cemmap
centre for microdata methods and practice

Linear panel regressions with two-way unobserved heterogeneity

Hugo Freeman
Martin Weidner

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP39/21

Economic and Social Research Council

Linear Panel Regressions with Two-Way Unobserved Heterogeneity*

Hugo Freeman† Martin Weidner§

November 2, 2021

Abstract

This paper studies linear panel regression models in which the unobserved error term is an unknown smooth function of two-way unobserved fixed effects. In standard additive or interactive fixed effect models the individual specific and time specific effects are assumed to enter with a known functional form (additive or multiplicative), while we allow for this functional form to be more general and unknown. We discuss two different estimation approaches that allow consistent estimation of the regression parameters in this setting as the number of individuals and the number of time periods grow to infinity. The first approach uses the interactive fixed effect estimator in Bai (2009), which is still applicable here, as long as the number of factors in the estimation grows asymptotically. The second approach first discretizes the two-way unobserved heterogeneity (similar to what Bonhomme, Lamadon and Manresa 2021 are doing for one-way heterogeneity) and then estimates a simple linear fixed effect model with additive two-way grouped fixed effects. For both estimation methods we obtain asymptotic convergence results, perform Monte Carlo simulations, and employ the estimators in an empirical application to UK house price data.

*We are grateful for feedback from conference participants at the 27th International Panel Data Conference, and the 2021 Bristol Econometric Study Group. We also thank Eric Auerbach, Stéphane Bonhomme, Arturas Juodis, Liangjun Su, and Andrei Zeleneev for useful comments and discussions. This research was supported by the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001, and by the European Research Council grants ERC-2014-CoG-646917-ROMIA and ERC-2018-CoG-819086-PANEDA.

†Department of Economics, University College London, Gower Street, London WC1E 6BT, UK. Email: hugo.freeman.16@ucl.ac.uk

§Department of Economics and Nuffield College, University of Oxford, Manor Road, Oxford OX1 3UQ, and Institute for Fiscal Studies, London, UK. Email: martin.weidner@economics.ox.ac.uk

1 Introduction

We consider the following panel data model for $i = 1, \dots, N$ cross-sectional units, and $t = 1, \dots, T$ time periods,

$$Y_{it} = X_{it}' \beta + u_{it}, \qquad u_{it} = h(\alpha_i, \gamma_t) + \varepsilon_{it}, \tag{1}$$

where Y_{it} is an observed dependent variable, $X_{it} = (X_{it,1}, \dots, X_{it,K})'$ is a K-vector of observed explanatory variables, and u_{it} is an unobserved error term. Within the unobserved error term, we have an unknown real-valued function $h(\cdot, \cdot)$ that depends on the (vector-valued) unobserved fixed effects $\alpha_i \in \mathbb{R}^{d_\alpha}$ and $\gamma_t \in \mathbb{R}^{d_\gamma}$, which are allowed to be arbitrarily correlated with the observed regressors X_{it}, while ε_{it} is a mean-zero error term that is uncorrelated with X_{it}. Our focus is on estimation of and inference on the parameter $\beta \in \mathbb{R}^K$ — the regression coefficient of X_{it} on Y_{it} when properly controlling for the unobserved α_i and γ_t.

The key model restrictions in (1) are the linearity in X_{it} as well as the additive separability between $X_{it}' \beta$ and u_{it}. If the unobserved error term u_{it} is of the more general form $u_{it} = g(\alpha_i, \gamma_t, \xi_{it})$, for some idiosyncratic errors ξ_{it} that are identically distributed across i and over t, then under appropriate regularity conditions we can define $h(\alpha_i, \gamma_t) = \mathbb{E}\left[u_{it} \,\middle|\, \alpha_i, \gamma_t\right]$ and $\varepsilon_{it} = u_{it} - h(\alpha_i, \gamma_t)$ to again obtain model (1). The additive separability between $h(\alpha_i, \gamma_t)$ and ε_{it} is therefore without loss of generality, as long as the distribution of u_{it} satisfies some homogeneity and regularity conditions. However, throughout this paper we take the representation of the model in (1) as the starting point for our analysis.

Analogous to the singular value decomposition of a *matrix*, there exists, under weak regularity conditions, the singular value decomposition of a *function* $h : \mathbb{R}^{d_\alpha} \times \mathbb{R}^{d_\gamma} \to \mathbb{R}$, which reads

$$h(\alpha, \gamma) = \sum_{r=1}^{\infty} \sigma_r \, \varphi_r(\alpha) \, \psi_r(\gamma), \tag{2}$$

for some functional singular values $\sigma_r > 0$, and appropriate normalized functions $\varphi_r : \mathbb{R}^{d_\alpha} \to \mathbb{R}$ and $\psi_r : \mathbb{R}^{d_\gamma} \to \mathbb{R}$, $r \in \{1, 2, 3, \dots\}$. Equation (2) allows us to rewrite model (1) as

$$Y_{it} = X_{it}' \beta + \sum_{r=1}^{\infty} \lambda_{ir} \, f_{tr} + \varepsilon_{it}, \tag{3}$$

with $\lambda_{ir} := \sigma_r \, \varphi_r(\alpha_i)$ and $f_{tr} := \psi_r(\gamma_t)$. Thus, our model can be viewed as a linear panel regression model with unobserved "factor structure" or "interactive fixed effects", but where the number of factors f_{tr} and corresponding factor loadings λ_{ir} is infinite. The same rewriting of a function $h(\alpha_i, \gamma_t)$ by an infinite sum $\sum_{r=1}^{\infty} \lambda_{ir} \, f_{tr}$ is used in Menzel (2021), but for a different model, and with the goal of analyzing the bootstrap for multidimensional data.

