A.1 and Assumption 4(ii) hold, then*

(i) $\psi_f^{-1}\tau_T\sqrt{L_{\mathcal{D}}}\max_{(i,j)\in\mathcal{Z}_0}\dot{\mu}_{ij}^{-1} = o_p(1)$.

(ii) $\psi_1^{-1}\tau_T\max_i\max_{l\in\mathcal{V}_i}\dot{w}_{il}^{-1} = o_p(1)$.

Proof of Lemma A.4.

$$\begin{aligned}
\psi_f^{-1}\tau_T\sqrt{L_{\mathcal{D}}}\max_{(i,j)\in\mathcal{Z}_0}\dot{\mu}_{ij}^{-1} &= \psi_f^{-1}\tau_T\sqrt{L_{\mathcal{D}}}\max_{(i,j)\in\mathcal{Z}_0}\left\|\dot{\boldsymbol{\delta}}_i-\dot{\boldsymbol{\delta}}_j\right\|^{\kappa_f}\\
&\leq \psi_f^{-1}\tau_T\sqrt{L_{\mathcal{D}}}\max_{(i,j)\in\mathcal{Z}_0}\left\{\left\|\dot{\boldsymbol{\delta}}_i-\boldsymbol{\delta}_i^0\right\|+\left\|\dot{\boldsymbol{\delta}}_j-\boldsymbol{\delta}_j^0\right\|\right\}^{\kappa_f}\\
&\lesssim \psi_f^{-1}\tau_T\sqrt{L_{\mathcal{D}}}\max_i\left\|\dot{\boldsymbol{\delta}}_i-\boldsymbol{\delta}_i^0\right\|^{\kappa_f}\\
&\leq \psi_f^{-1}\tau_T\sqrt{L_{\mathcal{D}}}\mathcal{O}_p\left(\widetilde{\varkappa}_{NT}^{\kappa_f}\right)\\
&= o_p(1),
\end{aligned} \tag{A.24}$$

where we apply triangle inequality in the second line, the fourth line invokes Lemma A.1(ii) and the last line follows from Assumption 4(ii), and (A.24) implies (i).

$$\begin{aligned}
\psi_1^{-2}\tau_T^2\max_i\max_{l\in\mathcal{V}_i}\dot{w}_{il}^{-2} = \psi_1^{-2}\tau_T^2\max_i\max_{l\in\mathcal{V}_i}\dot{\delta}_{il}^{2\kappa_1} &\leq \psi_1^{-2}\tau_T^2\max_i\left\|\dot{\boldsymbol{\delta}}_i-\boldsymbol{\delta}_i^0\right\|^{2\kappa_1}\\
&\leq \psi_1^{-2}\tau_T^2\mathcal{O}_p\left(\widetilde{\varkappa}_{NT}^{2\kappa_1}\right)\\
&= o_p(1),
\end{aligned} \tag{A.25}$$

where the second line invokes Lemma A.1(ii) and the last line follows from Assumption 4(ii), and (A.25) implies (ii). □

Proof of Theorem 2. In the proof, we first show the first statement in (ii) and (iii). Then we can leverage the results to show the uniform consistency of $\widehat{\boldsymbol{\delta}}_i$ in part (i), and the second statement in (ii) and (iii) directly follow.

Rewrite the objective function (9), with notation $\widehat{\Psi}_{NT}$, as

$$\widehat{\Psi}_{NT}(\boldsymbol{\theta},\boldsymbol{D}) = \frac{1}{N}\sum_{i=1}^{N}\left\{\widehat{g}_{i,T}(\boldsymbol{\theta},\boldsymbol{\delta}_i)'\boldsymbol{W}_{i,NT}\widehat{g}_{i,T}(\boldsymbol{\theta},\boldsymbol{\delta}_i)+\frac{\psi_f}{2N}\sum_{j=1}^{N}\dot{\mu}_{ij}\left\|\boldsymbol{\delta}_i-\boldsymbol{\delta}_j\right\|+\psi_1\sum_{l=1}^{L_{\mathcal{D}}}\dot{w}_{il}\left|\delta_{il}\right|\right\}. \tag{A.26}$$

