Blocked Clusterwise Regression[*]

Max Cytrynbaum[†]

December 5, 2021

Abstract

A recent literature in econometrics models unobserved cross-sectional heterogeneity in panel data by assigning each cross-sectional unit a one-dimensional, discrete latent type. Such models have been shown to allow estimation and inference by regression clustering methods. This paper is motivated by the finding that the clustered heterogeneity models studied in this literature can be misspecified, even when the panel has significant discrete cross-sectional structure. To address this issue, we generalize previous approaches to discrete unobserved heterogeneity by allowing each unit to have multiple, imperfectly-correlated latent variables that describe its *response-type* to different covariates. We give inference results for a k-means style estimator of our model and develop information criteria to jointly select the number clusters for each latent variable. Monte Carlo simulations confirm our theoretical results and give intuition about the finite-sample performance of estimation and model selection. We also contribute to the theory of clustering with an over-specified number of clusters and derive new convergence rates for this setting. Our results suggest that over-fitting can be severe in k-means style estimators when the number of clusters is over-specified.

[*] I wish to thank Alberto Abadie and Anna Mikusheva for their support during this project. This paper also benefited from conversations with Mert Demirer and David Hughes.

[†] Department of Economics, MIT. Correspondence: mcytryn@mit.edu

1 Introduction

Researchers often worry that, while their models impose a common parameter, there may actually be significant cross-sectional heterogeneity in the structural relationship between outcomes and observed covariates. Even with panel data, however, estimating distinct regression coefficients for each cross-sectional unit can be noisy or infeasible when the time dimension is small. Clusterwise regression methods (e.g. Lin and Ng (2012), Bonhomme and Manresa (2015)), which model individual heterogeneity as a function of a one-dimensional discrete latent type, have recently become popular as a viable *compromise* between the common parameter assumption and full heterogeneity. However, as we show, this discretization of heterogeneity can be misspecified even when the panel has significant cross-sectional structure.

This paper introduces panel data models with multiple, imperfectly correlated latent types, significantly enriching the set of panel structures that can be handled by clustering methods. In particular, our approach is motivated by a class of data-generating processes where units are clustered along multiple latent dimensions or "response-types" to distinct blocks of the covariate vector. We motivate this generalization with several examples from finance and production function estimation. The main contribution of this paper is to modify existing clustering methods for use in this larger family of models and show that they can likewise be used to perform inference on regression parameters.

Following Bonhomme and Manresa (2015) (henceforth BM), we establish consistency and asymptotic normality for a k-means style estimator in our setting. The estimation algorithm is iterative and alternates between (1) solving a least-squares problem to estimate cluster parameters for each block and (2) updating latent types for each cross-sectional unit based on a unit-wise predictive criterion. As in BM, our proof proceeds by establishing asymptotic equivalence with the oracle estimator where each unit's latent types are known. We extend the approach in Ando and Bai (2016) to give a C_p style information criterion to choose the number of clusters (types) for all the latent variables simultaneously.

In general, the true number of clusters in a given data set is unknown. Thus, the behavior of estimators with a misspecified number of clusters is important both for model selection theory as well as for our understanding the finite-sample properties of clustering estimators. Here, we make some contributions to the theory of models with an over-specified number of clusters, improving the convergence rates given in Liu et al. (2019) for the linear regression setting. In contrast to the well-specified case, difficulty obtaining the "fast rate" $O_p(\frac{1}{NT})$ when we over-specify the number of clusters suggests that over-fitting may be severe when the number of clusters is over-specified. We conjecture that $\sqrt{T}$-consistency may be optimal for over-specified models.

1.1 Motivating Example - Production Function Estimation

Consider panel data on firms' production levels and factor usage. We are interested in estimating the firm-specific production functions

$$y_{it} = \theta_{i1} L_{it} + \theta_{i2} K_{it} + \theta_{i3} M_{it} + \theta_{i4} \text{Elec}_{it} + e_{it} \tag{1.1}$$

where y_{it} is a measure of output and L_{it}, K_{it}, M_{it} are labor, capital and materials (all in logs), and Elec_{it} is a measure of electricity usage. Suppose that the heterogeneity in factor elasticities can be well approximated by

$$\theta_{i\ell} \in \{\theta_\ell^{low}, \theta_\ell^{mid}, \theta_\ell^{high}\} \quad 1 \leq \ell \leq 4$$

Ignoring endogeneity in input choice, we can estimate Equation 1.1 with using clusterwise regression, as in BM. The problem with this approach is readily apparent - although there are only $3 \cdot 4 = 12$ parameters, θ_i can take up to $3^4 = 81$ distinct values. Thus, estimating this model with clusterwise regression would require $k = 81$ clusters to be well-specified. For a panel of 200 firms, this would lead to estimation with approximately $N/81 \leq 3$ firms in each regression, in spite of significant cross-sectional homogeneity. However, with $k = 81$ clusters the model is also significantly over-parameterized. For instance, there will be 27 distinct clusters with each level of labor elasticity coefficient.

The problem is that current clustering models assume limited heterogeneity in the individual parameter vectors θ_i. In our example, however, cross-sectional heterogeneity takes the form of a few discrete elasticity levels for *each input factor*, while the support of θ_i itself is large. This suggests a model with multiple latent heterogeneity types. For instance

$$\theta_i = (\theta_1(c_{i1}), \theta_2(c_{i2}), \theta_3(c_{i3}), \theta_4(c_{i4}))$$

with latent type $c_{i\ell}$ for $1 \leq \ell \leq 4$ controlling the elasticity level of factor ℓ.

1.2 Related Literature and Outline

Early contributions to the econometric literature on clustering include Sun (2005) and Buchinsky et al. (2005). Linear panel data models with discrete unobservable heterogeneity have recently been studied in Lin and Ng (2012), Bonhomme and Manresa (2015), Su et al. (2016), Wang et al. (2016), Dzemski and Okui (2018). Our asymptotic normality results for the well-specified case closely follow the analysis pioneered in Bonhomme and Manresa (2015). Ando and Bai (2016) extends clustering methods to linear factor models and gives an information criterion for choosing the number of clusters. We develop a similar C_p-style criterion in our setting. Outside of the linear case, Zhang et al. (2019) and Chen et al. (2019) study clustered linear conditional quantile regression, and Bonhomme and Manresa (2019) considers discrete latent types as an approximation to continuous unobserved heterogeneity. Liu et al. (2019) studies clustering in M-estimation with an over-specified number of groups. We build on their techniques and significantly sharpen their rate results for the linear case. In contemporaneous work, Cheng et al. (2019) consider a clustering model with two latent types in a GMM setting. By contrast, we allow for $B > 1$ latent types in a linear model with individual fixed effects. Clusterwise regression was initially proposed in Späth (1979) as "Algorithm 39 - Clusterwise Linear Regression."

Further afield, this paper is related to a number of literatures in statistics and computer science, such as the literature on clustering functional data, e.g. Serban and Wasserman (2005), Yamamoto and Terada (2014), Vogt and Linton (2019), and subspace clustering, e.g. Candes and Soltanolkotabi (2012). In statistics, related methods include homogeneity pursuit, proposed in Ke et al. (2015). See also Ke et al. (2016) and Lian et al. (2019).

In the Bayesian literature, clusterwise regression is also known as multilevel regression; see Gelman and Hill (2007).

The rest of the paper is organized as follows: we introduce our model and estimator in section 2. Asymptotic properties of the estimator and consistency of model selection are given in section 3. Section 4 discusses models with an over-specified number of clusters. Monte Carlo simulations are given in section 5, and proofs in section A. Supplementary appendix B collects technical lemmas and other ancillary discussions.

2 Model and Estimation

2.1 Model

Let y_{it} and x_{it} denote repsonse and covariates for $t = 1, \dots, T$ time periods and $i = 1, \dots, N$ cross-sectional observations. The covariate vector $x_{it} \in \mathbb{R}^p$ is divided into $1 \leq \ell \leq B$ blocks x_{it}^{ℓ}, where $x_{it} = (x_{it}^1, \dots, x_{it}^B)$, and B denotes the total number of blocks. We let $k = (k_1, \dots, k_B)$, where k_ℓ denotes the number of distinct latent types (clusters) associated with the ℓ^{th} block. Possible cluster assignments are denoted $c = (c_1, \dots, c_B) \in \prod_\ell [k_\ell] \equiv \mathcal{C}$. For instance, a unit in cluster 1 in the first block and cluster 3 in the second block would have $c = (1, 3)$.

Each cross-sectional unit belongs to exactly one cluster for each block. We let $\gamma : [N] \to \prod_\ell [k_\ell]$ denote an assignment of cross-sectional units to cluster vectors, so that $\gamma(i) = c_i$. The set of all possible cluster assignments is denoted Γ. In the main specification, we assume that the response y_{it} is given by

$$y_{it} = x_{it}'\theta(c_i) + e_{it} \tag{2.1}$$

with the ℓ^{th} block parameter selected by the latent variable $c_{i\ell}$

$$\theta(c_i) = (\theta_1(c_{i1}), \dots, \theta_B(c_{iB})) \tag{2.2}$$

Thus, each covariate grouping ℓ is associated with k_ℓ parameter sub-vectors with $\theta_\ell(c_\ell) \in \mathbb{R}^{d_\ell}$. Our goal is to jointly estimate the true parameter θ^0 and true cluster assignments $\gamma^0(i) = (c_{i1}^0, \dots, c_{iB}^0)$ of each cross-sectional unit. In appendix C.1, we also give results for the model with individual fixed effects

$$y_{it} = x_{it}'\theta(c_i) + a_i + e_{it} \tag{2.3}$$

Relationship with Clusterwise Regression: The model above nests clusterwise regression (as in Lin and Ng (2012)) when $B = 1$. For $B > 1$, it is statistically equivalent to clusterwise regression when the conditional pdf $\mathbb{P}(c_{i(-\ell)}^0 | c_{i\ell}^0)$ is degenerate (perfectly correlated types). Similarly, clusterwise regression ($B = 1$) with exponentially many clusters $k = \prod_{\ell=1}^B k_\ell$ and exponentially many constraints nests our model. For instance, let $p = B$ and assume $\theta_\ell \in \{\pm 1\}$ for each ℓ. Then our model would be equivalent to a clusterwise regression model with 2^p clusters and $p2^{p-1}$ linear equality constraints.

Example - Exchange Rates: Consider financial market data with y_{it} the exchange rate against USD of country i at time t, p_t^{oil} the price of crude oil, b_{it} a measure of the country's business cycle and r_t the US discount rate, we model

$$y_{it} = \theta_{i1} p_t^{oil} + \theta_{i2} b_{it} + \theta_{i3} r_t + e_{it}$$

Due to differences in national industry composition, the magnitude and composition of foreign trade, financial openness and so on, we may expect heterogenous marginal responses $\theta_{i\ell}$ of y_{it} to each of the factors above. As in the introduction, we may model $\theta_{i\ell} \in \{\theta_\ell(1), \ldots, \theta_\ell(k_\ell)\}$, corresponding to k_ℓ different sensitivity levels to factor ℓ. However, we don't expect these unobserved types to be perfectly correlated across factors. For instance, we might expect both Venezuela and China to have large θ_{i1} but very different θ_{i3}. A factor error structure could be accommodated using the techniques in Ando and Bai (2016).

2.2 Estimator

We define our estimator of the parameter θ^0 and cluster assignment γ^0 as

$$(\widehat{\theta}, \widehat{\gamma}) = \underset{\gamma \in \Gamma, \theta \in \Theta}{\operatorname{argmin}} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - x_{it}' \theta(c_i))^2 \tag{2.4}$$

We let $\widehat{Q}(\theta, \gamma)$ denote the sample risk in 2.4. There are many algorithms available for the least squares partitioning problem above[1]. One benchmark approach, known as Lloyd's algorithm (Lloyd (1982)) in the setting of k-means clustering, takes a coordinate ascent approach to the problem in 2.4, alternately updating the parameters θ and assignments γ until convergence.

Lloyd's Algorithm - Fix a division of the covariate vector $x_{it} = (x_{it}^1, \ldots, x_{it}^B)$ into blocks with $x_{it}^\ell \in \mathbb{R}^{d_\ell}$ and fix the number of clusters $k = (k_1, \ldots, k_B)$ in each block. Our approach is a modification of Lloyd's algorithm for k-means clustering. We perform coordinate ascent on the sample objective $\widehat{Q}(\theta, \gamma)$ by alternating parameter updates and cluster assignment updates until convergence.

(1) Randomly initialize parameters θ^1 and cluster assignments γ^1.
(2) Given θ^s, set $\gamma^{s+1} = \operatorname{argmin}_{\gamma \in \Gamma} \widehat{Q}(\theta^s, \gamma)$.
(3) Given γ^{s+1}, update $\theta^s \to \theta^{s+1}$.
(4) Repeat (2) and (3) until convergence.

Since problem 2.4 is not generally convex, we repeat steps (1) through (4) from different random initializations (in parallel), and take the estimate that achieves the lowest sample risk $\widehat{Q}$. See appendix C.2 for more discussion of the implementation of this algorithm and related computational issues.

[1] See, for instance, the discussion in BM Appendix S1.

3 Asymptotic Properties

In this section, we investigate the asymptotic properties of the estimator $(\widehat{\theta}, \widehat{\gamma})$ defined above as $N, T \to \infty$. In what follows, we assume the data is generated from the model 2.1 with (θ^0, γ^0) the true slope parameters and cluster assignment function. We let $\| \cdot \|$ denote the usual Euclidean norm.

3.1 Consistency

Assumption 3.1. *We make the following assumptions*

(a) For each $\ell \in [B]$, the parameter space $\Theta_\ell \subset \mathbb{R}^{d_\ell \times k_\ell}$ is compact

(b) $\frac{1}{NT^2} \sum_{i=1}^N \sum_{t=1}^T \sum_{s=1}^T e_{it} e_{is} x_{it}' x_{is} \xrightarrow{p} 0$

(c) $\|\theta_{\ell a}^0 - \theta_{\ell b}^0\| \equiv d(\ell, a, b) > 0$ for each pair of clusters $a, b \in [k_\ell]$

(d) Define $M(c, c', \gamma) \equiv \frac{1}{NT} \sum_{i,t} x_{it} x_{it}' \mathbb{1}(c_i^0 = c') \mathbb{1}(c_i = c)$ and let $\rho(c, c', \gamma) \equiv \lambda_{min}(M(c, c', \gamma))$. Then there exists $\delta > 0$ such that $\inf_{c', \gamma} \max_c \rho(c, c', \gamma) \geq \delta - o_P(1)$ as $N, T \to \infty$.

Assumption 3.1.(a) is the usual parameter space compactness condition. Assumption 3.1.(b) can be seen as limiting the time-series dependence of errors and covariates, averaged over cross-sectional units. Condition 3.1.(c) ensures that the clusters within each grouping are non-identical. The final assumption 3.1.(d) is the analogue in our setting of assumption S2(a). in BM. This condition is used to ensure curvature of the sample risk function $\widehat{Q}$. If there is a common parameter ($B = 1$, $k = 1$), this is the usual non-collinearity condition for pooled panel regression. See section S4.2 in the supplementary appendix of BM for further discussion.

Cluster Label Ambiguity. The minimizer $\operatorname{argmin}_{\gamma \in \Gamma, \theta \in \Theta} \widehat{Q}(\theta, \gamma)$ is only unique up to permutations of the labels in $\mathcal{C}$ and their associated parameter vectors in θ. Thus, the $c \in \mathcal{C}$ used to label estimated clusters in each block is arbitrary, and to resolve this ambiguity we need to fix a correspondence $\sigma_\ell : [k_\ell] \to [k_\ell]$ between true and estimated cluster parameters for each ℓ[2]. Let

$$\sigma_\ell(a) \equiv \underset{b \in [k_\ell]}{\operatorname{argmin}} \|\widehat{\theta}_{\ell b} - \theta_{\ell a}^0\| \tag{3.1}$$

and define the estimator of the true parameter $\theta_{\ell a}^0$ to be $\widehat{\theta}_{\ell \sigma(a)}$. Note that this is infeasible without access to the true parameters θ^0.

Lemma 3.2. *Under the assumptions in 3.1, $\mathbb{P}(\sigma_\ell$ invertible$) \to 1$ as $N, T \to \infty$*

By the lemma, we can relabel the estimated clusters $\widehat{\theta}_{\ell \sigma(a)} \to \widehat{\theta}_{\ell a}$, and this is well-defined w.h.p as $N, T \to \infty$.

Theorem 3.3. *Under the assumptions in 3.1, for all groupings ℓ and $a \in [k_\ell]$, we have*

$$\|\theta_{\ell a}^0 - \widehat{\theta}_{\ell a}\| = o_P(1)$$

[2] Note that, for finite T, it can be the case that $\widehat{\theta}(\widehat{c}_i) \neq \widehat{\theta}(\widehat{c}_j)$, but $c_i^0 = c_j^0$, so the estimates $\widehat{\theta}(\widehat{c}_i)$ do not in general induce a well-defined estimator of any fixed cluster parameter $\theta_{\ell a}^0$.

equivalently

$$\min_{x\in\mathcal{C}} \|\widehat{\theta}(x) - \theta(c)\| = o_p(1) \quad \forall c \in \mathcal{C}$$

as $N, T \to \infty$.

See section A.1 for the proof of the theorem and lemma.

3.2 Asymptotic Equivalence

In this section, we establish asymptotic equivalence of $\widehat{\theta}$ to the infeasible oracle estimator with known clusters. We need the following assumptions in addition to those already stated in 3.1.

Assumption 3.4. *Make the following assumptions*

(a) $\frac{1}{NT^2}\sum_i \sum_{t,s} \|x_{it}\|^2 \|x_{is}\|^2 = O_P(1)$

(b) Define $M^c_{NT} = \frac{1}{NT}\sum_{i,t} \mathbb{1}(c^0_i = c) x_{it} x'_{it}$. *Then there exists* $\underline{\rho} > 0$ *such that for all* $a > 0$, *this sequence of matrices satisfies* $\min_{c\in\mathcal{C}} \lambda_{min}(M^c_{NT}) \xrightarrow{p} \underline{\rho}$ *as* $N, T \to \infty$.

(c) There exist constants $b > 0$ *and* $d_1 > 0$ *and sequence* $\alpha(t) \leq e^{-bt^{d_1}}$ *such that for all* $i \in [N]$ $\{x_{it}\}_t$ *and* $\{x_{it}e_{it}\}_t$ *are strongly mixing with coefficients* $\alpha(t)$.

(d) There exist constants $f > 0$ *and* $d_2 > 0$ *such that for all* $i \in [N]$ *and all* $z > 0$, *for all components* x^j_{it}, $x^{j'}_{it}$ *of the vector* x_{it} *we have* $\mathbb{P}(|x^j_{it}x^{j'}_{it} - E(x^j_{it}x^{j'}_{it})| > z)$ *and* $\mathbb{P}(|e_{it}x^j_{it} - Ee_{it}x^j_{it}| > z)$ *are bounded above by* $e^{1-(z/f)^{d_2}}$.

(e) The uniform limits $\max_{i\in[N]} \frac{1}{T}\sum_t E[e_{it}x_{it}] \to 0$ *and* $\min_{i\in[N]} \frac{1}{T}\sum_t \mathbb{E}(x'_{it}(\theta(c) - \theta(c')))^2 \to d(c,c')$ *hold as* $T \to \infty$, *and* $d(c,c') \geq d_{min} > 0$ *for* $c \neq c'$.

(f) There exists $M' > 0$ *such that for all* $a > 0$

$$\max_{i\in[N]} \mathbb{P}\left(\frac{1}{T}\sum_t \|x_{it}\|^2 > M'\right) = o(T^{-a})$$

We will show that $\widehat{\theta}$ is asymptotically equivalent to the infeasible oracle estimator where true cluster membership c^0_i is known for all i. Define the problem

$$\tilde{Q}(\theta) \equiv \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x'_{it}\theta(c^0_i))^2$$

$$\tilde{\theta} = \underset{\theta\in\Theta}{\operatorname{argmin}}\, \tilde{Q}(\theta) \tag{3.2}$$

The following theorem shows that $\widehat{\theta}$ and $\tilde{\theta}$ are asymptotically equivalent.

Theorem 3.5. *Let the assumptions in 3.1 and 3.4 hold. Then for any* $a > 0$ *and as* $N, T \to \infty$, *we have*

$$\widehat{\theta} = \tilde{\theta} + o_P(T^{-a}) \tag{3.3}$$

Moreover, individual cluster estimates satisfy

$$\mathbb{P}\left(\exists i \in [N]\, s.t. \, \widehat{c}_i \neq c^0_i\right) = o(1) + o(NT^{-a}) \tag{3.4}$$

See appendix A.2 for the proof. Because of this theorem, for asymptotic sequences with N growing at a sub-polynomial rate relative to T, it suffices to characterize the asymptotic distribution of the estimator $\tilde{\theta}$.

3.3 Inference

Notation - To aid the exposition, we start with a few definitions. For $A \in \mathbb{R}^{p\times q}$, let $\mathrm{vec}(A) \equiv ((A^1)', \dots, (A^q)')' \in \mathbb{R}^{pq}$. Thinking of $\theta = \{\theta_1, \dots, \theta_B\}$ as a collection of matrices $\theta_\ell \in \mathbb{R}^{d_\ell \times k_\ell}$, we denote $\mathrm{vec}(\theta) = (\mathrm{vec}(\theta_1)', \dots, \mathrm{vec}(\theta_B)')' \in \mathbb{R}^{d_\theta}$, where $d_\theta \equiv \sum_\ell k_\ell d_\ell$ is the total dimension of $\mathrm{vec}(\theta)$. For $1 \leq \ell \leq B$ and $a \in [k_\ell]$, we use the block index convention that $\mathrm{vec}(\theta)_{\ell a}$ refers to the d_ℓ dimensional sub-vector in the a^{th} position of the ℓ^{th} block. Using the notation above, for $1 \leq \ell, s \leq B$ and $a \in [k_\ell]$, $b \in [k_s]$ define $\widehat{M} \in \mathbb{R}^{d_\theta \times d_\theta}$ and $v \in \mathbb{R}^{d_\theta}$ by

$$\widehat{M}_{\ell a, sb} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it\ell}x'_{its}\mathbb{1}(c^0_{is} = b)\mathbb{1}(c^0_{i\ell} = a) \tag{3.5}$$

$$v_{\ell a} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}\mathbb{1}(c^0_{i\ell} = a)x_{it\ell} \tag{3.6}$$

$$w_{\ell a} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}\mathbb{1}(c^0_{i\ell} = a)x_{it\ell} \tag{3.7}$$

Proposition 3.6. *The solution $\tilde{\theta}$ to problem 3.2 satisfies*

$$\widehat{M}\,\mathrm{vec}(\tilde{\theta}) = v \tag{3.8}$$

The proof follows by taking the first order conditions of 3.2 and rearranging. Note that the first order conditions $\nabla_{\theta_{\ell a}}\tilde{Q}(\tilde{\theta}) = 0$ can potentially vary with all the other parameters θ_{sb} in the model (for $s \neq \ell$). Therefore, in contrast to the $B = 1$ case considered in the existing literature, the estimator $\tilde{\theta}$ is *not* equivalent to simply running k separate regressions over the partition of the cross-sectional units.