Within a panel regression context, most of the existing literature assumes that the number of unobserved factors is finite, which, from our perspective, corresponds to a truncation of the infinite sequence of factors in (3), that gives

$$Y_{it} = X_{it}'\,\beta + \sum_{r=1}^{R} \lambda_{ir}\, f_{tr} + e_{it}, \tag{4}$$

where $e_{it} := \varepsilon_{it} + \sum_{r=R+1}^{\infty} \lambda_{ir}\, f_{tr}$. For datasets with both N and T large, the two currently dominant estimation methods for the panel regression model in (4) are the common correlated effect (CCE) estimator of Pesaran (2006) and the least-squares (LS) estimator (also called quasi maximum likelihood estimator) in Bai (2009). Since those original papers by Pesaran and Bai, a large literature has emerged that has extended the CCE and LS estimation methods, and has analyzed the properties of those estimators in more general settings — see Chudik and Pesaran (2013), Bai and Wang (2016), and Karabiyik, Palm and Urbain (2019) for recent surveys. We follow that literature here by also considering panels with both N and T large, that is, for our asymptotic results we consider $N, T \to \infty$.[1]

The "conventional" interactive fixed effect model in (4) is a special case of our model (1), with $\alpha_i = \lambda_i = (\lambda_{i1}, \ldots, \lambda_{iR})'$, $\gamma_t = f_t = (f_{t1}, \ldots, f_{tR})'$, and $h(\alpha_i, \gamma_t) = \lambda_i' f_t$. The key question that we ask in this paper is what happens when the multiplicative factor structure $\lambda_i' f_t$ is replaced by a more general non-linear factor structure $h(\alpha_i, \gamma_t)$. However, we do maintain all other assumptions of model (4), in particular, the homogenous regression coefficient β, and the additive separability between $X_{it}'\,\beta$ and the unobserved factor structure.

The main challenge that we need to tackle when considering this extension is that, if the data generating process is given by (1), then the error term e_{it} in (4) will generally be correlated with X_{it}, because e_{it} contains the truncated part $\sum_{r=R+1}^{\infty} \lambda_{ir}\, f_{tr}$ of the infinite factor structure,[2] and $\lambda_{ir} = \varphi_r(\alpha_i)$ and $f_{tr} = \psi_r(\gamma_t)$ are functions of α_i and γ_t, which can be correlated with X_{it}. Once e_{it} is correlated with X_{it} in this way, then the existing results for the CCE and the LS estimator are not applicable anymore. The currently known results on the CCE and LS estimator in the presence of an infinite number of factors (e.g. Pesaran and Tosetti 2011, Chudik, Pesaran and Tosetti 2011b, and Westerlund and Urbain 2013) require that the "unaccounted" factors $\sum_{r=R+1}^{\infty} \lambda_{ir}\, f_{tr}$ are uncorrelated with the regressors, so that they can be considered part of the error term e_{it} without generating an endogeneity problem.

[1] There is of course also work on model (4) in the context of short T panels, for example, Holtz-Eakin, Newey and Rosen (1988), Ahn, Lee and Schmidt (2001, 2013), Sarafidis and Robertson (2009) Juodis and Sarafidis (2020, 2018, 2020), Westerlund, Petrova and Norkute (2019),

[2] Notice that the majority of these truncated factors will be "weak", see Onatski Onatski (2010, 2012) and Chudik, Pesaran and Tosetti Chudik, Pesaran, and Tosetti (2011a) for the distinction between "strong" and "weak" factors.

For the case that X_{it} and e_{it} are correlated, there exist instrumental variable (IV) generalizations of both the CCE and LS method (e.g. Harding and Lamarche 2011, Lee, Moon and Weidner 2012, Robertson and Sarafidis 2015, Moon, Shum and Weidner 2018, and Norkutė, Sarafidis, Yamagata and Cui 2021), but those require observed instruments Z_{it} that are uncorrelated with e_{it}. In our setting, this could only be reasonably justified if Z_{it} is independent of α_i, γ_t and ε_{it}. If such strong instruments would be available, then a standard pooled 2SLS estimator – completely ignoring $h(\alpha_i, \gamma_t)$ in model (1) – would be consistent under appropriate regularity conditions, but existence of such instruments is unlikely in most applications, and is not assumed in this paper.

The two main theoretical contributions of our paper are as follows: Firstly, we formally show that the LS estimator of Bai (2009) can still provide consistent estimates of β in model (1), as long as the number of factors $R = R_{NT}$ used in estimation grows to infinity jointly with N and T. Secondly, we suggest an alternative estimator for β, which we denote the *two-way group fixed-effect estimator* (generalizing ideas in Bonhomme, Lamadon and Manresa 2021 on the discretization of one-way heterogeneity), and we provide conditions under which this new estimator is $\sqrt{NT}$-consistent as $N, T \to \infty$. In addition, we also suggest inference procedures using both of these estimators, but we do not formally derive inference results in this paper. Instead, we study the properties of our suggested confidence intervals in Monte Carlo simulations. We also apply the estimators to an empirical application on UK house price data.

When employing the LS estimator with factors from Bai (2009) to model (1), we are effectively estimating a misspecified model — the DGP is given by (1), but the estimating equation by (4). Galvao and Kato (2014) and Juodis (2020) have recently studied linear panel regression models with additive fixed effects under misspecification. We consider interactive fixed effects for estimation here, and the type of misspecification we allow for is more restrictive. We therefore do not have to introduce any pseudo-true parameter, but we find that the LS estimator is still consistent for the true value of β under our assumptions.

In related work, allowing for the number of factors to grow with sample size has been considered in Li, Li and Shi (2017), where they explicitly detail a factor model with the number of factors growing with sample size. The difference to this paper is our model admits an infinite number of factors even in small samples and considers finite factor estimation as an approximation to the true data generating process.

There also exist other work on non-linear generalizations of the interactive fixed effect and factor model specification. Zeleneev (2020) considers the same model (1) in the context of network data, but in his baseline discussion, the outcome Y_{ij} (instead of Y_{it} here) is symmetric in i and j. The main difference to our work, however, is that Zeleneev estimates the model based on a strategy that identifies agents with similar fixed effect values based on the distribution of

their outcomes. His estimation method is accordingly also completely different to ours.

Bodelet and Shan (2020) also consider non-linear functions in place of the standard linear factor model. In our notation, their model assumes a series of smooth univariate functions of the form $\sum_{q=1}^{Q} h_{iq}(\gamma_{tq})$ for unobserved heterogeneity. Their approach models individual specific responses to structural shocks but is different to our approach, which uses a homogeneous bivariate function. Therefore, their approach allows for discontinuities across how individual effects are modelled whereas our assumption is more restrictive since variation across individuals, via α_i, must be smooth.