The Karush-Kuhn-Tucker (KKT) condition (with respect to δ_{il}) evaluated at $\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{D}}\right)$, scaled up by τ_T^{-1}, is

$$\begin{aligned} 0 &= 2\tau_T^{-1}\boldsymbol{\gamma}'_{L_{\mathcal{S}}+l}\boldsymbol{W}_{i,NT}\widehat{g}_{i,T}\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\delta}}_i\right) + \frac{\psi_f\tau_T^{-1}}{2N}\sum_{j\notin\mathcal{G}_{k(i)}}\dot{\mu}_{ij}\widehat{e}_{ij,l} + \frac{\psi_f\tau_T^{-1}}{2N}\sum_{j\in\mathcal{G}_{k(i)}}\dot{\mu}_{ij}\widehat{e}_{ij,l} + \psi_1\tau_T^{-1}\dot{w}_{il}\widehat{s}_{il} \\ &:= \Xi_{il,m} + \Xi_{il,\mathcal{Z}_1} + \Xi_{il,\mathcal{Z}_0} + \Xi_{il,1}, \end{aligned} \tag{A.27}$$

where $\widehat{\boldsymbol{e}}_{ij} = \frac{\widehat{\boldsymbol{\delta}}_i - \widehat{\boldsymbol{\delta}}_j}{\left\|\widehat{\boldsymbol{\delta}}_i - \widehat{\boldsymbol{\delta}}_j\right\|}$ if $\left\|\widehat{\boldsymbol{\delta}}_i - \widehat{\boldsymbol{\delta}}_j\right\| \neq 0$ and $\|\widehat{\boldsymbol{e}}_{ij}\| \leq 1$ otherwise; $\widehat{s}_{il} = \operatorname{sgn}\left(\widehat{\delta}_{il}\right)$ if $\widehat{\delta}_{il} \neq 0$ and $\widehat{s}_{il} \in [-1, 1]$ if $\widehat{\delta}_{il} = 0$; $\boldsymbol{\gamma}_{L_{\mathcal{S}}+l} \in \mathbb{R}^L$ is the vector with $(L_{\mathcal{S}} + l)$-th element equal to 1 and others being 0.

In the KKT condition (A.27):

$$\begin{aligned} |\Xi_{il,m}| &\leq 2\tau_T^{-1}\left(\|\boldsymbol{W}_i\| + \|\boldsymbol{W}_{i,NT} - \boldsymbol{W}_i\|\right)\left\|\widehat{g}_{i,T}\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\delta}}_i\right)\right\| \\ &\leq 2\tau_T^{-1}\left(C_w + o_p(1)\right)\left(C_\Gamma\left(\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right\|^2 + \left\|\widehat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i^0\right\|^2\right)\right)^{\frac{1}{2}} \\ &\leq 2\tau_T^{-1}\left(C_w + o_p(1)\right)\mathcal{O}_p(\tau_T) \\ &= \mathcal{O}_p(1), \end{aligned} \tag{A.28}$$

where we apply Cauchy-Schwartz inequality in the first inequality, Assumption 2(iv) and (v) as in (A.14) for the second inequality, and make use of the convergence rate derived in Theorem 1(i) to reach the result that $\Xi_{il,m}$ is stochastically bounded.

$$|\Xi_{il,\mathcal{Z}_1}| \leq \frac{\psi_f\tau_T^{-1}}{2N}\left(\max_{(i,j)\in\mathcal{Z}_1}\dot{\mu}_{ij}\right)\left|\sum_{j\notin\mathcal{G}_{k(i)}}\widehat{e}_{ij,l}\right| \leq \psi_f\tau_T^{-1}\left(\max_{(i,j)\in\mathcal{Z}_1}\dot{\mu}_{ij}\right)\frac{N - N_{k(i)}}{2N} \leq \mathcal{O}_p(1), \tag{A.29}$$

which is due to Assumption 3(ii) and Lemma A.1(iii).

To facilitate the analysis of $\Xi_{il,\mathcal{Z}_0}$ and $\Xi_{il,1}$, we introduce the following notations. Let $c_e \in (0, \frac{1}{3})$ be a constant. Denote

$$\widehat{\mathcal{Z}}_{i,0} = \left\{j \in \mathcal{G}_{k(i)} : \left\|\widehat{\boldsymbol{\delta}}_i - \widehat{\boldsymbol{\delta}}_j\right\| = 0\right\} \text{ and } \widehat{\mathcal{Z}}_{i,1} = \left\{j \in \mathcal{G}_{k(i)} : \left\|\widehat{\boldsymbol{\delta}}_i - \widehat{\boldsymbol{\delta}}_j\right\| > 0\right\}.$$

Define the events $\mathcal{E}_{G,i} = \left\{\frac{\left|\widehat{\mathcal{Z}}_{i,1}\right|}{N_{K(i)}} > c_e\right\}$ and $\mathcal{E}_{S,i} = \left\{\max_{l\in\mathcal{V}_i}\left|\widehat{\delta}_{il}\right| > 0\right\}$.

