A Simple and Computationally Trivial Estimator for Grouped Fixed Effects Models*

Martin Mugnier[†]

Abstract

This paper provides a new fixed effects estimator for linear panel data models with clustered time patterns of unobserved heterogeneity. The method combines smooth and convex nuclear norm regularization with a pairwise differencing argument that takes at most $O(N^3)$ elementary operations to agglomeratively cluster cross-sectional units. Asymptotic guarantees are established in a framework where T can grow at any power of N, as both N and T diverge to infinity. In contrast to existing approaches, the proposed estimator (i) is computationally straightforward, (ii) does not require a known upper bound on the number of groups, (iii) consistently estimates the number of groups, (iv) correctly classifies units into groups with probability tending to one uniformly across units, (v) is asymptotically equivalent to the infeasible least squares estimator that controls for the true group indicators, (vi) is asymptotically normal at parametric rates, (vii) and is free of the incidental parameters problem.

Keywords: panel data, time-varying unobserved heterogeneity, grouped fixed effects, agglomerative clustering

JEL Codes: C14, C23, C25.

---

*An earlier version of this paper has been circulated under the title "Make the Difference: Computationally Trivial Estimators for Grouped Fixed Effects Models". *First draft version*: March 14, 2022. *This version:* October 12, 2022. I thank Stéphane Bonhomme, Xavier D'Haultfœuille, Elena Manresa, Pauline Rossi, Ao Wang, Andrei Zeleneev, and seminar participants at CREST, University of Chicago, and 2022 Bristol Econometric Study Group Conference for helpful comments and discussions. This research benefitted from financial support from the research grants Otelo (ANR-17-CE26-0015-041) and ANR "Investissements d'avenir": EUR DATA EFM (ANR-18-EURE-0005).

[†]CREST, ENSAE, Institut Polytechnique de Paris, martin.mugnier@ensae.fr

1 Introduction

Suppose to observe a sample of longitudinal or panel data (y_{it}, x_{it}) and consider the grouped fixed effects model:

$$y_{it} = x_{it}'\beta + \alpha_{g_i t} + v_{it}, \quad i = 1, ..., N,\ t = 1, ..., T, \tag{1.1}$$

where i denotes cross-sectional units, t denotes time periods, $y_{it} \in \mathbb{R}$ is a dependent variable, $\beta \in \mathbb{R}^p$ is an unknown vector of interest, and $x_{it} \in \mathbb{R}^p$ is a vector of covariates contemporaneously uncorrelated with the zero-mean random variable $v_{it} \in \mathbb{R}$ but potentially correlated with the group-specific effect $\alpha_{g_i t} \in \mathbb{R}$. The number of groups $G \in \mathbb{N}^*$, the group membership variable $g_i \in \{1, ..., G\}$, and the vector of group-specific time effects $(\alpha_{1t}, ..., \alpha_{Gt})' \in \mathbb{R}^G$ are unrestricted and treated as parameters to estimate. This model, which was introduced in the pioneering work of Bonhomme and Manresa (2015), is an instance of Bai (2009)'s "interactive fixed effects" models. In particular, it has a *sparse* factor-analytic structure.[1]

To the best of my knowledge, the panel data literature does not mention the existence of an estimator for the $(1 + N + GT + p)$-dimensional parameter

$$\theta := (G, g_1, ..., g_N, \alpha_{11}, \alpha_{12}, ..., \alpha_{1T}, \alpha_{21}, \alpha_{22}, ..., \alpha_{2T}, ..., \alpha_{GT}, \beta')'$$

that, in an asymptotic framework where T can grow at any power of N, (i) is computationally feasible, (ii) does not require a known upper bound $G_{\max} \geq G$, (iii) consistently estimates G, (iv) correctly classifies units into groups with probability tending to one uniformly across units, (v) is asymptotically equivalent to the oracle OLS regression that controls for the true group indicators, (vi) is asymptotically normal at parametric rates, (vii) and is free of Neyman and Scott (1948)'s incidental parameters problem. Meeting such constraints is particularly difficult, as the number of possible allocations of N individuals into G groups grows exponentially with N.

In this paper, I first introduce a new and simple two-step estimator $\widehat{\theta}$ that enjoys properties (i)-(ii). It combines smooth and convex nuclear norm regularization in a first step with a pairwise differencing argument in the second step, which costs

[1] Note that $\alpha_{g_i t} = \lambda_i' F_t$ with $\lambda_i' := (\mathbf{1}\{g_i = 1\}, ..., \mathbf{1}\{g_i = G\})$ and $F_t' := (\alpha_{1t}, ..., \alpha_{Gt})$. The vector of factor loadings λ_i thus lies in the (finite) set of vertices of the unit simplex of $\mathbb{R}^G$.

only $O(N^3)$ elementary operations. Next, I derive sufficient conditions for $\widehat{\theta}$ to enjoy properties (iii)-(vii). To the best of my knowledge, this paper is the first to propose an estimator for model (1.1) with properties (i)-(vii). As a byproduct, it is also the first paper to show property (iii) for an estimator that also verifies property (ii) in a specific interactive fixed effects model. This remarkable property stands in sharp contrast with the general case, in which a known upper bound is generally required (see, e.g., Bai and Ng, 2002; Bai, 2009; Bai and Ng, 2019).

This paper does not address the question of conducting inference on groups. Dzemski and Okui (2018) provide pointwise valid inference methods given an estimator $\widehat{\beta}$ is available. One can use the estimator proposed in the present paper.

Related Literature There is a vast literature on estimating interactive fixed effects models (see, e.g., Bai, 2009; Moon and Weidner, 2015, 2017). The leading methods rely on non-convex least squares and Principal Component Analysis, which violate property (i) and are not known to verify the other properties. Bonhomme and Manresa (2015)'s grouped fixed-effects (GFE) estimator, an extension of k-means clustering to handle covariates, does not verify property (i) and is not known to verify (ii)-(iii). The same is true for recent extensions to grouped factor models (see, e.g., Ando and Bai, 2022). Chetverikov and Manresa (2021)'s spectral and post-spectral estimators are not known to verify properties (ii)-(iii). The authors also impose a grouped factor structure on the covariates. Moon and Weidner (2019)'s nuclear norm estimator is not known to verify properties (iv-vii). Pollard (1981, 1982) provides asymptotic theory for the solution to the population k-means sum of squares problem in the cross-sectional case, i.e., only for a pseudo-true value. In contrast, my results hold for the true population model parameter. The theory developed in this paper covers the estimates effectively reported by the researcher, which fundamentally differs from approaches relying on heurisitic approximating algorithms or local minima (e.g., Bonhomme and Manresa, 2015). The "testing" statistics used in the first step is similar (though different) to the $\widehat{d}^2_\infty$ empirical distance employed by Zeleneev (2020) and Auerbach (2022). It has already been employed in the mathematical statistics literature to study topological properties of the graphon (e.g., Zhang, Levina, and Zhu, 2017; Lovász, 2012). Dyad, triad, or tetrad comparisons have proven useful in a variety of different econometric contexts (see, e.g., Graham, 2017; Charbonneau, 2017; Jochmans, 2017; Zeleneev, 2020). Albeit close in spirit, the procedure is different from the binary segmentation algorithm developed in Wang and Su (2021), or the pairwise comparisons method proposed in Krasnokutskaya, Song, and Tang (2022). Some papers rely on spectral clustering (see, e.g., Ng, Jordan, and Weiss, 2002; von Luxburg, 2007; Chetverikov and Manresa, 2021; Brownlees, Guðmundsson, and Lugosi, 2022; Yu, Gu, and Volgushev, 2022). The operational research literature has proposed other agglomerative clustering approaches (e.g., DBSCAN) that, to the best of my knowledge, have not been applied to the model considered in this paper.

Section 2 introduces the two-step estimator. Section 3 presents the main asymptotic results which, among others, include uniform consistency of the clustering estimator and asymptotic normality at parametric rates. Section 4 contains a brief discussion and concludes. All proofs are in the Appendix. Additional material is in the Supplemental Material (Mugnier, 2022), with section numbers S.1, etc.