Other papers on unobserved two-way heterogeneity in panel or network models either make more parametric assumptions (e.g. Graham 2017, Dzemski 2019, Chen, Fernández-Val and Weidner 2020), or employ stochastic block or graphon models (e.g. Holland, Laskey and Leinhardt 1983, Wolfe and Olhede 2013, Gao, Lu, Zhou et al. 2015, Auerbach 2019), and are therefore less closely related to our paper.

There are also recent papers that use matrix completion methods for the purpose of treatment effect estimation in panel models with two-way heterogeneity, e.g. Athey, Bayati, Doudchenko, Imbens and Khosravi (2017) and Amjad, Shah and Shen (2018), Chernozhukov, Hansen, Liao and Zhu (2021), and Fernández-Val, Freeman and Weidner (2021). Those papers do not require the additive separability between the regressors and error term in (1), but as a result they also have to make stronger assumptions and employ more complicated estimation methods than we do here. The same is true for Freyberger (2017), who considers a non-separable model with interactive fixed effects. Alternative non-linear extensions of factor models are discussed, for example, in Cunha, Heckman and Schennach (2010) and Gunsilius and Schennach (2019).

The rest of the paper is organized as follows. Section 2 introduces our suggested estimators and inference methods. Section 3 and Section 4 provide asymptotic results for the LS estimator of Bai (2009) and for our new two-way group fixed-effect estimator, respectively. Section 5 discusses the practical implementation. Monte Carlo simulations are presented in Section 6, and an empirical application is worked out in Section 7.

2 Estimation approaches

In this section, we introduce the two estimation approaches that are afterwards analyzed and used in the rest of the paper.

2.1 Least-squares interactive fixed effect estimator

Following Bai (2009) we consider

$$\left(\widehat{\beta}_{\mathrm{LS}},\, \widehat{\lambda},\, \widehat{f}\right) = \underset{(\beta,\lambda,f)\in\mathbb{R}^{K+N\times R+T\times R}}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(Y_{it} - X_{it}'\,\beta - \sum_{r=1}^{R}\lambda_{ir}\,f_{tr}\right)^2. \tag{5}$$

This estimator was introduced for the exact factor model in equation (4), and Bai (2009) shows that it is $\sqrt{NT}$-consistent and asymptotically normally distributed for $N,T\to\infty$ when the true number of factors is fixed and known. Moon and Weidner (2015) extend this result to the case where the true number of factors is chosen too large in the estimation. To make the estimates $\widehat{\lambda}$ and $\widehat{f}$ in (5) unique, we choose the usual normalization $T^{-1}\widehat{f}'\widehat{f} = \mathbb{I}_R$, and $\widehat{\lambda}'\widehat{\lambda}$ to be a diagonal matrix.

As explained above, the model (1) that we consider in this paper can be rewritten as the factor model in (3) with an infinite number of factors in the true data generating process. This suggests that the least-squares estimators in (5) can still be consistent as long as the number of factors $R = R_{NT}$ used in the estimation is allowed to grow to infinity jointly with N and T. Estimation of $\left(\widehat{\beta}_{\mathrm{LS}},\, \widehat{\lambda},\, \widehat{f}\right)$ is done using an iterative scheme. That is, we start by initialising $\widehat{\beta}_{\mathrm{LS}}$, and then iterate between estimating the principal components of $Y - X\cdot\widehat{\beta}_{\mathrm{LS}}$ to obtain $\left(\widehat{\lambda},\, \widehat{f}\right)$ and least squares of $Y = X\cdot\beta + \widehat{\lambda}\widehat{f}' + e$ to obtain $\widehat{\beta}_{\mathrm{LS}}$. The convergence metric we use is the sum of squares in (5). However, this iteration scheme can converge to a local minimum, and it is therefore important to repeat the procedure with multiple starting values of β. For more details on the numerical computation of the estimator in (5) we refer to Bai (2009) and Moon and Weidner (2015).

This least-squares estimator of Bai (2009) is very well-established in the panel regression literature. It is used regularly both in empirical and in methodological papers, e.g. Su and Chen (2013), Kim and Oka (2014), Lu and Su (2016), Gobillon and Magnac (2016), Totty (2017), Su and Wang (2017), Moon and Weidner (2017), Giglio and Xiu (2021), to name just a few.

2.2 Group fixed effects estimator

Here, we introduce two-way grouped fixed effects estimator, which discretizes the unobserved heterogeneity that is parameterized by α_i and γ_t in the spirit of Bonhomme, Lamadon and Manresa (2021). We first describe the main idea of this estimator before explaining its practical implementation in more details.

2.2.1 Main Idea

We partition the set $\{1,\ldots,N\}$ of cross-sectional units into $G = G_{NT}$ groups such that individuals in the same group have similar values of α_i. Let $g_i \in \{1,\ldots,G\}$ denote the group

membership of individual i. Analogously, we partition the set $\{1,\dots,T\}$ of time periods into $C = C_{NT}$ groups such that time periods in the same group have similar values of γ_t. Let $c_t \in \{1,\dots,C\}$ denote the group membership of time period t. Details on how we construct those partitionings in practice are described below. Once we have obtained those groups, then we estimate β by applying pooled OLS to the linear fixed-effect model

$$Y_{it} = X_{it}'\,\beta + \delta_{i,c_t} + \nu_{t,g_i} + \epsilon_{it}, \tag{6}$$

where $\delta_{i,c_t} \in \mathbb{R}$ and $\nu_{t,g_i} \in \mathbb{R}$ are nuisance parameters that are jointly estimated with β, that is, the basic two-way grouped fixed effect estimator for β can be written as

$$\widehat{\beta}_{\mathrm{G}} = \underset{\beta\in\mathbb{R}^K}{\operatorname{argmin}} \min_{\delta\in\mathbb{R}^{N\times C}} \min_{\nu\in\mathbb{R}^{T\times G}} \sum_{i=1}^{N}\sum_{t=1}^{T} \left(Y_{it} - X_{it}'\,\beta - \delta_{i,c_t} - \nu_{t,g_i}\right)^2. \tag{7}$$

Notice that within each pair of groups for i and t, that is, for fixed values of c_i and g_t, the model in (6) is simply a standard additive two-way fixed effect model $Y_{it} = X_{it}'\,\beta+\delta_i+\nu_t+\epsilon_{it}$. However, as the group membership changes we allow the parameters δ_i and ν_t to change arbitrarily, as indicated by the additional subscripts c_t and g_i in (6). We could have written $\delta_{i,g_i,c_t} + \nu_{t,g_i,c_t}$ to indicate explicitly that both the individual and time effect are allowed to change across groups, but the notation in (6) of course already allows for that generality. The parameters δ therefore form an $N \times C$ matrix, while the parameters ν form a $T \times G$ matrix.