Consider the following assumptions that allow us to characterize the asymptotic distribution of the infeasible $\tilde{\theta}$.

Assumption 3.7. *We make the following assumptions*

(a) There is a matrix $\Omega \succ 0$ such that for all $\ell, s \in [B]$ and $1 \leq a \leq k_\ell$, $1 \leq b \leq k_s$

$$\frac{1}{NT}\sum_{i,j=1}^{N}\sum_{t,t'=1}^{T} E[e_{it}e_{jt'}\mathbb{1}(c^0_{i\ell} = a)\mathbb{1}(c^0_{js} = b)x_{it\ell}x'_{jt's}] \to \Omega_{\ell a, sb} \tag{3.9}$$

as $N, T \to \infty$.

(b) $\mathbb{E}[e_{it}\mathbb{1}(c^0_{i\ell} = a)x_{it\ell}] = 0$ for all ℓ, a.

(c) $\widehat{M} \xrightarrow{p} M$ as $N, T \to \infty$, with $M \succ 0$.

(d) $\sqrt{NT}w \xrightarrow{d} \mathcal{N}(0, \Omega)$ as $N, T \to \infty$.

Theorem 3.8. *Suppose that the assumptions in 3.7 are satisfied. Also suppose there is some $r > 0$ such that $\sqrt{N}T^{-r} = o(1)$ as $N, T \to \infty$. Then we have*

$$\sqrt{NT}(\text{vec}(\widehat{\theta} - \theta^0)) \xrightarrow{d} \mathcal{N}(0, M^{-1}\Omega M) \tag{3.10}$$

The proof of this theorem is given in appendix A.2.

Consider the case where cross-sectional units are independent, then under assumption 3.7.(b), the terms in 3.9 with $i \neq j$ vanish. In this case, we propose the HAC estimators

$$\begin{aligned}\widehat{\Omega}_{\ell a,sb} &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t,t'=1}^{T}\widehat{e}_{it}\widehat{e}_{it'}\mathbb{1}(\widehat{c}_{i\ell} = a)\mathbb{1}(\widehat{c}_{is} = b)x_{it\ell}x'_{it's}\\ \widehat{V} &= \widehat{M}^{-1}\widehat{\Omega}\widehat{M}^{-1}\end{aligned} \tag{3.11}$$

where $\widehat{M}$ is as in equation 3.5. Variance estimators of this form were originally proposed in Arellano (1987), and their asymptotic theory for $N, T \to \infty$ jointly was first analyzed in Hansen (2007). For further discussion on adapting the results of Hansen (2007) to our setting, see appendix A.4.

3.4 Model Selection

In this section we let $k^0 = (k^0_1, \dots, k^0_B)$ denote the true number of clusters and develop a Cp-like criterion to estimate k^0. We suppose that prior information can be used to bound the true number of clusters from above $k^0 \leq k_{max}$. So far, we have defined the sample risk $\widehat{Q}(\theta, \gamma)$ with domain the true parameter space i.e. $\theta \in \prod_\ell \mathbb{R}^{d_\ell \times k^0_\ell}$ and $\gamma : [N] \to \prod_\ell [k^0_\ell]$. However, note that $\widehat{Q} = \frac{1}{NT}\sum_{i,t}(y_{it} - x'_{it}\theta(\gamma(i)))^2$ only varies through $\theta(\gamma(i)) \in \mathbb{R}^p$. Thus, we can extend the domain of $\widehat{Q}$ to models with $k \neq k^0$, since $\theta(\gamma(i)) \in \mathbb{R}^p$ for any conformable (θ, γ).[3]

We slightly strengthen some assumptions above

Assumption 3.9. *Impose the following assumptions*

(a) For all $c \in \mathcal{C}^{k_0}$, $\frac{1}{NT}\sum_{i,t} e_{it}x_{it}\mathbb{1}(c^0_i = c) = O_p(1/\sqrt{NT})$

(b) With $\rho(c, c', \gamma)$ defined as in assumption 3.4.(b), there exists $\delta > 0$ such that for all $k \geq k^0$ we have

$$\rho^k_{NT} \equiv \min_{c' \in \mathcal{C}^{k_0}} \min_{\gamma \in \Gamma^k} \max_{c \in \mathcal{C}^k} \rho(c, c', \gamma) \geq \delta - o_p(1)$$

(c) As $N, T \to \infty$

$$\inf_{j \in [N]} \lambda_{min}\left(\frac{1}{T}\sum_t E[x_{jt}x'_{jt}]\right) \to \underline{\lambda} > 0$$

(d) For some $0 < \epsilon < \frac{1}{2} \wedge \frac{d_1 d_2}{d_1 + d_2}$ we have $\log N = o(T^\epsilon)$ as $N, T \to \infty$, where d_1, d_2 are the mixing and tail parameters defined in assumptions 3.4.(c) and 3.4.(d)

[3] Formally, let $\widehat{Q} : \bigcup_{k \geq 0}\{\prod_\ell \mathbb{R}^{d_\ell \times k_\ell} \times [[N] \to \prod_\ell [k^0_\ell]]\} \to \mathbb{R}_{\geq 0}$ with $\widehat{Q}(\theta, \gamma) = \frac{1}{NT}\sum_{i,t}(y_{it} - x'_{it}\theta(\gamma(i)))^2$

We can think of assumption 3.9.(a) as stating that a CLT holds for $e_{it}x_{it}\mathbb{1}(c_i^0 = c)$ for each $c \in \mathcal{C}^{k_0}$. This will be easiest to satisfy when $E[e_{it}x_{it}\mathbb{1}(c_i^0 = c)] = 0$, a stronger form of unconfoundedness. Assumption 3.9.(b) is the extension of assumption 3.1.(d) in our consistency proof to the case of models with misspecified number of clusters. See section **??** of the supplementary appendix for a discussion of this condition. In the stationary case with identically distributed cross-sectional units, having $E[x_{it}x_{it}']$ of full rank is sufficient for assumption 3.9.(c). Finally, assumption 3.9.(d) requires that $\log N$ is sub-polynomial in T as $N, T \to \infty$.

Information Criterion - Let $(\widehat{\theta}^k, \widehat{\gamma}^k)$ be the minimizer of $\widehat{Q}$ with $(k_i)_i$ clusters and denote $\widehat{Q}(k) = \widehat{Q}(\widehat{\theta}^k, \widehat{\gamma}^k)$. Then we define the C_p criterion

$$C_p(k) \equiv \widehat{Q}(k) + f(N,T)\sum_i k_i \tag{3.12}$$

$$\widehat{k} = \underset{k \leq k_{\max}}{\operatorname{argmin}}\, C_p(k)$$

We have the following result on consistency of model selection

Theorem 3.10. *Suppose the assumptions in 3.9 hold. Let $f(N,T)$ be such that $f(N,T) \to 0$ and for some ϵ as in assumption 3.9.(d), $f(N,T)T^{1-3\epsilon} \to \infty$ as $N, T \to \infty$. Then*

$$\mathbb{P}(\widehat{k} = k^0) \to 1$$

as $N, T \to \infty$

For the proof, see appendix A.3.

Remark 3.11 (Choice of f)**.** While any function $f(N,T)$ satisfying the conditions of the theorem will give asymptotically consistent model selection, the choice of f will significantly affects finite sample performance. To put $f(N,T)$ on the same scale as $\widehat{Q}(k)$, we use $f(N,T) = \widehat{\sigma}^2 g(N,T)$ in our simulations, where $\widehat{\sigma}^2$ is a consistent estimate of the long run variance $\lim_{N,T} \frac{1}{NT}\sum_{i,t} E[e_{it}^2]$. By lemma A.5 in the appendix, $\widehat{\sigma}^2 \equiv \widehat{Q}(k_{max})$ is such a consistent estimator. We find good performance with $g(N,T) = \frac{\log T}{T}$ in our simulations. Alternatively, e.g. $g(N,T) = \frac{\log T}{T^{1-\epsilon'}}$ for small ϵ' can be used to be technically consistent with the theory.

4 Overspecification of k

In this section, we report new results on the performance of k-means style estimators with an over-specified number of clusters. The work in this section builds on and sharpens the results in Liu et al. (2019) for the case of linear regression. Our proof of model selection consistency in Theorem 3.10 heavily relies on the following result.

Theorem 4.1. *Suppose the assumptions in 3.1 hold. Then*

$$\sup_{i\in[N]} \|\widehat{\theta}^k(\widehat{c}_i^k) - \theta^0(c_i^0)\|^2 = o_p(T^{-1+4\epsilon}) \tag{4.1}$$

$$\min_{x\in\mathcal{C}^k} \|\widehat{\theta}^k(x) - \theta^0(c)\|^2 = o_p(T^{-1+3\epsilon}) \tag{4.2}$$

$$\frac{1}{N}\sum_i (\widehat{\theta}^k(\widehat{c}_i^k) - \theta^0(c_i^0))^2 = o_p(T^{-1+3\epsilon}) \tag{4.3}$$

Proof. Follows from Proposition A.8, Proposition A.9, and Corollary A.7 in the appendix. □

Remark 4.2. The preceding result can be compared with Liu et al. (2019) Theorem 1 and Lemma 5.16, which give $o_p(1)$, $o_p(1)$, and $o_p(T^{\frac{-1}{2(1+d)}})$ rates (respectively) for each of the losses above, with $d = \frac{d_1 d_2}{d_1+d_2}$. Note that their setting is a more general model of clustered M-estimation. Our rate improvements come from (1) optimizing the Fuk-Nagaev inequality in Merlevede et al. (2011) for our purposes (see lemma B.3) and (2) an inductive strategy that allows us to "boost" $O_p(T^{-1/4})$ rates arbitrarily close to $O_p(T^{-1/2})$. For a description of this approach, see lemma A.6 as well as the propositions and corollary referenced above.

Remark 4.3. The rate established in equation 4.3 above is used to bound the magnitude of over-fitting for estimators with $k > k^0$, as in the second part of corollary A.7. A result of this form is necessary to determine the complexity penalty in 3.10. In particular, in contrast to the result in Liu et al. (2019), theorem 4.1 gives feasible rates for $f(T)$ that do not depend on mixing parameters and tail bounds of e_{it} and x_{it}, which may be difficult or impossible to estimate.

Remark 4.4. Difficulty obtaining the fast rate $\widehat{Q}(k) - \widehat{Q}^0 = O_p(\frac{1}{NT})$ for $k > k^0$ suggests that over-fitting may be severe under over-specification of k. Difficulty obtaining $\sqrt{NT}$-consistency of $\widehat{\theta}^k$ when $k > k^0$ suggests a type of incidental parameter problem. In fact, in the linear case it is known[4] that under $N \to \infty$, finite T asymptotics, estimators with $k > k^0$ can suffer a bias of order up to $\frac{1}{\sqrt{T}}$.

5 Monte Carlo Simulations

In this section, we describe the results of our Monte Carlo simulations. All tables are reported in section D of the appendix. Throughout, we denote

1. Param. MSE $= \frac{1}{N}\sum_{i=1}^N \|\widehat{\theta}(\widehat{c}_i) - \theta^0(c_i^0)\|^2$
2. Function MSE $= \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \|\widehat{\theta}(\widehat{c}_i)' x_{it} - \theta^0(c_i^0)' x_{it}\|^2$
3. Cluster Loss $= \frac{1}{N}\sum_{i=1}^N \mathbb{1}(\widehat{c}_i \neq c_i^0)$

We use two specifications for the joint distribution (x_{it}, y_{it}). (1) Specifications labeled *AR(1)* take $e_{it} \sim$ AR(1) and $x_{it} \sim$ VAR(1). The AR process e_{it} has normal innovations, and x_{it} has multivariate normal innovations with constant, diagonal covariance matrix. The respective autocorrelation parameters are $\rho_e = 0.3$, $\rho_x = 0.5$. (2) Specifications

[4] See the example in BM, appendix S3.1.

labeled *HK* use a heteroskedastic design inspired by Hansen (2007). With x_{it} as above, we use $e_{it} = \rho e_{it-1} + v_{it} \cdot \sqrt{\frac{1}{2} + \frac{\|x_{it}\|^2}{2p}}$, with independent normal innovations v_{it}. Innovation variances are normalized so that $\text{Var}(e_{it}) = 1$ for all designs and all (i, t). All simulations use 500 independent samples. See appendix C.2 for additional details on the computational specification.

5.1 Estimator Performance

Design 5.1 (Cluster Separation)**.** We let $p = 4$, $k = (2, 2)$, $B = 2$ and parameters

$$\theta_1^0 = \begin{pmatrix} 1 & \cos\alpha \\ 0 & \sin\alpha \end{pmatrix} \quad \theta_2^0 = \begin{pmatrix} 0 & -\sin\alpha \\ 1 & \cos\alpha \end{pmatrix}$$

where the columns of θ_ℓ list parameters of block $1 \leq \ell \leq B$. Thus, as $\alpha \to 0$, the cluster parameters in each block rotate towards each other. Cluster estimation accuracy radically worsens for small α. Coverage is around $80 - 90\%$ for well-separated clusters. As $\alpha \to 0$, our confidence intervals do not account for variation due to cluster estimation, and coverage is poor. Parameter loss is inverse U-shaped in cluster separation α. For small α, classification of c_i^0 becomes worse, giving large losses on some units. For α near 0, misclassification contributes less to parameter loss since the cluster centers are very close. Results are shown in Table 1.

Design 5.2 (Sample Size (N, T))**.** We use the specification in the simulation above with $\alpha = \frac{\pi}{2}$. Cluster loss is quite insensitive to N, in line with the theory. Increasing T has a much larger effect than N on parameter loss and coverage. For $T = 5$, we find coverage actually decreases with larger N, which could be an example of the over-fitting issue discussed in section 4. Results are shown in Table 2.

Design 5.3 (Number of Clusters)**.** Again with $p = 4$ and $B = 2$, we let $k = (k_1, k_2)$ vary. We define clusters $\theta_{1a} = (\cos(\frac{2\pi}{5} \cdot a), \sin(\frac{2\pi}{5} \cdot a))'$ for $1 \leq a \leq k_1$ and similarly for the second block. All performance measures decrease as the number of clusters increase. Results are shown in Table 3.

Design 5.4 (Misspecification)**.** In this simulation, we repeat the design above using $B = 1$ and $k = k_1 \cdot k_2$, the minimal number of clusters for consistent estimation using the single latent type assumption $(B = 1)$ considered in the literature. As expected, there is a significant power loss. Results are shown in Table 4.

Design 5.5 (Block Dimension Imbalance)**.** We let $p = 12$, $B = 2$ and $(k_1, k_2) = (2, 2)$. We vary the grouping of covariates, taking the first block to be $(x_{it1}, \ldots, x_{itm})$ for $m \in \{1, \ldots, 6\}$. Coverage-small denotes the average coverage for parameters belonging to the small block $(x_{it1}, \ldots, x_{itm})$ and conversely for Coverage-large. Cluster loss-large and Cluster loss-small are defined similarly. Classification and coverage are worse for the block of smaller dimension m when $m/(p - m)$ is very small, but quickly equalize as m gets larger. Results are shown in Table 5.

Design 5.6 (Covariate Dimension)**.** In this simulation, we take $B = p$ (one latent variable for each covariate) and study the effect of increasing p. We let $k_\ell = 2$ for $1 \leq \ell \leq p$ and clusters $\theta_{\ell a} = \pm 1$ $(a \in \{1, 2\})$. Performance only slightly deteriorates as p increases. Results are shown in Table 6

5.2 Model Selection

Design 5.7 (Model Selection - Number of Clusters)**.** We implement the C_p criterion and study its performance on the DGP in design 5.3 above. We use penalty sequence $f(N,T) = \widehat{\sigma}^2 \frac{\log T}{T}$, as in section 2. Model loss is calculated using average (over $k_1, \ldots k_B$) distance from the truth $\|\widehat{k} - k^0\|_1/B$. We set $k_{max} = (6,6)$ and use 200 independent samples. For $k^0 = (2,3)$, we estimate $\mathbb{E}\frac{\|\widehat{k}-k^0\|}{B} = 0.03$. For $k^0 = (4,4)$, we find $\mathbb{E}\frac{\|\widehat{k}-k^0\|}{B} = 0.73$, with all estimates $\widehat{k} = (3,3), (3,4)$, or $(4,3)$. We view this performance as reasonable given that the clusters are quite close in this design, though the results suggest we may be slightly over-penalizing.

6 Conclusion

Clustering methods have recently become popular as a way of modeling limited heterogeneity in panel data. This paper motivates a family of panel structures, nesting the standard regression clustering model, that have significant cross-sectional homogeneity but are nevertheless ill-suited to estimation by the clustering methods currently considered in the literature We propose a modified procedure that simultaneously clusters on multiple discrete latent types, significantly expanding the set of panel structures that can be accommodated by these methods. We employ Lloyd's algorithm to compute the estimator, and give consistency and asymptotic normality results for the resulting estimates.

A Proofs

Throughout the following proofs, unless otherwise specified $\max_i$, $\sum_i$, $\sum_t$, $\sum_c$ denote $\max_{i\in[N]}$, $\sum_{i\in[N]}$, $\sum_{t\in[T]}$, $\sum_{c\in\mathcal{C}}$ respectively.

A.1 Proof of Theorem 3.3

We define

$$\tilde{Q}(\theta,\gamma) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}'(\theta(c_i) - \theta^0(c_i^0))^2 + \sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}^2 \tag{A.1}$$

and recall that

$$\widehat{Q}(\theta,\gamma) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}'(\theta^0(c_i^0) - \theta(c_i)) + e_{it})^2$$

we begin by showing uniform convergence of risk surfaces

Lemma A.1. $\sup_{\theta\in\Theta,\gamma\in\Gamma}\left[\widehat{Q}(\theta,\gamma) - \tilde{Q}(\theta,\gamma)\right] \xrightarrow{p} 0$ *as* $N,T\to\infty$

Proof. Define $\Delta\theta_i = \theta^0(c_i^0) - \theta(c_i)$ and note that

$$\widehat{Q}(\theta,\gamma) - \tilde{Q}(\theta,\gamma) = \frac{2}{NT}\sum_{i,t} e_{it}x_{it}'\Delta\theta_i$$

Then we can compute

$$\begin{aligned}\left(\widehat{Q}(\theta,\gamma) - \tilde{Q}(\theta,\gamma)\right)^2 &= \left[\frac{1}{N}\sum_i \Delta\theta_i'\left(\frac{1}{T}\sum_t e_{it}x_{it}\right)\right]^2 \le \frac{1}{N}\sum_i \|\Delta\theta_i\|^2\frac{1}{T^2}\left\|\sum_t e_{it}x_{it}\right\|^2 \\ &\lesssim \frac{1}{NT^2}\sum_i\sum_{t,s} e_{it}e_{is}x_{it}'x_{is} = o_P(1)\end{aligned} \tag{A.2}$$

The first inequality follows from Jensen and Cauchy-Schwarz, the next uses assumption 3.1.(a) on compactness, and the final equality is assumption 3.1.(b). Taking $\sup_{\theta\in\Theta,\gamma\in\Gamma}$ on both sides of the inequality gives the statement of the lemma. □

Now we make the usual observation that $\tilde{Q}(\widehat{\theta},\widehat{\gamma}) - \tilde{Q}(\theta^0,\gamma^0) = o_P(1)$ since

$$\begin{aligned}\tilde{Q}(\widehat{\theta},\widehat{\gamma}) &= \widehat{Q}(\widehat{\theta},\widehat{\gamma}) + [\tilde{Q}(\widehat{\theta},\widehat{\gamma}) - \widehat{Q}(\widehat{\theta},\widehat{\gamma})] \le \widehat{Q}(\widehat{\theta},\widehat{\gamma}) + \sup_{\theta\in\Theta,\gamma\in\Gamma}\left[(\tilde{Q}-\widehat{Q})(\theta,\gamma)\right] \\ &= \widehat{Q}(\widehat{\theta},\widehat{\gamma}) + o_P(1) \le \widehat{Q}(\theta^0,\gamma^0) + o_P(1) = \tilde{Q}(\theta^0,\gamma^0) + o_P(1) \\ &\implies 0 \le \tilde{Q}(\widehat{\theta},\widehat{\gamma}) - \tilde{Q}(\theta^0,\gamma^0) \le o_P(1)\end{aligned}$$

The second equality follows from lemma A.1, and the third equality from the definition of the estimator. The next step is to show curvature of the auxiliary sample risk $\tilde{Q}$. The following curvature calculation is almost identical to the proof in appendix S6 in BM. For arbitrary $\theta \in \Theta$ and $\gamma \in \Gamma$, we have

$$\begin{aligned}\tilde{Q}(\theta,\gamma) - \tilde{Q}(\theta^0,\gamma^0) &= \frac{1}{NT}\sum_{i,t}(\theta(c_i) - \theta^0(c_i^0))' x_{it}x_{it}'(\theta(c_i) - \theta^0(c_i^0)) \qquad \text{(A.3)}\\ &= \frac{1}{NT}\sum_{i,t}\sum_{c\in\mathcal{C}}\sum_{c'\in\mathcal{C}}(\theta(c) - \theta^0(c'))' x_{it}x_{it}'(\theta(c) - \theta^0(c'))\mathbb{1}(c_i^0 = c')\mathbb{1}(c_i = c)\\ &= \sum_{c,c'}(\theta(c) - \theta^0(c'))'\left(\frac{1}{NT}\sum_{i,t} x_{it}x_{it}'\mathbb{1}(c_i^0 = c')\mathbb{1}(c_i = c)\right)(\theta(c) - \theta^0(c'))\end{aligned}$$

Let $M(c,c',\gamma) \equiv \frac{1}{NT}\sum_{i,t} x_{it}x_{it}'\mathbb{1}(c_i^0 = c')\mathbb{1}(c_i = c)$ and define $\rho(c,c',\gamma) \equiv \lambda_{min}(M(c,c',\gamma))$. Then the last line is bounded below by

$$\begin{aligned}\sum_{c,c'}\|\theta(c) - \theta^0(c')\|^2\rho(c,c',\gamma) &\geq \sum_{c,c'}\rho(c,c',\gamma)\inf_{x\in\mathcal{C}}\|\theta(x) - \theta^0(c')\|^2\\ &\geq \sum_{c'}\inf_{\tilde{c},\gamma}\max_{c}\rho(c,\tilde{c},\gamma)\inf_{x\in\mathcal{C}}\|\theta(x) - \theta^0(c')\|^2\\ &\geq \left(\inf_{\tilde{c},\gamma}\max_{c}\rho(c,\tilde{c},\gamma)\right)\max_{c'}\inf_{x\in\mathcal{C}}\|\theta(x) - \theta^0(c')\|^2\end{aligned}$$

Then we see that

$$\begin{aligned}o_P(1) = \tilde{Q}(\widehat{\theta},\widehat{\gamma}) - \tilde{Q}(\theta^0,\gamma^0) &\geq \left(\inf_{\tilde{c},\gamma}\max_{c}\rho(c,\tilde{c},\gamma)\right)\max_{c'}\inf_{x\in\mathcal{C}}\|\widehat{\theta}(x) - \theta^0(c')\|^2\\ &\geq \delta\max_{c}\inf_{x\in\mathcal{C}}\|\widehat{\theta}(x) - \theta^0(c)\|^2 - o_P(1)\\ &\implies \max_{c}\inf_{x\in\mathcal{C}}\|\widehat{\theta}(x) - \theta^0(c)\|^2 = o_P(1)\end{aligned}$$

So that $\max_c \inf_{x\in\mathcal{C}}\|\widehat{\theta}(x) - \theta^0(c)\|^2 = o_P(1)$. The second equality uses assumptions 3.1.(d) and 3.1.(a). As noted in the main text, the problem $\operatorname{argmin}_{\gamma\in\Gamma,\theta\in\Theta}\widehat{Q}(\theta,\gamma)$ is invariant to permutations of the labels in $\mathcal{C}$ and their associated parameter vectors in θ. The next step is resolve this degeneracy by giving a well-defined estimator of $\theta_{\ell a}^0$ for each $\ell \in [B]$, $a \in [k_\ell]$.