2 A Two-Step Estimator

In this section, I introduce the two-step estimator. For all $g \in \mathbb{N}^*$, define $\Theta_g := \{g\} \times \{1, ..., g\}^N \times \mathcal{A}^{gT} \times \mathcal{B}$, where $\mathcal{B} \subset \mathbb{R}^p$, $p \geq 1$ is fixed, and $\mathcal{A} \subset \mathbb{R}$. Let $\Theta := \bigcup_{g=1}^{+\infty} \Theta_g$. The goal is to estimate $\theta \in \Theta$ as N and T diverge jointly to infinity. With this purpose, I first introduce a useful preliminary nuclear norm regularized estimator. Let $\left\| \cdot \right\|_{\mathrm{F}}$ and $\left\| \cdot \right\|_1$ denote the Frobenius norm and the nuclear norm respectively. Additionally, let $Y := (y_{it})_{i=1,...,N;t=1,...,T} \in \mathbb{R}^{N\times T}$, $X_k := (x_{it,k})_{i=1,...,N;t=1,...,T} \in \mathbb{R}^{N\times T}$ for all $k \in \{1, .., p\}$, and $v \cdot X := \sum_{k=1}^{p} X_k v_k$ for all $v \in \mathbb{R}^p$. For $(\lambda, \beta')' \in (0, +\infty) \times \mathbb{R}^p$, I define

$$Q_\lambda(\beta) = \min_{\Gamma \in \mathbb{R}^{N\times T}} \left\{ \frac{1}{2NT} \left\| Y - \beta \cdot X - \Gamma \right\|_{\mathrm{F}}^2 + \frac{\lambda}{\sqrt{NT}} \left\| \Gamma \right\|_1 \right\} \tag{2.1}$$

and

$$\widehat{\beta}^1(\lambda) = \underset{\beta \in \mathbb{R}^p}{\arg\min}\, Q_\lambda(\beta). \tag{2.2}$$

$\widehat{\beta}^1(\lambda)$ is the nuclear norm regularized estimator (see, e.g., Moon and Weidner, 2019). Rgularization is needed because the true number of groups is unknown. The two-step estimator depends on two tuning parameters $\lambda_1, \lambda_2 > 0$. It is obtained through the following steps.

1. CLUSTERING STEP:

   (a) Compute the residuals $\widehat{v}_{it} = y_{it} - x_{it}'\widehat{\beta}^1(\lambda_1), \quad i = 1, ..., N,\ t = 1, ..., T$.

   (b) Compute the pairwise distance

   $$\widehat{d}_\infty^2(i,j) = \max_{k\in\{1,...,N\}\setminus\{i,j\}} \left| \frac{1}{T}\sum_{t=1}^{T}(\widehat{v}_{it} - \widehat{v}_{jt})\widehat{v}_{kt} \right|, \quad i = 1, ..., N,\ j = 1, ..., N.$$

   (c) Compute the adjency matrix $\widehat{W} \in \{0,1\}^{N\times N}$ where

   $$\widehat{W}_{ij} = \mathbf{1}\left\{\widehat{d}_\infty^2(i,j) \leq \lambda_2\right\}, \quad i = 1, ..., N,\ j = 1, ..., N.$$

   (d) Set $\widehat{g}_1 = 1$, $k = 1$, `stop`=False, and

   $$\widehat{C}_1 = \left\{i \in \{1, ..., N\} : \widehat{W}_{1j} = \widehat{W}_{ij} \quad \forall j \in \{1, ..., N\}\right\}.$$

   Set $\widehat{g}_i = 1$ for all $i \in \widehat{C}_1$.

   While not `stop`:

   - Let $i_k^* := \inf\left\{i \in \{1, ..., N\} : i \notin \cup_{\ell=1}^{k}\widehat{C}_\ell\right\}$;
   - If $i_k^* < \infty$, set

     $$\widehat{C}_{k+1} = \left\{i \in \{1, ..., N\} : \widehat{W}_{i_k^* j} = \widehat{W}_{ij} \quad \forall j \in \{1, ..., N\}\right\},$$

     $\widehat{g}_i = k + 1$ for all $i \in \widehat{C}_{k+1}$, and k<-k+1.

     Else, set `stop`=True.

   (e) Set $\widehat{G} = k$.

2. ESTIMATION STEP:

   Compute $\left(\widehat{\beta}', \widehat{\alpha}_{11}, ..., \widehat{\alpha}_{\widehat{G}T}\right)'$ by using a pooled OLS regression of y_{it} on x_{it} and the interactions of groups $(\mathbf{1}\{\widehat{g}_i = g\})_{1\leq g\leq\widehat{G}}$ and time $(\mathbf{1}\{t = s\})_{1\leq s\leq T}$ dummies:

   $$\left(\widehat{\beta}', \widehat{\alpha}_{11}, ..., \widehat{\alpha}_{\widehat{G}T}\right) \in \underset{\left(\beta', \alpha_{11}, ..., \alpha_{\widehat{G}T}\right)\in\mathcal{B}\times\mathcal{A}^{\widehat{G}T}}{\arg\max} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\beta - \alpha_{\widehat{g}_i t}\right)^2.$$

   Obtain $\widehat{\theta} := (\widehat{G}, \widehat{g}_1, ..., \widehat{g}_N, \widehat{\alpha}_{11}, ..., \widehat{\alpha}_{\widehat{G}T}, \widehat{\beta}')'$.

Remark 1: The parameter λ_2 determines the willingness to merge units i and j together based on the distance measure $\widehat{d}^2_\infty(i,j)$. Note that this procedure always produces $\widehat{G} \in \{1,...,N\}$ groups, which contain units with the same set of estimated links. If all random variables are continuous, then to the extreme where $\lambda_2 \to 0$, $\widehat{G} \to N$ and each group contains a single unit. To the extreme where $\lambda_2 \to +\infty$, $\widehat{G} \to 1$ and a single group contains all units. Between these two regimes, an entire regularization path can be reported by the researcher given the low CPU time of the method. In particular, the clustering step can be made all vectorized, greatly reducing computational burden compared to running loops.[2] The parameter λ_1 can be chosen by cross-validation (splitting samples at the cross-sectional level). Similarly, the parameter λ_2^* can be chosen such that, given λ_1, $\widehat{G}(\lambda_1, \lambda_2^*)$ locates on the largest plateau.

Remark 2: Given the low CPU time, it is possible to improve the finite sample performance by re-running the first step with $\widehat{v}_{it} = y_{it} - x_{it}'\widehat{\beta}$ in place of $\widehat{v}_{it} = y_{it} - x_{it}'\widehat{\beta}^1$ to obtain new $\widehat{g}_1, ..., \widehat{g}_N$, and then re-running the second step and iterating again until some convergence criterion is achieved. The asymptotic results will hold for any of the subsequent iterates.

Remark 3: When unobserved heterogeneity is assumed to be time-constant, the $O(N^3)$ computation cost can be reduced to $O(N^2)$ and the preliminary estimator can be replaced with any standard differencing fixed effects estimator such as Arellano and Bond (1991). See also Wooldridge (2010). This particular case, as well as some extensions of model (1.1), is discussed in the Supplemental Material S.2.

Remark 4: While the use of $\widehat{W}_{ij}$ to estimate the group memberships g_i^0 appears to be new in the literature, the statistics $\widehat{d}^2_\infty(i,j)$ has previously (though differently) been employed to study topological properties of the graphon in the mathematical statistics literature (e.g., Zeleneev, 2020; Zhang, Levina, and Zhu, 2017; Lovász, 2012; Auerbach, 2022).[3] Here, the intuition builds on similar reasoning as in p.14 in Zeleneev (2020). Asymptotically, $\widehat{v}_{it} \approx \alpha^0_{g_i^0 t}$ so that $g_i^0 = g_j^0$ implies, "uniformly" over

[2]MATLAB code is provided on the author's website: https://martinmugnier.github.io/research. A small Monte Carlo exercise and an empirical illustration are presented in sections S.3-4 of the Supplemental Material.