In the introduction, we explained how the LS-estimator with interactive effects can be justified for model (1) by a truncation of the functional singular value expansion in (2). In other words, a particular approximation of the function $h(\alpha_i, \gamma_t)$ naturally leads to the estimator in (5).

The grouped fixed effect estimator in (7) can be justified analogously by a different approximation of the function $h(\alpha_i, \gamma_t)$. Under appropriate regularity conditions, by a joint Taylor expansion in α_i and γ_t around the corresponding group means $\overline{\alpha}_{g_i} = \frac{\sum_{j=1}^{n} \mathbb{1}\{g_i=g_j\}\,\alpha_j}{\sum_{j=1}^{n} \mathbb{1}\{g_i=g_j\}}$ and $\overline{\gamma}_{c_t} = \frac{\sum_{s=1}^{T} \mathbb{1}\{c_t=c_s\}\,\gamma_s}{\sum_{s=1}^{T} \mathbb{1}\{c_t=c_s\}}$, we find that

$$h(\alpha_i, \gamma_t) = \delta_{i,c_t} + \nu_{t,g_i} + O\left(\|\alpha_i - \overline{\alpha}_{g_i}\|^2 + \|\gamma_t - \overline{\gamma}_{c_t}\|^2\right), \tag{8}$$

where

$$\delta_{i,c_t} := h(\overline{\alpha}_{g_i}, \overline{\gamma}_{c_t}) + \frac{\partial h(\overline{\alpha}_{g_i}, \overline{\gamma}_{c_t})}{\partial \alpha_i'}(\alpha_i - \overline{\alpha}_{g_i}), \qquad \nu_{t,g_i} := \frac{\partial h(\overline{\alpha}_{g_i}, \overline{\gamma}_{c_t})}{\partial \gamma_t'}(\gamma_t - \overline{\gamma}_{c_t}).$$

This shows that the leading order dependence of $h(\alpha_i, \gamma_t)$ on α_i and γ_t can be described by the additive specification $\delta_{i,c_t} + \nu_{t,g_i}$ used in (6). Since this two-way grouped fixed effect ignores the terms $O\left(\|\alpha_i - \overline{\alpha}_{g_i}\|^2 + \|\gamma_t - \overline{\gamma}_{c_t}\|^2\right)$ entirely, it is of course crucial to construct the groups such

that $\alpha_i - \overline{\alpha}_{g_i}$ and $\gamma_t - \overline{\gamma}_{c_t}$ are small. The clustering algorithm that we use to achieve that is described in Subsection 2.2.2 below.

Notice that a naive application of the ideas in Bonhomme, Lamadon and Manresa (2021) to our two-way fixed effect model would *not* result in our estimating equation (6) but in $Y_{it} = X_{it}'\beta + \chi_{g_i,c_t} + \epsilon_{it}$, where $\chi_{g,c}$ is a fixed effect specific to each pair of groups $(g,c) \in \{1,\dots,G\} \times \{1,\dots,C\}$. The analog of equation (8) for that alternative approach reads

$$h(\alpha_i,\gamma_t) = \underbrace{h(\overline{\alpha}_{g_i},\overline{\gamma}_{c_t})}_{=\chi_{g_i,c_t}} + O\left(\|\alpha_i - \overline{\alpha}_{g_i}\| + \|\gamma_t - \overline{\gamma}_{c_t}\|\right),$$

that is, the approximation error would be of linear order in the discrepancies $\alpha_i - \overline{\alpha}_{g_i}$ and $\gamma_t - \overline{\gamma}_{c_t}$ within groups. By contrast, for our estimating equation (6) the resulting approximation error in (8) is of quadratic order, which explains why we prefer that approach.

Finally, notice that if our original model would only contain individual specific fixed effects α_i, that is, $Y_{it} = X_{it}'\beta + h(\alpha_i) + \varepsilon_{it}$, then the analog of (6) is the standard additive fixed effect model $Y_{it} = X_{it}'\beta + \delta_i + \varepsilon_{it}$, which requires no grouping at all, and also entails no approximation error since we can set $\delta_i = h(\alpha_i)$. The way in which we generalize the grouping ideas in Bonhomme, Lamadon and Manresa (2021) is therefore quite specific to the two-way fixed effect model in (1).

2.2.2 Clustering algorithm

To make the two-way grouped fixed effect estimator in (7) operational we employ the following three-step algorithm:

A. Obtain the factor loading and factor estimates $\widehat{\lambda}$ and $\widehat{f}$ of the interactive fixed effect LS estimator in (5) for a relatively large number of factors R. Only keep the leading few R^* factor loading and factor estimates and denote those by $\widehat{\lambda}^* = (\widehat{\lambda}_{ir} : i = 1,\dots,N,\ r = 1,\dots,R^*)$ and $\widehat{f}^* = (\widehat{f}_{tr} : t = 1,\dots,T,\ r = 1,\dots,R^*)$.

B. Use the $\widehat{\lambda}_1^*,\dots,\widehat{\lambda}_N^*$ as inputs into the clustering algorithm in Table 1 to partition the set of individuals $\{1,\dots,N\}$. This algorithm returns the number G of chosen groups and the group membership $g_i \in \{1,\dots,G\}$ of each individual. Analogously, we use the inputs $\widehat{f}_1^*,\dots,\widehat{f}_T^*$ into the same algorithm to partition $\{1,\dots,T\}$, resulting in the number of groups C and the group membership $c_t \in \{1,\dots,C\}$ for each time period. Notice that the words partition, cluster, and group are used interchangeably in this paper.