Conditioning on $\mathcal{E}_{G,i}$, we have $\left|\left\{j \in \mathcal{G}_{k(i)} : \left|\widehat{\delta}_{il} - \widehat{\delta}_{jl}\right| > 0\right\}\right| > c_e \frac{N_{k(i)}}{\sqrt{L_{\mathcal{D}}}}$, and

$$\begin{aligned}
\left|\Xi_{il,\mathcal{Z}_0}\right| =& \psi_f \tau_T^{-1} \left(\min_{(i,j)\in\mathcal{Z}_0} \dot{\mu}_{ij}\right) \left|\frac{1}{2N} \sum_{j\in\mathcal{G}_{k(i)}} \widetilde{\mu}_{ij} \widehat{e}_{ij,l}\right| \\
\gtrsim& \psi_f \tau_T^{-1} L_{\mathcal{D}}^{-\frac{1}{2}} \left(\min_{(i,j)\in\mathcal{Z}_0} \dot{\mu}_{ij}\right) \frac{c_e N_{k(i)}}{N} \xrightarrow{p} \infty, \qquad \text{(A.30)}
\end{aligned}$$

where $\widetilde{\mu}_{ij} := \frac{\dot{\mu}_{ij}}{\min_{(i,j)\in\mathcal{Z}_0} \dot{\mu}_{ij}} \geq 1$ for $(i,j) \in \mathcal{Z}_0$, the inequality holds since the system $\sum_{j\in\mathcal{G}_{k(i)}} \widetilde{\mu}_{ij}\widehat{e}_{ij,l} = 0$, with $\|\widehat{\boldsymbol{e}}_{ij}\| = 1$ if $\left\|\widehat{\boldsymbol{\delta}}_i - \widehat{\boldsymbol{\delta}}_j\right\| > 0$, for $i \in \mathcal{G}_k$ for some $k \in [K^0]$ and $l \in [L_{\mathcal{D}}]$, is over-determined in terms of $\widehat{e}_{ij,l}$ and does not vanish, and the probability limit follows from Assumption 4(iii) that $\lim_{N\to\infty} N_{k(i)}/N > \pi_{\min}$ and Lemma A.4(i). As a result, $\left|\Xi_{il,\mathcal{Z}_1}\right|$ is asymptotically explosive.

Conditional on $\mathcal{E}_{S,i}$, we have

$$\left|\Xi_{il,1}\right| \geq \tau_T^{-1}\psi_1 \min_i \min_{l\in\mathcal{V}_i} \dot{w}_{il} \xrightarrow{p} \infty \text{ if } l \in \mathcal{V}_i \text{ and } \left|\widehat{\delta}_{il}\right| > 0, \qquad \text{(A.31)}$$

by Lemma A.4(ii).

Combine the KKT condition (A.27) and (A.28), (A.29), (A.30), (A.31), and triangle inequality, for each $i \in [N]$,

$$\Pr\left(\mathcal{E}_{G,i} \bigcup \mathcal{E}_{S,i}\right) \leq \Pr\left(\left|\left|\Xi_{il,\mathcal{Z}_0}\right| - \left|\Xi_{il,1}\right|\right| \leq \left|\Xi_{il,\mathcal{Z}_1}\right| + \left|\Xi_{il,m}\right|, \mathcal{E}_{G,i} \bigcup \mathcal{E}_{S,i}\right) \to 0 \qquad \text{(A.32)}$$

as $(N,T) \to \infty$, since $\left|\Xi_{il,m}\right| + \left|\Xi_{il,\mathcal{Z}_1}\right| = \mathcal{O}_p(1)$ while $\left|\left|\Xi_{il,\mathcal{Z}_0}\right| - \left|\Xi_{il,1}\right|\right| \xrightarrow{p} \infty$ with suitable choice of κ_1 and κ_f that guarantees $\Xi_{il,\mathcal{Z}_0}$ and $\Xi_{il,1}$ do not coincide conditional on $\mathcal{E}_{G,i} \bigcap \mathcal{E}_{S,i}$.

Then we turn to the desired uniform results. Denote the event

$$\mathcal{E}_G = \left\{\max_{(i,j)\in\mathcal{Z}_0} \left\|\widehat{\boldsymbol{\delta}}_i - \widehat{\boldsymbol{\delta}}_j\right\| > 0\right\} = \left\{\exists \left(i^\star, j^\star\right) \in \mathcal{Z}_0 \text{ s.t. } \left\|\widehat{\boldsymbol{\delta}}_{i^\star} - \widehat{\boldsymbol{\delta}}_{j^\star}\right\| > 0\right\}.$$