Lemma A.2. *Define $\sigma(c) \equiv \operatorname{argmin}_{x\in\mathcal{C}}\|\widehat{\theta}(x) - \theta^0(c)\|^2$. The map $\sigma_\ell(x) \equiv \sigma(c)_\ell$ for any c such that $c_\ell = x$ is well defined.*

Proof. Existence of the map is clear; we show it is a function. Note that $\min_{x\in\mathcal{C}}\|\widehat{\theta}(x) - \theta^0(c)\|^2 = \sum_{\ell=1}^{B}\min_{x_\ell\in[k_\ell]}\|\widehat{\theta}_\ell(x_\ell) - \theta_\ell^0(c_\ell)\|^2$. Then for any $c \in \mathcal{C}$, we have $\sigma(c)_\ell = f(c_\ell, \theta^0, \widehat{\theta})$, so $c_\ell = c_\ell' \implies \sigma(c)_\ell = \sigma(c')_\ell$. □

In fact, since $\min_{x\in\mathcal{C}}\|\widehat{\theta}(x) - \theta^0(c)\|^2 = o_P(1)$, the proof of the lemma above shows that $\|\theta_{\ell a}^0 - \widehat{\theta}_{\ell\sigma_\ell(a)}\| = o_P(1)$ for all groupings ℓ and each cluster $a \in [k_\ell]$, completing the main statement of the theorem.

We show that for each ℓ, σ_ℓ is a bijection w.h.p. Since $\sigma_\ell : [k_\ell] \to [k_\ell]$, it suffices to show injection. Let $a, b \in [k_\ell]$, then we have

$$\|\theta^0_{\ell a} - \theta^0_{\ell b}\| \leq \|\theta^0_{\ell a} - \widehat{\theta}_{\ell\sigma_\ell(a)}\| + \|\widehat{\theta}_{\ell\sigma_\ell(a)} - \widehat{\theta}_{\ell\sigma_\ell(b)}\| + \|\widehat{\theta}_{\ell\sigma_\ell(b)} - \theta^0_\ell(b)\| \leq \|\widehat{\theta}_{\ell\sigma_\ell(a)} - \widehat{\theta}_{\ell\sigma_\ell(b)}\| + X_{N,T}$$

Where $X_{N,T} = o_P(1)$ as $N, T \to \infty$. Then $\{\sigma_\ell(a) = \sigma_\ell(b) \implies a = b\} \subset \{X_{N,T} < d(\ell, a, b)\}$ by assumption 3.1.(c), and the latter event has probability going to 1 as $N, T \to \infty$. Since $\cap_\ell\{\sigma_\ell \, \text{injective}\}$ is an intersection of finitely many events of the form above, we have

$$\mathbb{P}(\sigma_\ell \, \text{injective} \, \forall \ell) \to 1$$

as $N, T \to \infty$.

A.2 Proof of Theorem 3.5

The proof closely follows the strategy used in Bonhomme and Manresa (2015). We define the problem

$$\widehat{Q}(\theta) = \inf_{\gamma \in \Gamma} \widehat{Q}(\theta, \gamma) \tag{A.4}$$

And let $\widehat{c}_i(\theta)$ denote the cluster assignments that minimize the RHS of A.4. Thus, $\widehat{Q}$ is the original problem from 2.4 with the cluster assignments concentrated out. The proof of theorem 3.5 crucially relies on the following lemma

Lemma A.3. *For $\eta > 0$, define $\mathcal{N}_\eta = \{\theta \in \Theta : \max_{c \in \mathcal{C}} \|\theta(c) - \theta^0(c)\| < \eta\}$. Then there exists $\eta > 0$ such that for all $a > 0$*

$$\sup_{\theta \in \mathcal{N}_\eta} \frac{1}{N} \sum_i^N \mathbb{1}(\widehat{c}_i(\theta) \neq c^0_i) = o_P(T^{-a})$$

Proof. First note that for each $i \in [N]$, we may write $\mathbb{1}(\widehat{c}_i(\theta) \neq c^0_i)$ as

$$\begin{aligned}
&\sum_{c \neq c^0_i} \mathbb{1}(\widehat{c}_i(\theta) = c) \leq \sum_{c \neq c^0_i} \mathbb{1}\left(\frac{1}{T}\sum_{t=1}^T (y_{it} - x'_{it}\theta(c))^2 \leq \frac{1}{T}\sum_{t=1}^T (y_{it} - x'_{it}\theta(c^0_i))^2\right) \\
&= \sum_{c \neq c^0_i} \mathbb{1}\left(\frac{1}{T}\sum_{t=1}^T (x'_{it}(\theta^0(c^0_i) - \theta(c)) + e_{it})^2 \leq \frac{1}{T}\sum_{t=1}^T (x'_{it}(\theta^0(c^0_i) - \theta(c^0_i)) + e_{it})^2\right) \\
&\leq \sum_{c \in \mathcal{C}} \max_{c' \neq c} \mathbb{1}\left(\frac{1}{T}\sum_{t=1}^T (x'_{it}(\theta^0(c') - \theta(c)) + e_{it})^2 \leq \frac{1}{T}\sum_{t=1}^T (x'_{it}(\theta^0(c') - \theta(c')) + e_{it})^2\right) \\
&\equiv \sum_{c \in \mathcal{C}} \max_{c' \neq c} Z_{ic}(c', \theta)
\end{aligned}$$

We can rewrite inequality inside the indicator as (for $c \in \mathcal{C}$) as

$$B^T_i(\theta) \equiv \frac{1}{T}\sum_{t=1}^T 2e_{it}x'_{it}(\theta(c') - \theta(c)) + [x'_{it}(\theta^0(c') - \theta(c))]^2 - [x'_{it}(\theta^0(c') - \theta(c'))]^2 \leq 0$$

Then we calculate

$$\begin{aligned}|B_i^T(\theta) - B_i^T(\theta^0)| \le& \left|\frac{1}{T}\sum_{t=1}^{T} 2e_{it}x'_{it}(\theta(c') - \theta^0(c') + \theta^0(c) - \theta(c))\right| \\ &+ \left|\frac{1}{T}\sum_{t=1}^{T}[x'_{it}(\theta^0(c') - \theta(c))]^2 - [x'_{it}(\theta^0(c') - \theta(c'))]^2 - [x'_{it}(\theta^0(c') - \theta^0(c))]^2\right| \end{aligned} \tag{A.5}$$

The second term is bounded above by

$$\begin{aligned}&\left|\frac{1}{T}\sum_{t=1}^{T}[x'_{it}(\theta^0(c') - \theta(c'))]^2\right| + \left|\frac{1}{T}\sum_{t=1}^{T}[x'_{it}(\theta^0(c') - \theta^0(c) + \theta^0(c) - \theta(c))]^2 - [x'_{it}(\theta^0(c') - \theta^0(c))]^2\right| \\ &\le \left|\frac{1}{T}\sum_{t=1}^{T}[x'_{it}(\theta^0(c') - \theta(c'))]^2\right| + \left|\frac{1}{T}\sum_{t=1}^{T}[x'_{it}(\theta^0(c) - \theta(c))]^2\right| \\ &+ 2\left|\frac{1}{T}\sum_{t=1}^{T}[x'_{it}(\theta^0(c') - \theta^0(c))][x'_{it}(\theta^0(c) - \theta(c))]\right|\end{aligned}$$

Using $\theta \in \mathcal{N}_\eta$ and applying the triangle inequality, Cauchy-Schwarz, and assumption 3.1.(a), the last expression can be bounded above by

$$2\eta^2\frac{1}{T}\sum_{t}\|x_{it}\|^2 + 2M\eta\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2 = 2\eta(M+\eta)\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2 \le 4\eta M\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2$$

Similarly, one can show that the first term in A.5 is bounded by $4\eta\left\|\frac{1}{T}\sum_{t=1}^{T} e_{it}x_{it}\right\|$. This shows that for any $c \neq c'$

$$\begin{aligned}&\sup_{\theta\in\mathcal{N}_\eta} Z_{ic}(c',\theta) \le \sup_{\theta\in\mathcal{N}_\eta} \mathbb{1}(B_i^T(\theta^0) \le |B_i^T(\theta) - B_i^T(\theta^0)|) \\ &\le \mathbb{1}\left(B_i^T(\theta^0) \le 4\eta M\frac{1}{T}\sum_{t}\|x_{it}\|^2 + 4\eta\left\|\frac{1}{T}\sum_{t=1}^{T}e_{it}x_{it}\right\|\right) \\ &= \mathbb{1}\left(\frac{1}{T}\sum_{t=1}^{T}2e_{it}x'_{it}(\theta^0(c') - \theta^0(c)) + [x'_{it}(\theta^0(c') - \theta^0(c))]^2 \le 4\eta M\frac{1}{T}\sum_{t}\|x_{it}\|^2 + 4\eta\left\|\frac{1}{T}\sum_{t=1}^{T}e_{it}x_{it}\right\|\right) \\ &\le \mathbb{1}\left(\frac{1}{T}\sum_{t=1}^{T}[x'_{it}(\theta^0(c') - \theta^0(c))]^2 \le 4\eta M\frac{1}{T}\sum_{t}\|x_{it}\|^2 + (4\eta + 2M)\left\|\frac{1}{T}\sum_{t=1}^{T}e_{it}x_{it}\right\|\right)\end{aligned}$$

Where $\mathrm{Diam}(\Theta) \le M$ by assumption 3.1.(a). Let M' be the constant from 3.4.(f) Then taking expectations

$$\begin{aligned}
&\max_i \mathbb{E} \sup_{\theta\in\mathcal{N}_\eta} Z_{ic}(c',\theta)\\
&\leq \max_i \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}[x_{it}'(\theta^0(c')-\theta^0(c))]^2 \leq 4\eta M M' + (4\eta+2M)\eta\right)\\
&+\max_i \mathbb{P}\left(\frac{1}{T}\sum_t \|x_{it}\|^2 > M'\right) + \max_i \mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T} e_{it}x_{it}\right\| > \eta\right) \qquad \text{(A.6)}
\end{aligned}$$

To bound these terms we will use Lemma B.5 from BM, which is an application of Rio (2017), on concentration of strongly mixing sequences. We restate the lemma here

Lemma A.4 (BM Lemma B.5)**.** *Let z_t be a strongly mixing process with zero mean, with strong mixing coefficients $\alpha(t)$ satisfying 3.4.(c) and tails $\mathbb{P}(|z_t| > z) \leq e^{1-(z/f)^{d_2}}$. Then for all $a > 0$ and $z > 0$, we have as $T \to \infty$*

$$T^a \cdot \mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} z_t\right| > z\right) \leq r(T) = o(1)$$

Moreover, the function r only depends on the constants b, f, d_1, d_2 from assumption 3.4.(c) and 3.4.(d).

We want to apply this result to the terms in A.6 above. Observe that if $\{x_{it}\}$ is strongly mixing with mixing coefficients $\alpha(t)$ then $\{(x_{it}'(\theta^0(c)-\theta^0(c')))^2\}$ is also strongly mixing with coefficients uniformly bounded above by $\alpha(t)$. This follows because continuous transformations can only decrease the mixing coefficients. For completeness, we can show that the tail assumptions in 3.4.(d) imply that that $z_t \equiv (x_{it}'(\theta(c)-\theta(c')))^2 - E(x_{it}'(\theta(c)-\theta(c')))^2$ also satisfies the tail bound required in the lemma. Let $\Delta\theta \equiv \theta^0(c) - \theta^0(c')$ and recall $p = \dim(x_{it})$, then

$$\begin{aligned}
\mathbb{P}((x_{it}'(\theta(c)-\theta(c')))^2 - E(x_{it}'(\theta(c)-\theta(c')))^2 > z) &= \mathbb{P}(\Delta\theta'(x_{it}x_{it}' - Ex_{it}x_{it}')\Delta\theta > z)\\
&= \mathbb{P}\left(\sum_{j,j'} \Delta\theta^j \Delta\theta^{j'}(x_{it}^k x_{it}^{k'} - Ex_{it}^k x_{it}^{k'}) > z\right)\\
&\leq \sum_{k,k'} \mathbb{P}\left(|(x_{it}^k x_{it}^{k'} - Ex_{it}^k x_{it}^{k'})| > \frac{z}{p^2(M')^2}\right) \qquad \text{(A.7)}
\end{aligned}$$

Note that $\mathbb{P}(|Z| > z) \leq e^{1-(z/f)^{d_2}}$ does not imply that $C\cdot\mathbb{P}(|Z| > z)$ satisfies a tail bound of the same form (possibly with different constants f, d_2) if $C > 1$. However, a calculation shows that for any $C > 1$, there exist f', d_2' such for all $z > 0$, $\min(1, Ce^{1-(z/f)^{d_2}}) \leq \min(1, e^{1-(z/f')^{d_2'}})$, so this is not a problem. This shows that the final term in A.7 above satisfies a tail bound of the required form.

We now apply the lemma to each of the terms in equation A.6. Choose η such that $4\eta M M' + (4\eta + 2M)\eta < \frac{1}{3}d_{min}$. Let $g_{it} \equiv E(x_{it}'(\theta^0(c')-\theta^0(c)))^2$ and T' such that $\min_i \frac{1}{T}\sum_{t=1}^{T'} g_{it} > (1/2)d_{min}$, using assumption 3.4.(e). Then for $T > T'$, the first term in

A.6 is

$$\max_i \mathbb{P}\left(\frac{1}{T}\sum_t \left([x_{it}'(\theta^0(c')-\theta^0(c))]^2 - g_{it}\right) \leq 4\eta M M' + (4\eta+2M)\eta - \frac{1}{T}\sum_t g_{it}\right)$$
$$\leq \max_i \mathbb{P}\left(\left|\frac{1}{T}\sum_t [x_{it}'(\theta^0(c')-\theta^0(c))]^2 - g_{it}\right| \geq \frac{1}{6}d_{min}\right) = o(T^{-a})$$

where the last line follows from applying lemma A.4 with $z_{it} = [x_{it}'(\theta^0(c')-\theta^0(c))]^2 - g_{it}$. A similar argument using assumptions 3.4.(c), 3.4.(d), 3.4.(e) on the process $\{e_{it}x_{it}\}_t$ shows that the second term in equation *A*.6 is also $o(T^{-a})$, and the final term is just as assumption 3.4.(f).

Then for $\epsilon > 0$, the Markov inequality gives

$$\mathbb{P}\left(T^a \sup_{\theta\in\mathcal{N}_\eta} \frac{1}{N}\sum_{i=1}^N \mathbb{1}(\widehat{c}_i(\theta) \neq c_i^0) > \epsilon\right) \leq T^a \frac{1}{\epsilon}\mathbb{E}\sup_{\theta\in\mathcal{N}_\eta}\frac{1}{N}\sum_{i=1}^N\sum_{c\in\mathcal{C}}\max_{c'\neq c} Z_{ic}(c',\theta)$$
$$\leq T^a\frac{1}{\epsilon}\frac{1}{N}\sum_{i=1}^N\sum_{c\in\mathcal{C}}\sum_{c'\neq c}\mathbb{E}\sup_{\theta\in\mathcal{N}_\eta} Z_{ic}(c',\theta)$$
$$\leq T^a\frac{1}{\epsilon}\frac{1}{N}\sum_{i=1}^N |\mathcal{C}|^2 \max_{c\neq c'}\max_{i\in[N]}\mathbb{E}\sup_{\theta\in\mathcal{N}_\eta} Z_{ic}(c',\theta) = o(1)$$

This completes the proof of the lemma. □

In what follows, we let η satisfy the conditions posited in A.3. Recall the sample risk with oracle cluster membership $\tilde{Q} \equiv \widehat{Q}(\theta,\gamma^0)$. We show that for every $a > 0$, $\sup_{\theta\in\mathcal{N}_\eta}(\widehat{Q} - \tilde{Q})(\theta) = o_P(T^{-a})$. For any $\theta \in \mathcal{N}_\eta$, we can write

$$|(\widehat{Q}-\tilde{Q})(\theta)| = \left|\frac{1}{NT}\sum_{i,t}[y_{it} - x_{it}'\theta(\widehat{c}_i(\theta))]^2 - [y_{it} - x_{it}'\theta(c_i^0)]^2\right| \leq \left|\frac{1}{NT}\sum_{i,t} 2e_{it}x_{it}'(\theta(c_i^0) - \theta(\widehat{c}_i(\theta)))\right|$$
$$+\left|\frac{1}{NT}\sum_{i,t}[x_{it}'(\theta^0(c_i^0)-\theta(c_i^0))]^2 - [x_{it}'(\theta^0(c_i^0)-\theta(\widehat{c}_i(\theta)))]^2\right| \tag{A.8}$$

The first term on the right hand side is bounded above by

$$\begin{aligned}\frac{1}{N}\sum_i \left|(\theta(c_i^0)-\theta(\widehat{c}_i(\theta)))'\frac{1}{T}\sum_t 2e_{it}x_{it}\mathbb{1}(\widehat{c}_i(\theta)\neq c_i^0)\right| &\leq \frac{2M}{N}\sum_i \mathbb{1}(\widehat{c}_i(\theta)\neq c_i^0)\left\|\frac{1}{T}\sum_t e_{it}x_{it}\right\| \\ &\leq \left(\frac{1}{N}\sum_i \mathbb{1}(\widehat{c}_i(\theta)\neq c_i^0)\right)^{1/2}\left(\frac{1}{N}\sum_i\left\|\frac{1}{T}\sum_t e_{it}x_{it}\right\|^2\right)^{1/2} \\ &= o_P(T^{-(2a)/2})\left(\frac{1}{NT^2}\sum_i\sum_{t,s} e_{it}e_{is}x_{it}'x_{is}\right)^{1/2} \\ &= o_P(T^{-(2a)/2})o_P(1) = o_P(T^{-a})\end{aligned}$$

where the last line follows by lemma A.3 and assumption 3.1.(b). The second term in equation A.8 can be expanded as

$$\begin{aligned}&\left|\frac{1}{NT}\sum_{i,t}[x_{it}'(\theta^0(c_i^0)-\theta(\widehat{c}_i(\theta))+\theta(\widehat{c}_i(\theta))-\theta(c_i^0))]^2-[x_{it}'(\theta^0(c_i^0)-\theta(\widehat{c}_i(\theta)))]^2\right| \\ \leq &\left|\frac{1}{NT}\sum_{i,t}2x_{it}'(\theta^0(c_i^0)-\theta(\widehat{c}_i(\theta)))x_{it}'(\theta(\widehat{c}_i(\theta))-\theta(c_i^0))\right|+\left|\frac{1}{NT}\sum_{i,t}(x_{it}'(\theta(\widehat{c}_i(\theta))-\theta(c_i^0)))^2\right|\end{aligned}$$

For instance, the second term can be rewritten

$$\begin{aligned}\left|\frac{1}{N}\sum_i\frac{1}{T}\sum_t(x_{it}'(\theta(\widehat{c}_i(\theta))-\theta(c_i^0)))^2\mathbb{1}(\widehat{c}_i(\theta)\neq c_i^0)\right| &\leq \left|\frac{M^2}{N}\sum_i\mathbb{1}(\widehat{c}_i(\theta)\neq c_i^0)\frac{1}{T}\sum_t\|x_{it}\|^2\right| \\ \leq M^2\left(\frac{1}{N}\sum_i\mathbb{1}(\widehat{c}_i(\theta)\neq c_i^0)\right)^{\frac{1}{2}}&\left(\frac{1}{N}\sum_i\left(\frac{1}{T}\sum_t\|x_{it}\|^2\right)^2\right)^{\frac{1}{2}} \leq o_P(T^{-a})\end{aligned} \tag{A.9}$$

where the last inequality uses lemma A.3 and assumption 3.4.(a). It follows that

$$\sup_{\theta\in\mathcal{N}_\eta}|(\widehat{Q}-\tilde{Q})(\theta)| = o_P(T^{-a}) \tag{A.10}$$

We claim that $\tilde{\theta}-\theta^0 = o_P(1)$. Note that since $\tilde{Q}(\theta)=\widehat{Q}(\theta,\gamma^0)$, it suffices to check that the assumptions in 3.1 hold for $\Gamma'\equiv\{\gamma^0\}$. The only thing we need to check is assumption 3.1.(d), which is clear since $\{\gamma^0\}\subset\Gamma$ implies $\inf_{c',\gamma\in\{\gamma^0\}}\max_c\rho(c,c',\gamma)\geq\inf_{c',\gamma\in\Gamma}\max_c\rho(c,c',\gamma)\geq\delta-o_P(1)$ as $N,T\to\infty$ by assumption 3.1.(d). This shows $\tilde{\theta}-\theta^0=o_P(1)$.