[3]I thank Andrei Zeleneev for pointing this to me as the previous version of the paper was based

k,

$$\frac{1}{T}\sum_{t=1}^{T}(\widehat{v}_{it}-\widehat{v}_{jt})\widehat{v}_{kt}\approx 0 \implies \max_{k\in\{1,...,N\}\setminus\{i,j\}}\left|\frac{1}{T}\sum_{t=1}^{T}(\widehat{v}_{it}-\widehat{v}_{jt})\widehat{v}_{kt}\right|\approx 0.$$

Reciprocally, if

$$\max_{k\in\{1,...,N\}\setminus\{i,j\}}\left|\frac{1}{T}\sum_{t=1}^{T}(\widehat{v}_{it}-\widehat{v}_{jt})\widehat{v}_{kt}\right|\approx 0, \tag{2.3}$$

then necessarily $g_i^0 = g_j^0$. To see it, note that if $\mathbf{1}\left\{g_i^0 \neq g_j^0\right\}$, and provided each group has at least 2 units asymptotically (which is weak), then there exist $k^*, l^* \in \{1,...,N\}\setminus\{i,j\}$ such that $g_{k^*}^0 = g_i^0$ and $g_{l^*}^0 = g_j^0$. Equation (2.3) implies in turn

$$\frac{1}{T}\sum_{t=1}^{T}(\widehat{v}_{it}-\widehat{v}_{jt})\widehat{v}_{k^*t}\approx 0 \tag{2.4}$$

$$\frac{1}{T}\sum_{t=1}^{T}(\widehat{v}_{it}-\widehat{v}_{jt})\widehat{v}_{l^*t}\approx 0. \tag{2.5}$$

Differencing (2.4)-(2.5) yields

$$\frac{1}{T}\sum_{t=1}^{T}(\alpha^0_{g_i^0t}-\alpha^0_{g_j^0t})^2\approx 0,$$

a contradiction if groups are well separated, e.g., if for all $(g,\widetilde{g})\in\left\{1,...,G^0\right\}^2$, $g\neq\widetilde{g}$, there exists $c_{g,\widetilde{g}}>0$ such that

$$\frac{1}{T}\sum_{t=1}^{T}(\alpha^0_{gt}-\alpha^0_{\widetilde{g}t})^2\geq c_{g,\widetilde{g}}>0.$$

The next section formalizes the identification result.

---

instead on the tetrad statistics:

$$\widetilde{d}^2_\infty(i,j) := \max_{(k,l)\in\{\{1,...,N\}\setminus\{i,j\}\}^2}\left|\frac{1}{T}\sum_{t=1}^{T}(\widehat{v}_{it}-\widehat{v}_{jt})(\widehat{v}_{kt}-\widehat{v}_{lt})\right|,$$

and $\widetilde{W}_{ij} = \mathbf{1}\left\{\widetilde{d}^2_\infty(i,j)\leq\lambda_2\right\}$. Asymptotic properties of $\widehat{W}_{ij}$ and $\widetilde{W}_{ij}$ are exactly the same but computing the matrix $\widehat{D}_\infty := \left\{\widehat{d}^2_\infty(i,j):(i,j)\in\{1,..,N\}^2\right\}$ costs only $O(N^3)$ operations compared to $O(N^4)$ for $\widetilde{D}_\infty := \left\{\widetilde{d}^2_\infty(i,j):(i,j)\in\{1,..,N\}^2\right\}$, which can yield significant differences in very large data sets.

3 Large Sample Properties

Consider the following data generating process:

$$y_{it} = x_{it}'\beta^0 + \alpha^0_{g_i^0 t} + v_{it}, \quad i = 1, ..., N, \ t = 1, ..., T,$$

where $g_i^0 \in \{1..., G^0\}$ denotes group membership, and where the 0 superscripts refer to true parameter values. I assume for now that the number of groups $G = G^0$ is fixed (relative to N, T) but unknown, and I defer the discussion on the case of an increasing sequence $G^0 = G^0_{NT}$ to the Supplemental Material S.1.[4]
For any set $\mathcal{I} \subset \mathbb{N}^*$, for all $k \in \mathbb{N}^*$, let $\mathscr{P}_k(\mathcal{I})$ denote the set of subsets of $\mathcal{I}$ with cardinal k. Consider the following assumption.

Assumption 1 *There exist constants $a, b, d_1, d_2, c > 0$ and a sequence $\tau(t) \leq e^{-at^{d_1}}$ such that:*

(a) *$\mathcal{A}$ is a compact subset of $\mathbb{R}$.*

(b) *For all $(i, t) \in \{1, ..., N\} \times \{1, ..., T\}$: $\Pr(|v_{it}| > m) \leq e^{1-(m/b)^{d_2}}$ for all $m > 0$ and $\mathbb{E}(v_{it}) = 0$.*

(c) *For all $(g, \widetilde{g}) \in \{1, ..., G^0\}^2$ such that $g \neq \widetilde{g}$: $\text{plim}_{T\to\infty} T^{-1} \sum_{t=1}^{T} (\alpha^0_{gt} - \alpha^0_{\widetilde{g}t})^2 = c_{g,\widetilde{g}} \geq c$.*

(d) *For all $(i, j, k, g, \widetilde{g}) \in \mathscr{P}_3(\{1, ..., N\}) \times \{1, ..., G^0\}^2$ such that $g \neq \widetilde{g}$, $\{v_{it}\}_t$, $\{(v_{it} - v_{jt})v_{kt}\}_t$, $\left\{\alpha^0_{gt} - \alpha^0_{\widetilde{g}t}\right\}_t$, and $\left\{(\alpha^0_{gt} - \alpha^0_{\widetilde{g}t})v_{it}\right\}_t$ are strongly mixing processes with mixing coefficients $\tau(t)$. Moreover, $\mathbb{E}((\alpha^0_{gt} - \alpha^0_{\widetilde{g}t})v_{it}) = \mathbb{E}(v_{it}v_{jt}) = 0$.*

(e) $\lim_{N\to\infty} \Pr(\min_{g\in\{1,...,G^0\}} \sum_{i=1}^{N} \mathbf{1}\{g_i^0 = g\} \geq 2) = 1$.

Assumptions 1(a), 1(b), and 1(d) collect standard moment, tail, and dependence conditions. They do not impose homoscedasticity but only require uniform bounds on the unconditional variances. Assumption 1(c) requires groups to be well-separated. Assumption 1(e) allows for asymptotically negligible groups but requires that each group has at least two members with probability approaching one. Define $\gamma^0 =$

[4] In particular, when $\theta^0 = 0$ is known, the number of groups can increase with sample size at any rate bounded by $N/2$.

$(\mathbf{1}\{g_i = g\})_{i=1,...,N;g=1,...,G^0} \in \{0,1\}^{N\times G^0}$, $\alpha^0 = (\alpha^0_{gt})_{t=1,...,T;g=1,...,G^0} \in \mathcal{A}^{T\times G^0}$, $x_k = \text{vec}(X_k)$, and $x = (x_1, ..., x_k)$.

Assumption 2

(a) *There exists $(\nu, \kappa) \in (0, +\infty) \times (0, 1/2)$ such that, as N and T tend to infinity: $N = o(T^\nu)$, $\lambda_1 = o(1)$, $\sqrt{\min(N,T)}\lambda_1 \to \infty$, $\lambda_2 = o(1)$, $\lambda_2/\lambda_1^2 \to \infty$, and $\lambda_2 \gtrsim 1/T^\kappa$, where $a_n \gtrsim b_n$ if and only if $\exists (C, n_0) \in (0, +\infty) \times \mathbb{N}, \forall n \geq n_0, a_n \geq C b_n$.*

(b) *There exists a constant $M > 0$ such that, as N, T tend to infinity:*

$$\sup_{i\in\{1,...,N\}} \Pr\left(\frac{1}{T}\sum_{t=1}^{T} \left\|x_{it}\right\|^2 \geq M\right) = o(T^{-\delta}) \text{ for all } \delta > 0,$$

where $\left\|\cdot\right\|$ denotes the Euclidean norm.