By (A.32), we have

$$\begin{aligned}
\Pr\left(\mathcal{E}_G\right) <& \Pr\left(\mathcal{E}_G \bigcap \left\{\frac{\left|\widehat{\mathcal{Z}}_{i^\star,0}\right|}{N_{k(i^\star)}} > c_e\right\}\right) + \Pr\left(\mathcal{E}_G \bigcap \left\{\frac{\left|\widehat{\mathcal{Z}}_{j^\star,0}\right|}{N_{k(j^\star)}} > c_e\right\}\right) \\
&+ \Pr\left(\mathcal{E}_G \bigcap \left\{\frac{\left|\mathcal{G}_k \setminus \left(\widehat{\mathcal{Z}}_{i^\star,0} \bigcup \widehat{\mathcal{Z}}_{j^\star,0}\right)\right|}{N_{k(i^\star)}} \geq 1 - 2c_e\right\}\right) \\
\leq& \Pr\left(\mathcal{E}_{G,j^\star}\right) + 2\Pr\left(\mathcal{E}_{G,i^\star}\right) \to 0, \qquad \text{(A.33)}
\end{aligned}$$

as $(N,T)\to\infty$, which completes the proof of the first statement in part (ii).

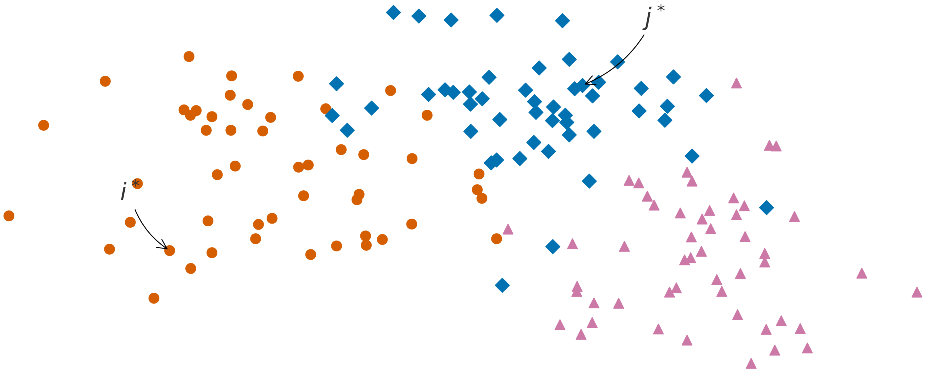


Figure A.1: Illustration of the proof arguments in (A.33).

Conditioning on $\left\{\bigcup_{i=1}^{N}\mathcal{E}_{S,i}\right\}$, there exist $i^\star\in[N]$ such that $\max_{l\in\mathcal{V}_{i^\star}}\left|\widehat{\delta}_{i^\star l}\right|>0$. Note that (A.31) holds uniformly across i, and the above arguments can go through for $i^\star$, which leads

$$\Pr\left(\max_{i}\max_{l\in\mathcal{V}_i}\left|\widehat{\delta}_{il}\right|>0\right)\to 0$$

as $(N,T)\to\infty$ and the first statement in part (iii) is established.

Next, we turn to the proof of the uniform convergence result in part (i). Note that it suffices to show

$$\Pr\left(\max_{i}\left\|\widehat{\boldsymbol{\delta}}_i-\boldsymbol{\delta}_i^0\right\|>\varkappa_{NT}\;\middle|\;\max_{(i,j)\in\mathcal{Z}_0}\left\|\widehat{\boldsymbol{\delta}}_i-\widehat{\boldsymbol{\delta}}_j\right\|=0\right)=o\left(1\right), \tag{A.34}$$

given that $\Pr\left(\max_{(i,j)\in\mathcal{Z}_0}\left\|\widehat{\boldsymbol{\delta}}_i-\widehat{\boldsymbol{\delta}}_j\right\|=0\right)\to 1$ as $(N,T)\to\infty$. Rewrite that the penalized GMM objective in (A.26) as

$$\widehat{\Psi}_{NT}\left(\boldsymbol{\theta},\boldsymbol{D}\right)=\widehat{Q}_{NT}\left(\boldsymbol{\theta},\boldsymbol{D}\right)+\frac{1}{N}\sum_{i=1}^{N}\frac{\psi_f}{2N}\sum_{j=1}^{N}\dot{\mu}_{ij}\left\|\boldsymbol{\delta}_i-\boldsymbol{\delta}_j\right\|+\frac{\psi_1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L_{\mathcal{D}}}\dot{w}_{il}\left|\delta_{il}\right|, \tag{A.35}$$

and denote

$$\Pi_{NT}=\left\{\boldsymbol{D}\in\Theta_{\delta}^{N}:\max_{1\le i\le N}\left\|\boldsymbol{\delta}_i-\boldsymbol{\delta}_i^0\right\|>\varkappa_{NT},\ \text{and}\ \max_{(i,j)\in\mathcal{Z}_0}\left\|\boldsymbol{\delta}_i-\boldsymbol{\delta}_j\right\|=0\right\}.$$

It is then desired to show that, w.p.a.1,

$$\inf_{\boldsymbol{D}\in\Pi_{NT}}\widehat{\Psi}_{NT}\left(\widehat{\boldsymbol{\theta}},\boldsymbol{D}\right)>\widehat{\Psi}_{NT}\left(\boldsymbol{\theta}^0,\boldsymbol{D}^0\right), \tag{A.36}$$

which implies that the minimizer $\widehat{\boldsymbol{D}} \notin \Pi_{NT}$ w.p.a.1.

