Working Paper in Economics No. 790

Convergence rate of estimators of clustered panel models with misclassification

Andreas Dzemski and Ryo Okui

Department of Economics, August 2020

ISSN 1403-2473 (Print)
ISSN 1403-2465 (Online)

UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

Convergence rate of estimators of clustered panel models with misclassification*

Andreas Dzemski[†] and Ryo Okui[‡]

August 11, 2020

Abstract

We study *kmeans* clustering estimation of panel data models with a latent group structure and N units and T time periods under long panel asymptotics. We show that the group-specific coefficients can be estimated at the parametric root NT rate even if error variances diverge as $T \to \infty$ and some units are asymptotically misclassified. This limit case approximates empirically relevant settings and is not covered by existing asymptotic results.

Keywords: Panel data, latent grouped structure, clustering, kmeans, convergence rate, misclassification.

JEL codes: C23, C33, C38

Mathematical Subjects Classification (2010): 62H30, 62H12

Declarations of interest: none

1 Introduction

Panel models can account for unobserved heterogeneity by dividing units into a finite number of latent groups and allowing a unit's coefficients to be group-specific (Bonhomme and Manresa 2015; Su, Shi, and Phillips 2016; Vogt and Linton 2017; Wang, Phillips, and Su 2018; Okui and Wang 2020). Estimators of such models simultaneously estimate group memberships and group-specific coefficients. For example, Bonhomme and Manresa (2015) propose a *kmeans*-type estimator and Su, Shi, and Phillips (2016) propose the CLasso estimator that is based on solving a penalized regression program. These two and other related estimators are justified under a long panel asymptotic framework that sends both the number of units N and the number of

[*] The research presented in this paper started as Appendix G of a previous version of Dzemski and Okui (2019) and has been substantially extended. The relevant results are removed from Dzemski and Okui (2019). Okui gratefully acknowledges the financial support of the School of Social Sciences and a New Faculty Startup Grant at Seoul National University and from the Housing and Commercial Bank Economic Research Fund in the Institute of Economic Research at Seoul National University. Dzemski gratefully acknowledges financial support from Jan Wallanders och Tom Hedelius samt Tore Browaldhs stiftelse grant P19-0079. A part of this research was done while Okui was at NYU Shanghai.

[†] Department of Economics, University of Gothenburg, P.O. Box 640, SE-405 30 Gothenburg, Sweden. Email: andreas.dzemski@economics.gu.se

[‡] Department of Economics and the Institute of Economic Research, Seoul National University, Building 16, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, South Korea. Department of Economics, University of Gothenburg, P.O. Box 640, SE-405 30 Gothenburg, Sweden. Email: okuiryo@snu.ac.kr;

time periods T to infinity. Existing theoretical results show that coefficients that are group-specific and time invariant can be estimated at a root NT rate, i.e., at the parametric rate. In this paper we show that the parametric rate can be obtained even if some units have a positive probability of being misclassified in the limit. This limit case is highly relevant in practice since it is common to misclassify at least some units in empirical applications (Bonhomme, Lamadon, and Manresa 2019). However, existing results do not apply in such settings.

Existing asymptotic results for linear panel models assume that the variance of the error term is universally bounded. From this assumption, it can then be shown that group memberships can be estimated uniformly consistently, i.e., the probability of misclassifying one or more units vanishes as $N, T \to \infty$. This implies that the rate at which group-specific coefficients can be estimated is the same as under a known group structure and is therefore equal to the parametric rate.

However, the assumption of a universal bound on the variance of the error term may not reflect real circumstances. It implies that the asymptotic limit as $T \to \infty$ prescribes that, for each unit, the level of statistical noise is negligible when compared to the number of observed time periods. This is not characteristic of typical empirical applications. The number of observed time periods is often rather small and, at least for some units, statistical noise plays an important role in determining the outcome.

In this paper, we extend previous theoretical results to a heteroscedastic setting in which units are endowed with unit-specific error variances $\sigma_1^2, \ldots, \sigma_N^2$. A unit i with small σ_i is easy to classify, whereas a unit i with large σ_i is difficult to classify. The individual error variances may depend on N and T and may diverge as $T \to \infty$. We expect our asymptotic framework to be a more faithful approximation of the finite sample behavior of the estimators than the conventional framework.

For *kmeans*-estimation, we show that uniform consistency of group memberships holds provided that the unit-specific error variances do not diverge too fast. Units i for which σ_i diverges too fast are potentially misclassified in the limit. However, if the proportion of such potentially misclassified units is sufficiently small then it is still possible to estimate the group-specific coefficients at a root NT rate.

Pollard (1981), Pollard (1982), and Bonhomme and Manresa (2015) consider panel models with fixed T and estimate cluster-specific coefficients. They show that the cluster-specific coefficients converge to a pseudo-true value at rate root N even though units are misclassified in the limit with positive probability. Their setting and results are distinct from ours. We consider long panel asymptotics under which true rather than pseudo-true cluster-specific coefficients can be identified and estimated at a root NT rate.

We prove our results for a simple linear panel model with group-specific intercepts and focus on estimation by least squares (equivalent to *kmeans*). By focusing on this simple model we are able to derive our results under interpretable and intuitive conditions on the structure of heteroscedasticity. While we think that our argument can be extended to more general regression models with group-specific coefficients, we believe that such an exercise would impose more involved assumptions and would not be as instructive about the mechanisms that allow root NT-consistency to arise despite of diverging error variances and possibly misclassified units.

Bonhomme and Manresa (2015) conduct a simulation experiment that is calibrated to their empirical application. They find that the group-specific coefficients are estimated precisely, even though it is likely that one or more units are misclassified. Existing theoretical results about the rate of consistency of the group-specific coefficients cannot explain this phenomenon as they do not apply in the presence of misclassification. We fill this gap in the literature by showing that uniform consistency is sufficient but not necessary for precise estimation of the group-specific coefficients.