Next, we will show that for any $a>0$

$$\tilde{Q}(\widehat{\theta})-\tilde{Q}(\tilde{\theta}) = o_P(T^{-a}) \tag{A.11}$$

Let $a > 0$ and $\epsilon > 0$. Define the event $E_T \equiv \{T^a(\tilde{Q}(\widehat{\theta}) - \tilde{Q}(\tilde{\theta})) > \epsilon\}$.

$$\mathbb{P}(E_T) \leq \mathbb{P}(E_T \cap \{\widehat{\theta}, \tilde{\theta} \in \mathcal{N}_\eta\}) + \mathbb{P}(\widehat{\theta} \notin \mathcal{N}_\eta \text{ or } \tilde{\theta} \notin \mathcal{N}_\eta) = \mathbb{P}(E_T \cap \{\widehat{\theta}, \tilde{\theta} \in \mathcal{N}_\eta\}) + o(1)$$

The final equality follows from a union bound and consistency of $\widehat{\theta}$ and $\tilde{\theta}$. On the event $E_T \cap \{\widehat{\theta}, \tilde{\theta} \in \mathcal{N}_\eta\}$, we have

$$\begin{aligned} 0 \leq \tilde{Q}(\widehat{\theta}) - \tilde{Q}(\tilde{\theta}) &= (\tilde{Q}(\widehat{\theta}) - \widehat{Q}(\widehat{\theta})) + (\widehat{Q}(\widehat{\theta}) - \widehat{Q}(\tilde{\theta})) + (\widehat{Q}(\tilde{\theta}) - \tilde{Q}(\tilde{\theta})) \\ &\leq 2 \sup_{\theta \in \mathcal{N}_\eta} |(\tilde{Q} - \widehat{Q})(\theta)| \end{aligned}$$

where we used that $(\widehat{Q}(\widehat{\theta}) - \widehat{Q}(\tilde{\theta})) \leq 0$ by the definition of $\widehat{\theta}$. Then using the inequality above, apparently

$$\mathbb{P}(E_T) \leq \mathbb{P}\left(T^a \cdot 2 \sup_{\theta \in \mathcal{N}_\eta} |(\tilde{Q} - \widehat{Q})(\theta)| > \epsilon\right) + o(1) = o(1)$$

by equation A.10. This completes the proof of A.11. We now show a curvature lower bound for $\tilde{Q}$. For every $1 \leq \ell \leq G$ and each $x \in [k_\ell]$, $\tilde{\theta} \in \operatorname{argmin}_{\theta \in \Theta} \tilde{Q}(\theta)$ implies

$$0 = \nabla_{\theta_{\ell x}} \tilde{Q}(\tilde{\theta}) = \frac{2}{NT} \sum_{i: c^0_{i\ell} = x} \sum_{t=1}^{T} (y_{it} - x'_{it}\tilde{\theta}(c^0_i)) x^\ell_{it} \tag{A.12}$$

Define $\tilde{e}_{it} \equiv (y_{it} - x'_{it}\tilde{\theta}(c^0_i)$ and compute

$$\begin{aligned} \tilde{Q}(\widehat{\theta}) - \tilde{Q}(\tilde{\theta}) &= \frac{1}{NT}\sum_{i,t}(y_{it} - x'_{it}\widehat{\theta}(c^0_i))^2 - \frac{1}{NT}\sum_{i,t}(y_{it} - x'_{it}\tilde{\theta}(c^0_i))^2 \\ &= \frac{1}{NT}\sum_{i,t}(y_{it} - x'_{it}\tilde{\theta}(c^0_i) + x'_{it}[\tilde{\theta}(c^0_i) - \widehat{\theta}(c^0_i)])^2 - \frac{1}{NT}\sum_{i,t}(y_{it} - x'_{it}\tilde{\theta}(c^0_i))^2 \\ &= \frac{1}{NT}\sum_{i,t}(x'_{it}[\tilde{\theta}(c^0_i) - \widehat{\theta}(c^0_i)])^2 + \frac{1}{NT}\sum_{i,t}\tilde{e}_{it}x'_{it}[\tilde{\theta}(c^0_i) - \widehat{\theta}(c^0_i)] \end{aligned}$$

We claim that the second term is identically zero. Define a map[5] $F : \Theta - \Theta \to \mathbb{R}$ by $F(\theta) = \sum_{i,t} \tilde{e}_{it}x'_{it}\theta(c^0_i)$. Note that for any $1 \leq \ell \leq G$, we can write

$$\begin{aligned} F(\theta) &= \sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{e}_{it}x'_{it}\theta(c^0_i) = \sum_{t=1}^{T}\sum_{x\in[k_\ell]}\sum_{i:c^0_{i\ell}=x}\tilde{e}_{it}x'_{it}\theta(c^0_i) = \sum_{t=1}^{T}\sum_{x\in[k_\ell]}\sum_{i:c^0_{i\ell}=x}\sum_{\tilde{\ell}}\tilde{e}_{it}\langle x^{\tilde{\ell}}_{it}, \theta^{\tilde{\ell}}(c^0_i)\rangle \\ &= \sum_{t=1}^{T}\sum_{x\in[k_\ell]}\sum_{i:c^0_{i\ell}=x}\sum_{\tilde{\ell}\neq\ell}\tilde{e}_{it}\langle x^{\tilde{\ell}}_{it}, \theta^{\tilde{\ell}}(c^0_i)\rangle + \sum_{x\in[k_\ell]}\sum_{i:c^0_{i\ell}=x}\sum_{t=1}^{T}\tilde{e}_{it}\langle x^{\ell}_{it}, \theta^{\ell}(c^0_i)\rangle \\ &= \sum_{t=1}^{T}\sum_{x\in[k_\ell]}\sum_{i:c^0_{i\ell}=x}\sum_{\tilde{\ell}\neq\ell}\tilde{e}_{it}\langle x^{\tilde{\ell}}_{it}, \theta^{\tilde{\ell}}(c^0_i)\rangle \end{aligned}$$

where we have used that $\sum_{i:c^0_{i\ell}=x}\sum_{t=1}^{T}\tilde{e}_{it}(x^\ell_{it})'\theta^\ell(c^0_i) = 0$ for each x by the first order

[5] For S_1 and S_2 subsets of the same vector space, we define $S_1 - S_2 \equiv \{s_1 - s_2 : s_i \in S_i, i = 1, 2\}$.

condition A.12. Since the last expression doesn't involve θ^{ℓ}, we conclude that for any $\theta \in \mathrm{Dom}(F)$, the equality $F(\theta^{\ell}, \theta^{-\ell}) = F(0, \theta^{-\ell})$ holds. Applying this fact inductively, we find that $F = F(0) = 0$ identically. In particular, $F(\tilde{\theta} - \widehat{\theta}) = 0$, which is what we needed to show. Then similar to the proof of 3.3, we calculate

$$\begin{aligned}
\tilde{Q}(\widehat{\theta}) - \tilde{Q}(\tilde{\theta}) &= \frac{1}{NT}\sum_{i,t}(x_{it}'[\tilde{\theta}(c_i^0) - \widehat{\theta}(c_i^0)])^2 \\
&= \sum_{c \in \mathcal{C}} (\tilde{\theta}(c) - \widehat{\theta}(c))' \left(\frac{1}{NT} \sum_{i,t} \mathbb{1}(c_i^0 = c) x_{it} x_{it}' \right)' (\tilde{\theta}(c) - \widehat{\theta}(c)) \\
&\geq \sum_{c \in \mathcal{C}} \|\tilde{\theta}(c) - \widehat{\theta}(c)\|^2 \lambda_{min}(M_{NT}^c) \geq \sum_{c \in \mathcal{C}} \|\tilde{\theta}(c) - \widehat{\theta}(c)\|^2 \min_{c' \in \mathcal{C}} \lambda_{min}(M_{NT}^{c'})
\end{aligned}$$

Define $W_{NT} \equiv \min_{c' \in \mathcal{C}} \lambda_{min}(M_{NT}^{c'})$, so that $W_{NT}^c \geq 0$ by positive semi-definiteness of M_{NT}^c for all N, T, c. Also denote $E_{NT} = \{W_{NT} > \underline{\rho}/2\}$. Then by assumption 3.4.(b), $\mathbb{P}(E_{NT}) = o(1)$. We have

$$\begin{aligned}
\sum_{c \in \mathcal{C}} \|\tilde{\theta}(c) - \widehat{\theta}(c)\|^2 \min_{c' \in \mathcal{C}} \lambda_{min}(M_{NT}^{c'}) &= \sum_{c \in \mathcal{C}} \|\tilde{\theta}(c) - \widehat{\theta}(c)\|^2 (\underline{\rho}/2 + (W_{NT} - \underline{\rho}/2)) \\
&\geq \sum_{c \in \mathcal{C}} \|\tilde{\theta}(c) - \widehat{\theta}(c)\|^2 (\underline{\rho}/2 + (W_{NT} - \underline{\rho}/2)\mathbb{1}(W_{NT} < \underline{\rho}/2)) \\
&= \sum_{c \in \mathcal{C}} \|\tilde{\theta}(c) - \widehat{\theta}(c)\|^2 \underline{\rho}/2 + o_P(T^{-a})
\end{aligned}$$

In the last line we used the compactness assumption 3.1.(a), the fact that $|W_{NT} - \underline{\rho}/2| \leq \underline{\rho}/2$ on E_{NT}^c, and $T^a \mathbb{1}(E_{NT}^c) = o_P(1)$ for any $a > 0$ since $\mathbb{P}(E_{NT}) \to 1$ by assumption 3.4.(b). Combining this with equation A.11 shows that $\sup_{c \in \mathcal{C}} \|\tilde{\theta}(c) - \widehat{\theta}(c)\| = o_P(T^{-a})$, which completes the proof of part 3.3 of the theorem.

For the second part of the theorem 3.4 on cluster assignment, note that for η satisfying the conditions in lemma A.3, using the bounds developed in the proof of the lemma we find that

$$\begin{aligned}
\mathbb{P}(\exists i : \widehat{c}_i \neq c_i^0) &\leq \mathbb{P}(\exists i : \exists \theta \in \mathcal{N}_\eta : \widehat{c}_i(\theta) \neq c_i^0 \text{ and } \widehat{\theta} \in \mathcal{N}_\eta) + \mathbb{P}(\widehat{\theta} \notin \mathcal{N}_\eta) \\
&\leq \sum_i \mathbb{P}(\exists \theta \in \mathcal{N}_\eta : \widehat{c}_i(\theta) \neq c_i^0) + o(1) = \sum_i \mathbb{E}[\sup_{\theta \in \mathcal{N}_\eta} \mathbb{1}(\widehat{c}_i(\theta) \neq c_i^0)] + o(1) \\
&\leq \sum_i \mathbb{E} \sup_{\theta \in \mathcal{N}_\eta} \sum_{c \in \mathcal{C}} \sum_{c' \neq c} Z_{ic}(c', \theta) + o(1) \leq \sum_i \sum_{c \in \mathcal{C}} \sum_{c' \neq c} \mathbb{E} \sup_{\theta \in \mathcal{N}_\eta} Z_{ic}(c', \theta) + o(1) \\
&= o(NT^{-a}) + o(1)
\end{aligned}$$

This completes the proof of the theorem.

A.3 Proof of Theorem 3.10

In this section, we prove consistency of model selection for the Cp criterion defined in the main text. The assumptions of theorem 3.10 (stated in assumption 3.9) are imposed everywhere in this section. First we need some additional definitions. Let $\Theta^k = \prod_\ell \mathbb{R}^{d_\ell \times k_\ell}$ be the parameter space for a model with $k = (k_1, \dots, k_B)$ clusters. Let $\mathcal{C}_k = \prod_i [k_i]$ and $\Gamma_k = [\,[N] \to \mathcal{C}_k\,]$ denote the set of possible cluster labels and cluster labelings of the cross-sectional units, where we may have $k \neq k^0$, the true number of clusters in each group.

Define $\widehat{Q}^0 = \widehat{Q}(\theta^0, \gamma^0) = \frac{1}{NT}\sum_{i,t} e_{it}^2$ to be the sample risk evaluated at the true model. We begin with the following lemma on the sample risk of different models.

Lemma A.5. *The following hold*

- *(i) If $k = k^0$, then $\widehat{Q}(k) - \widehat{Q}^0 = O_p(\frac{1}{NT})$*
- *(ii) If $k > k^0$, then $\widehat{Q}(k) - \widehat{Q}^0 = o_p(T^{-1+3\epsilon})$*
- *(iii) If k is such that $k_i < k_i^0$ for some i, then $\widehat{Q}(k) - \widehat{Q}^0 = \Omega(1) + o_p(1)$*

Proof of (i) and (iii). Statement (i) follows from lemma B.2 in the supplemental appendix. We note that if $k = k_0$, then $\widehat{\theta}$ satisfies the conditions of lemma B.2 by our inference result theorem 3.8 and lemma A.3 above on the convergence of average classification risk.

For the proof of part (iii), first define $\Delta\widehat{\theta}_i^k = \theta^0(c_i^0) - \widehat{\theta}^k(\widehat{c}^k)$ and recall that

$$\widehat{Q}(k) - \widehat{Q}^0 = \frac{1}{NT}\sum_{i,t}(x_{it}'\Delta\widehat{\theta}_i^k)^2 + \frac{1}{NT}\sum_{i,t} e_{it}x_{it}'\Delta\widehat{\theta}_i^k \tag{A.13}$$

The expression $\frac{1}{NT}\sum_{i,t} e_{it}x_{it}'\Delta\widehat{\theta}_i$ was already shown to be $o_p(1)$ uniformly over $\Delta\widehat{\theta}_i \in \Theta$ in equation A.2 in the consistency proof. Similarly, the first term was analyzed in equation A.3. The exact same argument as before shows that for arbitrary $(\theta^k, \gamma^k) \in \Theta^k \times \Gamma^k$

$$\begin{aligned}
&\frac{1}{NT}\sum_{i,t}(x_{it}'(\theta^0(c_i^0) - \theta^k(c_i^k))^2 \\
&= \frac{1}{NT}\sum_{i,t}\sum_{c\in\mathcal{C}^k}\sum_{c'\in\mathcal{C}^{k_0}} (\theta^k(c) - \theta^0(c'))' x_{it}x_{it}'(\theta^k(c) - \theta^0(c'))\mathbb{1}(c_i^0 = c')\mathbb{1}(c_i^k = c) \\
&= \sum_{c\in\mathcal{C}^k}\sum_{c'\in\mathcal{C}^{k_0}} (\theta^k(c) - \theta^0(c'))' \left(\frac{1}{NT}\sum_{i,t} x_{it}x_{it}'\mathbb{1}(c_i^0 = c')\mathbb{1}(c_i^k = c)\right)(\theta^k(c) - \theta^0(c')) \\
&\geq \sum_{c\in\mathcal{C}^k}\sum_{c'\in\mathcal{C}^{k_0}} \|\theta^k(c) - \theta^0(c')\|^2 \rho(c, c', \gamma) \\
&\geq \sum_{c\in\mathcal{C}^k}\sum_{c'\in\mathcal{C}^{k_0}} \rho(c, c', \gamma) \max_{x\in\mathcal{C}^k} \|\theta^k(x) - \theta^0(c')\|^2 \\
&\geq \sum_{c'\in\mathcal{C}^{k_0}} \min_{\tilde{c}\in\mathcal{C}^{k_0}} \min_{\gamma^k\in\Gamma^k} \max_{c\in\mathcal{C}^k} \rho(c, \tilde{c}, \gamma) \min_{x\in\mathcal{C}^k} \|\theta^k(x) - \theta^0(c')\|^2 \\
&\geq (\delta - o_P(1)) \sum_{c'\in\mathcal{C}^{k_0}} \min_{x\in\mathcal{C}^k} \|\theta^k(x) - \theta^0(c')\|^2
\end{aligned}$$

We claim that $\max_{c'\in\mathcal{C}^{k_0}} \min_{x\in\mathcal{C}^k} \|\widehat{\theta}^k(x) - \theta^0(c')\|^2 = \Omega(1)$. Let $1 \leq \ell \leq B$ be such that $k_\ell < k_\ell^0$ and define $\sigma(j) = \text{argmin}_i \|\widehat{\theta}_{\ell i}^k - \theta_{\ell j}^0\|$. Since $k_\ell < k_\ell^0$, by the pigeonhole principle

$\sigma(j) = \sigma(i)$ for some $i, j \in [k_\ell^0]$. Then by cluster separation (assumption 3.1)

$$0 < d_{min} \leq \|\theta_{\ell j}^0 - \theta_{\ell i}^0\| \leq \|\theta_{\ell j}^0 - \widehat{\theta}_{\ell\sigma(j)}^k\| + \|\widehat{\theta}_{\ell\sigma(j)}^k - \widehat{\theta}_{\ell\sigma(i)}^k\| + \|\widehat{\theta}_{\ell\sigma(i)}^k - \theta_{\ell i}^0\|$$

Since the middle term on the RHS is 0, $\max(\|\theta_{\ell j}^0 - \widehat{\theta}_{\ell\sigma(j)}^k\|, \|\theta_{\ell i}^0 - \widehat{\theta}_{\ell\sigma(i)}^k\|) > d_{min}/2$. Without loss suppose the max is achieved at i. Then for any c' with $c'_\ell = i$, we have $\min_{x\in\mathcal{C}^k} \|\widehat{\theta}^k(x) - \theta^0(c')\|^2 \geq (d_{min}/2)^2$. Plugging in $(\widehat{\theta}^k, \widehat{\gamma}^k)$ into our uniform bound above, we find

$$\frac{1}{NT}\sum_{i,t}(x'_{it}(\theta^0(c_i^0) - \widehat{\theta}^k(\widehat{c}_i^k))^2 \geq (\delta - o_P(1)) \sum_{c'\in\mathcal{C}^{k_0}} \min_{x\in\mathcal{C}^k} \|\widehat{\theta}^k(x) - \theta^0(c')\|^2 \geq \delta(d_{min}/2)^2 - o_p(1)$$

where we have used compactness of Θ in the final line. Then we have shown that $\widehat{Q}(k) - \widehat{Q}^0 \geq \delta(d_{min}/2)^2 + o_p(1)$, which completes the proof of (ii). □

For the proof of part (i), we need to develop some extra machinery. In this section, we denote $m = (\theta, \gamma) \in \Theta \times \Gamma$, and let m^k and m^{k_0} be parameter, cluster label pairs in $\Theta^k \times \Gamma^k$ and $\Theta^{k_0} \times \Gamma^{k_0}$ respectively. We denote $\widehat{m}^k = (\widehat{\theta}^k, \widehat{\gamma}^k)$ and $m^{k_0} = (\theta^0, \gamma^0)$. Define

$$d(m, m') \equiv \frac{1}{N}\sum_i (\theta(c_i) - \theta'(c'_i))^2$$

The following key lemma forms the backbone of our inductive approach for establishing (near) $\sqrt{T}$-consistency for over-specified estimators.