(c) *Let $\mathbb{C} = \left\{A \in \mathbb{R}^{N\times T} : \left\|M_{\gamma^0} A M_{\alpha^0}\right\|_1 \leq 3\left\|A - M_{\gamma^0} A M_{\alpha^0}\right\|_1\right\}$, where $M_B := I - B(B'B)^\dagger B$, I is the identity matrix of appropriate dimensions, and † refers to the Moore-Penrose generalized inverse. There exists $\mu > 0$, independent from N and T, such that for any $a \in \mathbb{R}^{NT}$ with $\text{mat}(a) \in \mathbb{C}$ we have $a' M_x a \geq \mu a' a$, for N, T sufficiently large.*

(d) *$\left\|(v_{it})_{i=1,...,N;t=1,...,T}\right\|_\infty = O_p(\sqrt{\max(N,T)})$, where $\left\|\cdot\right\|_\infty$ denotes the spectral norm.*

(e) *$\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T x_{it} x_{it}' \to \Sigma > 0$ and $\frac{1}{\sqrt{NT}}\sum_{i=1}^N \sum_{t=1}^T v_{it} x_{it} = O_p(1)$.*

Assumption 2(a) allows T to grow considerably more slowly than N (when $\nu \gg 1$). Assumption 2(b) is a slight reinforcement of Assumption 2(e) discussed in Bonhomme and Manresa (2015). It holds if covariates have bounded support or if they satisfy dependence and tail conditions similar to v_{it}. Assumptions 2(c)-(e) will imply $\left\|\widehat{\beta}^1 - \beta^0\right\| = o_p(1)$ using results in Moon and Weidner (2019). Assumption 2(c) is a restricted eigenvalue condition, common in high-dimensional modeling (see, e.g., Bickel, Ritov, and Tsybakov, 2009). Sufficient conditions for Assumption 2(d) are given in the Supplementary Appendix S.2 of Moon and Weidner (2017). All results below are understood up to group relabeling.

Proposition 3.1 (Sup-Norm Classification Consistency) *Let Assumptions 1-2 hold. Then, as N and T tend to infinity,*

$$\left\|\widehat{\beta}^1 - \beta^0\right\| = O_p(\lambda_1), \tag{3.1}$$

$$\widehat{G} - G^0 = o_p(1), \tag{3.2}$$

and

$$\sup_{i\in\{1,...,N\}} \left|\widehat{g}_i - g_i^0\right| = o_p(1). \tag{3.3}$$

The following assumption is useful to establish the asymptotic distribution of $\widehat{\beta}$ and $\widehat{\alpha}_{gt}$.

Assumption 3

(a) *For all $g \in \{1, ..., G^0\}$:* $\operatorname{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^N \mathbf{1}\{g_i^0 = g\} = \pi_g > 0.$

(b) *For all $(g,t) \in \{1, ..., G^0\} \times \{1, ..., T\}$:*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \mathbb{E}\left(\mathbf{1}\left\{g_i^0 = g\right\} \mathbf{1}\left\{g_j^0 = g\right\} v_{it}v_{jt}\right) = \omega_{gt} > 0.$$

(c) *For all $(g,t) \in \{1, ..., G^0\} \times \{1, ..., T\}$: as N and T tend to infinity,*

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N \mathbf{1}\left\{g_i^0 = g\right\} v_{it} \xrightarrow{d} \mathcal{N}\left(0, \omega_{gt}\right),$$

where $\xrightarrow{d}$ denotes convergence in distribution.

(d) *For all $(i,j,t) \in \{1, ..., N\}^2 \times \{1, ..., T\}$:* $\mathbb{E}(x_{jt}v_{it}) = 0.$

(e) *There exist positive definite matrices Σ_β and Ω_β such that*

$$\Sigma_\beta = \operatorname*{plim}_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \overline{x}_{g_i^0 t})(x_{it} - \overline{x}_{g_i^0 t})',$$
$$\Omega_\beta = \operatorname*{plim}_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^N \sum_{j=1}^N \sum_{t=1}^T \sum_{s=1}^T \mathbb{E}\left[v_{it}v_{js}(x_{it} - \overline{x}_{g_i^0 t})(x_{js} - \overline{x}_{g_j^0 s})'\right],$$

where $\overline{x}_{gt} := \left(\sum_{i=1}^N \mathbf{1}\{g_i^0 = g\}\right)^{-1} \sum_{i=1}^N \mathbf{1}\{g_i^0 = g\} x_{it}.$

(f) *As N and T tend to infinity:* $\frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \overline{x}_{g_i^0 t})v_{it} \xrightarrow{d} \mathcal{N}(0, \Omega_\beta).$

Corollary 3.2 (Asymptotic Distribution) *Let Assumptions 1-3 hold. Then, as N and T tend to infinity,*

$$\sqrt{NT}(\widehat{\beta} - \beta^0) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_\beta^{-1}\Omega_\beta\Sigma_\beta^{-1}\right), \tag{3.4}$$

and, for all t:

$$\sqrt{N}(\widehat{\alpha}_{gt} - \alpha_{gt}^0) \xrightarrow{d} \mathcal{N}\left(0, \frac{\omega_{gt}}{\pi_g^2}\right), \; g = 1, ..., G^0, \tag{3.5}$$

where $\Sigma_\beta, \Omega_\beta, \omega_{gt}$, and π_g are defined in Assumption 3.

Consistent plug-in estimates of the asymptotic variances can easily be constructed (see, e.g., Supplemental Material in Bonhomme and Manresa, 2015).

4 Discussion and Conclusion

Grouped fixed effects models are plagued with an underlying difficult combinatorial classification problem, rendering estimation and inference difficult. In this paper, I propose a novel constructive identification argument for all the model parameters including the number of groups. The corresponding two-step estimator has polynomial computational cost and is straightforward to implement (only smooth convex optimization and elementary arithmetic operations are required). It is based on thresholding a suitable pairwise differencing transformation of the regression equation and a preliminary off-the-shelf consistent estimator of the slope. Mild conditions are given under which the proposed estimator is uniformly consistent for the latent grouping structure and asymptotically normal as both dimensions diverge jointly. Importantly, the number of groups is consistently estimated without any prior knowledge, and the time-dimension can grow much more slowly than the cross-sectional dimension. The paper leaves a few questions unanswered. First, could the approach be fruitful to build a test for the grouping assumption? Second, how does the new estimator perform relative to alternative methods that require the number of groups to be known? Third, could similar differencing ideas be applied to more general nonlinear models? I leave these questions for further research.

Appendix

A Proofs of the Results

A.1 Proof of Proposition 3.1

(3.1) immediately follows from an application of Theorem 2 in Moon and Weidner (2019), after noticing the factor-analytic structure of model (1.1) (see footnote 1). Let $W^0 := (\mathbf{1}\left\{g_i^0 = g_j^0\right\})_{i=1,\ldots,N;j=1,\ldots,N}$ and $\widehat{W} := (\widehat{W}_{ij})_{i=1,\ldots,N;j=1,\ldots,N}$, and define the matrix norm $\left\|\cdot\right\|_{\max}$ such that $\left\|A\right\|_{\max} = \max_{i=1,\ldots,n;j=1,\ldots,m}|a_{ij}|$ for any $A = (a_{ij})_{i,j} \in \mathbb{R}^{n\times m}$. (3.2) and (3.3) are immediate corollaries of Lemma A.1 below.