2 Setting

The units $i = 1, \ldots, N$ are partitioned into G groups. The set of all groups is $\mathbb{G} = \{1, \ldots, G\}$ and unit i belongs to group $g_i^0 \in \mathbb{G}$. For units in group $g \in \mathbb{G}$ the mean outcome in each period is given by μ_g. At time $t = 1, \ldots, T$ we observe the scalar outcome y_{it} generated by

$$y_{it} = \mu_{g_i^0} + \sigma_i v_{it},$$

where v_{it} is a noise term with variance one. Let Γ denote the space of possible group assignments $\mathbf{g} = (g_1, \ldots, g_N)$ and let $\mathcal{M}$ denote the space of possible group-specific means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_G)$. The true group assignment $\mathbf{g}^0 \in \Gamma$ and the true group-specific mean $\boldsymbol{\mu}^0 \in \mathcal{M}$ are unknown parameters and are estimated.

We consider *kmeans*-type estimation as suggested in Bonhomme and Manresa (2015). The objective function for estimation is defined on $\Gamma \times \mathcal{M}$ and is given by

$$Q_{N,T}(\mathbf{g}, \boldsymbol{\mu}) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - \mu_{g_i}\right)^2 .$$

The estimator is defined as $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{g}}) = \arg\min_{\boldsymbol{\mu} \in \mathcal{M}, \boldsymbol{g} \in \Gamma} Q_{N,T}(\mathbf{g}, \boldsymbol{\mu})$. In practice, the estimator is computed by the iterative *kmeans* procedure. We start with an initial group membership structure $\mathbf{g}^{(0)}$ and then iterate $\boldsymbol{\mu}$ and $\mathbf{g}$ such that the s-th iteration sets $\boldsymbol{\mu}^{(s)} = \arg\min_{\boldsymbol{\mu} \in \mathcal{M}} Q_{N,T}(\mathbf{g}^{(s-1)}, \boldsymbol{\mu})$ and $\mathbf{g}^{(s)} = \arg\min_{\boldsymbol{g} \in \Gamma} Q_{N,T}(\mathbf{g}, \boldsymbol{\mu}^{(s)})$ until convergence. Since the iteration may converge to a local minimum we re-start the procedure from many initial values for $\mathbf{g}$.

3 Main results

We consider asymptotic sequences under which $N, T \to \infty$ and

$$\frac{(\log T)\sqrt{\log N}}{\sqrt{T}} = o(1). \tag{1}$$

We treat $(\sigma_1, \ldots, \sigma_N)$ and $\mathbf{g}^0$ as unobserved deterministic parameters.

We first state sufficient conditions for consistent estimation of $\boldsymbol{\mu}^0$.

Assumption 1. *i) $\{v_{it}\}_{t=1}^T$ is an independent sequence with $\mathbb{E}v_{it} = 0$ and $\mathbb{E}v_{it}^2 = 1$.*

ii) The average error variance satisfies $N^{-1} \sum_{i=1}^N \sigma_i^2 = o(T)$.

iii) There is a bounded set $\mathcal{M} \subset \mathbb{R}^G$ such that $\boldsymbol{\mu}^0 \in \mathcal{M}$.

iv) There is a positive constant M_G such that

$$\min_{g \in \mathbb{G}} \min_{h \in \mathbb{G} \setminus \{g\}} \left| \mu_g^0 - \mu_h^0 \right| > M_G.$$

v) For all $g \in \mathbb{G}$, $N^{-1} \sum_{i=1}^N 1(g_i^0 = g) \geq q_{\min}$.

Part i) imposes independence of the error term over time. Using this assumption we obtain asymptotic results under simply conditions on between-unit heteroscedasticity. The assumption can be relaxed to allow for weak serial correlation at the expense of conditions on heteroscedasticity that are more difficult to interpret. Part ii) states that the average error variance increases at a slower rate than T. This assumption ensures that, as $T \to \infty$, the additional information from observing more time periods is not undone by an increased noisiness of the signal. Part iii) is a standard regularity assumption. Part iv) requires that the group-specific means are distinct (group separation). Part v) ensures that the effective sample size that can be used to estimate the group-specific mean grows at the same asymptotic rate for all groups.

Assumption 1 does not restrict cross-sectional dependence. Assumption 3 below limits the amount of cross-sectional dependence and is required for our result on NT-convergence of the group-specific parameters, but not any of our intermediate results.

The grouped model is invariant to a relabeling of the groups and the vector of group-specific means $\boldsymbol{\mu}^0$ is therefore only identified up to a re-ordering of its components. The following result states that the identified set is consistently estimated.

Lemma 1 (Consistency of group-specific means)**.** *Suppose that Assumption 1 holds. Then, there is a (possibly random) permutation function $\pi : \mathbb{G} \to \mathbb{G}$ such that for all $\epsilon > 0$*

$$\lim_{N,T \to \infty} P\left(\max_{g \in \mathbb{G}} \left| \hat{\mu}_{\pi(g)} - \mu_g^0 \right| > \epsilon \right) = 0.$$

Similarly to related results in the literature (e.g. Bonhomme and Manresa 2015), proving this result does not require establishing that group memberships are consistently estimated for all units. In Theorem 1 below, we strengthen the result to root NT convergence under weaker assumptions on heteroscedasticity than are commonly assumed in the literature.

The subsets of units for which we can guarantee that group memberships are uniformly consistently estimated is given by

$$\mathcal{I}_{N,T} = \left\{ i \in \{1, \ldots, N\} : \sigma_i \leq \frac{M_G}{140} \sqrt{\frac{T}{\log N}} \right\}. \tag{2}$$

For the units in $\mathcal{I}_{N,T}$ the error variances are allowed to diverge but only at rate $\sqrt{T / \log N}$. Controlling the rate of divergence is necessary to ensure that observing additional time periods adds enough information to estimate group memberships precisely. What rates of divergence are permissible is determined by bounds on the tail of the error distribution. The error term of our panel model is given by $\sigma_i v_{it}$. We assume that v_{it} is a sub-exponential random variable.