Lemma A.6. *Let $k \geq k_0$ and $b_T \equiv T^{-\frac{1}{2}+\epsilon}$. Then for any sequence $a_T = o(1)$, we have*

$$d(\widehat{m}^k, m^{k_0}) = O_p(a_T) \implies d(\widehat{m}^k, m^{k_0}) = o_p(a_T^{1/2} b_T) \tag{A.14}$$

Proof. In what follows, let $(\widehat{\theta}^k, \widehat{\gamma}^k) = \operatorname{argmin}_{\theta\in\Theta^k,\gamma\in\Gamma^k} Q(\theta, \gamma)$ and again let $\Delta\widehat{\theta}_i^k = (\widehat{\theta}^k(\widehat{c}_i^k) - \theta^0(c_i^0))$. With $\tilde{Q}$ defined as in our consistency proof in equation A.1, we have

$$\begin{aligned}
|\tilde{Q}(\widehat{\theta}^k, \widehat{\gamma}^k) - \widehat{Q}(\widehat{\theta}^k, \widehat{\gamma}^k)| &= \left|-2\frac{1}{N}\sum_i \left\langle \Delta\widehat{\theta}_i^k, \left(\frac{1}{T}\sum_t e_{it}x_{it}\right)\right\rangle\right| \lesssim \frac{1}{N}\sum_i \|\Delta\widehat{\theta}_i^k\| \left\|\frac{1}{T}\sum_t e_{it}x_{it}\right\| \\
&\leq \left(\frac{1}{N}\sum_i \|\Delta\widehat{\theta}_i^k\|^2\right)^{1/2} \left(\frac{1}{N}\sum_i \left\|\frac{1}{T}\sum_t e_{it}x_{it}\right\|^2\right)^{1/2} \\
&\leq \left(\frac{1}{N}\sum_i \|\Delta\widehat{\theta}_i^k\|^2\right)^{1/2} \left(\sup_{i\in[N]} \left\|\frac{1}{T}\sum_t e_{it}x_{it}\right\|\right) \\
&= O_p(a_T^{1/2})o_p(b_T) = o_p(a_T^{1/2} b_T)
\end{aligned}$$

The second to last equality holds by our assumption and applying lemma B.3. Now we reason

$$\begin{aligned}\tilde{Q}(\widehat{\theta}^k,\widehat{\gamma}^k) &= \widehat{Q}(\widehat{\theta}^k,\widehat{\gamma}^k) + [\tilde{Q}(\widehat{\theta}^k,\widehat{\gamma}^k) - \widehat{Q}(\widehat{\theta}^k,\widehat{\gamma}^k)] \leq \widehat{Q}(\widehat{\theta}^k,\widehat{\gamma}^k) + |\tilde{Q}(\widehat{\theta}^k,\widehat{\gamma}^k) - \widehat{Q}(\widehat{\theta}^k,\widehat{\gamma}^k)| \\ &= \widehat{Q}(\widehat{\theta}^k,\widehat{\gamma}^k) + o_p(a_T^{1/2}b_T) \\ &\leq \widehat{Q}(\theta^0,\gamma^0) + o_p(a_T^{1/2}b_T) = \tilde{Q}(\theta^0,\gamma^0) + o_p(a_T^{1/2}b_T)\end{aligned}$$

The inequality holds because $k \geq k^0 \implies (\theta^0,\gamma^0)$ is in the parameter space of the misspecified estimator.[6] This shows that $0 \leq \tilde{Q}(\widehat{\theta}^k,\widehat{\gamma}^k) - \tilde{Q}(\theta^0,\gamma^0) \leq o_P(a_T^{1/2}b_T)$. Then by above we have

$$\begin{aligned}o_p(a_T^{1/2}b_T) &\geq \tilde{Q}(\widehat{\theta}^k,\widehat{\gamma}^k) - \tilde{Q}(\theta^0,\gamma^0) = \frac{1}{NT}\sum_{i,t}(x_{it}'\Delta\widehat{\theta}_i^k)^2 = \frac{1}{N}\sum_i(\Delta\widehat{\theta}_i^k)'\left(\frac{1}{T}\sum_t x_{it}x_{it}'\right)\Delta\widehat{\theta}_i^k \\ &= \frac{1}{N}\sum_i(\Delta\widehat{\theta}_i^k)'\left(\frac{1}{T}\sum_t E[x_{it}x_{it}']\right)\Delta\widehat{\theta}_i^k + \frac{1}{N}\sum_i(\Delta\widehat{\theta}_i^k)'\left(\frac{1}{T}\sum_t(x_{it}x_{it}' - E[x_{it}x_{it}'])\right)\Delta\widehat{\theta}_i^k \\ &\geq \frac{1}{N}\sum_i(\Delta\widehat{\theta}_i^k)'\left(\frac{1}{T}\sum_t E[x_{it}x_{it}']\right)\Delta\widehat{\theta}_i^k - \left|\frac{1}{N}\sum_i(\Delta\widehat{\theta}_i^k)'\left(\frac{1}{T}\sum_t(x_{it}x_{it}' - E[x_{it}x_{it}'])\right)\Delta\widehat{\theta}_i^k\right|\end{aligned}$$

Now again applying the triangle inequality, Cauchy-Schwarz, and the definition of an operator norm we have

$$\begin{aligned}&\left|\frac{1}{N}\sum_i(\Delta\widehat{\theta}_i^k)'\left(\frac{1}{T}\sum_t(x_{it}x_{it}' - E[x_{it}x_{it}'])\right)\Delta\widehat{\theta}_i^k\right| \leq \frac{1}{N}\sum_i\|\Delta\widehat{\theta}_i^k\|^2\left\|\frac{1}{T}\sum_t(x_{it}x_{it}' - E[x_{it}x_{it}'])\right\| \\ &\leq \frac{1}{N}\sum_i\|\Delta\widehat{\theta}_i^k\|^2\sup_{j\in[N]}\left\|\frac{1}{T}\sum_t(x_{jt}x_{jt}' - E[x_{jt}x_{jt}'])\right\| = O_p(a_T)o_p(b_T) = o_p(a_T\cdot b_T)\end{aligned}$$

where the last equality uses lemma B.3. Then continuing the chain of inequalities above we have

$$o_p(a_T^{1/2}b_T) \geq \tilde{Q}(\widehat{\theta}^k,\widehat{\gamma}^k) - \tilde{Q}(\theta^0,\gamma^0) \geq \frac{1}{N}\sum_i\|\Delta\widehat{\theta}_i^k\|^2\min_{j\in[N]}\lambda_{min}\left(\frac{1}{T}\sum_t E[x_{jt}x_{jt}']\right) - o_p(a_T\cdot b_T)$$

By assumption $a_T = o(1)$, so collecting the o_p terms on the LHS and defining λ_{NT} to be the eigenvalue term on the RHS, we have

$$\begin{aligned}o_p(a_T^{1/2}b_T) &\geq \frac{1}{N}\sum_i\|\Delta\widehat{\theta}_i^k\|^2\lambda_{NT} \geq \frac{1}{N}\sum_i\|\Delta\widehat{\theta}_i^k\|^2(\underline{\lambda}/2 + (\lambda_{NT} - \underline{\lambda}/2)\mathbb{1}(\lambda_{NT} \leq \underline{\lambda}/2)) \\ &\geq \underline{\lambda}/2\frac{1}{N}\sum_i\|\Delta\widehat{\theta}_i^k\|^2 - \left|\frac{1}{N}\sum_i\|\Delta\widehat{\theta}_i^k\|^2(\lambda_{NT} - \underline{\lambda}/2)\mathbb{1}(\lambda_{NT} \leq \underline{\lambda}/2))\right| \\ &\geq \underline{\lambda}/2\frac{1}{N}\sum_i\|\Delta\widehat{\theta}_i^k\|^2 - (\underline{\lambda}/2)M^2\mathbb{1}(\lambda_{NT} \leq \underline{\lambda}/2)) \geq \underline{\lambda}/2\frac{1}{N}\sum_i\|\Delta\widehat{\theta}_i^k\|^2 - o(a_T^{1/2}b_T)\end{aligned}$$

The second to last inequality follows by assumption 3.9.(b) and compactness. The final

[6] Specifically, there exist $\theta^k \in \Theta^k$ and $\gamma^k \in \Gamma^k$ such that the $N \times p$ matrix with ith row $\theta^0(\gamma^0(i)) = \theta^k(\gamma^k(i))$ for all $i \in [N]$

inequality holds because indicator functions that converge to 0 do so at arbitrary rate. This completes the proof of the lemma. □

Corollary A.7. *For any $k \geq k_0$*

$$d(\widehat{m}^k, m^{k_0}) = o_p(T^{-1+3\epsilon}) \tag{A.15}$$

$$\widehat{Q}(k) - \widehat{Q}^0 = o_p(T^{-1+3\epsilon}) \tag{A.16}$$

Proof. We claim that for all $r \geq 0$, we have $d(\widehat{m}^k, m^{k_0}) = O_p\left(T^{c_r}\right)$, where $c_r = -(1 - \frac{1}{2^r}) + \epsilon \sum_{j=0}^{r} 2^{-j}$. The proof is by induction. The base case $c_0 = \epsilon$ is immediate since $d(\widehat{m}^k, m^{k_0}) = O_p(1)$ by compactness of Θ. Assume the statement is true for all $0 \leq m \leq r$, then by lemma A.6

$$d(\widehat{m}^k, m^{k_0}) = O_p(T^{c_r}) \implies d(\widehat{m}^k, m^{k_0}) = o_p\left(T^{c_r/2} \cdot T^{-1/2+\epsilon}\right)$$

and $c_r/2 - 1/2 + \epsilon = -(1/2 - \frac{1}{2^{r+1}}) + \epsilon \sum_{j=0}^{r} 2^{-j-1} - 1/2 + \epsilon = -(1 - \frac{1}{2^{r+1}}) + \epsilon \sum_{j=0}^{r+1} 2^{-j}$, which completes the induction. In particular, the first statement of the corollary holds as soon as $2^{-r} \leq \epsilon$. For the second statement of the corollary, recall that

$$\widehat{Q}(k) - \widehat{Q}^0 = \frac{1}{NT} \sum_{i,t} (x_{it}' \Delta \widehat{\theta}_i^k)^2 + \frac{1}{NT} \sum_{i,t} e_{it} x_{it}' \Delta \widehat{\theta}_i^k$$

The proof of lemma A.6, showed that $d(\widehat{m}^k, m^{k_0}) = o_p(a_T) \implies \frac{1}{NT} \sum_{i,t} e_{it} x_{it}' \Delta \widehat{\theta}_i^k = o_p(a_T^{1/2} b_T)$. Under the same conditions

$$\begin{aligned} \frac{1}{NT} \sum_{i,t} (x_{it}' \Delta \widehat{\theta}_i^k)^2 &\leq \frac{1}{NT} \sum_{i,t} \|x_{it}\|^2 \|\Delta \widehat{\theta}_i^k\|^2 \leq \frac{1}{N} \sum_i \|\Delta \widehat{\theta}_i^k\|^2 \sup_{j \in [N]} \frac{1}{T} \sum_t \|x_{jt}\|^2 \\ &\leq o_p(a_T) O_p(1) = o_p(a_T) \end{aligned}$$

That $\sup_{i \in [N]} \frac{1}{T} \sum_t \|x_{it}\|^2$ is $O_p(1)$ can easily be shown by a union bound in combination with assumption 3.4.(f) (as long as $NT^{-a} = o(1)$ for some $a > 0$). Putting this together, we get that $\widehat{Q}(k) - \widehat{Q}^0 = o_p(T^{-1+3\epsilon}) + o_p(T^{-1/2+3\epsilon/2-1/2+\epsilon}) = o_p(T^{-1+3\epsilon})$. This completes the proof of the corollary and of the first part of lemma A.5. □

Proposition A.8. *For any $k \geq k^0$*

$$\forall c \in \mathcal{C}^{k_0} \quad \min_{x \in \mathcal{C}^k} \|\widehat{\theta}^k(x) - \theta^0(c)\|^2 = o_p(T^{-1+3\epsilon}) \tag{A.17}$$

Proof. Applying corollary A.7, we find that

$$\begin{aligned} o_p(T^{-1+3\epsilon}) &\geq \tilde{Q}(\widehat{\theta}^k, \widehat{\gamma}^k) - \tilde{Q}(\theta^0, \gamma^0) = \frac{1}{NT} \sum_{i,t} (x_{it}'(\theta^0(c_i^0) - \widehat{\theta}^k(\widehat{c}_i{}^k))^2 \\ &\geq \min_{\tilde{c} \in \mathcal{C}^{k_0}} \min_{\gamma^k \in \Gamma^k} \max_{c \in \mathcal{C}^k} \rho(c, \tilde{c}, \gamma) \max_{c' \in \mathcal{C}^{k_0}} \min_{x \in \mathcal{C}^k} \|\theta^k(x) - \theta^0(c')\|^2 \\ &= \max_{c' \in \mathcal{C}^{k_0}} \min_{x \in \mathcal{C}^k} \|\theta^k(x) - \theta^0(c')\|^2 (\delta/2 + (\rho_{NT}^k - \delta/2) \mathbb{1}(\rho_{NT}^k - \delta/2 \leq 0)) \\ &= (\delta/2) \cdot \max_{c' \in \mathcal{C}^{k_0}} \min_{x \in \mathcal{C}^k} \|\theta^k(x) - \theta^0(c')\|^2 - o_p(T^{-1+3\epsilon}) \\ &\implies \max_{c \in \mathcal{C}^{k_0}} \min_{x \in \mathcal{C}^k} \|\widehat{\theta}^k(x) - \theta^0(c)\|^2 = o_p(T^{-1+3\epsilon}) \end{aligned}$$

The final equality again follows by compactness of Θ, positivity of ρ^k_{NT}, and because indicator functions that converge to 0 (in probability) do so at arbitrary rates. Since the square norm above is additively separable in the norms of each block of the covariate vector, for any $c, c' \in \mathcal{C}^{k_0}$ with $c_\ell = c'_\ell$, we must have $(\text{argmin}_{x\in\mathcal{C}^k} \|\widehat{\theta}^k(x) - \theta^0(c)\|^2)_\ell = (\text{argmin}_{x\in\mathcal{C}^k} \|\widehat{\theta}^k(x) - \theta^0(c')\|^2)_\ell$. This shows that setting $\sigma_\ell(a) = (\text{argmin}_{x\in\mathcal{C}^k} \|\widehat{\theta}^k(x) - \theta^0(c)\|^2)_\ell$ for any $c \in \mathcal{C}^{k_0}$ with $c_\ell = a$ is well-defined. □

The following proposition is our analogue of Theorem 3.2 in Liu et al. (2019). We use a recursive argument to give a faster rate for the worst case cross-sectional unit error in our setting.

Proposition A.9. *For any $k \geq k^0$*

$$\sup_{i\in[N]} \|\widehat{\theta}^k(\widehat{c}^k_i) - \theta^0(c^0_i)\| = o_p(T^{-\frac{1}{2}+2\epsilon}) \tag{A.18}$$

Proof. Define $\widehat{Q}_i(\theta, c_i) = \frac{1}{T}\sum_t (y_{it} - x'_{it}\theta(c_i))^2$ and $\tilde{Q}_i(\theta, c_i) = \frac{1}{T}\sum_t (x'_{it}(\theta^0(c^0_i) - \theta(c_i)))^2 + \frac{1}{T}\sum_t e^2_{it}$. Recall the random cluster mapping $\sigma : \mathcal{C}^{k_0} \to \mathcal{C}^k$ defined above. Then since $\widehat{c}_i$ is the optimal cluster choice given estimated parameters $\widehat{\theta}$,

$$\begin{aligned}
\widehat{Q}_i(\widehat{\theta}^k, \widehat{c}^k_i) \leq \widehat{Q}_i(\widehat{\theta}^k, \sigma(c^0_i)) \implies \tilde{Q}_i(\widehat{\theta}^k, \widehat{c}^k_i) &\leq \tilde{Q}_i(\widehat{\theta}^k, \sigma(c^0_i)) \\
+ (\widehat{Q}_i - \tilde{Q}_i)(\widehat{\theta}^k, \sigma(c^0_i)) &+ (\tilde{Q}_i - \widehat{Q}_i)(\widehat{\theta}^k, \widehat{c}^k_i) \\
\leq \tilde{Q}_i(\widehat{\theta}^k, \sigma(c^0_i)) &+ |\widehat{Q}_i - \tilde{Q}_i|(\widehat{\theta}^k, \sigma(c^0_i)) + |\tilde{Q}_i - \widehat{Q}_i|(\widehat{\theta}^k, \widehat{c}^k_i)
\end{aligned}$$

The second term above has

$$\begin{aligned}
\sup_i |\widehat{Q}_i - \tilde{Q}_i|(\widehat{\theta}^k, \sigma(c^0_i)) &= \sup_i \left| (\theta^0(c^0_i) - \widehat{\theta}^k(\sigma(c^0_i)))' \frac{1}{T}\sum_t e_{it}x_{it} \right| \\
&\leq \max_{c\in\mathcal{C}^{k_0}} \min_{x\in\mathcal{C}^k} \|\widehat{\theta}^k(x) - \theta^0(c)\| \sup_{i\in[N]} \left\| \frac{1}{T}\sum_t e_{it}x_{it} \right\| \leq o_p(T^{-1+\frac{5}{2}\epsilon})
\end{aligned}$$

where we apply proposition A.8 and lemma B.3. Similarly, the third term is

$$\begin{aligned}
\sup_i |\tilde{Q}_i - \widehat{Q}_i|(\widehat{\theta}^k, \widehat{c}^k_i) &\leq \sup_{i\in[N]} \|\theta^0(c^0_i) - \widehat{\theta}^k(\widehat{c}^k_i)\| \sup_{j\in[N]} \left\| \frac{1}{T}\sum_t e_{jt}x_{jt} \right\| \\
&= O_p\left(\sup_{i\in[N]} \|\theta^0(c^0_i) - \widehat{\theta}^k(\widehat{c}^k_i)\| \right) o_p(T^{-\frac{1}{2}+\epsilon})
\end{aligned}$$

Moreover, we reason

$$\begin{aligned}
&\sup_{i\in[N]} |\tilde{Q}_i(\widehat{\theta}^k, \sigma(c^0_i)) - \tilde{Q}_i(\theta^0, c^0_i)| = \sup_{i\in[N]} \frac{1}{T}\sum_t (x'_{it}(\theta^0(c^0_i) - \widehat{\theta}^k(\sigma(c^0_i))))^2 \\
&\leq \sup_{i\in[N]} \frac{1}{T}\sum_t \|x_{it}\|^2 \|\Delta\widehat{\theta}^k(c^0_i, \sigma(c^0_i))\|^2 \leq \sup_{i\in[N]} \frac{1}{T}\sum_t \|x_{it}\|^2 \sup_{j\in[N]} \|\Delta\widehat{\theta}^k(c^0_j, \sigma(c^0_j))\|^2 \\
&\leq \sup_{i\in[N]} \frac{1}{T}\sum_t \|x_{it}\|^2 \max_{c\in\mathcal{C}^{k_0}} \min_{x\in\mathcal{C}^k} \|\widehat{\theta}^k(x) - \theta^0(c)\|^2 = O_p(1)o_p(T^{-1+3\epsilon})
\end{aligned}$$

That $\sup_{i\in[N]} \frac{1}{T}\sum_t \|x_{it}\|^2$ is $O_p(1)$ can easily be shown by a union bound in combination with assumption 3.4.(f) (as long as $NT^{-a} = o(1)$ for some $a > 0$). Putting this all together, we have

$$\begin{aligned}
0 &\leq \sup_{i\in[N]} [\tilde{Q}_i(\widehat{\theta}^k, \widehat{c}_i^k) - \tilde{Q}_i(\theta^0, c_i^0)] \\
&\leq \sup_{i\in[N]} [\tilde{Q}_i(\widehat{\theta}^k, \sigma(c_i^0)) + \sup_{j\in[N]} [\tilde{Q}_i(\widehat{\theta}^k, \widehat{c}_j^k) - \tilde{Q}_j(\widehat{\theta}^k, \sigma(c_j^0))] - \tilde{Q}_i(\theta^0, c_i^0)] \\
&= o_p(T^{-1+3\epsilon}) + O_p\left(\sup_{i\in[N]} \|\theta^0(c_i^0) - \widehat{\theta}^k(\widehat{c}_i^k)\|\right) o_p(T^{-\frac{1}{2}+\epsilon})
\end{aligned}$$

where the first $o_p(1)$ is from work above and the second by lemma B.3. Now

$$\begin{aligned}
\sup_{i\in[N]} |\tilde{Q}_i(\widehat{\theta}^k, \widehat{c}_i^k) - \tilde{Q}_i(\theta^0, c_i^0)| &= \sup_{i\in[N]} |\Delta\widehat{\theta}_i^k(\widehat{c}_i^k, c_i^0)' \frac{1}{T}\sum_t x_{it}x_{it}' \Delta\widehat{\theta}_i^k(\widehat{c}_i^k, c_i^0)| \\
&\geq \sup_{i\in[N]} |\Delta\widehat{\theta}_i^k(\widehat{c}_i^k, c_i^0)' \frac{1}{T}\sum_t E[x_{it}x_{it}'] \Delta\widehat{\theta}_i^k(\widehat{c}_i^k, c_i^0)| \\
&\quad - \sup_{i\in[N]} |\Delta\widehat{\theta}_i^k(\widehat{c}_i^k, c_i^0)' \frac{1}{T}\sum_t (x_{it}x_{it}' - E[x_{it}x_{it}'])\Delta\widehat{\theta}_i^k(\widehat{c}_i^k, c_i^0)| \\
&\geq \sup_{i\in[N]} \|\Delta\widehat{\theta}_i^k(\widehat{c}_i^k, c_i^0)\|^2 \inf_{j\in[N]} \lambda_{min}\left(\frac{1}{T}\sum_t E[x_{jt}x_{jt}']\right) - C_{NT}
\end{aligned}$$

where similar arguments show that

$$C_{NT} = O_p\left(\sup_{i\in[N]} \|\Delta\widehat{\theta}_i^k(\widehat{c}_i^k, c_i^0)\|^2\right) o_p(T^{-\frac{1}{2}+\epsilon})$$

The indicator function trick used in the proof of lemma A.6 above then shows that

$$\begin{aligned}
\sup_{i\in[N]} \|\Delta\widehat{\theta}_i^k(\widehat{c}_i^k, c_i^0)\|^2 &= O_p\left(\sup_{i\in[N]} \|\Delta\widehat{\theta}_i^k(\widehat{c}_i^k, c_i^0)\|^2\right) o_p(T^{-\frac{1}{2}+\epsilon}) + o_p(T^{-1+3\epsilon}) \\
&\quad + O_p\left(\sup_{i\in[N]} \|\Delta\widehat{\theta}_i^k(\widehat{c}_i^k, c_i^0)\|\right) o_p(T^{-\frac{1}{2}+\epsilon}) \qquad \text{(A.19)}
\end{aligned}$$

The remainder of the proof follows by induction. For the base case, using compactness in the expression above shows that $\sup_{i\in[N]} \|\Delta\widehat{\theta}_i^k(\widehat{c}_i^k, c_i^0)\|^2 = o_p(T^{-\frac{1}{2}+\epsilon})$. The inductive step follows from the recursion in equation A.19. This completes the proof. □

A.3.1 Bias

We also need the following lemma on the order of our proposed bias correction

Lemma A.10. *The following are true*

(i) If $k \geq k^0$, then $\widehat{b}(k) = O_p(\frac{1}{NT})$

(ii) If k is such that $k_i < k_i^0$ for some i, then $\widehat{b}(k) = o_p(1)$

Proof. TBD, current C_p criterion uses non bias-corrected sample risk. □

We are now ready to complete the proof of model selection consistency using our C_p criterion. For completeness, suppose that we choose $\widehat{k}$ uniformly (independently) at random in the case of a tie. Denote $k > k'$ if $k_i \geq k'_i$ for all i and strictly for some index.