Lemma A.1 *Let Assumptions 1-2 hold. Then, as N and T tend to infinity,*

$$\left\|\widehat{W} - W^0\right\|_{\max} = \max_{(i,j)\in\{1,\ldots,N\}^2}\left|\widehat{W}_{ij} - W_{ij}^0\right| = o_p(1). \tag{A.1}$$

Proof of Lemma A.1: Let $\epsilon > 0$. By (3.1), there exists $K > 0$ such that, letting $\mathcal{E}_{1NT} = \left\{\left\|\widehat{\beta}^1 - \beta^0\right\| > K\lambda_1\right\}$, $\Pr(\mathcal{E}_{1NT}) < \epsilon$ for N, T sufficiently large. Define $Z_{1NT}(i,j) = \widehat{W}_{ij}(1 - W_{ij}^0)$, $Z_{2NT}(i,j) = (1 - \widehat{W}_{ij})W_{ij}^0$, and the probability events $\mathcal{E}_{2N} = \left\{\min_{g\in\{1,\ldots,G^0\}}\sum_{i=1}^N \mathbf{1}\left\{g_i^0 = g\right\} \geq 2\right\}$ and $\mathcal{E}_{NT} = \mathcal{E}_{1NT}^c \cap \mathcal{E}_{2N}$. By the union bound,

$$\begin{aligned}
&\Pr\left(\max_{(i,j)\in\{1,\ldots,N\}^2}\left|\widehat{W}_{ij} - W_{ij}^0\right| > 0\right)\\
&\quad\leq \Pr\left(\mathcal{E}_{1NT}\right) + \Pr\left(\mathcal{E}_{2NT}^c\right) + \sum_{(i,j)\in\{1,\ldots,N\}^2}\Pr\left(\left|\widehat{W}_{ij} \neq W_{ij}^0\right|, \mathcal{E}_{NT}\right)\\
&\quad= \epsilon + o(1) + \sum_{(i,j)\in\{1,\ldots,N\}^2}\Pr\left(Z_{1NT}(i,j) = 1, \mathcal{E}_{NT}\right) + \Pr\left(Z_{2NT}(i,j) = 1, \mathcal{E}_{NT}\right),
\end{aligned} \tag{A.2}$$

where I have used $\lim_{N\to\infty}\Pr\left(\mathcal{E}_{2N}^c\right) = 0$ by Assumption 1(e) to obtain the equality. Below, I prove that, for $\ell \in \{1,2\}$ and as N and T tend to infinity,

$$\max_{(i,j)\in\{1,\ldots,N\}^2}\Pr\left(Z_{\ell NT}(i,j) = 1, \mathcal{E}_{NT}\right) = o(N^2T^{-\delta}) \text{ for all } \delta > 0. \tag{A.3}$$

Equation (A.1) then follows by combining (A.2)-(A.3) and Assumption 2(a), and because ϵ is unrestricted.

1. I first show (A.3) for $\ell = 1$.[5] Let $(i,j) \in \{1,...,N\}^2$ and $\delta > 0$.

$$Z_{1NT}(i,j) = \mathbf{1}\left\{ \max_{k \in \{1,\dots,N\} \setminus \{i,j\}} \left| \frac{1}{T} \sum_{t=1}^{T} (\widehat{v}_{it} - \widehat{v}_{jt}) \widehat{v}_{kt} \right| \leq \lambda_2 \right\} \mathbf{1}\left\{ g_i^0 \neq g_j^0 \right\}.$$

Since almost surely $Z_{1NT}(i,j) = 0$ if $G^0 = 1$, I assume w.l.o.g. that $G^0 \geq 2$. Then,

$$\begin{aligned}
&\mathbf{1}\left\{ Z_{1NT}(i,j) = 1, \mathcal{E}_{NT} \right\} \\
&\quad = \mathbf{1}\left\{ \mathcal{E}_{NT} \right\} \times \\
&\qquad \sum_{\substack{(g,\widetilde{g}) \in \{1,\dots,G^0\}^2 \\ g \neq \widetilde{g}}} \mathbf{1}\left\{ g_i^0 = g \right\} \mathbf{1}\left\{ g_j^0 = \widetilde{g} \right\} \mathbf{1}\left\{ \max_{k \in \{1,\dots,N\} \setminus \{i,j\}} \left| \frac{1}{T} \sum_{t=1}^{T} (\widehat{v}_{it} - \widehat{v}_{jt}) \widehat{v}_{kt} \right| \leq \lambda_2 \right\}.
\end{aligned}$$

If $\mathbf{1}\left\{ \mathcal{E}_{NT} \right\} \mathbf{1}\left\{ g_i^0 \neq g_j^0 \right\} = 1$, there exists a pair $\left(k^*(i,j,g_i^0), l^*(i,j,g_j^0) \right) \in \mathscr{P}_2\left(\{1,...,N\} \setminus \{i,j\} \right)$ such that $g^0_{k^*(i,j,g_i^0)} = g_i^0$ and $g^0_{l^*(i,j,g_j^0)} = g_j^0$. It follows that

$$\begin{aligned}
&\mathbf{1}\left\{ Z_{1NT}(i,j) = 1, \mathcal{E}_{NT} \right\} \\
&\quad \leq \mathbf{1}\left\{ \mathcal{E}_{NT} \right\} \times \\
&\qquad \sum_{\substack{(g,\widetilde{g}) \in \{1,\dots,G^0\}^2 \\ g \neq \widetilde{g}}} \mathbf{1}\left\{ g_i^0 = g \right\} \mathbf{1}\left\{ g_j^0 = \widetilde{g} \right\} \mathbf{1}\left\{ \left| \frac{1}{T} \sum_{t=1}^{T} (\widehat{v}_{it} - \widehat{v}_{jt}) \widehat{v}_{k^*(i,j,g_i^0)t} \right| \leq \lambda_2 \right\} \times \\
&\qquad \mathbf{1}\left\{ \left| \frac{1}{T} \sum_{t=1}^{T} (\widehat{v}_{it} - \widehat{v}_{jt}) \widehat{v}_{l^*(i,j,g_j^0)t} \right| \leq \lambda_2 \right\} \\
&\quad \leq \mathbf{1}\left\{ \mathcal{E}_{NT} \right\} \times \\
&\qquad \sum_{\substack{(g,\widetilde{g}) \in \{1,\dots,G^0\}^2 \\ g \neq \widetilde{g}}} \mathbf{1}\left\{ g_i^0 = g \right\} \mathbf{1}\left\{ g_j^0 = \widetilde{g} \right\} \mathbf{1}\left\{ \left| \frac{1}{T} \sum_{t=1}^{T} (\widehat{v}_{it} - \widehat{v}_{jt}) (\widehat{v}_{k^*(i,j,g_i^0)t} - \widehat{v}_{l^*(i,j,g_j^0)t}) \right| \leq 2\lambda_2 \right\}.
\end{aligned}$$

There is most one pair $(g,\widetilde{g}) \in \{1,..,G^0\}^2$ such that $g \neq \widetilde{g}$ and $\mathbf{1}\left\{ g_i^0 = g \right\} \mathbf{1}\left\{ g_j^0 = \widetilde{g} \right\} = 1$. Hence, by developing the product and using $\mathbf{1}\left\{ |a| \leq b \right\} \leq \mathbf{1}\left\{ a \leq b \right\}$ for any

[5] Actually, I show the weaker result that the supremum is $o(T^{-\delta})$.

$(a,b) \in \mathbb{R} \times \mathbb{R}^*$, I obtain

$$
\begin{aligned}
&\mathbf{1}\left\{Z_{1NT}(i,j)=1,\mathcal{E}_{NT}\right\} \\
&\quad \leq \mathbf{1}\left\{\mathcal{E}_{NT}\right\} \times \\
&\qquad \max_{\substack{(g,\widetilde{g})\in\left\{1,\ldots,G^0\right\}^2 \\ g\neq\widetilde{g}}} \mathbf{1}\Bigg\{\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)^2+\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)\left(v_{it}-v_{jt}+v_{k^*(i,j,g)t}-v_{l^*(i,j,\widetilde{g})t}\right) \\
&\qquad +\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)\left(\beta^0-\widehat{\beta}^1\right)'\left(x_{it}-x_{jt}+x_{k^*(i,j,g)t}-x_{l^*(i,j,\widetilde{g})t}\right) \\
&\qquad +\frac{1}{T}\sum_{t=1}^{T}(v_{it}-v_{jt})\left(v_{k^*(i,j,g)t}-v_{l^*(i,j,\widetilde{g})t}\right) \\
&\qquad +\frac{1}{T}\sum_{t=1}^{T}\left(\beta^0-\widehat{\beta}^1\right)'(x_{it}-x_{jt})\left(\beta^0-\widehat{\beta}^1\right)'\left(x_{k^*(i,j,g)t}-x_{l^*(i,j,\widetilde{g})t}\right) \\
&\qquad +\frac{1}{T}\sum_{t=1}^{T}(v_{it}-v_{jt})\left(\beta^0-\widehat{\beta}^1\right)'\left(x_{k^*(i,j,g)t}-x_{l^*(i,j,\widetilde{g})t}\right) \\
&\qquad +\frac{1}{T}\sum_{t=1}^{T}(v_{k^*(i,j,g)t}-v_{l^*(i,j,\widetilde{g})t})\left(\beta^0-\widehat{\beta}^1\right)'(x_{it}-x_{jt})\leq 2\lambda_2\Bigg\} \\
&\quad = \mathbf{1}\left\{\mathcal{E}_{NT}\right\}\times \max_{\substack{(g,\widetilde{g})\in\left\{1,\ldots,G^0\right\}^2 \\ g\neq\widetilde{g}}} \mathbf{1}\left\{A_T(i,j,g,\widetilde{g})\leq 2\lambda_2\right\}, \qquad \text{(A.4)}
\end{aligned}
$$