Under this assumption, new observations add information at the usual parametric rate root T and the price of uniformity is root $\log N$.

Assumption 2 (Sub-exponential errors)**.** *There are positive constants ν, α such that*

$$\max_{1\leq i\leq N} \max_{1\leq t\leq T} \mathbb{E}\exp(\lambda\,|v_{it}|) \leq \exp\left(\frac{\lambda^2\nu^2}{2}\right) \qquad \text{for all } \lambda > 0 \text{ such that } \lambda < \frac{1}{\alpha}.$$

In addition to errors that are Gaussian and sub-Gaussian (conditional on σ_i) this assumption allows also for certain "fat-tailed" distributions such as Poisson or chi-squared. It is possible to relax this assumption and allow for distributions with even heavier tails, but only at the expense of a different rate condition in (3) that is more difficult to state and to interpret. In our setting, misclassification can occur even for moderate realizations of v_{it} if σ_i is sufficiently large. Therefore, misclassification does not hinge on heavy tails of v_{it} and is not ruled out or limited by Assumption 2.

The following lemma states that group membership is estimated consistently uniformly over all units in $\mathcal{I}_{N,T}$.

Lemma 2. *Suppose that Assumptions 1 and 2 hold. Then, there exists a (possibly random) permutation function $\pi : \mathbb{G} \to \mathbb{G}$ such that*

$$\lim_{N,T\to\infty} P\left(\sup_{i\in\mathcal{I}_{N,T}} \left|\pi(\hat{g}_i) - g_i^0\right| > 0\right) \to 0.$$

This lemma extends existing results in the literature that are derived under the assumption that $\max_{1\leq i\leq N}\sigma_i^2$ is bounded in which case $\mathcal{I}_{N,T} = \{1,\ldots,N\}$ eventually. Lemma 2 shows that uniform consistency over all units can be obtained even if the error variance σ_i^2 diverges for some or all units. In this case, all unit-specific error variances must diverge at most at the rate given in (2) and the average error variance must diverge at most at the rate given in Assumption 1ii).

We study the asymptotic behavior of $\hat{\boldsymbol{\mu}}$ without requiring that all units are contained in $\mathcal{I}_{N,T}$ and therefore guaranteed to be estimated consistently. The idea of Theorem 1 below is that units that are not in $\mathcal{I}_{N,T}$ do not affect the asymptotic distribution provided that there are sufficiently few of them.

Let $\mathcal{I}_{N,T}^{\mathsf{c}} = \{1,\ldots,N\}\setminus\mathcal{I}_{N,T}$ and write $\#A$ to denote the cardinality of a set A. We assume

$$\frac{\#\mathcal{I}_{N,T}^{\mathsf{c}}}{N}\max\left\{\sqrt{NT}, \sqrt{N\frac{1}{\#\mathcal{I}_{N,T}^{\mathsf{c}}}\sum_{i\in\#\mathcal{I}_{N,T}^{\mathsf{c}}}\sigma_i^2}\right\} = o(1). \tag{3}$$

Existing theoretical results cover only settings under which no units are potentially misclassified in the asymptotic limit, i.e., $\#\mathcal{I}_{N,T}^{\mathsf{c}} = 0$. In this case (3) is trivially satisfied. Our result allows $\#\mathcal{I}_{N,T}^{\mathsf{c}} \neq 0$ provided that the proportion of possibly misclassified units $\#\mathcal{I}_{N,T}^{\mathsf{c}}/N$ vanishes at a sufficiently fast rate. The rate in the first component of the max ensures that units in $\mathcal{I}_{N,T}^{\mathsf{c}}$ asymptotically do not affect the mean of $\hat{\boldsymbol{\mu}}$. The rate of the second component in the max ensures that units in $\mathcal{I}_{N,T}^{\mathsf{c}}$ asymptotically do not affect the variance of $\hat{\boldsymbol{\mu}}$. By (2), the second component satisfies

$$\sqrt{N\frac{1}{\#\mathcal{I}^{\mathsf{c}}_{N,T}}\sum_{i\in\#\mathcal{I}^{\mathsf{c}}_{N,T}}\sigma_i^2} > \sqrt{NT}\frac{M_G}{140\sqrt{\log N}}.$$

This shows that the first component can dominate the second component at most at a root log N rate. Therefore, replacing the max in (3) by the second component gives a good approximation (up to order root $\log N$) of the required rate condition.

To state the assumption for asymptotic normality of $\hat{\mu}_g$, $g \in \mathbb{G}$, let $\mathcal{I}_{N,T}(g) = \{i \in \mathcal{I}_{N,T} : g_i^0 = g\}$ and

$$\tilde{N}_g = \#\{i \in \mathcal{I}_{N,T} : g_i^0 = g\}, \quad N_g = \#\{i \in 1,\dots,N : g_i^0 = g\}, \quad \hat{N}_g = \#\{i \in 1,\dots,N : \hat{g}_i = g\}.$$