Proof of Theorem 3.10. We reason that

$$\begin{aligned}\mathbb{P}(\widehat{k} \neq k^0) &\leq \mathbb{P}(\exists k \neq k^0 \, s.t. \, C_p(k) \leq C_p(k^0)) \\ &\leq \sum_{\substack{k:\exists k_i < k_i^0 \\ k \leq k_{\max}}} \mathbb{P}(C_p(k) \leq C_p(k^0)) + \sum_{\substack{k > k^0 \\ k \leq k_{\max}}} \mathbb{P}(C_p(k) \leq C_p(k^0))\end{aligned}$$

For k in the first summation (with $k_i < k_i^0$ for some i), we have

$$\begin{aligned}\mathbb{P}(C_p(k) \leq C_p(k^0)) &= \mathbb{P}\left(\widehat{Q}(k) - \widehat{Q}(k^0) + \widehat{b}(k) - \widehat{b}(k^0) \leq \sum_i (k_i^0 - k_i) f(N,T)\right) \\ &= \mathbb{P}\left([\widehat{Q}(k) - \widehat{Q}^0] - [\widehat{Q}(k^0) - \widehat{Q}^0] + \widehat{b}(k) - \widehat{b}(k^0) \leq \sum_i (k_i^0 - k_i) f(N,T)\right) \\ &= \mathbb{P}\left(\Omega(1) + o_p(1) + O_p(1/NT) + o_p(1) \leq o(1)\right) = \mathbb{P}(\Omega(1) \leq o_p(1)) \to 0\end{aligned}$$

Where we have applied lemmas A.5 and A.10 in the final line. Similarly, for k in the second summation (with $k \geq k^0$ and $k_j > k_j^0$ for some j)

$$\begin{aligned}\mathbb{P}(C_p(k) \leq C_p(k^0)) &= \mathbb{P}\left([\widehat{Q}(k) - \widehat{Q}^0] - [\widehat{Q}(k^0) - \widehat{Q}^0] + \widehat{b}(k) - \widehat{b}(k^0) \leq \sum_i (k_i^0 - k_i) f(N,T)\right) \\ &\leq \mathbb{P}\left(2 \cdot O_p(1/NT) + o_p(T^{-1+3\epsilon}) \leq -f(N,T)\right) \\ &= \mathbb{P}\left(O_p\left(\frac{1}{NT^{3\epsilon}}\right) + o_p(1) \leq -T^{1-3\epsilon} f(N,T)\right) \to 0\end{aligned}$$

since $T^{1-3\epsilon} f(N,T) \to \infty$ by assumption. Because each of the sums above is finite, this shows that $\mathbb{P}(\widehat{k} \neq k^0) = o(1)$, which completes the proof. □

A.4 Variance Estimator Consistency

In this section, we sketch how to adapt Hansen (2007)'s proof of the consistency of the Arellano (1987) HAC variance estimator to the estimator proposed in equation 3.11 under asymptotics where $N, T \to \infty$ jointly. To use Hansen's proof, we impose the following assumptions, as well as compactness 3.1.(a) and mixing conditions 3.4.(c). Throughout, we assume that $1 \in x_{it}$.

Assumption A.11. *Impose the following assumptions*

(a) *(e_{it}, x_{it}, c_i^0) are cross-sectionally independent*

(b) *There exists $c < \infty$ and $\delta > 0$ such that for all i, t, and components x_{itq} of x_{it}, we have $\mathbb{E}|x_{itq}|^{4+\delta} < c$ and $\mathbb{E}|e_{it}|^{4+\delta} < c$*

First note that for any $a \in [k_\ell]$ and $b \in [k_s]$ and $\epsilon > 0$ and $\delta > 0$, we have

$$\begin{aligned}\mathbb{P}(\|\widehat{\Omega}_{\ell a,sb}(\widehat{c}) - \Omega_{\ell a,sb}\| > \epsilon) &\leq \mathbb{P}(\|\widehat{\Omega}_{\ell a,sb}(c^0) - \Omega_{\ell a,sb}\| > \epsilon) + \mathbb{P}\left(\exists i \in [N]\, s.t.\, \widehat{c}_i \neq c_i^0\right) \\ &= \mathbb{P}(\|\widehat{\Omega}_{\ell a,sb}(c^0) - \Omega_{\ell a,sb}\| > \epsilon) + o(1) + O(NT^{-\delta}) \qquad \text{(A.20)}\end{aligned}$$

So for consistency it suffices to focus on the estimator $\widehat{\Omega}(c^0)$ defined by 3.11 evaluated at the true cluster membership matrix.

Let $Z_i = \mathbb{1}(c^0_{i\ell} = a)\mathbb{1}(c^0_{is} = b) \in \{0,1\}$. Define $\Delta\widehat{\theta}(c) \equiv \theta^0(c) - \widehat{\theta}(c)$ and $\Delta\widehat{\theta}_i \equiv \Delta\widehat{\theta}(c_i^0)$. Using this notation, we have

$$\begin{aligned}\widehat{\Omega}_{\ell a,sb} &= \frac{1}{NT}\sum_{i,t,t'} \widehat{e}_{it}\widehat{e}_{it'} x_{it\ell} x'_{it's} Z_i \\ &= \frac{1}{NT}\sum_{i,t,t'} (e_{it}e_{it'} + e_{it}x'_{it'}\Delta\widehat{\theta}_i + x'_{it}\Delta\widehat{\theta}_i e_{it'} + \Delta\widehat{\theta}'_i x_{it'} x'_{it}\Delta\widehat{\theta}_i) x_{it\ell} x'_{it's} Z_i\end{aligned}$$

We focus on just one term in the $d_\ell \times d_s$ matrix $x_{it\ell}x'_{it's}$, which we denote $x_{itp}x_{it'q}$. Then, for instance, the second term in the preceding expansion can be written as

$$\begin{aligned}&\frac{1}{NT}\sum_i Z_i \left(\sum_t e_{it}x_{itp}\right)\left(\sum_{t'} x_{it'}x_{it'q}\right)' \Delta\widehat{\theta}_i \\ &\qquad = \frac{1}{NT}\sum_i Z_i \sum_{c\in\mathcal{C}} \mathbb{1}(c_i^0 = c)\left(\sum_t e_{it}x_{itp}\right)\left(\sum_{t'} x_{it'}x_{it'q}\right)' \Delta\widehat{\theta}_i \\ &\qquad = \frac{1}{NT}\sum_{c\in\mathcal{C}}\sum_i Z_i \mathbb{1}(c_i^0 = c)\left(\sum_t e_{it}x_{itp}\right)\left(\sum_{t'} x_{it'}x_{it'q}\right)' \Delta\widehat{\theta}(c)\end{aligned}$$

Note that $\mathcal{C}$ is finite. Each term inside the sum $\sum_{c\in\mathcal{C}}$ above

$$\frac{1}{NT}\sum_i Z_i \mathbb{1}(c_i^0 = c)\left(\sum_t e_{it}x_{itp}\right)\left(\sum_{t'} x_{it'}x_{it'q}\right)' \Delta\widehat{\theta}(c)$$

has the form of equation (O.2) in the supplementary appendix of Hansen (2007), up to the extra term $Z_i\mathbb{1}(c_i^0 = c)$. However, since $E\|Z_i\mathbb{1}(c_i^0 = c)v\| \leq E\|v\|$ for any vector v, these extra terms preserve the moment bounds needed for application of the Markov LLN (Hansen, Lemma A.2).

To show the moment bound $E\|(\sum_t e_{it}x_{itp})(\sum_t' x_{it'}x_{it'q})\|^{1+\delta}$ needed for the Markov LLN, Hansen's Theorem 3 and Lemma A.4 assume decay rates on the mixing coefficients $\alpha(t)$ of (x_{it}, e_{it}). Our exponential mixing assumption 3.4.(c) is already sufficient for the polynomial rate used in his proof. Thus, Hansen's results apply to show that $\frac{1}{NT}\sum_{i,t,t'} e_{it}x'_{it'}\Delta\widehat{\theta}_i x_{it\ell}x'_{it's} = O_p(\frac{1}{\sqrt{n}})$. The arguments from Hansen's proof similarly show that under the conditions in assumption A.11 the third term in equation A.20 is $O_p(\frac{1}{\sqrt{n}})$, the fourth term is $O_p(\frac{1}{n})$, and the first term converges to the limit postulated in 3.7.(a).

B Supplementary Lemmas

In the following lemma, we show that $\max_i \operatorname{Var}(\overline{x}_i'\Delta\theta) = o(1)$, needed for the proof of C.4. The proof is an application of methods developed in Rio (1993). Also see Rio (2017) for a more complete exposition of covariance inequalities for strongly-mixing processes.

Lemma B.1. *Under the fixed effects assumptions C.3,* $\max_i \operatorname{Var}(\overline{x}_i'\Delta\theta) \to 0$ *as* $T \to \infty$.

Proof. For a sequence of mixing coefficients $\{\alpha(t)\}_{t\geq 0}$ define $\alpha^{-1}(u) = \sum_{t\geq 0} \mathbb{1}(\alpha(t) > u)$ for $0 \leq u \leq 1$. Also, for scalar random variable X we let $Q_X(u) \equiv \inf\{t \geq 0 : P(|X| > t) \leq u\}$ be the reversed quantile function of $|X|$. First note that

$$\begin{aligned}\operatorname{Var}(\overline{x}_i'\Delta\theta) &= \operatorname{Var}\left(\sum_{k=1}^{p} \Delta\theta_k \overline{x}_{ik}\right) \leq \sum_{k,j} \operatorname{Var}(\overline{x}_{ik})^{1/2}\operatorname{Var}(\overline{x}_{ij})^{1/2}|\Delta\theta_k||\Delta\theta_j| \\ &\leq M^2\left(\sum_{k} \operatorname{Var}(\overline{x}_{ik})^{1/2}\right)^2 \leq pM^2\sum_{k=1}^{p}\operatorname{Var}(\overline{x}_{ik})\end{aligned}$$

The first inequality is from Cauchy-Schwarz, the second from compactness, and the final from Jensen's inequality. Then apparently it suffices to prove that $\max_i \operatorname{Var}(\overline{x}_{ik}) \to 0$ as $T \to \infty$ for each k. Thus, in what follows we assume that x_{it} is a scalar random variable. Corollary 1.1 of Rio (2017) gives the bound

$$\operatorname{Var}(\overline{x}_i) \leq \frac{4}{T^2}\sum_{t\geq 0}\int_0^1 \alpha^{-1}(u)Q_{x_{it}}^2(u)du \tag{B.1}$$

We claim that for any random variable $x \in L^1$, the inequality $Q_x(u) \leq |Ex| + Q_{x-Ex}(u)$ holds. Note that for $t \geq 0$,

$$\mathbb{P}(|x - Ex| > t) \leq u \implies \mathbb{P}(|x| > t + |Ex|) \leq \mathbb{P}(||x| - |Ex|| > t) \leq \mathbb{P}(|x - Ex| > t) \leq u$$

by the reverse triangle inequality. Then we have shown that

$$\begin{aligned}&\{t + |Ex| : t \geq 0, \mathbb{P}(|x - Ex| > t) \leq u\} \subset \{t \geq 0 : \mathbb{P}(|x| > t) \leq u\} \\ &\implies \inf\{t \geq 0 : \mathbb{P}(|x| > t) \leq u\} \leq |Ex| + \inf\{t \geq 0 : \mathbb{P}(|x - Ex| > t) \leq u\} \\ &\iff Q_x(u) \leq |Ex| + Q_{x-Ex}(u) \leq E|x| + Q_{x-Ex}(u)\end{aligned}$$

In what follows, denote $z_{it} = x_{it} - E(x_{it})$. Then using this inequality in B.1, we get the bound

$$\begin{aligned}T^2\operatorname{Var}(\overline{x}_i) \leq &\sum_{t=1}^{T} E|x_{it}|^2\int_0^1 \alpha^{-1}(u)du + 2\sum_{t=1}^{T} E|x_{it}|\int_0^1 \alpha^{-1}(u)Q_{z_{it}}(u)du \\ &+ \sum_{t=0}^{T}\int_0^1 \alpha^{-1}(u)Q_{z_{it}}(u)^2du\end{aligned} \tag{B.2}$$

where we applied Jensen's inequality to reduce $(E|x|)^2 \leq E|x|^2$. For the first term, note that $\int_0^1 \alpha^{-1}(u)du = \sum_{s\geq 0}\int_0^1 \mathbb{1}(\alpha(s) \geq u)du = \sum_{s\geq 0}\alpha(s)$. For the second term, we need a bound on the function $Q_{z_{it}}(u)$ for each i, t. Note that from assumption C.3.(d), we have

$\mathbb{P}(|x_{it} - Ex_{it}| > t) \le e^{1-(t/f)^{d_2}}$, giving

$$\begin{aligned}\sup_{i,t} Q_{z_{it}}(u) &= \sup_{i,t} \inf\{t \ge 0 : \mathbb{P}(|x_{it} - Ex_{it}| > t) \le u\} \le \inf\{t \ge 0 : e^{1-(t/f)^{d_2}} \le u\} \\ &= f(1 - \log(u))^{1/d_2}\end{aligned}$$

where the last line is just the inverse of the tail bound. We claim that for all $a > 1$ and $u \in (0, 1]$, we have $1 - \log(u) \le a u^{-1/a}$. Note that $1 - log(1) = 1 \le a(1)^{-1/a} = a$. Moreover, for all $u \in (0, 1]$, $-\frac{\partial}{\partial u}(1 - \log u) = 1/u \le u^{-1/a-1} = -\frac{\partial}{\partial u} a u^{-1/a}$. This proves the claim. Let $a > 2/d_2 \vee 1$, then our work shows $(1 - \log(u))^{2/d_2} \le \frac{1}{u^{1-\epsilon(d_2)}}$, for some $\epsilon(d_2) \in (0, 1)$

$$\begin{aligned}\int_0^1 \alpha^{-1}(u) Q_{z_{it}}(u)^2 du = \sum_{s\ge0} \int_0^{\alpha(s)} Q_{z_{it}}(u)^2 &\le \sum_{s\ge0} \int_0^{\alpha(s)} f^2 a(d_2)^2 u^{-1+\epsilon(d_2)} du \\ &\le c(f, d_2) \sum_{s\ge0} \alpha(s)^{\epsilon(d_2)} du \le c(f, d_2) \sum_{s\ge0} e^{-b\epsilon(d_2) s^{d_1}} \\ &\equiv K(b, f, d_1, d_2) < \infty\end{aligned}$$

where the final sum can easily be shown to converge by standard methods. Moreover, since d_2 is arbitrary, writing $(1 - \log(u))^{1/d_2} = (1 - \log(u))^{2/\tilde{d}_2}$, the same proof shows that $\int_0^1 \alpha^{-1}(u) Q_{z_{it}}(u) du < K'(b, f, d_1, d_2)$. Our work showed that $\sum_{s\ge0} \alpha(s)^\epsilon = K(b, f, d_1, d_2) < \infty$ for some $\epsilon \in (0, 1)$. Then apparently $\sum_{s\ge0} \alpha(s) \le K''(b, f, d_1, d_2) < \infty$ for some different constant K'', because the sequence spaces ℓ_p are nested and increasing in $p > 0$.

The work above in equation C.9 shows that $\sup_i \frac{1}{T} \sum_t E|x_{it}|^p = O(1)$ for $p = 1, 2$. Then the decomposition in B.2 above becomes

$$\begin{aligned}\operatorname{Var}(\bar{x}_i) \le{}& \frac{1}{T^2} O(T) O(1) + \frac{2}{T^2} \left(\sum_{t=1}^T E|x_{it}|^2 \right)^{1/2} \left(\sum_{t=0}^T \left(\int_0^1 \alpha^{-1}(u) Q_{z_{it}}(u) du \right)^2 \right)^{1/2} \\ &+ \frac{1}{T^2} \sum_{t=0}^T K(b, f, d_1, d_2) \\ &= O(1/T) + 2 \left(\frac{1}{T} \sum_{t=1}^T E|x_{it}|^2 \right)^{1/2} \left(\frac{1}{T^3} \sum_{t=0}^T K''(b, f, d_1, d_2)^2 \right)^{1/2} \\ &= O(1/T) + O(1) O(1/T)\end{aligned}$$

where all the order statements above hold uniformly in i. This completes the proof. □

B.1 Model Selection Lemmas

The following lemma gives conditions under which the sample risk deviation from the irreducible sample risk $\widehat{Q}(\widehat{\theta}, \widehat{\gamma}) - \widehat{Q}^0$ converges at the rate needed for theorem 3.10.

Lemma B.2. *Suppose that $(\widehat{\theta}, \widehat{\gamma}) \in \Theta^{k_0} \times \Gamma^{k_0}$ has rate $\widehat{\theta} - \theta^0 = O_p(r_{NT})$ and satisfies $\widehat{c_i} = \widehat{c_i}(\widehat{\theta})$ for all i (as defined in lemma A.3). Also, suppose there exists a neighborhood*

$\mathcal{N}$ of θ^0 such that

$$\sup_{\theta\in\mathcal{N}} \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\widehat{c}_i(\theta) \neq c_i^0) = o_p(T^{-a}) \tag{B.3}$$

for any $a > 0$. Then $\widehat{Q}(\widehat{\theta}, \widehat{\gamma}) = O_p(\frac{r_{NT}}{\sqrt{NT}}) + O_p(r_{NT}^2) + o_p(T^{-a})$.

Proof. In what follows, denote $\Delta\widehat{\theta}(c) = (\theta^0(c) - \widehat{\theta}(c))$ and $\Delta\widehat{\theta}(c, c') = (\theta^0(c) - \widehat{\theta}(c'))$. By definition, we have

$$\widehat{Q}(\widehat{\theta}, \widehat{\gamma}) - \widehat{Q}^0 = \frac{1}{NT}\sum_{i,t} e_{it}x_{it}'\Delta\widehat{\theta}(c_i^0, \widehat{c}_i) + \frac{1}{NT}\sum_{i,t}(x_{it}'\Delta\widehat{\theta}(c_i^0, \widehat{c}_i))^2 \tag{B.4}$$

We consider each term separately. The first term is

$$\begin{aligned}
\frac{1}{NT}\sum_{i,t} e_{it}x_{it}'\Delta\widehat{\theta}(c_i^0, \widehat{c}_i) &= \frac{1}{NT}\sum_{i,t} e_{it}x_{it}'\Delta\widehat{\theta}(c_i^0)\mathbb{1}(\widehat{c}_i = c_i^0) + \frac{1}{NT}\sum_{i,t} e_{it}x_{it}'\Delta\widehat{\theta}(c_i^0, \widehat{c}_i)\mathbb{1}(\widehat{c}_i \neq c_i^0) \\
&= \frac{1}{NT}\sum_{i,t} e_{it}x_{it}'\Delta\widehat{\theta}(c_i^0)\mathbb{1}(\widehat{c}_i = c_i^0) + \frac{1}{NT}\sum_{i,t} e_{it}x_{it}'\Delta\widehat{\theta}(c_i^0, \widehat{c}_i)\mathbb{1}(\widehat{c}_i \neq c_i^0) \\
&- \frac{1}{NT}\sum_{i,t} e_{it}x_{it}'\Delta\widehat{\theta}(c_i^0)(1 - \mathbb{1}(\widehat{c}_i = c_i^0)) \\
&= \frac{1}{NT}\sum_{i,t} e_{it}x_{it}'\Delta\widehat{\theta}(c_i^0) + \frac{1}{NT}\sum_{i,t} e_{it}x_{it}'(\Delta\widehat{\theta}(c_i^0, \widehat{c}_i) - \Delta\widehat{\theta}(c_i^0))\mathbb{1}(\widehat{c}_i \neq c_i^0)
\end{aligned}$$

Consider the first term in the final line. This can be written

$$\begin{aligned}
\frac{1}{NT}\sum_{i,t} e_{it}x_{it}'\Delta\widehat{\theta}(c_i^0) &= \sum_{c\in\mathcal{C}^{k_0}} \frac{1}{NT}\sum_{i,t} e_{it}x_{it}'\Delta\widehat{\theta}(c)\mathbb{1}(c_i^0 = c) \\
&= \sum_{c\in\mathcal{C}^{k_0}} \Delta\widehat{\theta}(c)'\left(\frac{1}{NT}\sum_{i,t} e_{it}x_{it}\mathbb{1}(c_i^0 = c)\right) \\
&= \sum_{c\in\mathcal{C}^{k_0}} O_p(r_{NT})'O_p(1/\sqrt{NT}) = O_p(\frac{r_{NT}}{\sqrt{NT}})
\end{aligned}$$

where we used assumption 3.9.(a) in the final line. Using Cauchy-Schwarz and recalling

$M = \text{Diam}(\Theta)$, the second term can be written

$$\begin{aligned}
&\frac{1}{NT}\sum_{i,t} e_{it}x_{it}'(\Delta\widehat{\theta}(c_i^0, \widehat{c}_i) - \Delta\widehat{\theta}(c_i^0))\mathbb{1}(\widehat{c}_i \neq c_i^0) \\
&\quad\leq \frac{1}{NT}\left(\sum_i \|\widehat{\theta}(c_i^0) - \widehat{\theta}(\widehat{c}_i)\|^2 \mathbb{1}(\widehat{c}_i \neq c_i^0)\right)^{1/2} \left(\sum_i \left\|\sum_t e_{it}x_{it}\right\|^2\right)^{1/2} \\
&\quad\leq M^2 \left(\frac{1}{N}\sum_i \mathbb{1}(\widehat{c}_i \neq c_i^0)\right)^{1/2} \left(\frac{1}{NT^2}\sum_{i,t,s} e_{it}e_{is}x_{it}'x_{is}\right)^{1/2} \\
&\quad\leq o_p(1)\left(\frac{1}{N}\sum_i \mathbb{1}(\widehat{c}_i(\widehat{\theta}) \neq c_i^0)\right)^{1/2} (\mathbb{1}(\widehat{\theta} \in \mathcal{N}) + \mathbb{1}(\widehat{\theta} \notin \mathcal{N})) \\
&\quad\leq o_p(1)\left(\sup_{\theta\in\mathcal{N}} \frac{1}{N}\sum_i \mathbb{1}(\widehat{c}_i(\theta) \neq c_i^0)\right)^{1/2} + o_p(1)\mathbb{1}(\widehat{\theta} \notin \mathcal{N}) \\
&\quad\leq o_p(1)o_p(T^{-a}) + o_p(1)o_p(T^{-a}) = o_p(T^{-a})
\end{aligned}$$

The third inequality uses assumption 3.1, and the fourth uses equation B.3. For the final inequality, note that by consistency of $\widehat{\theta}$, the indicator $\mathbb{1}(\widehat{\theta} \notin \mathcal{N})$ converges in probability to 0 at arbitrary rate. We deal with the second term in equation B.4 similarly. Note that

$$\begin{aligned}
\frac{1}{NT}\sum_{i,t}(x_{it}'\Delta\widehat{\theta}(c_i^0, \widehat{c}_i))^2 &= \frac{1}{NT}\sum_{i,t}(x_{it}'\Delta\widehat{\theta}(c_i^0))^2\mathbb{1}(\widehat{c}_i = c_i^0) + \frac{1}{NT}\sum_{i,t}(x_{it}'\Delta\widehat{\theta}(c_i^0, \widehat{c}_i))^2\mathbb{1}(\widehat{c}_i \neq c_i^0) \\
&= \frac{1}{NT}\sum_{i,t}(x_{it}'\Delta\widehat{\theta}(c_i^0))^2 \\
&\quad + \frac{1}{NT}\sum_{i,t}((x_{it}'\Delta\widehat{\theta}(c_i^0, \widehat{c}_i))^2 - (x_{it}'\Delta\widehat{\theta}(c_i^0))^2)(\mathbb{1}(\widehat{c}_i \neq c_i^0))
\end{aligned}$$

Again, we argue the first term above is

$$\sum_{c\in\mathcal{C}^{k_0}} \frac{1}{NT}\sum_{i,t}(x_{it}'\Delta\widehat{\theta}(c))^2 \leq \sum_{c\in\mathcal{C}^{k_0}} \|\Delta\widehat{\theta}(c)\|^2 \frac{1}{NT}\sum_{i,t}\|x_{it}\|^2 = O_p(r_{NT}^2)O_p(1)$$

where we have used the tail assumption 3.4.(d) and $\Delta\widehat{\theta}(c) = O_p(1/\sqrt{NT})$ for all c in the final equality. Now, for instance, we can break the second term into parts and compute

$$\begin{aligned}
\frac{1}{NT}\sum_{i,t}(x_{it}'\Delta\widehat{\theta}(c_i^0))^2\mathbb{1}(\widehat{c}_i \neq c_i^0) &\leq \frac{1}{NT}\sum_{i,t}\|x_{it}\|^2\|\Delta\widehat{\theta}(c_i^0)\|^2\mathbb{1}(\widehat{c}_i \neq c_i^0) \\
&\leq M^2\left(\frac{1}{N}\sum_i \mathbb{1}(\widehat{c}_i \neq c_i^0)\right)^{1/2}\left(\frac{1}{NT^2}\sum_i\sum_{t,s}\|x_{it}\|^2\|x_{is}\|^2\right)^{1/2} \\
&\leq O_p(1)\left(\frac{1}{N}\sum_i \mathbb{1}(\widehat{c}_i(\widehat{\theta}) \neq c_i^0)\right)^{1/2} \leq O_p(1)o_p(T^{-a})
\end{aligned}$$

In the final line we have used assumption 3.4.(a) as well as our analysis of the sum of indicator functions above. An identical proof shows that $\frac{1}{NT}\sum_{i,t}((x'_{it}\Delta\widehat{\theta}(c_i^0,\widehat{c}_i))^2 = o_p(T^{-a})$. Putting this all together gives the claimed result. □

The following lemma is needed to establish rates of convergence for strongly mixing sequences.