We first establish an upper bound for $\widehat{\Psi}_{NT}\left(\boldsymbol{\theta}^0, \boldsymbol{D}^0\right)$. (A.3) together with Lemma A.1(i) implies

$$\widehat{Q}_{NT}\left(\boldsymbol{\theta}^0, \boldsymbol{D}^0\right) = \mathcal{O}_p\left(\widetilde{\varkappa}_{NT}^2\right). \tag{A.37}$$

Following the similar arguments as in (A.4) and (A.5), together with Lemma A.1(iii)-(iv) and Assumption 4(i), we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\frac{\psi_f}{2N}\sum_{j=1}^{N}\dot{\mu}_{ij}\left\|\boldsymbol{\delta}_i^0-\boldsymbol{\delta}_j^0\right\| &\lesssim \max_{1\leq i\leq N}\frac{\psi_f}{N}\sum_{j\notin\mathcal{G}_{k(i)}}\dot{\mu}_{ij}\left\|\boldsymbol{\delta}_i^0-\boldsymbol{\delta}_j^0\right\| \\ &\leq \psi_f\sqrt{L_{\mathcal{D}}}\left(\max_{(i,j)\in Z_1}\dot{\mu}_{ij}\right) = \psi_f\sqrt{L_{\mathcal{D}}}\mathcal{O}_p\left(\rho_{NT}^{-\kappa_f}\right) = \mathcal{O}_p\left(\widetilde{\varkappa}_{NT}^2\right),\end{aligned} \tag{A.38}$$

$$\frac{\psi_1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L_{\mathcal{D}}}\dot{w}_{il}\left|\delta_{il}^0\right| \leq \psi_1\max_{1\leq i\leq N}\left\|\dot{\boldsymbol{w}}_{i,\mathcal{I}_i}\right\| = \psi_1\sqrt{L_{\mathcal{D}}}\mathcal{O}_p\left(\zeta_{NT}^{-\kappa_1}\right) = \mathcal{O}_p\left(\widetilde{\varkappa}_{NT}^2\right). \tag{A.39}$$

Combine (A.37) - (A.39), we establish the upper bound for $\widehat{\Psi}_{NT}\left(\boldsymbol{\theta}^0, \boldsymbol{D}^0\right)$ as

$$\widehat{\Psi}_{NT}\left(\boldsymbol{\theta}^0, \boldsymbol{D}^0\right) = \mathcal{O}_p\left(\widetilde{\varkappa}_{NT}^2\right). \tag{A.40}$$

Next, we investigate the L.H.S. of (A.36). Denote $\Pi_{k,NT} = \left\{\boldsymbol{D} \in \Theta_\delta^N : \left\|\boldsymbol{\delta}_i - \boldsymbol{\delta}_i^0\right\| > \varkappa_{NT},\ \forall i \in \mathcal{G}_k\right\}$.

$$\begin{aligned}\inf_{\boldsymbol{D}\in\Pi_{NT}}\widehat{\Psi}_{NT}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{D}\right) &\geq \inf_{1\leq k\leq K^0}\inf_{\boldsymbol{D}\in\Pi_{k,NT}}\widehat{Q}_{i,NT}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{D}\right) \geq \inf_{1\leq k\leq K^0}\inf_{\boldsymbol{D}\in\Pi_{k,NT}}\frac{1}{N}\sum_{i\in\mathcal{G}_k}\widehat{Q}_{i,NT}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\delta}_i\right) \\ &\gtrsim \inf_{1\leq k\leq K^0}\inf_{\boldsymbol{D}\in\Pi_{k,NT}}\frac{1}{N}\sum_{i\in\mathcal{G}_k}\left\|\widehat{m}_{\mathcal{D},i,T}\left(\widehat{\boldsymbol{\theta}}\right) - \boldsymbol{\delta}_i\right\|^2 \\ &\geq \inf_{1\leq k\leq K^0}\inf_{\boldsymbol{D}\in\Pi_{k,NT}}\frac{1}{N}\sum_{i\in\mathcal{G}_{k(i^\star)}}\left|\left\|\widehat{m}_{\mathcal{D},i,T}\left(\widehat{\boldsymbol{\theta}}\right) - \boldsymbol{\delta}_i^0\right\| - \left\|\boldsymbol{\delta}_i - \boldsymbol{\delta}_i^0\right\|\right|^2\end{aligned}$$