Assumption 3. *i) Condition* (3) *is satisfied.*

ii) For each $g \in \mathbb{G}$ there are positive constants δ_g and q_g such that $N_g/N \to q_g$ and

$$\frac{1}{\tilde{N}_g}\sum_{i\in\mathcal{I}_{N,T}(g)}\sigma_i^2 + \frac{1}{\tilde{N}_g}\sum_{\substack{i,j\in\mathcal{I}_{N,T}(g)\\ i\neq j}}\sigma_i\sigma_j\operatorname{cov}(v_{i1},v_{j1}) \to \delta_g.$$

iii) We have

$$\frac{1}{\#\mathcal{I}_{N,T}}\sum_{i\in\mathcal{I}_{N,T}}\sigma_i^2 = O(\sqrt{T}) \quad \textit{and} \quad \frac{1}{\#\mathcal{I}_{N,T}}\sum_{i\in\mathcal{I}_{N,T}}\sigma_i^4 = O(NT).$$

iv) In addition,

$$\begin{aligned}\sum_{\substack{i,j,k\in\mathcal{I}_{N,T}\\ \{i\}\cap\{j\}\cap\{k\}=\emptyset}}\sigma_i\sigma_j\sigma_k\mathbb{E}[v_{i1}^2v_{j1}v_{k1}] &=O(N^2T),\\ \sum_{\substack{i,j,k,\ell\in\mathcal{I}_{N,T}\\ \{i\}\cap\{j\}\cap\{k\}\cap\{\ell\}=\emptyset}}\sigma_i\sigma_j\sigma_k\sigma_\ell\mathbb{E}[v_{i1}v_{j1}v_{k1}v_{\ell1}] &=O(N^2T).\end{aligned}$$

Part ii) ensures that the asymptotic variance of $\hat{\mu}_g$ converges. Part iii) imposes two conditions on the rate of divergence of the L_2 and the L_4 norm of $\{\sigma_i : i \in \mathcal{I}_{N,T}\}$. Under cross-sectional independence the first condition is implied by ii). The second condition is satisfied if $N\log^2 N/T \to \infty$. Part iv) limits the amount of cross-sectional dependence.

The following theorem guarantees root NT-consistency and asymptotic normality of $\hat{\mu}_g$.

Theorem 1. *Suppose that Assumptions 1–3 hold. Then, for $g \in \mathbb{G}$ as $N, T \to \infty$*

$$\sqrt{NT}\left(\hat{\mu}_{\pi(g)} - \mu_g^0\right) \xrightarrow{d} \mathcal{N}(0, q_g^{-1}\delta_g).$$

This result shows that root NT-consistency can be obtained even if some units are potentially misclassified in the limit. In addition, the error variance for the units that are consistently estimated need not be bounded. For root NT-consistency we require a stronger assumption on the average error variance than for the result on consistent estimation of group memberships in Lemma 2. Assumption 3ii) implies that the average error variance is bounded. In contrast, Lemma 2 allows the average error variance to diverge at a controlled rate.

4 Conclusion

We have shown that uniformly consistent estimation of group memberships is not a necessary condition of root NT estimation of time invariant group-specific parameters. The simple model with group-specific intercepts served our purpose of providing an example of a grouped panel model in which a root NT rate can be obtained even under misclassification in the limit. We are confident that similar results can be obtained for general linear panel regression, albeit under more involved conditions that may not be straightforward to interpret. We leave such extensions to future research. For scenarios where the amount of misclassification permitted by our assumption (3) is exceeded by only a sufficiently small margin, our proofs suggest that it is possible to obtain a convergence rate that is slower than root NT but faster than root N. This suggests a negative relationship between the difficulty of classifying individual units and the precision of the estimator of the vector of group-specific coefficients.

A Appendix: Mathematical proofs

Lemma 3. *Let $\mathcal{P}$ denote a class of probability measures that satisfy Assumption 2. Then*

$$\sup_{P\in\mathcal{P}} P\left(\max_{1\le i\le N}\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{it}\right| > 14\sqrt{\log N}\right) \le 3N^{-1}.$$

Proof. Fix a probability measure $P \in \mathcal{P}$ and let $\nu, \alpha > 0$ denote the parameters from Assumption 2. Let $\lambda^* > 0$ large enough that $\lambda^* < 1/\pi$ and $\exp(\nu^2(\lambda^*)^2/2) \le 2$. Define the Orlicz norm

$$\|v_{it}\|_{\psi_1} = \inf\left\{\eta > 0 : \mathbb{E}\left[\psi_1(|v_{it}|/\eta)\right] \le 1\right\}$$

with $\psi_1(t) = \exp(t) - 1$. By Assumption 2,

$$\max_{1\le i\le N}\max_{1\le t\le T}\mathbb{E}\exp(\lambda^*|v_{it}|) \le \exp\left(\nu^2(\lambda^*)^2/2\right) \le 2.$$

Defining $K = 1/\lambda^*$ this implies for all $1 \le i \le N$ and $1 \le t \le T$

$$\mathbb{E}\left[\exp\left(\frac{|v_{it}|}{K}\right) - 1\right] \le 1$$

and therefore

$$\|v_{it}\|_{\psi_1} = \inf\left\{\eta > 0 : \mathbb{E}\left[\exp\left(\frac{|v_{it}|}{\eta}\right) - 1\right] \le 1\right\} \le K.$$

Hence, $\max_{1\le i\le N}\max_{1\le t\le T}\|v_{it}\|_{\psi_1} \le K$. Applying Theorem 3.4 in Kuchibhotla and Chakrabortty

(2018) with $\alpha = 1$, $K_{n,q} = K$, $\Gamma_{n,q} = 1$ and $t = \log N$ yields

$$P\left(\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T} v_{it}\right| > 7\sqrt{\frac{2\log N}{T}} + \frac{C_1 K \log(2T)(2\log N)}{T}\right) \leq 3N^{-1}.$$

By Assumption 2,

$$\frac{C_1 K \log(2T)(2\sqrt{\log N})}{\sqrt{T}} = o(1)$$

and therefore

$$14\sqrt{\frac{\log N}{T}} > 7\sqrt{\frac{2\log N}{T}} + \frac{C_1 K \log(2T)(2\log N)}{T}.$$