Lemma B.3. *Impose the mixing and tail assumptions in 3.4.(c) and 3.4.(d), and suppose also that $\log N = o(T^{\epsilon})$ for some ϵ with $d/2 > \epsilon > 0$. Then*

$$\sup_{i\in[N]}\left\|\frac{1}{T}\sum_{t=1}(x_{it}x'_{it} - E[x_{it}x'_{it}])\right\|_{2,2} = o_p(T^{-\frac{1}{2}+\epsilon})$$

$$\sup_{i\in[N]}\left\|\frac{1}{T}\sum_{t=1}e_{it}x_{it}\right\| = o_p(T^{-\frac{1}{2}+\epsilon})$$

where the first line uses the standard operator norm.

Proof. For $(z_t)_{t\geq 0}$ a mean zero-process satisfying the assumptions in 3.4.(d) and 3.4.(c), let $s(z)^2 = \sup_t Ez_t^2 + 2\sum_{s>t}|Ez_tz_s| < \infty$. Let $d \equiv \frac{d_1d_2}{d_1+d_2}$. Then setting $\lambda = \frac{Tz}{4}$ in equation (1.7) in Merlevede et al. (2011), for any $r \geq 1$ we have

$$\mathbb{P}\left(\frac{1}{T}\left|\sum_t z_t\right| > z\right) \leq 4\left(1+\frac{Tz^2}{16rs(z)^2}\right)^{-r/2} + \frac{16C}{z}\exp\left(-b\frac{(Tz)^d}{(4fr)^d}\right)$$

Where C is a constant only depending on the mixing and tail parameters b, f, d_1, d_2. In particular, plugging in $z = xT^{-\frac{1}{2}+\epsilon}$ and $r = T^{\epsilon}$ gives

$$\mathbb{P}\left(\frac{1}{T^{\frac{1}{2}+\epsilon}}\left|\sum_t z_t\right| > x\right) \leq 4\left(1+\frac{T^{\epsilon}x^2}{16s(z)^2}\right)^{-T^{\epsilon}/2} + \frac{16CT^{\frac{1}{2}-\epsilon}}{x}\exp\left(-b\frac{T^{d/2}x^d}{(4f)^d}\right) \equiv f(T,x,s(z))$$

Let $v \equiv \sup_{i,q} s((e_{it}x_{itq})_t)$. Then, for instance, applying this to the second expression above we get for any $x > 0$

$$\begin{aligned}\mathbb{P}\left(T^{\frac{1}{2}-\epsilon}\sup_{i\in[N]}\left\|\frac{1}{T}\sum_{t=1}e_{it}x_{it}\right\| > x\right) &\leq \mathbb{P}\left(T^{\frac{1}{2}-\epsilon}\sup_{i\in[N]}\left\|\frac{1}{T}\sum_{t=1}e_{it}x_{it}\right\|_1 > x\right)\\ &\leq Np\sup_{q\in[p]}\sup_{i\in[N]}\mathbb{P}\left(\frac{1}{T^{\frac{1}{2}+\epsilon}}\left|\sum_t e_{it}x_{itq}\right| > x/p\right)\\ &\lesssim Nf(T,x/p,v) \to 0\end{aligned}$$

as $N, T \to \infty$ if $\log N = o(T^{\epsilon\wedge(d/2)})$ and $\sup_{i,q} s((e_{it}x_{itq})_t)) < \infty$. Note that covariance inequalities from Rio (2017) can be used to show that

$$\sup_i \sup_{1\leq a\leq p} s((e_{it}x_{ita})_t)) < \infty \quad \text{and} \quad \sup_i \sup_{1\leq a,b\leq p} s((x_{ita}x_{itb} - E[x_{ita}x_{itb}])_t) < \infty$$

under the assumptions 3.4.(d) and 3.4.(c) in our setting, as noted in BM. This completes the proof for the second term. For the first term, note that by equivalence of finite-dimensional vector space norms, there is a constant $c(p)$ depending only on the dimension such that

$$\left\|\frac{1}{T}\sum_{t=1}(x_{it}x_{it}' - E[x_{it}x_{it}'])\right\|_{2,2} \leq c(p)\left\|\frac{1}{T}\sum_{t=1}(x_{it}x_{it}' - E[x_{it}x_{it}'])\right\|_{1}$$

where $\|A\|_1 \equiv \sum_{i,j}|a_{ij}|$ for a matrix $A \in \mathbb{R}^{p\times p}$. The first statement of the lemma then follows by exactly the same argument, substituting $x_{it}x_{it}' - E[x_{it}x_{it}']$ for $e_{it}x_{it}$. □

Corollary B.4. *The following rates hold*

$$\frac{1}{N}\sum_{i=1}\left\|\frac{1}{T}\sum_{t=1}(x_{it}x_{it}' - E[x_{it}x_{it}'])\right\| = o_p(T^{-\frac{1}{2}+\epsilon})$$

$$\frac{1}{N}\sum_{i=1}\left\|\frac{1}{T}\sum_{t=1}e_{it}x_{it}\right\| = o_p(T^{-\frac{1}{2}+\epsilon})$$

Proof. Immediate by lemma B.3, noting that for any positive real numbers $(a_i)_{i=1}^N$ we have $\frac{1}{N}\sum_i a_i \leq \sup_{i\in[N]} a_i$. □

C Extensions and Supplementary Material

C.1 Fixed Effects Model

In this section, we consider an extension of the main specification with individual fixed effects.

$$y_{it} = x_{it}'\theta(c_i^0) + a_i + e_{it} \tag{C.1}$$

We propose to estimate equation C.1 by (1) de-meaning the time series for each cross-sectional unit followed by (2) applying Lloyd's Algorithm to the de-meaned data. In other words, defining $\tilde{z}_{it} \equiv z_{it} - \frac{1}{T}\sum_t z_{it} = z_{it} - \overline{z}_i$ for any variable z_{it}, we apply Lloyd's algorithm to the model $\tilde{y}_{it} = \tilde{x}_{it}'\theta(c_i^0) + \tilde{e}_{it}$.

The main challenge in extending our results to this setting is that the differencing operation changes the autocorrelation structure of the data, so that the mixing conditions in assumption 3.4.(c) may no longer be satisfied. In the remainder of this section, we overload notation and let $\widehat{\theta}$ and $\widehat{\gamma}$ refer to the fixed effects estimates defined by

$$(\widehat{\theta},\widehat{\gamma}) = \underset{\gamma\in\Gamma,\theta\in\Theta}{\text{argmin}}\ \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{y}_{it} - \tilde{x}_{it}'\theta(c_i))^2 \tag{C.2}$$

Letting the cluster permutation σ_ℓ be defined analagously to equation 3.1 in the main text, we have

Lemma C.1. *Under the assumptions in 3.1 with* (x_{it}, e_{it}) *replaced by* $(\tilde{x}_{it}, \tilde{e}_{it})$, $\mathbb{P}(\sigma_\ell$ *invertible*$) \to 1$ *as* $N, T \to \infty$

Relabeling $\widehat{\theta}_{\ell\sigma(a)} \to \widehat{\theta}_{\ell a}$ (which is well-defined w.h.p. as $N, T \to \infty$ by the lemma), we have

Theorem C.2. *Under the assumptions in 3.1 with (x_{it}, e_{it}) replaced by $(\tilde{x}_{it}, \tilde{e}_{it})$, for all blocks ℓ and $a \in [k_\ell]$, we have $\|\theta^0_{\ell a} - \widehat{\theta}_{\ell a}\| = o_P(1)$ as $N, T \to \infty$.*

Proof. Immediate from lemma 3.2 and theorem 3.3 applied to the the de-meaned data $(\tilde{y}, \tilde{x}, \tilde{e})_{it}$. □

For the analogue of theorem 3.5, we modify assumption 3.4 to the following

Assumption C.3. *Consider the following assumptions*

(a) $\max_i \frac{1}{T^2} \sum_{t,s} E(\|x_{it}\|^2 \|x_{is}\|^2) = O(1)$ *as* $T \to \infty$

(b) Let 3.4.(b) hold with $(e, x)_{it}$ replaced by $(\tilde{e}, \tilde{x})_{it}$ Also, let assumptions 3.4.(c) and 3.4.(f) on mixing conditions of $x_{it}e_{it}$, and large deviations of $\sum_t \|x_{it}\|^2$ hold exactly as in assumption 3.4 from the main theorem

(c) The uniform limits $\max_{i \in [N]} \frac{1}{T} \sum_t E[e_{it} x_{it}] \to 0$ and $\min_{i \in [N]} \frac{1}{T} \sum_t \mathbb{E}(\tilde{x}'_{it}(\theta(c) - \theta(c')))^2 \to \tilde{d}(c, c')$ hold as $T \to \infty$, and $\tilde{d}(c, c') \geq \tilde{d}_{min} > 0$ for $c \neq c'$.

(d) There exist constants f and d_2 such that for all $i \in [N]$ and all $z > 0$, for all components x^j_{it}, $x^{j'}_{it}$ of the vector x_{it} we have $\mathbb{P}(|x^j_{it} x^{j'}_{it} - E(x^j_{it} x^{j'}_{it})| > z)$, $\mathbb{P}(|e_{it} x^j_{it} - E e_{it} x^j_{it}| > z)$ and $\mathbb{P}(|x^j_{it} - E x^j_{it}| > z)$ are bounded above by $e^{1-(z/f)^{d_2}}$

(e) The covariate vector x_{it} contains a constant

Then, analogously letting $\tilde{\theta}$ be the infeasible estimator that minimizes C.2 with the true cluster identities c^0_i plugged in, we have

Theorem C.4. *Let the assumptions needed for consistency (assumption 3.1) hold with $(x, e)_{it}$ replaced by $(\tilde{x}, \tilde{e})_{it}$, and let the assumptions in C.3 hold. Then for any $a > 0$ and as $N, T \to \infty$, we have the following theorem*

$$\widehat{\theta} = \tilde{\theta} + o_P(T^{-a}) \tag{C.3}$$

Moreover, individual cluster estimates satisfy

$$\mathbb{P}\left(\exists i \in [N] \, s.t. \, \widehat{c}_i \neq c^0_i\right) = o(1) + o(NT^{-a}) \tag{C.4}$$

The analogue of theorem 3.8 follows immediately from theorem C.4 by replacing y_{it}, x_{it} and e_{it} with the appropriate de-meaned variables. Specifically, define

$$\widehat{M}_{\ell a, sb} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it\ell} \tilde{x}'_{its} \mathbb{1}(c_{is} = b) \mathbb{1}(c_{i\ell} = a)$$

$$v_{\ell a} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{e}_{it} \mathbb{1}(c_{i\ell} = a) \tilde{x}_{it\ell}$$

and let

$$\frac{1}{NT} \sum_{i,j=1}^{N} \sum_{t,t'=1}^{T} E[(e_{it} - \bar{e}_i)(e_{jt'} - \bar{e}_j) \mathbb{1}(c^0_{i\ell} = a) \mathbb{1}(c^0_{js} = b)(x_{it\ell} - \bar{x}_{i\ell})(x_{jt's} - \bar{x}_{js})'] \to \Omega_{\ell a, sb}$$

as N, T goes to infinity (we assume the limit exists), then the analogue of 3.8 is

Theorem C.5. *Suppose that the assumptions in 3.7 are satisfied with $(x,e)_{it}$ replaced by $(\tilde{x},\tilde{e})_{it}$ and the matrices $\widehat{M}$ and Ω as defined above. Also let there $r>0$ such that $\sqrt{NT}^{-r}=o(1)$. Then*

$$\sqrt{NT}(\mathrm{vec}(\widehat{\theta}-\theta^0)) \xrightarrow{d} \mathcal{N}(0, M^{-1}\Omega M) \tag{C.5}$$

We propose to use the HAC estimator defined in equation 3.11, with $\widehat{e}_{it}$ replaced by the residuals from the fixed effects problem C.2.

Proof of Theorem C.4. Following the same arguments as in the proof of theorem 3.5, equation A.6 becomes

$$\begin{aligned}\max_i \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}[\tilde{x}_{it}'(\theta^0(c')-\theta^0(c))]^2 \le 4\eta MK + (4\eta+2M)\eta\right)\\ +\max_i \mathbb{P}\left(\frac{1}{T}\sum_t \|\tilde{x}_{it}\|^2 > M'\right) + \max_i \mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\tilde{e}_{it}\tilde{x}_{it}\right\| > \eta\right)\end{aligned} \tag{C.6}$$

where we have replaced M' with an arbitrary positive constant K, to be determined below. For the second term, note that by assumption 3.4.(f), we have $\max_i \mathbb{P}(\sum_t \|x_{it}\|^2 > M') = o(T^{-a})$ for each $a>0$. Note that for any $C>0$, since $z\le z^2+1$ on $\mathbb{R}_{\ge 0}$

$$\mathbb{P}\left(\frac{1}{T}\sum_t \|x_{it}\| > C\right) \le \mathbb{P}\left(\frac{1}{T}\sum_t (\|x_{it}\|^2+1) > C\right) \le \mathbb{P}\left(\frac{1}{T}\sum_t \|x_{it}\|^2 > C-1\right)$$

Also note that $\|x_{it}-\overline{x}_i\|^2 \le 2(\|x_{it}\|^2+\|\overline{x}_i\|^2)$, and $\mathbb{P}(\|\overline{x}_i\|^2 > C_1) \le \mathbb{P}(\frac{1}{T}\sum_t \|x_{it}\| > C_1^{1/2})$ by Cauchy-Schwarz. Putting this all together, we find that

$$\begin{aligned}\mathbb{P}\left(\frac{1}{T}\sum_t \|x_{it}-\overline{x}_i\|^2 > K\right) &\le \mathbb{P}\left(\frac{1}{T}\sum_t \|x_{it}\|^2 > K/4\right) + \mathbb{P}\left(\|\overline{x}_i\|^2 > K/4\right)\\ &\le \mathbb{P}\left(\frac{1}{T}\sum_t \|x_{it}\|^2 > K/4\right) + \mathbb{P}\left(\frac{1}{T}\sum_t \|x_{it}\|^2 > \sqrt{K}/2-1\right)\\ &= o(T^{-a})\end{aligned}$$

The first inequality follows from $\|a+b\|^2 \le 2(\|a\|^2+\|b\|^2)$ and a union bound. Moreover, the $o(T^{-a})$ statement holds uniformly over i as long as $K\ge 4(M'+1)^2$, where M' is as in the uniform large deviations bound on $\|x_{it}\|^2$ in assumption 3.4.(f).

For the third term, note that $\frac{1}{T}\sum_t \tilde{e}_{it}\tilde{x}_{it} = \frac{1}{T}\sum_t (e_{it}-\overline{e}_i)(x_{it}-\overline{x}_i) = \frac{1}{T}\sum_t e_{it}x_{it} - \overline{e}_i\overline{x}_i$. The term $\frac{1}{T}\sum_t e_{it}x_{it}$ is uniformly $o_p(T^{-a})$ for any $a>0$ by assumption C.3.(b), as shown in the proof of 3.5. For $c>0$, the second term $\overline{e}_i\overline{x}_i$ has

$$\begin{aligned}\mathbb{P}(\|\overline{e}_i\overline{x}_i\| > c) &\le \mathbb{P}(\|\overline{x}_i\| > M'+1) + \mathbb{P}\left(|\overline{e}_i| > \frac{c}{M'+1}\right)\\ &\le \mathbb{P}(\frac{1}{T}\sum_t \|x_{it}\|^2 > M') + \mathbb{P}\left(|\overline{e}_i| > \frac{c}{M'+1}\right)\end{aligned}$$

The first term is uniformly $o(T^{-a})$ (in the sense of equation A.6) by assumption, and the second term is uniformly $o(T^{-a})$ by the same type of argument in the main proof using assumptions C.3.(e), 3.4.(d), and 3.4.(e) to invoke lemma A.4 on tail bounds for strongly mixing processes.

Let $K = 4(M'+1)^2$ and η such that $4\eta MK + (4\eta + 2M)\eta < \tilde{d}_{min}/2$. With $g_{it} = E((x'_{it}(\theta^0(c) - \theta^0(c')))^2)$ and T' such that $\frac{1}{T'}\sum_{t=1}^{T'} g_{it} \geq \frac{1}{3}\tilde{d}_{min}$. Then for $T > T'$, the first term is

$$\begin{aligned}\mathbb{P}\left(\frac{1}{T}\sum_t \left([\tilde{x}'_{it}(\theta^0(c') - \theta^0(c))]^2 - g_{it}\right) \leq (1/3)\tilde{d}_{min} - \frac{1}{T}\sum_t g_{it}\right)\\ \leq \mathbb{P}\left(\left|\frac{1}{T}\sum_t [\tilde{x}'_{it}(\theta^0(c') - \theta^0(c))]^2 - g_{it}\right| \geq \frac{1}{6}\tilde{d}_{min}\right)\end{aligned}$$

Setting $\Delta\theta \equiv \theta^0(c') - \theta^0(c)$, we can expand each term in the sum on the left hand side as

$$\begin{aligned}[\tilde{x}'_{it}(\theta^0(c') - \theta^0(c))]^2 - g_{it} &= ((\tilde{x}'_{it}\Delta\theta)^2 - E(\tilde{x}'_{it}\Delta\theta)^2) = (x'_{it}\Delta\theta)^2 - E(x'_{it}\Delta\theta)^2\\ &\quad - 2(\overline{x}'_i\Delta\theta\Delta\theta' x_{it} - E\overline{x}'_i\Delta\theta\Delta\theta' x_{it}) + ((\overline{x}'_i\Delta\theta)^2 - E(\overline{x}'_i\Delta\theta)^2)\\ &\equiv B^1_{iT} + B^2_{iT} + ((\overline{x}'_i\Delta\theta)^2 - E(\overline{x}'_i\Delta\theta)^2)\end{aligned}$$

Fix $C > 0$. The first term has $\mathbb{P}(\frac{1}{T}\sum_t (x'_{it}\Delta\theta)^2 - E(x'_{it}\Delta\theta)^2 > C) = o(T^{-a})$ uniformly over i by applying lemma A.4 exactly as in the proof of the main theorem. For the second term, note that $\frac{1}{T}\sum_t (\overline{x}'_i\Delta\theta\Delta\theta' x_{it} - \frac{1}{T}\sum_t E\overline{x}'_i\Delta\theta\Delta\theta' x_{it}) = (\overline{x}_i - E\overline{x}_i)'\Delta\theta\Delta\theta'\overline{x}_i$, and

$$|(\overline{x}_i - E\overline{x}_i)'\Delta\theta\Delta\theta'\overline{x}_i| \leq \|(\overline{x}_i - E\overline{x}_i)\|\|\Delta\theta\Delta\theta'\overline{x}_i\| \leq \|(\overline{x}_i - E\overline{x}_i)\|\|\Delta\theta\|^2\|\overline{x}_i\|$$

Then we have

$$\begin{aligned}&\mathbb{P}(\|(\overline{x}_i - E\overline{x}_i)\|\|\Delta\theta\|^2\|\overline{x}_i\| > C) &\text{(C.7)}\\ &\leq \mathbb{P}(\|\Delta\theta\|^2\|\overline{x}_i\| > M^2(M'+1)) + \mathbb{P}\left(\|(\overline{x}_i - E\overline{x}_i)\| > \frac{C}{M^2(M'+1)}\right)\\ &\leq \mathbb{P}(\frac{1}{T}\sum_t \|x_{it}\|^2 > M') + \mathbb{P}\left(\|(\overline{x}_i - E\overline{x}_i)\| > \frac{C}{M^2(M'+1)}\right)\\ &= o(T^{-a}) &\text{(C.8)}\end{aligned}$$

uniformly in i, where the second inequality follows by assumption 3.1.(*a*) and the same algebra used above to bound $\|\overline{x}_i\|$ in probability using assumption 3.4.(f). That the final term is uniformly $o(T^{-a})$ follows by lemma A.4, using the tail conditions and strong mixing assumed in C.3.(b). The final term above is

$$\begin{aligned}(\overline{x}'_i\Delta\theta)^2 - E(\overline{x}_i\Delta\theta)^2 &= (\overline{x}'_i\Delta\theta)^2 - \mathrm{Var}(\overline{x}'_i\Delta\theta) - (E\overline{x}'_i\Delta\theta)^2\\ &= -\mathrm{Var}(\overline{x}'_i\Delta\theta) + (\overline{x}'_i\Delta\theta - E(\overline{x}_i\Delta\theta))(\overline{x}'_i\Delta\theta + E(\overline{x}'_i\Delta\theta))\end{aligned}$$

Note that $|E(\overline{x}'_i\Delta\theta)| \leq ME\|\overline{x}_i\| \leq ME\frac{1}{T}\sum_t \|x_{it}\| \leq ME\frac{1}{T}\sum_t \|x_{it}\|^2$ by Cauchy-Schwarz, compactness of Θ, and monotonicity of Lp norms, respectively. Let $E_{iT} = \mathbb{1}(\frac{1}{T}\sum_t \|x_{it}\|^2 >$

M'), then

$$\sup_i E\left(\frac{1}{T}\sum_t \|x_{it}\|^2\right) \leq \sup_i E\left(\mathbb{1}(E_{it})\frac{1}{T}\sum_t \|x_{it}\|^2\right) + M'$$
$$\leq M' + \sup_i \mathbb{P}(E_{iT})^{1/2}\left(\frac{1}{T^2}\sum_{t,s} E\|x_{it}\|^2\|x_{is}\|^2\right)^{1/2}$$
$$= M' + o(T^{-a})O(1) = O(1) \tag{C.9}$$

where the second inequality uses Cauchy-Schwarz, and the last line uses assumptions 3.4.(f) and C.3.(a). Noting that $\overline{x}_i'\Delta\theta - E(\overline{x}_i'\Delta\theta = o_p(T^{-a})$ by mixing and tail assumptions on x_{it}, compactness, and lemma A.4, the product term can now be shown to have $\mathbb{P}[(\overline{x}_i'\Delta\theta - E(\overline{x}_i\Delta\theta))(\overline{x}_i'\Delta\theta + E(\overline{x}_i'\Delta\theta) > C] = o(T^{-a})$ using the same type of argument as in equation C.8. Lemma B.1 in the supplemental appendix shows that $\max_i \text{Var}(\overline{x}_i'\Delta\theta) = O(1/T)$ under our assumptions. In particular, we can choose T'' such that $\max_i \text{Var}(\overline{x}_i'\Delta\theta) < (\tilde{d}_{min}/12)$ for all $T > T''$.