C. Calculate the two-way grouped fixed effect estimator $\widehat{\beta}_{\mathrm{G}}$ via pooled OLS according to equation (7).

Algorithm

1: Input $\widehat{\lambda}_i^* \in \mathbb{R}^{R^*}$ for all $i = 1, \dots N$. Calculate all pairwise Euclidean distances $A_{ij} = \left\| \widehat{\lambda}_i^* - \widehat{\lambda}_j^* \right\|$, for $i \neq j$, and set $A_{ii} = \infty$. Initialize $\mathcal{P} = \{\{1\}, \{2\}, \dots, \{N\}\}$ as a partition of $\{1, \dots, N\}$.

2: **if** $\exists \mathcal{C}_* \in \mathcal{P}$ with $|\mathcal{C}_*| = 4$ **then** for that $\mathcal{C}_*$

3: Find the solution to

$$\min_{\{i,j,l,m \in \mathcal{C}_* : \mathcal{C}_* = \{i,j,l,m\}\}} A_{ij} + A_{lm},$$

and split $\mathcal{C}_*$ into $\{i, j\}$ and $\{l, m\}$, updating the partition $\mathcal{P}$.

4: **else if** $\exists \mathcal{C} \in \mathcal{P}$ with $|\mathcal{C}| = 1$ **then**

5: Find the solution to

$$\min_{\left\{i \in \bigcup_{\{\mathcal{C} \in \mathcal{P} : |\mathcal{C}| = 1\}}\right\}} \min_{\left\{j \in \bigcup_{\{\mathcal{C} \in \mathcal{P} : |\mathcal{C}| \leq 3\}}\right\}} A_{ij},$$

and merge the clusters containing i and j into a single cluster, updating the partition $\mathcal{P}$.

6: **end if**

7: Repeat 2-6 until $\{|\mathcal{C}| : \mathcal{C} \in \mathcal{P}\} \subset \{2, 3\}$.

Table 1: Hierarchical clustering with minimum single linkage.

It is constructive to briefly describe our algorithm from Table 1 in words before we discuss features of this whole procedure. Step 1 defines the proxy variable to cluster on ($\widehat{\lambda}_i^*$ in this instance) and sets the distance metric we wish to use, Euclidean distance, which could easily be changed to another norm or metric. Then, we initialise each individual into their own cluster. Steps 2 and 3 then splits any groups of four into two groups of two, since we want groups of no larger than three in our final output. The optimisation in Step 3 looks at all combinations of two by two splits within this group of four and takes the smallest sum of distances. This type of optimisation is only suitable for very small groups of individuals because it is a combinatorially hard problem.

Steps 4 and 5 then finds the solitary individual with the smallest distance to any other existing cluster and merges it to that cluster. Combined with Steps 2 and 3 we create an iteration that merges single clusters one at a time to groups of one, two or three, then splits any groups of four as and when they occur. This means Step 2 can only ever return one group of four. Doing this iteration one at a time is important so that we may split these groups of four immediately and have a larger choice set in Step 5 for each unmatched individual. Also, splitting groups of four into two by two groups rather than groups of one and three avoids infinite iterations. The repetition of Steps 2-5 is guaranteed to converge, and delivers a partition of $\{1, \dots, N\}$ into groups of size two or three.

Now to discuss the procedure as a whole. The choice of R in step A here is not too important since we only need this to generate proxy variables for clustering and otherwise dispose of β estimates from this initial LS step. The important hyperparameter is the number of proxies per

observation, R^*, which we choose equal to two to five. We discuss the theoretical properties of the hyperparameter in Section 4.1 but here outline a heuristic approach to this choice. Choosing R^* to be more than one is important to capture cases when α_i and γ_t have higher dimension or when the function $h(.,.)$ admits eigenfunctions that are not individually injective maps from α_i or γ_t. The aim is that a linear combination of non-injective maps provides a better mapping to the closeness of the primitives α_i and γ_t. An archetypal example of this is discussed in Griebel and Harbrecht (2014) where they show that the first few eigenfunctions of the exponential kernel are individually clearly not injective maps.

It is also important to not use too many proxies so as to avoid clustering on noise. This can make for poor matches that result in large deviations between α_i and α_j, respectively γ_t and γ_s, that show up in the leading $O\left(\|\alpha_i - \alpha_j\|^2\right)$ and $O\left(\|\gamma_t - \gamma_s\|^2\right)$ remainder terms in (8). Maintaining closeness in these primitives when clustering is key to any argument using Taylor's theorem, however, optimising this proxy hyperparameter is still rough and does require further development. We defer discussion about the presence of noise in factors with relation to the LS estimator to Section 3.

There are, of course, other choices for proxies such as the cross-sectional moments employed in Bonhomme, Lamadon and Manresa (2021). However, as displayed in (2) and formulated in Griebel and Harbrecht (2014), using the eigenfunctions from the singular value decomposition are a more natural choice since these are direct functions of the primitives α_i and γ_t and should in theory lead to closer proximity between these. Since we require cross-sectional and time-dependent clusters for our method, these eigenfunctions also provide a convenient means to find these. If one truly believes that other proxy variables have more precise injectivity with these primitives then they could always make those the the input to Step 1 in our clustering algorithm.

Another divergence from the existing literature is the use of clusters of size two or three, rather than letting these cluster sizes grow with sample size. Our motivation for using these small cluster sizes comes directly from the within-group β estimation, i.e. that we do not need consistent estimates of δ or ν since these are treated as nuisance parameters that are simply differenced out. Hence, for our purposes, it is more useful to have small groups that are very similar rather than to have large groups that have better central tendency estimates. This very conveniently removes one choice for the analyst, namely the setting of group sizes G or C.