where the second line follows from Assumption 2(iv) and the last line follows from the triangle inequality. Note that

$$\begin{aligned}\max_{1\leq i\leq N}\left\|\widehat{m}_{\mathcal{D},i,T}\left(\widehat{\boldsymbol{\theta}}\right) - \boldsymbol{\delta}_i^0\right\| \leq& \max_{1\leq i\leq N}\left\|\widehat{m}_{\mathcal{D},i,T}\left(\widehat{\boldsymbol{\theta}}\right) - \overline{m}_{\mathcal{D},i}\left(\widehat{\boldsymbol{\theta}}\right)\right\| + \left(\max_{1\leq i\leq N}\left\|\Gamma_{\mathcal{D},i}\left(\widetilde{\boldsymbol{\theta}}\right)\right\|\right)\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right\| \\ \leq& \mathcal{O}_p\left(\widetilde{\varkappa}_{NT} + \tau_T\right) = \mathcal{O}_p\left(\widetilde{\varkappa}_{NT}\right),\end{aligned}$$

where $\widetilde{\boldsymbol{\theta}}$ is between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^0$, the first inequality follows from the triangle inequality and mean value theorem, the second inequality is due to Assumption 2(v), Theoreom 1(i) and Lemma A.1(i), and the last equality follows from Lemma A.1(i) because the uniform rate of convergence must be at least as slow as the rate of convergence of the sample moments for each i. Meanwhile, for $\boldsymbol{D} \in \Pi_{k,NT}$ and $i \in \mathcal{G}_k$, $\|\boldsymbol{\delta}_i - \boldsymbol{\delta}_i^0\| > \varkappa_{NT}$. Then for sufficiently large (N,T), we have

$$\inf_{\boldsymbol{D}\in\Pi_{NT}} \widehat{\Psi}_{NT}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{D}\right) > \left(\inf_{1\le k\le K^0} \frac{N_k}{N}\right)\left(\varkappa_{NT} - \max_{1\le i\le N}\left\|\widehat{m}_{\mathcal{D},i,T}\left(\widehat{\boldsymbol{\theta}}\right) - \boldsymbol{\delta}_i^0\right\|\right)^2 \gtrsim \varkappa_{NT}^2, \quad (\text{A.41})$$

where we apply Assumption 4(iii). Combine (A.40) and (A.41), we reach (A.36) and complete the proof part (i).

The second statement in both parts (ii) and (iii) are directly implied by (i). For sufficiently large (N,T),

$$\begin{aligned}\Pr\left(\min_{(i,j)\in\mathcal{Z}_1}\left\|\widehat{\boldsymbol{\delta}}_i - \widehat{\boldsymbol{\delta}}_j\right\| > 0\right) &\ge \Pr\left(\rho_{NT} - \max_{(i,j)\in\mathcal{Z}_1}\left\{\left\|\widehat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i^0\right\| + \left\|\widehat{\boldsymbol{\delta}}_j - \boldsymbol{\delta}_j^0\right\|\right\} > 0\right)\\ &\ge \Pr\left(\rho_{NT} - 2\max_i\left\|\widehat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i^0\right\| > 0\right)\\ &\ge \Pr\left(\max_i\left\|\widehat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i^0\right\| < \varkappa_{NT}\right)\\ &= 1 - o(1) \qquad (\text{A.42})\end{aligned}$$

by the uniform convergence result in (i) and the rate condition Assumption 3(i).

For sufficiently large (N,T), we have

$$\begin{aligned}\Pr\left(\min_{1\le i\le N}\min_{l\in\mathcal{I}_i}\left|\widehat{\delta}_{il}\right| > 0\right) &\ge \Pr\left(\min_i\min_{l\in\mathcal{I}_i}\left[\left|\delta_{il}^0\right| - \left|\widehat{\delta}_{il} - \delta_{il}^0\right|\right] > 0\right)\\ &\ge \Pr\left(\zeta_{NT} - \max_i\max_{l\in\mathcal{I}_i}\left|\widehat{\delta}_{il} - \delta_{il}^0\right| > 0\right)\\ &\ge \Pr\left(\zeta_{NT} - \max_i\left\|\widehat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i^0\right\| > 0\right) \ge \Pr\left(\max_i\left\|\widehat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i^0\right\| < \varkappa_{NT}\right)\\ &= 1 - o(1) \qquad (\text{A.43})\end{aligned}$$

by triangle inequality, the uniform convergence result in (i) and the rate condition Assumption 3(i).