Finally, define $B^3_{iT} = (\overline{x}_i'\Delta\theta - E(\overline{x}_i\Delta\theta))(\overline{x}_i'\Delta\theta + E(\overline{x}_i'\Delta\theta)$ and $B^k_T \equiv \frac{1}{T}\sum_t B^k_{iT}$. Then for $T > \max(T', T'')$, for all i we have

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_t [\tilde{x}_{it}'(\theta^0(c') - \theta^0(c))]^2 - g_{it}\right| \geq \frac{1}{6}\tilde{d}_{min}\right) \leq \mathbb{P}\left(B^1_T + B^2_T + B^T_3 + \text{Var}(\overline{x}_i'\Delta\theta) \geq \frac{1}{6}\tilde{d}_{min}\right)$$
$$\leq \mathbb{P}\left(B^1_T + B^2_T + B^T_3 + \geq \frac{1}{12}\tilde{d}_{min}\right) \leq \sum_{k=1}^{3}\mathbb{P}\left(B^k_T > \frac{1}{36}\tilde{d}_{min}\right) = o(T^{-a})$$

The first inequality follows from the triangle inequality, the second from $T > T''$ and the final inequality from a union bound. The $o(T^{-a})$ holds uniformly in i by the arguments above. This completes the proof that equation C.6 is $o(T^{-a})$ for any $a > 0$.

One final issue is the use of assumption 3.4.(a) in equation A.9. For the fixed effects case, we replaced this assumption with assumption *C.3.(a)*; however, one can show that $\frac{1}{T^2}\sum_{t,s}\|\tilde{x}_{it}\|^2\|\tilde{x}_{is}\|^2 \leq \frac{16}{T^2}\sum_{t,s}\|x_{it}\|^2\|x_{is}\|^2$, so that $\max_i \frac{1}{T^2}\sum_{t,s} E(\|\tilde{x}_{it}\|^2\|\tilde{x}_{is}\|^2) = O(1)$. Then $\frac{1}{NT^2}\sum_i\sum_{t,s}\|\tilde{x}_{it}\|^2\|\tilde{x}_{is}\|^2 = O_P(1)$ by the Markov inequality. The remainder of the proof follows exactly as in the proof of theorem 3.5, substituting $(\tilde{x}, \tilde{e})_{it}$ for $(x, e)_{it}$. □

C.2 Computation

As described in section 2, to solve problem 2.1 we primarily rely on Lloyd's algorithm, which performs coordinate descent on $\Theta \times \Gamma$. It is well known that this problem may be nonconvex, so in general coordinate descent will only yield a local minimum. To mitigate this issue, we rely on multiple random initializations. In our simulations we choose initial $\theta \sim \mathcal{N}(0, \sigma^2 I)$ and each $c_{i\ell} \sim \text{Unif}([k_\ell])$ independently.

Convergence Over Initializations - In this section, we give some evidence on the convergence of our algorithm for different data-generating processes. Given $1 \leq v \leq S$ random initializations, let $(\widehat{\theta}_v, \widehat{\gamma}_v)$ be the estimator achieved on the v^{th} initialization. Define $\widehat{Q}_s^{opt} \equiv \min_{1\leq v\leq s}\widehat{Q}(\widehat{\theta}_v, \widehat{\gamma}_v)$ and $\widehat{\theta}_s^{opt}$ to be the estimator that achieves $\widehat{Q}_s^{opt}$ (out of the first s initializations). Define the mean relative errors

$$r_Q(s) = \mathbb{E}\left[\frac{\widehat{Q}_s^{opt} - \widehat{Q}_S^{opt}}{\widehat{Q}_S^{opt}}\right] \qquad r_\theta(s) = \mathbb{E}\left[\frac{\|\widehat{\theta}_s^{opt} - \widehat{\theta}_S^{opt}\|}{\|\widehat{\theta}_S^{opt}\|}\right] \tag{C.10}$$

where each expectation is taken over the joint distribution of (x_{it}, y_{it}) and the sequence of random initializations $(\theta, \gamma)_s^{init}$. Monte Carlo approximations of the paths $r_Q(s)$ and $r_\theta(s)$ for DGP's mirroring those used in section 5 are shown in figures 1, 2, and 3. Note in particular that "angle" refers to a measure of cluster separation, as in the simulation design in section 5. In table 7, we show the number of initializations required to achieve 0.1% relative error for each DGP. Each simulation reports results up to $S = 200$, calculated using 200 independent sample paths.

The results show that problem 2.1 becomes significantly easier as (N, T) increases. Problems with fewer, well-separated clusters also converge more quickly to a stable solution (specifically, no improvements with additional random initializations up to S). In particular, $r_Q(50) \leq 0.01\%$ for all DGP's. Thus, we use $S = 50$ for our Monte Carlo simulations, reported in section 5. There is a large literature in computer science on heuristics for the least-squares partitioning problem, as well as some recent work on exact methods. See BM Appendix S1 and the references therein for more details.

Algorithm Hyperparameters - The hyperparameters for our implementation are given by $(S, tol, itermax)$, where $(tol, itermax)$ define a stopping rule for coordinate descent. With j denoting the number of update cycles ((2) and (3) in our algorithm), we stop if either $j > itermax$ or $\|\widehat{\theta}_j - \widehat{\theta}_{j-1}\| < tol$. We use $tol = 1 \cdot 10^{-8}$ and $itermax = 400$. We found solutions to be very insensitive to both hyperparameters for $tol > 1 \cdot 10^{-6}$ and $itermax > 100$. We use $S = 50$, as described above.

D Tables and Figures

Table 1: Effect of Cluster Separation

	Coverage		Bootstrap Coverage		Param. MSE		Cluster Loss	
Angle (α)	AR(1)	HK	AR(1)	HK	AR(1)	HK	AR(1)	HK
1.57	0.90	0.89	0.88	0.87	0.047	0.053	0.044	0.050
1.26	0.86	0.84	0.83	0.81	0.055	0.061	0.074	0.083
0.94	0.75	0.72	0.73	0.69	0.059	0.067	0.13	0.14
0.63	0.53	0.50	0.51	0.49	0.058	0.064	0.22	0.23
0.31	0.25	0.25	0.24	0.24	0.052	0.059	0.37	0.38
0.16	0.20	0.19	0.20	0.19	0.048	0.054	0.44	0.44

Notes: N=150, T=10

Table 2: Effect of Sample Size (N, T)

		Param. MSE					Coverage (Analytical)					Cluster Loss				
Errors	T	N=50	100	150	200	250	50	100	150	200	250	50	100	150	200	250
	5	0.154	0.140	0.137	0.134	0.133	0.761	0.763	0.772	0.768	0.756	0.134	0.125	0.125	0.123	0.123
AR(1)	10	0.051	0.047	0.047	0.046	0.046	0.873	0.889	0.898	0.899	0.901	0.046	0.044	0.044	0.044	0.044
	15	0.021	0.019	0.018	0.018	0.018	0.909	0.927	0.939	0.927	0.926	0.018	0.017	0.017	0.017	0.017
	20	0.009	0.008	0.008	0.008	0.007	0.916	0.939	0.935	0.935	0.935	0.007	0.007	0.007	0.007	0.007
	25	0.005	0.004	0.004	0.004	0.003	0.930	0.937	0.936	0.941	0.943	0.003	0.003	0.003	0.003	0.003
	5	0.160	0.146	0.144	0.143	0.140	0.718	0.748	0.738	0.738	0.736	0.137	0.130	0.130	0.130	0.128
HK	10	0.059	0.055	0.053	0.052	0.051	0.856	0.875	0.886	0.888	0.887	0.053	0.051	0.050	0.049	0.049
	15	0.027	0.024	0.023	0.023	0.022	0.905	0.914	0.916	0.920	0.917	0.023	0.022	0.021	0.022	0.021
	20	0.013	0.011	0.011	0.010	0.010	0.915	0.927	0.930	0.936	0.941	0.010	0.010	0.010	0.009	0.009
	25	0.007	0.006	0.005	0.005	0.005	0.924	0.936	0.946	0.946	0.942	0.004	0.004	0.005	0.004	0.004

Table 3: Effect of Number of Clusters (k_1, k_2)

# Clusters	Coverage		Bootstrap Coverage		Param. MSE		Function MSE		Cluster Loss	
(k1, k2)	AR(1)	HK	AR(1)	HK	AR(1)	HK	AR(1)	HK	AR(1)	HK
(1, 2)	0.90	0.88	0.88	0.86	0.026	0.030	0.026	0.094	0.035	0.040
(2, 2)	0.86	0.83	0.83	0.81	0.054	0.063	0.054	0.184	0.074	0.085
(2, 3)	0.84	0.81	0.82	0.79	0.070	0.078	0.070	0.218	0.090	0.100
(3, 3)	0.81	0.78	0.80	0.76	0.085	0.095	0.085	0.258	0.106	0.119
(3, 4)	0.80	0.76	0.78	0.74	0.099	0.109	0.099	0.281	0.118	0.131
(4, 4)	0.77	0.73	0.76	0.72	0.114	0.123	0.114	0.306	0.132	0.143

Notes: N=150, T=10

Table 4: Effect of Misspecified Blocking of Covariates - Estimation with $B^0 = 1$ and $B = 2$

Errors	# Clusters	Param. MSE		Function MSE	
	(k1, k2)	B=1	B=2	B=1	B=2
AR(1)	(1, 2)	0.027	0.026	0.077	0.026
	(2, 2)	0.058	0.054	0.157	0.054
	(2, 3)	0.079	0.070	0.206	0.070
	(3, 3)	0.108	0.085	0.273	0.085
	(3, 4)	0.137	0.099	0.334	0.099
	(4, 4)	0.163	0.114	0.383	0.114
		B=1	B=2	B=1	B=2
HK	(1, 2)	0.031	0.030	0.097	0.094
	(2, 2)	0.068	0.063	0.200	0.184
	(2, 3)	0.090	0.078	0.252	0.218
	(3, 3)	0.122	0.095	0.335	0.258
	(3, 4)	0.149	0.109	0.393	0.281
	(4, 4)	0.174	0.123	0.441	0.306

Notes: N=150, T=10

Table 5: Effect of Dimension Imbalance

dim	Coverage-large		Coverage-small		Cluster loss-small		Cluster loss-large		Param. MSE	
(m, p-m)	AR(1)	HK	AR(1)	HK	AR(1)	HK	AR(1)	HK	AR(1)	HK
(1, 11)	0.921	0.918	0.599	0.538	0.156	0.170	0.000	0.000	0.009	0.009
(2, 10)	0.930	0.931	0.841	0.840	0.069	0.060	0.001	0.001	0.008	0.007
(3, 9)	0.932	0.931	0.917	0.910	0.029	0.028	0.001	0.001	0.006	0.006
(4, 8)	0.935	0.930	0.922	0.916	0.016	0.019	0.001	0.001	0.005	0.005
(5, 7)	0.931	0.934	0.927	0.934	0.006	0.007	0.003	0.002	0.005	0.005
(6, 6)	0.936	0.938	0.939	0.935	0.003	0.003	0.004	0.004	0.004	0.004

Notes: N=150, T=10

Table 6: Effect of Growing Model Dimension

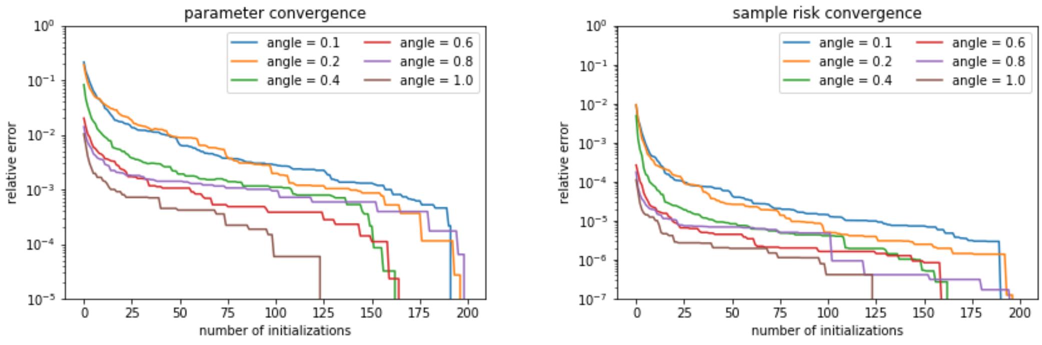
dim (p)	Coverage		Param. MSE		Function MSE		Cluster Loss	
Error	indep	AR(1)	indep	AR(1)	indep	AR(1)	indep	AR(1)
1	0.93	0.92	0.046	0.078	0.023	0.065	0.011	0.019
2	0.94	0.91	0.047	0.060	0.045	0.077	0.012	0.015
3	0.92	0.93	0.051	0.059	0.067	0.095	0.012	0.014
4	0.91	0.92	0.058	0.063	0.098	0.122	0.014	0.015
5	0.92	0.90	0.064	0.068	0.127	0.150	0.015	0.016

Notes: N=150, T=10

Table 7: Number of Initializations for 0.1% Rel. Error

(N, T)	s_θ	s_q	Angle	s_θ	s_q
(20, 10)	84	9	1	14	0
(50, 10)	65	1	0.8	98	0
(100, 10)	28	0	0.6	56	0
(150, 10)	9	0	0.4	109	2
(250, 10)	7	0	0.2	139	4
			0.1	157	5
(N, T)	s_θ	s_q	K	s_θ	s_q
(50, 5)	163	10	3	14	1
(50, 15)	17	0	4	48	0
(50, 20)	9	1	5	153	1
(50, 25)	1	0	6	130	4
			7	163	4

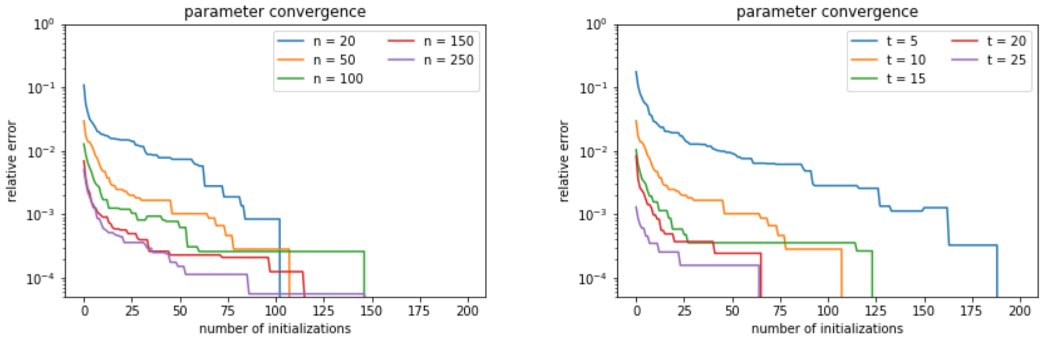
Figure 1: Algorithm Convergence and Cluster Separation



(a) $r_\theta(s)$

(b) $r_Q(s)$

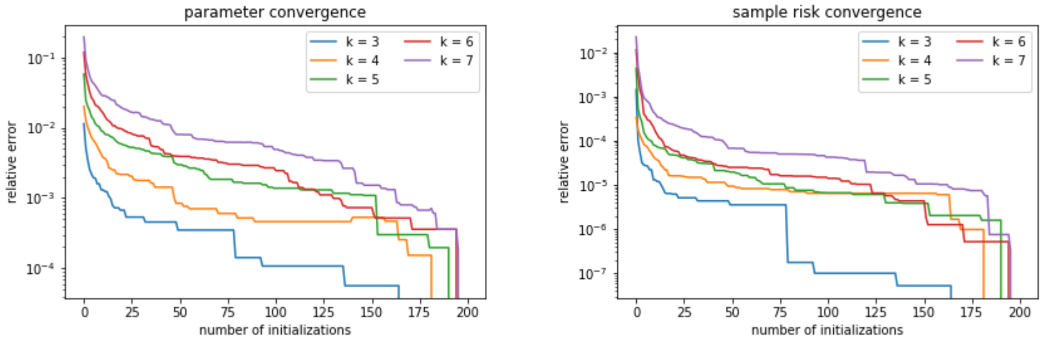
Figure 2: Algorithm Convergence and (N, T)



(a) $r_\theta(s)$

(b) $r_\theta(s)$

Figure 3: Algorithm Convergence and Number of Clusters $k = k_1 + k_2$



(a) $r_\theta(s)$

(b) $r_Q(s)$

References

Ando, T. and J. Bai (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics 31*, 163–191.

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*.

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

Bonhomme, S. and E. Manresa (2019). Discretizing unobserved heterogeneity. *Manuscript*.

Buchinsky, Hahn, and Hotz (2005). Cluster analysis: A tool for preliminary structural analysis. *Manuscript*.

Candes, E. J. and M. Soltanolkotabi (2012). A geometric analysis of subspace clustering with outliers. *The Annals of Statistics 40*(4), 2195–2238.

Chen, H., X. Leng, and W. Wang (2019). Latent group structures with heterogeneous distributions: Identification and estimation. *Manuscript*.

Cheng, X., F. Schorfheide, and P. Shao (2019). Clustering for multi-dimensional heterogeneity. *Manuscript*.

Dzemski, A. and R. Okui (2018). Confidence set for group membership. *Manuscript*.

Gelman, A. and J. Hill (2007). Data analysis using regression and multilevel/hierarchical models. *Cambridge University Press*.

Hansen, C. B. (2007). Asymptotic properties of a robust variance matrix estimator for panel data when t is large. *Journal of Econometrics 141*, 597–620.

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

Ke, Z. T., J. Fan, and Y. Wu (2015). Homogeneity pursuit. *Journal of the American Statistical Association 110*(509), 175–194.

Lian, H., X. Qiao, and W. Zhang (2019). Homogeneity pursuit in single index models based on panel data analysis. *Manuscript*.

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.

Liu, R., Z. Shang, Y. Zhang, and Q. Zhou (2019). Identification and estimation in panel models with overspecified number of groups. *Journal of Econometrics*.

Lloyd, S. P. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory 28*(2).

Merlevede, F., M. Peligrad, and E. Rio (2011). A bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields 151*, 435–474.

Rio, E. (1993). Covariance inequalities for strongly mixing processes. *Annales de l'I.H.P.*.

Rio, E. (2017). Asymptotic theory of weakly dependent random processes. *Springer-Verlag*.

Serban, N. and L. Wasserman (2005). Cats: Clustering after transformation and smoothing. *Journal of the American Statistical Association 100*(471).

Späth, H. (1979). Algorithm 39 clusterwise linear regression. *Computing 22*(367-373).

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

Sun, Y. (2005). Estimation and inference in panel structure models. *Manuscript*.

Vogt, M. and O. Linton (2019). Multiscale clustering of nonparametric regression curves. *Manuscript*.

Wang, W., P. C. B. Phillips, and L. Su (2016). Homogeneity pursuit in panel data models: theory and application. *Journal of Applied Econometrics*.

Yamamoto, M. and Y. Terada (2014). Functional factorial k-means analysis. *Computational Statistics and Data Analysis 79*, 133–148.

Zhang, Y., H. J. Wang, and Z. Zhu (2019). Quantile-regression-based clustering for panel data. *Journal of Econometrics 213*, 54–67.