2.2.3 Split-sample version of the estimator

As explained above, we estimate the group memberships g_i and c_t that enter into the estimator for β in (7) via a clustering method applied to $\widehat{\lambda}^*$ and $\widehat{f}^*$. However, clustering in this way creates dependence across i and t through $\widehat{\lambda}^*$ and $\widehat{f}^*$. This dependence creates technical difficulties when establishing asymptotic convergence results. To mitigate this dependence we augment

the clustering estimator by a simple sample splitting method. The resulting group fixed effect estimator with sample splitting is given by

$$\widehat{\beta}_{\mathrm{GS}} = \underset{\beta\in\mathbb{R}^K}{\operatorname{argmin}} \min_{\delta} \min_{\nu} \sum_{i=1}^{N}\sum_{t=1}^{T}\left[Y_{it} - X_{it}'\,\beta - \sum_{s=1}^{S} \mathbb{1}\left\{(i,t)\in\mathcal{O}_s\right\}\left(\delta^{(s)}_{i,c_t^{(s)}} + \nu^{(s)}_{t,g_i^{(s)}}\right)\right]^2, \tag{9}$$

where S is the number of partitions, and the sets $\mathcal{O}_s$, $s = 1,\ldots,S$, are the partitions of the sample space $\{1,\ldots,N\}\times\{1,\ldots,T\}$, that is, the observation (i,t) is a member of the s'th partition if and only if $(i,t)\in\mathcal{O}_s$. Compared to the original group fixed effect estimator in (6), the group membership indicators $g_i^{(s)}$ and $c_t^{(s)}$ and the group fixed effect $\delta^{(s)}_{i,c_t^{(s)}}$ and $\nu^{(s)}_{t,g_i^{(s)}}$ are all specific to the partition s. For the purpose of this paper, we choose the number of partitions to be $S = 4$ and we split the sample space into four blocks as follows:

$$\begin{aligned}
\mathcal{O}_1 &= \{1,\ldots,\lfloor N/2\rfloor\}\times\{1,\ldots,\lfloor T/2\rfloor\},\\
\mathcal{O}_2 &= \{1,\ldots,\lfloor N/2\rfloor\}\times\{\lfloor T/2\rfloor+1,\ldots,T\},\\
\mathcal{O}_3 &= \{\lfloor N/2\rfloor+1,\ldots,N\}\times\{1,\ldots,\lfloor T/2\rfloor\},\\
\mathcal{O}_4 &= \{\lfloor N/2\rfloor+1,\ldots,N\}\times\{\lfloor T/2\rfloor+1,\ldots,T\},
\end{aligned} \tag{10}$$

where $\lfloor\cdot\rfloor$ is the floor function.

We still need to explain how the group memberships $g_i^{(s)}$ and $c_t^{(s)}$ are obtained here. The aim of the sample splitting is to avoid any stochastic dependence between $g_i^{(s)}$ and $c_t^{(s)}$ and the idiosyncratic noise ε_{it}. For each partition $s = 1,\ldots,S$, we therefore construct the group memberships $g_i^{(s)}$ and $c_t^{(s)}$ without using outcomes Y_{it} for observations (i,t) of that partition $\mathcal{O}_s$. For that purpose, we define the sets

$$\begin{aligned}
\mathcal{O}_1^* &= \{1,\ldots,N\}\times\{1,\ldots,\lfloor T/2\rfloor\},\\
\mathcal{O}_2^* &= \{1,\ldots,N\}\times\{\lfloor T/2\rfloor+1,\ldots,T\},\\
\mathcal{O}_3^* &= \{1,\ldots,\lfloor N/2\rfloor\}\times\{1,\ldots,T\},\\
\mathcal{O}_4^* &= \{\lfloor N/2\rfloor+1,\ldots,N\}\times\{1,\ldots,T\},
\end{aligned} \tag{11}$$

and for $\tilde{s} = 1,\ldots,4$, we define the corresponding least-squares factor and loading estimates

$$\left(\widehat{\lambda}^{(\tilde{s})},\,\widehat{f}^{(\tilde{s})}\right) = \underset{(\lambda,f)\in\mathbb{R}^{N_{\tilde{s}}^*\times R+T_{\tilde{s}}^*\times R}}{\operatorname{argmin}} \min_{\beta\in\mathbb{R}^K} \sum_{(i,t)\in\mathcal{O}_{\tilde{s}}^*}\left(Y_{it} - X_{it}'\,\beta - \sum_{r=1}^{R}\lambda_{ir}\,f_{tr}\right)^2, \tag{12}$$

which is simply the LS estimator in (5) applied only to the $N_{\tilde{s}}^*\times T_{\tilde{s}}^*$ subpanel of observations $(i,t)\in\mathcal{O}_{\tilde{s}}^*$, and we also impose the same normalization on the factors and loadings explained after (5).[3] Now, for the original partition $\mathcal{O}_s$, $s = 1,\ldots 4$, we construct the group membership

[3]Notice that factor model proxies can only be used to compare observations from the same factor estimation sample space. This is because factors are only identified up to rotations, where these rotations may differ across estimation samples.

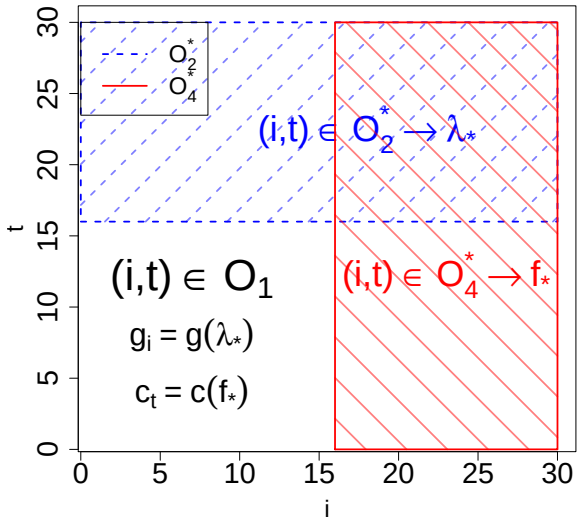

Figure 1: Sample split for partition 1

$g_i^{(s)}$ of unit i by applying the clustering algorithm in Table 1 to the loading estimates $\widehat{\lambda}_i^{(\tilde{s})}$ obtained from the subpanel $\mathcal{O}_{\tilde{s}}$ with $\tilde{s} = \tilde{s}(s)$ given by

$$\tilde{s} = \begin{cases} 2 & \text{for } s = 1, \\ 1 & \text{for } s = 2, \\ 2 & \text{for } s = 3, \\ 1 & \text{for } s = 4. \end{cases}$$