□

Proof of Theorem 4. In the proof, we drop the superscript "post" for notational simplicity. By Corollary 3, we have $\widehat{\mathcal{G}}_k = \mathcal{G}_{(k)}$ and $\widehat{\mathcal{I}}_k = \mathcal{I}_{(k)}$ w.p.a.1. for $k \in [K^0]$, so $\widehat{\boldsymbol{\beta}}_k$ has the same asymptotic distribution as the oracle estimator $\overline{\boldsymbol{\beta}}$, assuming the group structure and the set of invalid moment conditions are known, defined as

$$\left(\overline{\boldsymbol{\theta}}', \overline{\boldsymbol{\alpha}}'_{k,\mathcal{I}_k}\right)' = \underset{\boldsymbol{\theta}\in\Theta, \boldsymbol{\alpha}_{k,\mathcal{I}_k}\in\Theta_\delta^{|\mathcal{I}_k|}}{\arg\min} \left(\frac{1}{N_k}\sum_{i\in\mathcal{G}_k}\widetilde{g}^{(k)}_{i,T}\left(\boldsymbol{\theta},\boldsymbol{\alpha}_{k,\mathcal{I}_k}\right)\right)' \boldsymbol{W}_{k,NT} \left(\frac{1}{N_k}\sum_{i\in\mathcal{G}_k}\widetilde{g}^{(k)}_{i,T}\left(\boldsymbol{\theta},\boldsymbol{\alpha}_{k,\mathcal{I}_k}\right)\right), \tag{A.44}$$

and we let $\overline{\boldsymbol{\beta}}_k = \left(\overline{\boldsymbol{\theta}}', \overline{\boldsymbol{\alpha}}'_{k,\widehat{\mathcal{I}}_k}\right)'$. The proof directly follows from Theorem 3.3 in Cheng and Liao (2015). □

A.2 Convex Optimization Formulation

Recall the optimization problem (9),

$$\underset{\boldsymbol{\theta}\in\Theta,\, \boldsymbol{D}\in\Theta_\delta{}^N}{\arg\min}\ \widehat{Q}_{NT}\left(\boldsymbol{\theta},\boldsymbol{D}\right) + P_{\psi_1,\psi_f}\left(\boldsymbol{D}\right), \tag{A.45}$$

where

$$\widehat{Q}_{NT}\left(\boldsymbol{\theta},\boldsymbol{D}\right) = \frac{1}{N}\sum_{i=1}^{N}\widehat{g}_{i,T}\left(\boldsymbol{\theta},\boldsymbol{\delta}_i\right)'\boldsymbol{W}_{i,T}\widehat{g}_{i,T}\left(\boldsymbol{\theta},\boldsymbol{\delta}_i\right), \tag{A.46}$$

$$P_{\psi_1,\psi_c}\left(\boldsymbol{D}\right) = \frac{\psi_1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L_\mathcal{D}}\dot{w}_{il}\left|\delta_{il}\right| + \frac{\psi_f}{N^2}\sum_{1\le i<j\le N}\dot{\mu}_{ij}\left\|\boldsymbol{\delta}_i-\boldsymbol{\delta}_j\right\|, \tag{A.47}$$

and $\boldsymbol{W}_{i,T}$ is a positive definite weighting matrix.