□

Lemma 4. *Suppose that Assumption 1i)–iii) holds. Then, for all* $\epsilon > 0$

$$\lim_{N,T\to\infty} P\left(\sup_{\boldsymbol{g}\in\Gamma,\boldsymbol{\mu}\in\mathcal{M}}\left|Q_{N,T}(\mathbf{g},\boldsymbol{\mu}) - \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} u_{it}^2 + \frac{1}{N}\sum_{i=1}^{N}\left(\mu^0_{g_i^0} - \mu_{g_i}\right)^2\right| > \epsilon\right) = 0.$$

Proof. This proof is very similar to the proof of Lemma A.1 in Bonhomme and Manresa (2015). Expanding $Q_{N,T}$ gives

$$\begin{aligned} Q_{N,T}(\mathbf{g},\boldsymbol{\mu}) =& \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} u_{it}^2 + \frac{1}{N}\sum_{i=1}^{N}\left(\mu^0_{g_i^0} - \mu_{g_i}\right)^2 \\ &+ \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma_i v_{it}\left(\mu^0_{g_i^0} - \mu_{g_i}\right). \end{aligned}$$

By Cauchy-Schwarz

$$\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma_i v_{it}\left(\mu^0_{g_i^0} - \mu_{g_i}\right)\right|^2 \leq C_{\mathcal{M}}\frac{1}{N}\sum_{i=1}^{N}\left\{\left(\frac{\sigma_i^2}{T}\right)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{it}\right)^2\right\},$$

where $C_{\mathcal{M}}$ is a constant that depends on a bound on $\mathcal{M}$. Under the assumptions of the lemma,

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\{\left(\frac{\sigma_i^2}{T}\right)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{it}\right)^2\right\} = o(1).$$

Therefore, by Markov's inequality,

$$P\left(\frac{1}{N}\sum_{i=1}^{N}\left\{\left(\frac{\sigma_i^2}{T}\right)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{it}\right)^2\right\} > \epsilon\right) = o(1).$$

The conclusion follows. □

Lemma 5. *Suppose that Assumption 1i)–iii) holds. For each $\epsilon > 0$*

$$\lim_{N,T\to\infty} P\left(\frac{1}{N}\sum_{i=1}^{N}\left(\mu^0_{g_i^0} - \hat{\mu}_{\hat{g}_i}\right)^2 > \epsilon\right) = 0.$$

Proof. By definition,

$$Q_{N,T}(\hat{\mathbf{g}}, \hat{\boldsymbol{\mu}}) \leq Q_{N,T}(\mathbf{g}^0, \boldsymbol{\mu}^0).$$

Let $W_{N,T}$ denote a random variable such that for each $\epsilon > 0$

$$\lim_{N,T\to\infty} P(|W_{N,T}| > \epsilon) = 0.$$

Applying Lemma 4 to both sides of the inequality yields

$$\frac{1}{N}\sum_{i=1}^{N}\left(\mu^0_{g_i^0} - \hat{\mu}_{\hat{g}_i}\right)^2 \leq \frac{1}{N}\sum_{i=1}^{N}\left(\mu^0_{g_i^0} - \mu^0_{g_i^0}\right)^2 + W_{N,T}$$

and the conclusion follows. □

Proof of Lemma 1. This proof is very similar to the proof of Lemma B.3 in Bonhomme and Manresa (2015). By Lemma 5

$$\frac{1}{N}\sum_{i=1}^{N}\left(\mu^0_{g_i^0} - \hat{\mu}_{\hat{g}_i}\right)^2 = o_p(1).$$

Suppose that there is a constant $\epsilon > 0$ and $g \in \mathbb{G}$ such that for $N, T \to \infty$ satisfying (1)

$$\limsup_{N,T\to\infty} P\left(\min_{h\in\mathbb{G}} \left|\hat{\mu}_h - \mu^0_g\right| > \frac{\epsilon}{q_{\min}}\right) \geq \epsilon. \tag{4}$$

Under $\min_h \left|\hat{\mu}_h - \mu^0_g\right| > \epsilon/q_{\min}$ we have

$$\frac{1}{N}\sum_{i=1}^{N}\left(\mu^0_{g_i^0} - \hat{\mu}_{\hat{g}_i}\right)^2 > \frac{1}{N}\sum_{\substack{i=1,\ldots,N\\ g^0(i)=g}} \frac{\epsilon}{q_{\min}} \geq \epsilon$$

and therefore

$$\limsup_{N,T\to\infty} P\left(\frac{1}{N}\sum_{i=1}^{N}\left(\mu^0_{g_i^0} - \hat{\mu}_{\hat{g}_i}\right)^2 > \epsilon\right) \geq \epsilon.$$

This contradicts Lemma 5. Therefore (4) does not hold and for all $\epsilon > 0$

$$\lim_{N,T\to\infty} P\left(\max_{g\in\mathbb{G}}\min_{h\in\mathbb{G}} \left|\hat{\mu}_h - \mu^0_g\right| > \epsilon\right) \leq \sum_{g\in\mathbb{G}} \lim_{N,T\to\infty} P\left(\min_{h\in\mathbb{G}} \left|\hat{\mu}_h - \mu^0_g\right| > \epsilon\right) = 0.$$

This result implies that, for any constant $0 < \epsilon < M_G/2$ and

$$\limsup_{N,T\to\infty} P\left(\max_{g\in\mathbb{G}}\min_{h\in\mathbb{G}}\left|\hat{\mu}_h - \mu_g^0\right| \geq \epsilon\right) < \epsilon.$$