where $A_T(i,j,g,\widetilde{g})$ is defined implicitely. I now define

$$
\begin{aligned}
B_T(i,j,g,\widetilde{g}) = \Bigg| & A_T(i,j,g,\widetilde{g})-\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)^2 \\
& -\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)\left(v_{it}-v_{jt}+v_{k^*(i,j,g)t}-v_{l^*(i,j,\widetilde{g})t}\right) \\
& -\frac{1}{T}\sum_{t=1}^{T}(v_{it}-v_{jt})\left(v_{k^*(i,j,g)t}-v_{l^*(i,j,\widetilde{g})t}\right)\Bigg|.
\end{aligned}
$$

Let $\overline{a} = \sup_{a \in \mathcal{A}} |a|$. It is easy to show using the Cauchy-Schwarz inequality that

$$\begin{aligned}
B_T(i,j,g,\widetilde{g}) \leq \left\|\widehat{\beta}^1 - \beta^0\right\| \Bigg\{ & \frac{2\overline{a}}{T} \sum_{t=1}^{T} \left(\left\|x_{it}\right\| + \left\|x_{jt}\right\| + \left\|x_{k^*(i,j,g)t}\right\| + \left\|x_{l^*(i,j,\widetilde{g})t}\right\| \right) \\
& + \frac{4\left\|\widehat{\beta}^1 - \beta^0\right\|}{T} \sum_{t=1}^{T} \left(\left\|x_{it}\right\|^2 + \left\|x_{jt}\right\|^2 + \left\|x_{k^*(i,j,g)t}\right\|^2 + \left\|x_{l^*(i,j,\widetilde{g})t}\right\|^2 \right) \\
& + \left(\sqrt{\frac{1}{T}\sum_{t=1}^{T} v_{it}^2} + \sqrt{\frac{1}{T}\sum_{t=1}^{T} v_{jt}^2} \right) \sqrt{\frac{1}{T}\sum_{t=1}^{T} \left\|x_{k^*(i,j,g)t}\right\|^2 + \left\|x_{l^*(i,j,\widetilde{g})t}\right\|^2} \\
& + \left(\sqrt{\frac{1}{T}\sum_{t=1}^{T} v_{k^*(i,j,g)t}^2} + \sqrt{\frac{1}{T}\sum_{t=1}^{T} v_{l^*(i,j,\widetilde{g})t}^2} \right) \sqrt{\frac{1}{T}\sum_{t=1}^{T} \left\|x_{it}\right\|^2 + \left\|x_{jt}\right\|^2} \Bigg\}.
\end{aligned}$$

Because of Assumption 1(b), there exists $M^* > 0$ such that $\mathbb{E}(v_{it}^2) \leq M^*$ for all i, t. Let $\widetilde{M} > \max(M, \max(M^*, 1))$, where M is defined in Assumption 2(b) and let $\eta > 0$ such that

$$\eta \leq \min\left(1, \frac{c}{24\left(2\overline{a}4\sqrt{\widetilde{M}} + 8\widetilde{M} + 4\sqrt{2}\widetilde{M}\right)}\right). \tag{A.5}$$

Since $\lambda_1 \to 0$ as $N, T \to \infty$, for N, T sufficiently large, $\left\|\widehat{\beta}^1 - \beta^0\right\| \leq \eta$ on $\mathcal{E}_{NT}$. Using the Cauchy-Schwarz inequality and since $\eta \leq 1$, I obtain

$$\begin{aligned}
\mathbf{1}\left\{\mathcal{E}_{NT}\right\} B_T(i,j,g,\widetilde{g}) \leq \eta \Bigg\{ & 2\overline{a} \sqrt{\frac{1}{T}\sum_{t=1}^{T} \left\|x_{it}\right\|^2 + \left\|x_{jt}\right\|^2 + \left\|x_{k^*(i,j,g)t}\right\|^2 + \left\|x_{l^*(i,j,\widetilde{g})t}\right\|^2} \\
& + \frac{4}{T} \sum_{t=1}^{T} \left(\left\|x_{it}\right\|^2 + \left\|x_{jt}\right\|^2 + \left\|x_{k^*(i,j,g)t}\right\|^2 + \left\|x_{l^*(i,j,\widetilde{g})t}\right\|^2 \right) \\
& + \left(\sqrt{\frac{1}{T}\sum_{t=1}^{T} v_{it}^2} + \sqrt{\frac{1}{T}\sum_{t=1}^{T} v_{jt}^2} \right) \sqrt{\frac{1}{T}\sum_{t=1}^{T} \left\|x_{k^*(i,j,g)t}\right\|^2 + \left\|x_{l^*(i,j,\widetilde{g})t}\right\|^2} \\
& + \left(\sqrt{\frac{1}{T}\sum_{t=1}^{T} v_{k^*(i,j,g)t}^2} + \sqrt{\frac{1}{T}\sum_{t=1}^{T} v_{l^*(i,j,\widetilde{g})t}^2} \right) \sqrt{\frac{1}{T}\sum_{t=1}^{T} \left\|x_{it}\right\|^2 + \left\|x_{jt}\right\|^2} \Bigg\} \\
=: {} & C_T(i,j,g,\widetilde{g}).
\end{aligned}$$

Plugging this bound into (A.4), I obtain

$$
\begin{aligned}
&\mathbf{1}\left\{Z_{1NT}(i,j)=1,\mathcal{E}_{NT}\right\} \\
&\leq \max_{\substack{(g,\widetilde{g})\in\left\{1,\ldots,G^0\right\}^2 \\ g\neq\widetilde{g}}} \mathbf{1}\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)^2\right. \\
&+\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)\left(v_{it}-v_{jt}+v_{k^*(i,j,g)t}-v_{l^*(i,j,\widetilde{g})t}\right) \\
&\left.+\frac{1}{T}\sum_{t=1}^{T}(v_{it}-v_{jt})\left(v_{k^*(i,j,g)t}-v_{l^*(i,j,\widetilde{g})t}\right)\leq 2\lambda_2+C_T(i,j,g,\widetilde{g})\right\}.
\end{aligned}
$$

By the Cauchy-Schwarz inequality and because $\widetilde{M}\geq 1$, note the implication

$$
\frac{1}{T}\sum_{t=1}^{T}\left\|x_{it}\right\|^2\leq\widetilde{M}\implies\frac{1}{T}\sum_{t=1}^{T}\left\|x_{it}\right\|\leq\sqrt{\widetilde{M}}\leq\widetilde{M}.
$$

Using this result, the union bound, and some probability algebra, I obtain

$$
\begin{aligned}
&\Pr\left(Z_{1NT}(i,j)=1,\mathcal{E}_{NT}\right) \\
&\leq \sum_{\substack{(g,\widetilde{g})\in\left\{1,\ldots,G^0\right\}^2 \\ g\neq\widetilde{g}}} \Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)v_{it}\leq-\frac{c}{12}+2\lambda_2+\eta\left(2\overline{a}4\sqrt{\widetilde{M}}+8\widetilde{M}+4\sqrt{2}\widetilde{M}\right)\right) \\
&\quad+4G^0\left(G^0-1\right)\left[\sup_{g\neq\widetilde{g}}\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)^2\leq\frac{c}{2}\right)+\sup_{i\in\{1,\ldots,N\}}\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left\|x_{it}\right\|^2\geq\widetilde{M}\right)\right. \\
&\quad+\sup_{i\in\{1,\ldots,N\},g\neq\widetilde{g}}\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)v_{it}\right|\geq\frac{c}{12}\right)+\sup_{i\in\{1,\ldots,N\}}\Pr\left(\frac{1}{T}\sum_{t=1}^{T}v_{it}^2\geq\widetilde{M}\right) \\
&\quad\left.+\sup_{(i,j,k)\in\mathscr{P}_3(\{1,\ldots,N\})}\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}(v_{it}-v_{jt})v_{kt}\right|\geq\frac{c}{12}\right)\right],
\end{aligned}
\tag{A.6}
$$