To begin with, we deal with the PAFL penalty using similar techniques as in Gao and Shi (2021) for the C-Lasso penalty. Let $\boldsymbol{\phi}_{ij} = \boldsymbol{\delta}_i - \boldsymbol{\delta}_j$ for $1 \le i < j \le N$. By introducing an auxiliary variable v_{ij} and conic constraints $\|\boldsymbol{\phi}_{ij}\| \le v_{ij}$, the PAFL penalty in the objective function can be equivalently expressed as a linear component $\boldsymbol{\mu}'\boldsymbol{v} = \sum_{1\le i<j\le N}\mu_{ij}v_{ij}$ where $\mu_{ij} = \frac{\psi_f}{N^2}\dot{\mu}_{ij}$, $\boldsymbol{v} = \left(v'_{12}, v'_{13}, \ldots, v'_{(N-1)N}\right)' \in \mathbb{R}^{\frac{N(N-1)}{2}\times 1}$ and $\boldsymbol{\mu} = \left(\mu_{12},\mu_{13},\ldots,\mu_{(N-1)N}\right)' \in \mathbb{R}^{\frac{N(N-1)}{2}\times 1}$. For the adaptive Lasso penalty, we introduce $\boldsymbol{\delta}_i^+ = \left(\max\left\{0,\delta_{il}\right\}\right)_{l=1}^{L_\mathcal{D}}$ and $\boldsymbol{\delta}_i^- = \left(\max\left\{0,-\delta_{il}\right\}\right)_{l=1}^{L_\mathcal{D}}$ for $i \in [N]$ which satisfies $\boldsymbol{\delta}_i = \boldsymbol{\delta}_i^+ - \boldsymbol{\delta}_i^-$. The adaptive Lasso penalty can be equivalently expressed as a linear component $\boldsymbol{\omega}'\left(\boldsymbol{\delta}^+ + \boldsymbol{\delta}^-\right)$ where $\omega_{il} = \frac{\psi_1}{N}\dot{\omega}_{il}$, $\boldsymbol{\delta}^+ = \left(\boldsymbol{\delta}_1^{+\prime}, \boldsymbol{\delta}_2^{+\prime}, \ldots, \boldsymbol{\delta}_N^{+\prime}\right)' \in \mathbb{R}^{NL_\mathcal{D}\times 1}$ and $\boldsymbol{\delta}^- = \left(\boldsymbol{\delta}_1^{-\prime}, \boldsymbol{\delta}_2^{-\prime}, \ldots, \boldsymbol{\delta}_N^{-\prime}\right)' \in \mathbb{R}^{NL_\mathcal{D}\times 1}$. If $g_i\left(\boldsymbol{\theta},\boldsymbol{\delta}_i\right)$ is linear in $\boldsymbol{\theta}_i$, then $\widehat{g}_{i,T}\left(\boldsymbol{\theta},\boldsymbol{\delta}_i\right)$ is also linear in $\boldsymbol{\theta}_i$ and of the form $\boldsymbol{\gamma}_i - \Gamma_i\boldsymbol{\theta}$ where $\boldsymbol{\gamma}_i$ is a known vector and Γ_i is a known matrix. Let $\boldsymbol{W}_{i,T} = \boldsymbol{H}_i'\boldsymbol{H}_i$, then the quadratic term in the objective can be written as $\sum_{i=1}^{N}\frac{t_i}{N}$ where $t_i \ge \|\widetilde{\boldsymbol{g}}_i\|_2^2$ and $\widetilde{\boldsymbol{g}}_i = \boldsymbol{H}_i\boldsymbol{g}_i$. Note that the $\|\widetilde{\boldsymbol{g}}_i\|_2^2 \le t_i$ is equivalent to $\left\|\left(\widetilde{\boldsymbol{g}}_i, \frac{t_i-1}{2}\right)\right\|_2 \le \frac{t_i+1}{2}$, which is a standard second-order conic constraint.

As a result, the convex optimization problem is

$$
\begin{aligned}
\min_{\boldsymbol{g},\boldsymbol{\theta},\boldsymbol{\delta}^{+},\boldsymbol{\delta}^{-},\boldsymbol{\phi},\boldsymbol{v}} \ & \sum_{i=1}^{N} \frac{t_i}{N} + \boldsymbol{\omega}' \left(\boldsymbol{\delta}^{+} + \boldsymbol{\delta}^{-}\right) + \boldsymbol{\mu}'\boldsymbol{v} \\
\text{s.t. } & \boldsymbol{g}_i + \boldsymbol{\Gamma}_i \theta + \left(\mathbf{0}'_{(L-L_D)\times 1}, \boldsymbol{\delta}_i^{+\prime} - \boldsymbol{\delta}_i^{-\prime}\right)' = \boldsymbol{\gamma}_i \quad \forall i \in [N] \\
& \boldsymbol{\delta}_i^{+} - \boldsymbol{\delta}_j^{+} - \boldsymbol{\delta}_i^{-} + \boldsymbol{\delta}_j^{-} - \boldsymbol{\phi}_{ij} = \mathbf{0}_{L_D \times 1} \quad \forall 1 \le i < j \le N \\
& \|\boldsymbol{\phi}_{ij}\| \le v_{ij} \quad \forall 1 \le i < j \le N \\
& \widetilde{\boldsymbol{g}}_i = \boldsymbol{H}_i \boldsymbol{g}_i \quad \forall i \in [N] \\
& \left\| \left(\widetilde{\boldsymbol{g}}_i, \frac{t_i - 1}{2} \right) \right\|_2 \le \frac{t_i + 1}{2} \quad \forall i \in [N] \\
& v_{ij} \ge 0 \quad \forall 1 \le i < j \le N \\
& \boldsymbol{\delta}_i^{+} \ge \mathbf{0}_{L_D \times 1}, \boldsymbol{\delta}_i^{-} \ge \mathbf{0}_{L_D \times 1} \quad \forall i \in [N]
\end{aligned}
$$

which is readily solved using standard convex optimization solvers such as `MOSEK`.