If

$$\max_{g\in\mathbb{G}}\min_{h\in\mathbb{G}}\left|\hat{\mu}_h - \mu_g^0\right| < \epsilon$$

then there exists, to each $g \in \mathbb{G}$, a non-empty set $H_g \subset \mathbb{G}$ such that $\left|\hat{\mu}_h - \mu_g^0\right| < \epsilon$ for all $h \in H_g$. We now prove $H_g \cap H_{g'} = \emptyset$ for $g, g' \in \mathbb{G}$ with $g \neq g'$. Suppose $h \in H_g$. Then

$$\left|\hat{\mu}_h - \mu_{g'}^0\right| = \left|\hat{\mu}_h - \mu_g^0 + \mu_g^0 - \mu_{g'}^0\right| \geq \left|\mu_{g'}^0 - \mu_g^0\right| - \left|\hat{\mu}_h - \mu_g^0\right| \geq M_G - \epsilon > \epsilon.$$

Therefore $h \neq H_{g'}$ and $H_g \cap H_{g'} = \emptyset$. Since $H_g \neq \emptyset$ this implies that all sets H_g, $g \in \mathbb{G}$ are singletons. Define the function $\pi : \mathbb{G} \to \mathbb{G}$ that maps each group g to the unique h such that $\left|\hat{\mu}_h - \mu_g^0\right| < \epsilon$. The function π is a bijection and hence a permutation function. For any given $h \in \mathbb{G}$ setting $g = \pi^{-1}(h)$ guarantees $|\hat{\mu}_h - \mu_g^0| < \epsilon$. Therefore,

$$\limsup_{N,T\to\infty} P\left(\max_{h\in\mathbb{G}}\left|\hat{\mu}_{\pi(g)} - \mu_g^0\right| \geq \epsilon\right) \leq \epsilon.$$

□

Proof of Lemma 2. Let $\pi : \mathbb{R} \to \mathbb{R}$ denote the permutation function from Lemma 1. For $i = 1, \ldots, N$, we have $\hat{g}_i \neq \pi(g_i^0)$ only if there is $g \in \mathbb{G} \setminus \{\pi(g_i^0)\}$ such that

$$\sum_{t=1}^{T}\left(y_{it} - \hat{\mu}_{\pi(g_i^0)}\right)^2 \geq \sum_{t=1}^{T}\left(y_{it} - \hat{\mu}_g\right)^2.$$

Plugging in $y_{it} = \mu_{g_i^0}^0 + \sigma_i v_{it}$ and rewriting the inequality yields

$$\operatorname{sign}(\hat{\mu}_g - \hat{\mu}_{\pi(g_i^0)})\frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{it} \geq \frac{\sqrt{T}}{2\sigma_i}\left|\hat{\mu}_g - \hat{\mu}_{\pi(g_i^0)}\right| - \operatorname{sign}(\hat{\mu}_g - \hat{\mu}_{\pi(g_i^0)})\frac{\sqrt{T}}{\sigma_i}(\mu_{g_i^0} - \hat{\mu}_{\pi(g_i^0)}).$$

Let $\mathcal{E}_{N,T}$ denote the event

$$\mathcal{E}_{N,T} = \{\max_{g\in\mathbb{G}}\left|\hat{\mu}_{\pi(g)} - \mu_g^0\right| > M_G/5\}.$$

On $\mathcal{E}_{N,T}$,

$$\begin{aligned}&\frac{\sqrt{T}}{2\sigma_i}\left|\hat{\mu}_g - \hat{\mu}_{\pi(g_i^0)}\right| - \operatorname{sign}(\hat{\mu}_g - \hat{\mu}_{\pi(g_i^0)})\frac{\sqrt{T}}{2\sigma_i}(\mu_{g_i^0} - \hat{\mu}_{\pi(g_i^0)})\\ \geq&\frac{\sqrt{T}}{2\sigma_i}\left(\left|\mu_{\pi^{-1}(g)}^0 - \mu_{g_i^0}^0\right| - \left|\hat{\mu}_g - \mu_{\pi^{-1}(g)}^0\right| - 3\left|\hat{\mu}_{\pi(g_i^0)} - \mu_{g_i^0}^0\right|\right) \geq \frac{\sqrt{T}}{10\sigma_i}M_G.\end{aligned}$$

Therefore,

$$\begin{aligned}
&P\left(\max_{i\in\mathcal{I}_{N,T}}\left|\hat{g}_i-\pi(g_i^0)\right|>0\right)\\
\leq &P\left(\text{there exists } i\in\mathcal{I}_{N,T} \text{ such that } \operatorname{sign}(\hat{\mu}_g-\hat{\mu}_{\pi(g_i^0)})\frac{1}{\sqrt{T}}\sum_{t=1}^{T}v_{it}\geq\frac{\sqrt{T}}{10\sigma_i}M_G\right)+P\left(\mathcal{E}_{N,T}\right)\\
\leq &P\left(\max_{1\leq i\leq N}\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}v_{it}\right|\geq 14\sqrt{\log N}\right)+P\left(\mathcal{E}_{N,T}\right),
\end{aligned}$$

where the last inequality follows since

$$\frac{\sqrt{T}}{10\sigma_i}M_G\geq 14\sqrt{\log N}$$

for all $i\in\mathcal{I}_{N,T}$ and $\mathcal{I}_{N,T}\subset\{1,\dots,N\}$. By Lemma 1 and Lemma 3,

$$\lim_{N,T\to\infty}\left[P\left(\max_{1\leq i\leq N}\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}v_{it}\right|\geq 14\sqrt{\log N}\right)+P\left(\mathcal{E}_{N,T}\right)\right]=0.$$