First, I bound the terms with a supremum. By Assumption 1(c), it holds that $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0)^2]=c_{g,\widetilde{g}}>c$. So for T large enough, I have

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)^2\right]\geq\frac{2c}{3}.
$$

Applying Lemma B.5 in Bonhomme and Manresa (2015) to $z_t=(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0)^2-\mathbb{E}[(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0)^2]$, which satisfies appropriate mixing and tail conditions by Assumption 1(b) and (d), and taking $z=c/6$ yields, as T tends to infinity,

$$
\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)^2\leq\frac{c}{2}\right)=o(T^{-\delta}),
\tag{A.7}
$$

uniformly across g and $\widetilde{g}$. Similarly, applying Lemma B.5 to $z_t = v_{it}^2 - \mathbb{E}(v_{it})^2$ and taking $z = \widetilde{M} - M^*$ yields

$$\Pr\left(\frac{1}{T}\sum_{t=1}^{T} v_{it}^2 \geq \widetilde{M}\right) = o(T^{-\delta}), \tag{A.8}$$

uniformly across units i. Note that $\{v_{it}^2\}_t$ is strongly mixing as $\{v_{it}\}$ is strongly mixing by Assumption 1(c). By Assumption 1(d), the process $\left\{(\alpha_{\widetilde{g}t}^0 - \alpha_{gt}^0)v_{it}\right\}_t$ has zero mean, and is strongly mixing with faster-than-polynomial decay rate. Moreover, for all i, t and $m > 0$,

$$\Pr\left(\left|\left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right) v_{it}\right| > m\right) \leq \Pr\left(|v_{it}| > \frac{m}{2\overline{a}}\right),$$

so $\left\{(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0)v_{it}\right\}_t$ also satisfies the tail condition of Assumption 1(b), albeit with a different constant $b' > 0$ instead of $b > 0$. Lastly, applying Lemma B.5 from Bonhomme and Manresa (2015) again with $z_t = (\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0)v_{it}$ and taking $z = c/12$ yields

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right) v_{it}\right| \geq \frac{c}{12}\right) = o(T^{-\delta}) \tag{A.9}$$

uniformly across i, g, and $\widetilde{g}$. An analogous reasoning yields

$$\sup_{(i,j,k)\in\mathscr{P}_3(\{1,\ldots,N\})} \Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}(v_{it} - v_{jt})v_{kt}\right| \geq \frac{c}{12}\right) = o(T^{-\delta}). \tag{A.10}$$

Finally, for T sufficiently large, $\lambda_2 \leq c/72$ and a similar reasoning yields

$$\begin{aligned}
&\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right) v_{it} \leq -\frac{c}{12} + 2\lambda_2 + \eta\left(2\overline{a}4\sqrt{\widetilde{M}} + 8\widetilde{M} + 4\sqrt{2}\widetilde{M}\right)\right) \\
&\leq \Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right) v_{it} \leq -\frac{c}{72}\right) \\
&= o(T^{-\delta}),
\end{aligned} \tag{A.11}$$

uniformly across $g, \widetilde{g}$, where I have used the value of η given in (A.5). Combining (A.6)-(A.11) and using Assumption 2(b) yields

$$\sup_{(i,j)\in\{1,\ldots,N\}^2} \Pr\left(Z_{1NT}(i,j) = 1, \mathcal{E}_{NT}\right) = G^0(1 - G^0) \times o_p(T^{-\delta}) = o_p(T^{-\delta}),$$

i.e., (A.3) for $\ell = 1$ holds.

2. Second, I show (A.3) for $\ell = 2$. I now have

$$
\begin{aligned}
&\mathbf{1}\left\{Z_{2NT}(i,j)=1,\mathcal{E}_{NT}\right\} \\
&\quad = \mathbf{1}\left\{\mathcal{E}_{NT}\right\}\mathbf{1}\left\{\max_{k\in\{1,\dots,N\}\setminus\{i,j\}}\left|\frac{1}{T}\sum_{t=1}^{T}(\widehat{v}_{it}-\widehat{v}_{jt})\widehat{v}_{kt}\right|>\lambda_2\right\}\mathbf{1}\left\{g_i^0=g_j^0\right\} \\
&\quad \leq \mathbf{1}\left\{\mathcal{E}_{NT}\right\}\mathbf{1}\left\{\max_{k\in\{1,\dots,N\}\setminus\{i,j\}}\left|\frac{1}{T}\sum_{t=1}^{T}(v_{it}-v_{jt})v_{kt}+\frac{1}{T}\sum_{t=1}^{T}(v_{it}-v_{jt})\alpha_{kt}^0\right.\right. \\
&\qquad +\frac{1}{T}\sum_{t=1}^{T}\left(\beta^0-\widehat{\beta}^1\right)'(x_{it}-x_{jt})\left(\beta^0-\widehat{\beta}^1\right)'x_{kt} \\
&\qquad +\frac{1}{T}\sum_{t=1}^{T}(v_{it}-v_{jt})\left(\beta^0-\widehat{\beta}^1\right)'x_{kt}+\frac{1}{T}\sum_{t=1}^{T}\alpha_{kt}^0\left(\beta^0-\widehat{\beta}^1\right)'(x_{it}-x_{jt}) \\
&\qquad \left.\left.+\frac{1}{T}\sum_{t=1}^{T}v_{kt}\left(\beta^0-\widehat{\beta}^1\right)'(x_{it}-x_{jt})\right|>\lambda_2\right\}.
\end{aligned}
$$

By the union bound, the triangle inequality, and the Cauchy-Schwarz inequality,

$$
\begin{aligned}
&\Pr\left(Z_{2NT}(i,j)=1,\mathcal{E}_{NT}\right) \\
&\quad \leq (N-2)\sup_{(i,j,k)\in\{1,\dots,N\}^3}\left\{\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}(v_{it}-v_{jt})v_{kt}\right|>\frac{\lambda_2}{10}\right)\right. \\
&\qquad +\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left\|x_{it}\right\|^2+\left\|x_{jt}\right\|^2+\left\|x_{kt}\right\|^2>\frac{\lambda_2}{10\times 4K^2\lambda_1^2}\right) \\
&\qquad +4\Pr\left(\sqrt{\frac{1}{T}\sum_{t=1}^{T}v_{it}^2}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left\|x_{kt}\right\|^2}>\frac{\lambda_2}{10K\lambda_1}\right) \\
&\qquad +2\Pr\left(\sqrt{\frac{1}{T}\sum_{t=1}^{T}v_{it}^2}>\frac{\lambda_2}{10K\lambda_1\times\overline{a}}\right) \\
&\qquad \left.+2\Pr\left(\sqrt{\frac{1}{T}\sum_{t=1}^{T}v_{it}^2}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left\|x_{kt}\right\|^2}>\frac{\lambda_2}{10K\lambda_1\times\overline{a}}\right)\right\}.
\end{aligned}
$$

Under the strong mixing and tail conditions given by Assumptions 1(b) and 1(d), and the rate conditions on the ratios λ_2/λ_1^2 and λ_2/λ_1 given by Assumptions 2(a), all noninitial probabilities in the above expression can be shown to be $o(T^{-\delta})$ for all $\delta > 0$ by similar arguments as in Step 1. For the first probability, a close inspection of the proof of Lemma B.5 in Bonhomme and Manresa (2015) reveals that, by taking $z_t = (v_{it} - v_{jt})v_{kt}$ and $z = \lambda_2/6$, and because $\lambda_2 \gtrsim T^{-\kappa}$ as $T \to \infty$, for T sufficiently large,

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}(v_{it}-v_{jt})v_{kt}\right|\geq\frac{\lambda_2}{10}\right)\leq 4\left(1+\frac{T^{1/2-2\kappa}}{C_1}\right)^{-(1/2)T^{1/2}} + C_2T^{\kappa}\exp\left(-C_3\left(T^{(1/2-\kappa)/C_4}\right)\right), \tag{A.12}$$

where C_1, C_2, C_3, and C_4 are positive constants that do not depend on i. Since $\kappa < 1/2$, the upper bound is $o_p(T^{-\delta})$ for all $\delta > 0$. This shows (A.3) for $\ell = 2$. $\square$

The proof of Proposition 3.1 is complete.