Analogously, for the partition $\mathcal{O}_s$, $s = 1, \ldots 4$, we construct the group membership $c_t^{(s)}$ of time period t by applying the clustering algorithm in Table 1 to the factor estimates $\widehat{f}_t^{(\tilde{s})}$ obtained from the subpanel $\mathcal{O}_{\tilde{s}}$ with $\tilde{s} = \tilde{s}(s)$ given by

$$\tilde{s} = \begin{cases} 4 & \text{for } s = 1, \\ 4 & \text{for } s = 2, \\ 3 & \text{for } s = 3, \\ 3 & \text{for } s = 4. \end{cases}$$

Figure 1 details an example of this sample splitting technique for clustering within partition $\mathcal{O}_1$. Here we see clearly how the partitions for proxy estimation $\mathcal{O}_2^*$ and $\mathcal{O}_4^*$ do not overlap with the partition we are grouping within, $\mathcal{O}_1$. This guarantees that we do not introduce any dependence between the group functions and the noise term by making sure grouping within each partition is not a function of the independent noise term, ε_{it}, from observations within that

partition. This becomes important in our derivations in Section 4.2, where we require that the process $X_{it}\,\varepsilon_{it}$ remains zero mean and independently distributed after group means are projected out.

With these cluster assignments it then becomes straightforward to estimate (9) by first doing within-cluster mean-differences on oberservations Y and X_k for each k partition-by-partition, then simply using OLS on the transformed variables.

Notice that by allowing the partitioning in (11) used to estimate proxy variables to extend over the whole sample of either N or T, we get better estimates than just using the original partition (10). As discussed earlier, it is crucial to avoid poor initial estimates of proxy variables to better approximate the residual terms in the Taylor expansion in expression (8).

3 Asymptotic results for the least squares estimator

Here, we derive convergence rate results for the least-squares estimator (5) for a data generating process given by (1). Thus, we generalize the consistency results in Bai (2009) and Moon and Weidner (2015) to the case where the underlying panel regression model does not satisfy the factor model in (4). However, as explained in the introduction, the factor model in (4) can be viewed as an approximation of (1), and this approximation idea can be formalized asymptotically, as long as we allow the number of factors $R = R_{NT}$ used in the least-squares estimator (5) to grow with N and T.

3.1 Consistency and convergence rate

From now on, we denote the true parameter β that generates the data by β^0. We rewrite model (1) as

$$Y_{it} = X_{it}'\,\beta^0 + \Gamma_{it} + \varepsilon_{it}, \tag{13}$$

where both Γ_{it} and ε_{it} are unobserved. Our main convergence rate results in Theorem 1 actually holds for any $N \times T$ matrix $\Gamma = (\Gamma_{it})$ that satisfies Assumption 4 below, but ultimately we are of course interested in the case $\Gamma_{it} = h(\alpha_i, \gamma_t)$. Arbitrary dependence between X_{it} and Γ_{it} is admitted such that there is a potential endogeneity problem.

Remember that the components of the K-vector X_{it} are denoted by $X_{it,k}$, $k = 1, \ldots, K$. Let $X_k = (X_{it,k})$ and $\varepsilon = (\varepsilon_{it})$ be $N \times T$ matrices. For a matrix A we denote r'th largest singular value by $\sigma_r(A)$, that is, $\sigma_r^2(A)$ is equal to the r'th largest eigenvalue of AA'. Furthermore, for matrices we denote the spectral norm by $\|\cdot\|$, and for vectors the norm $\|\cdot\|$ denotes the Euclidean norm. We impose the following assumptions.

Assumption 1 (Bounded norms of X_k and ε).

(i) $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it,k})^2 = O_P(1), \quad$ *for* $k = 1, \ldots, K$.

(ii) $\|\varepsilon\| = O_P\left(\sqrt{\max\{N,T\}}\right)$.

Assumption 2 (**Weak Exogeneity of X_k**). $\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it,k}\varepsilon_{it} = O_P(\sqrt{NT})$, *for* $k = 1, \ldots, K$.

Assumption 3 (**Non-Collinearity of X_k**). *Consider linear combinations* $\delta \cdot X := \sum_{k=1}^{K} \delta_k X_k$ *of the regressors* X_k *with vectors* $\delta \in \mathbb{R}^K$ *such that* $\|\delta\| = 1$. *Assume that there exists a constant* $b > 0$ *such that*

$$\min_{\{\delta\in\mathbb{R}^K,\, \|\delta\|=1\}} \sum_{r=2R_{NT}+1}^{\min(N,T)} \sigma_r^2\left[\frac{(\delta \cdot X)}{\sqrt{NT}}\right] \geq b\,, \qquad \textit{wpa1.}$$

Assumption 4 (**Singular value decay**). *There exists a constant* $\rho > 3/2$ *such that*

$$\frac{1}{NT}\sum_{r=R_{NT}+1}^{\min(N,T)} \sigma_r^2(\Gamma) = O_P\left(R_{NT}^{1-2\rho}\right).$$

Here, $R = R_{NT}$ is the number of factors that is chosen in the computation of the least-squares estimator $\widehat{\beta}_{\mathrm{LS}}$ in (5). We require $R_{NT} \to \infty$ as $N, T \to \infty$ to obtain consistency of $\widehat{\beta}_{\mathrm{LS}}$.

Lemma 1 below justifies Assumption 4 for our main case of interest $\Gamma_{it} = h(\alpha_i, \gamma_t)$, and we therefore postpone the discussion of that assumption until that we discuss that lemma. The Assumptions 1-3 are very similar to the assumptions used in Bai (2009) and Moon and Weidner (2015) to show consistency of $\widehat{\beta}_{\mathrm{LS}}$,[4] and the following discussion of those assumptions will, accordingly, be brief.