□

Proof of Theorem 1. Throughout the proof we omit the N,T subscripts and write $\mathcal{I}$, $\mathcal{I}(g)$ and $\mathcal{I}^{\mathsf{c}}$ instead of $\mathcal{I}_{N,T}$, $\mathcal{I}_{N,T}(g)$ and $\mathcal{I}^{\mathsf{c}}_{N,T}$. Assumption 3i) implies

$$\frac{\#\mathcal{I}^{\mathsf{c}}_{N,T}}{N}=o(1).$$

Hence, for $g\in\mathbb{G}$,

$$1\leq\frac{\tilde{N}_g}{N_g}\leq\frac{N_g}{N_g}+\frac{\#\mathcal{I}^{\mathsf{c}}_{N,T}}{N_g}\leq 1+(1+o(1))\frac{q_g\#\mathcal{I}^{\mathsf{c}}_{N,T}}{N}\leq 1+o(1)$$

and therefore

$$\left|\frac{\tilde{N}_g}{N_g}-1\right|=o(1).$$

For $g\in\mathbb{G}$,

$$\frac{\hat{N}_h}{\tilde{N}_g}=\frac{1}{\tilde{N}_g}\sum_{i\in\mathcal{I}^{\mathsf{c}}}1(\pi(\hat{g}_i)\neq g)+\frac{1}{\tilde{N}_g}\sum_{i\in\mathcal{I}}1\left(\pi(\hat{g})=g\right).$$

By Lemma 2

$$\lim_{N,T\to\infty}P\left(\frac{1}{\tilde{N}_g}\sum_{i\in\mathcal{I}}1\left(\pi(\hat{g})=g\right)\neq 1\right)=0.$$

Moreover,

$$\frac{1}{\tilde{N}_g}\sum_{i\in\mathcal{I}^\mathsf{c}} 1(\pi(\hat{g}_i)\neq g) \leq (1+o(1))\frac{\#\mathcal{I}^\mathsf{c}_{N,T}}{q_g N} \leq o(1)$$

and therefore for all $\epsilon > 0$

$$\lim_{N,T\to\infty} P\left(\left|\frac{\hat{N}_h}{\tilde{N}_g}-1\right| > \epsilon\right) = 0.$$

For all $g \in \mathbb{G}$ we can bound

$$\begin{aligned}
&\left|\frac{1}{\hat{N}_g T}\sum_{i\in\mathcal{I}^\mathsf{c}}\sum_{t=1}^T 1\left(\pi(\hat{g}_i)=g\right) y_{it}\right| \\
\leq &\frac{1}{q_g N}\left(1+o_p(1)\right)\left(\sum_{i\in\mathcal{I}^\mathsf{c}} 1\left(\pi(\hat{g}_i)=g\right)\left|\mu_i^0\right| + \frac{1}{\sqrt{T}}\sum_{i\in\mathcal{I}^\mathsf{c}} 1\left(\pi(\hat{g}_i)=g\right)\sigma_i\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T v_{it}\right)\right) \\
\leq &\frac{1}{q_g N}\left(1+o_p(1)\right)\left(\#\mathcal{I}^\mathsf{c}\sup_{\boldsymbol{\mu}\in\mathcal{M}}\|\boldsymbol{\mu}\|_{\max} + \frac{\#\mathcal{I}^\mathsf{c}}{\sqrt{T}}\sqrt{\frac{1}{\#\mathcal{I}^\mathsf{c}}\sum_{i\in\mathcal{I}^\mathsf{c}}\sigma_i^2}\sqrt{\frac{1}{\#\mathcal{I}^\mathsf{c}}\sum_{i\in\mathcal{I}^\mathsf{c}}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T v_{it}\right)^2}\right),
\end{aligned}$$

where $\|\cdot\|_{\max}$ is the max norm in $\mathbb{R}^G$. By independence over time and $\mathbb{E}v_{it}^2 = 1$ we have

$$\mathbb{E}\frac{1}{\#\mathcal{I}^\mathsf{c}}\sum_{i\in\mathcal{I}^\mathsf{c}}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T v_{it}\right)^2 = 1$$

and hence by the Markov inequality

$$\frac{1}{\#\mathcal{I}^\mathsf{c}}\sum_{i\in\mathcal{I}^\mathsf{c}}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T v_{it}\right)^2 = O_p(1).$$

In addition, $\sup_{\boldsymbol{\mu}\in\mathcal{M}}\|\boldsymbol{\mu}\|_{\max}$ is bounded by Assumption 1iii). Therefore

$$\begin{aligned}
&\left|\frac{1}{\hat{N}_g T}\sum_{i\in\mathcal{I}^\mathsf{c}}\sum_{t=1}^T 1\left(\pi(\hat{g}_i)=g\right) y_{it}\right| \\
\leq &O(1)\left(1+o_p(1)\right)\frac{\#\mathcal{I}^\mathsf{c}}{N}\left(1+(1+O_p(1))T^{-1/2}\sqrt{\frac{1}{\#\mathcal{I}^\mathsf{c}}\sum_{i\in\mathcal{I}^\mathsf{c}}\sigma_i^2}\right) = o_p\left(\frac{1}{\sqrt{NT}}\right),
\end{aligned} \tag{5}$$

where the last equality follows from Assumption 3i). We will now apply the Lindeberg-Feller CLT to show

$$\frac{1}{\sqrt{\tilde{N}_g}}\sum_{i\in\mathcal{I}(g)}\left\{\sigma_i\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T v_{it}\right)\right\} \xrightarrow{d} \mathcal{N}(0,\delta_g). \tag{6}$$

The variance of the term is given by

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{\sqrt{\tilde{N}_g}}\sum_{i\in\mathcal{I}(g)}\sigma_i v_{it}\right)^2\right]=\frac{1}{\tilde{N}_g}\sum_{i\in\mathcal{I}(g)}\sigma_i^2+\frac{1}{\tilde{N}_g}\sum_{\substack{i,j\in\mathcal{I}(g)\\ i\neq j}}\sigma_i\sigma_j\operatorname{cov}(v_{i1},v_{j1})\to\delta_g$$