A.2 Proof of Corollary 3.2

Let $\widetilde{\beta}$ and $(\widetilde{\alpha}_{11}, ..., \widetilde{\alpha}_{G^0T})'$ denote the infeasible oracle estimators computed using the pooled OLS regression of y_{it} on x_{it} and the interactions of group and time indicators $\mathbf{1}\{g_i^0 = 1\}$, ..., $\mathbf{1}\{g_i^0 = G^0\}$, $\mathbf{1}\{t = 1\}$, ..., $\mathbf{1}\{t = T\}$. By the same reasoning as in section S.A.1. in Bonhomme and Manresa (2015)'s Supplemental Material, I have

$$\sqrt{NT}(\widetilde{\beta}-\beta^0)\xrightarrow{d}\mathcal{N}\left(0,\Sigma_\beta^{-1}\Omega_\beta\Sigma_\beta^{-1}\right), \tag{A.13}$$

and, for all $(g, t) \in \{1, ..., G^0\} \times \{1, ..., T\}$,

$$\sqrt{N}(\widetilde{\alpha}_{gt}-\alpha_{gt}^0)\xrightarrow{d}\mathcal{N}\left(0,\frac{\omega_{gt}}{\pi_g^2}\right). \tag{A.14}$$

Without loss of generality, I assume that the chosen labels match the true group labeling. By Proposition 3.1, for all $(g, t) \in \{1, ..., G^0\} \times \{1, ..., T\}$,

$$\begin{aligned}\Pr\left(\{\widehat{\alpha}_{gt}\neq\widetilde{\alpha}_{gt}\}\cup\left\{\widehat{\beta}\neq\widetilde{\beta}\right\}\right)&\leq\Pr\left(\widehat{G}\neq G^0\right)+\Pr\left(\max_{i\in\{1,...,N\}}\left|\widehat{g}_i-g_i^0\right|>0\right)\\&=o(1)+o(1)\\&=o(1).\end{aligned}$$

Then, (3.5) follows from

$$\begin{aligned}&\left|\Pr\left(\sqrt{N}(\widehat{\alpha}_{gt}-\alpha_{gt}^0)\leq a\right)-\Pr\left(\sqrt{N}(\widetilde{\alpha}_{gt}-\alpha_{gt}^0)\leq a\right)\right|\\&\leq\left|\Pr\left(\sqrt{N}(\widehat{\alpha}_{gt}-\alpha_{gt}^0)\leq a,\sqrt{N}(\widetilde{\alpha}_{gt}-\alpha_{gt}^0)>a\right)\right|\\&\quad+\left|\Pr\left(\sqrt{N}(\widehat{\alpha}_{gt}-\alpha_{gt}^0)>a,\sqrt{N}(\widetilde{\alpha}_{gt}-\alpha_{gt}^0)\leq a\right)\right|\\&\leq\quad\Pr\left(\widehat{\alpha}_{gt}\neq\widetilde{\alpha}_{gt}\right)+\Pr\left(\widehat{\alpha}_{gt}\neq\widetilde{\alpha}_{gt}\right)=o(1).\end{aligned}$$

for any $a > 0$. (3.4) follows from a similar argument.

References

Ando, T., and J. Bai (2022): “Large-scale generalized linear longitudinal data models with grouped patterns of unobserved heterogeneity,” Discussion paper.

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.

Auerbach, E. (2022): “Identification and Estimation of a Partially Linear Regression Model Using Network Data,” *Econometrica*, 90(1), 347–365.

Bai, J. (2009): “Panel Data Models With Interactive Fixed Effects,” *Econometrica*, 77(4), 1229–1279.

Bai, J., and S. Ng (2002): “Determining the Number of Factors in Approximate Factor Models,” *Econometrica*, 70(1), 191–221.

——— (2019): “Rank regularized estimation of approximate factor models,” *Journal of Econometrics*, 212(1), 78–96, Big Data in Dynamic Predictive Econometric Modeling.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): “Simultaneous analysis of Lasso and Dantzig selector,” *The Annals of Statistics*, 37(4), 1705 – 1732.

Bonhomme, S., and E. Manresa (2015): “Grouped Patterns of Heterogeneity in Panel Data,” *Econometrica*, 83(3), 1147–1184.

Brownlees, C., G. S. Guðmundsson, and G. Lugosi (2022): “Community Detection in Partial Correlation Network Models,” *Journal of Business & Economic Statistics*, 40(1), 216–226.

Charbonneau, K. B. (2017): “Multiple fixed effects in binary response panel data models,” *The Econometrics Journal*, 20(3), S1–S13.

Chetverikov, D., and E. Manresa (2021): “Spectral and Post-Spectral Estimators for Grouped Panel Data Models,” Discussion paper.

Dzemski, A., and R. Okui (2018): "Confidence set for group membership," .

Graham, B. S. (2017): "An Econometric Model of Network Formation With Degree Heterogeneity," *Econometrica*, 85(4), 1033–1063.

Jochmans, K. (2017): "Two-Way Models for Gravity," *The Review of Economics and Statistics*, 99(3), 478–485.

Krasnokutskaya, E., K. Song, and X. Tang (2022): "Estimating unobserved individual heterogeneity using pairwise comparisons," *Journal of Econometrics*, 226(2), 477–497.

Lovász, L. (2012): *Large Networks and Graph Limits.*, vol. 60 of *Colloquium Publications.* American Mathematical Society.

Moon, H., and M. Weidner (2017): "Dynamic Linear Panel Regression Models With Interactive Fixed Effects," *Econometric Theory*, 33(1), 158–195.

Moon, H. R., and M. Weidner (2015): "Linear Regression for Panel With Unknown Number of Factors as Interactive Fixed Effects," *Econometrica*, 83(4), 1543–1579.

——— (2019): "Nuclear Norm Regularized Estimation of Panel Regression Models," .

Mugnier, M. (2022): "Supplemental Material for A Simple and Computationally Trivial Estimator for Grouped Fixed Effects Models," https://martinmugnier.github.io/research.

Neyman, J., and E. L. Scott (1948): "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16(1), 1–32.

Ng, A., M. Jordan, and Y. Weiss (2002): "On Spectral Clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, ed. by T. Dietterich, S. Becker, and Z. Ghahramani, vol. 14. MIT Press.

Pollard, D. (1981): "Strong Consistency of K-Means Clustering," *The Annals of Statistics*, 9(1), 135 – 140.

——— (1982): "A Central Limit Theorem for k-Means Clustering," *The Annals of Probability*, 10(4), 919 – 926.

von Luxburg, U. (2007): "A Tutorial on Spectral Clustering," .

Wang, W., and L. Su (2021): "Identifying Latent Group Structures in Nonlinear Panels," *Journal of Econometrics*, 220(2), 272–295, Annals Issue: Celebrating 40 Years of Panel Data Analysis: Past, Present and Future.

Wooldridge, J. (2010): *Econometric Analysis of Cross Section and Panel Data, second edition*, Econometric Analysis of Cross Section and Panel Data. MIT Press.

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

Zeleneev, A. (2020): "Identification and Estimation of Network Models with Nonparametric Unobserved Heterogeneity," Discussion paper.

Zhang, Y., E. Levina, and J. Zhu (2017): "Estimating network edge probabilities by neighbourhood smoothing," *Biometrika*, 104(4), 771–783.