To verify the Lindeberg condition it suffices to show that

$$\mathbb{E}\left[T^{-1/2}\sum_{t=1}^{T}z_{N,t}\right]^4\leq K \tag{7}$$

eventually, where

$$z_{N,t}=\frac{1}{\sqrt{\tilde{N}_g}}\sum_{i\in\mathcal{I}(g)}\sigma_i v_{it}$$

and K is a constant that does not depend on N and T. By independence across time periods

$$\mathbb{E}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}z_{N,t}\right]^4=\frac{\binom{4}{2}}{2!}\frac{1}{T^2}\sum_{s=1}^{T}\sum_{t\neq s}\mathbb{E}[z_{N,s}^2]\mathbb{E}[z_{N,t}^2]+\frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}[z_{N,t}^4]=3\delta_g^2+\frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}[z_{N,t}^4]+o\left(1\right).$$

To bound the right-hand side write for $t=1,\ldots,T$

$$\begin{aligned}\mathbb{E}\left[\sqrt{\tilde{N}_g}z_{N,t}\right]^4&=\mathbb{E}\left[\sum_{i\in\mathcal{I}(g)}\sigma_i v_{it}\right]^4\\ =\sum_{i\in\mathcal{I}(g)}\sigma_i^4\mathbb{E}[v_{it}^4]&+\frac{\binom{4}{2}}{2!}\sum_{\substack{i,j\in\mathcal{I}(g)\\ i\neq j}}\sigma_i^2\sigma_j^2\mathbb{E}[v_{it}^2v_{jt}^2]+\frac{\binom{4}{2}}{2!}\sum_{\substack{i,j,k\in\mathcal{I}\\ \{i\}\cap\{j\}\cap\{k\}=\emptyset}}\sigma_i^2\sigma_j\sigma_k\mathbb{E}[v_{it}^2v_{jt}v_{kt}]\\ &+\sum_{\substack{i,j,k,\ell\in\mathcal{I}\\ \{i\}\cap\{j\}\cap\{k\}\cap\{\ell\}=\emptyset}}\sigma_i\sigma_j\sigma_k\sigma_\ell\mathbb{E}[v_{it}v_{jt}v_{kt}v_{\ell t}]=I_{1,t}+I_{2,t}+I_{3,t}+I_{4,t}.\end{aligned}$$

To show (7) it suffices to show $\sum_{t=1}^{T}I_{k,t}=O(N^2T^2)$ for $k=1,\ldots,4$. Assumption 3i) implies

$$\left|\frac{\#\mathcal{I}}{N}-1\right|=o(1).$$

Moreover, by Assumption 2 there is a finite constant M_4 independent of N and T such that $\max_{1\leq t\leq T}\mathbb{E}[v_{it}^4]\leq M_4$. Therefore,

$$\sum_{t=1}^{T}I_{1,t}\leq M_4NT(1+o(1))\left(\frac{1}{\#\mathcal{I}}\sum_{i\in\mathcal{I}}\sigma_i^4\right)=O(N^2T^2).$$

and

$$\sum_{t=1}^{T}I_{2,T}\leq 3M_4(1+o(1)(N^2T)\left\{\frac{1}{\#\mathcal{I}}\sum_{i\in\mathcal{I}}\sigma_i^2=O(N^2T^2)\right\}^2.$$

Moreover, Assumption 3iv) yields $\sum_{t=1}^{T} I_{k,t} = O(N^2T^2)$ for $k = 1, 2$. This proves (7). For $g \in \mathbb{G}$

$$\begin{aligned}\hat{\mu}_{\pi(g)} =& \frac{1}{\hat{N}_g T} \sum_{i \in \mathcal{I}^c} \sum_{t=1}^{T} 1\left(\pi(\hat{g}_i) = g\right) y_{it} + \frac{1}{\hat{N}_g T} \sum_{i \in \mathcal{I} \setminus \mathcal{I}(g)} \sum_{t=1}^{T} 1\left(\pi(\hat{g}_i) = g\right) y_{it} \\ &+ (1 + o_p(1)) \frac{1}{\tilde{N}_g T} \sum_{i \in \mathcal{I}(g)} \sum_{t=1}^{T} 1\left(\pi(\hat{g}_i) = g\right) \left(\mu_{g_i^0} + \sigma_i v_{it}\right)\end{aligned}$$

The first term on the right-hand side is $o_p\left((NT)^{-1/2}\right)$ by (5). The second term is $o_p\left((NT)^{-1/2}\right)$ by Lemma 2. The third term converges to a centered normal with variance δ_g by (6) and Slutzky's lemma. □

References

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa (2019). "Discretizing unobserved heterogeneity". Working paper.

Bonhomme, Stéphane and Elena Manresa (2015). "Grouped patterns of heterogeneity in panel data". In: *Econometrica* 83.3, pp. 1147–1184.

Dzemski, Andreas and Ryo Okui (2019). "Confidence set fo group membership". mimeo.

Kuchibhotla, Arun Kumar and Abhishek Chakrabortty (2018). "Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression". In: *arXiv preprint arXiv:1804.02605.*

Okui, Ryo and Wendun Wang (2020). "Heterogeneous structural breaks in panel data models". In: *Journal of Econometrics.* forthcoming.

Pollard, David (1981). "Strong consistency of k-means clustering". In: *The Annals of Statistics*, pp. 135–140.

— (1982). "A central limit theorem for *k*-means clustering". In: *The Annals of Probability* 10.4, pp. 919–926.

Su, Liangjun, Zhentao Shi, and Peter Phillips (2016). "Identifying latent structures in panel data". In: *Econometrica* 84.6, pp. 2215–2264.

Vogt, Michael and Oliver Linton (2017). "Classification of nonparametric regression functions in heterogeneous panels". In: *Journal of the Royal Statistical Society: Series B* 79 (1), pp. 5–27.

Wang, Wuyi, Peter C. B. Phillips, and Liangjun Su (2018). "Homogeneity pursuit in panel data models: theory and applications". In: *Journal of Applied Econometrics* 33, pp. 797–815.