Assumption 1(i) follows from Markov's inequality as long as the second moment of $X_{it,k}$ is uniformly bounded. Assumption 1(ii) follows, for example, from the inequality in Latala (2005) if ε_{it} has mean zero, uniformly bounded fourth moment, and is independent across i and t. However, the assumption still holds if ε_{it} is weakly correlated across i and over t, see Moon and Weidner (2015). Assumption 2 is satisfies as long as $X_{it}\varepsilon_{it}$ has zero mean, uniformly bounded second moment, and is weakly correlated across i and over t.

[4] Compared to the assumptions imposed in the consistency Theorem 4.1 of Moon and Weidner (2015), the only two differences are that we allow for R_{NT} to grow asymptotically, and that Assumption 1(i) requires a bound on the Frobenius norm $\|X_k\|_F := \left(\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it,k}^2\right)^2$ instead of a bound on the spectral norm $\|X_k\|$. Since $\|X_k\| \leq \|X_k\|_F$, our assumption here is technically stronger, but in practice, one will almost certainly justify any bound on $\|X_k\|$ by computing $\|X_k\|_F$ anyway.

To understand Assumption 3, notice first that for $R_{NT} = 0$ we have

$$\sum_{r=1}^{\min(N,T)} \sigma_r^2 \left[\frac{(\delta \cdot X)}{\sqrt{NT}} \right] = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\delta \cdot X)_{it}^2.$$

Thus, for $R_{NT} = 0$, the assumption is just a standard non-collinearity assumption on the regressors, which demands that every non-trivial linear combination $\delta \cdot X$ of the regressors has sufficient variation. Next, for $R_{NT} > 0$ we have

$$\sum_{r=2R_{NT}+1}^{\min(N,T)} \sigma_r^2 \left[\frac{(\delta \cdot X)}{\sqrt{NT}} \right] = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\delta \cdot X)_{it}^2 - \sum_{r=1}^{2R_{NT}} \sigma_r^2 \left[\frac{(\delta \cdot X)}{\sqrt{NT}} \right],$$

that is, the assumption demands that the variation in the linear combination $\delta \cdot X$ does not only come from the leading $2R_{NT}$ singular values of this linear combination.

Of course, if $\text{rank}(\delta \cdot X) \leq 2R_{NT}$, then for $r > R_{NT}$ all the singular values $\sigma_r(\delta \cdot X)$ are equal to zero and the assumption is violated. Thus, as necessary condition for Assumption 3 is that $\text{rank}(\delta \cdot X) > 2R_{NT}$, that is, any linear combination of the regressors needs to be a "high-rank matrix". For example, a constant regressor $X_{it,1} = 1$ violates this assumption (it constitutes a rank one matrix, which could be easily absorbed into the unobserved Γ_{it}), but if the regressors are drawn from a DGP with random variation across both i and t, then they typically have full rank. Again, we refer to the existing papers on the least-squares estimator with interactive fixed effects for further discussion of this generalized non-collinearity condition on the regressors.

Theorem 1 (Consistency of $\widehat{\beta}_{\text{LS}}$). *Let the Assumptions 1 – 4 hold, and furthermore assume that $R_{NT} = o(\min\{N,T\})$ as $N,T \to \infty$. Then we have*

$$\widehat{\beta}_{\text{LS}} - \beta^0 = O_P\Big(R_{NT}^{(3-2\rho)/2}\Big) + O_P\Big(R_{NT}\,(\min\{N,T\})^{-1/2}\Big). \tag{14}$$

Therefore, by choosing $R_{NT} \propto (\min\{N,T\})^{\frac{1}{2\rho-1}}$ we obtain that

$$\widehat{\beta}_{\text{LS}} - \beta^0 = O_P\Big(\min\{N,T\}^{\frac{3-2\rho}{2(2\rho-1)}}\Big).$$

Here, the first term on the right-hand side of (14) is due to the approximation error of Γ (which can have large rank) with only a finite number of factors (of rank only R_{NT}). Notice that Assumption 4 demands $\rho > 3/2$, which implies that the exponent $(3-2\rho)/2$ in that term is negative. Thus, this approximation error term is small if we choose R_{NT} to be large.

The second term on the right-hand side (14) reflects a potential endogeneity problem introduced by the least-squares estimation procedure. Namely, the estimated loadings $\widehat{\lambda}$ and factors and $\widehat{f}$ in (5) can be correlated with both the regressors X_k and the error term ε, and by controlling for such endogenous variables $\widehat{\lambda}$ and $\widehat{f}$ we therefore potentially reduce the convergence rate of the pooled estimator for β from $1/\sqrt{NT}$ to $R_{NT}\,(\min\{N,T\})^{-1/2}$.

In this first part of Theorem 1 we allow for general R_{NT}, subject only to the condition $R_{NT} = o(\min\{N,T\})$. However, we only obtain the desired consistency result $\widehat{\beta}_{\text{LS}} - \beta^0 = o_P(1)$ if we let $R_{NT} \to \infty$ (to guarantee $R_{NT}^{(3-2\rho)/2} \to 0$) such that $R_{NT} = o((\min\{N,T\})^{1/2})$ (to make sure the last term in (14) is $o_P(1)$). The second part of Theorem 1 then chooses the rate for R_{NT} that optimally balances the two terms on the right hand side of (14).

Even for very large values of ρ (corresponding to very smooth functions $h(\cdot,\cdot)$) we never obtain convergence rates for $\widehat{\beta}_{\text{LS}}$ faster than $(\min\{N,T\})^{-1/2}$. This, however, should not be surprising, since even in the standard interactive fixed effect model we do not obtain convergence rates for $\widehat{\beta}_{\text{LS}}$ faster than $(\min\{N,T\})^{-1/2}$ unless we impose additional assumptions. Specifically, in the interactive fixed effect model of Bai (2009) one also requires the so-called "strong factor assumption" to obtain a convergence rate of $\widehat{\beta}_{\text{LS}}$ faster than $(\min\{N,T\})^{-1/2}$. Indeed, the example in Section 4.3 of Moon and Weidner (2015) provides a concrete data generating process for which $\widehat{\beta}_{\text{LS}}$ converges only at the rate $(\min\{N,T\})^{-1/2}$, and not faster, because the "strong factor assumption" (with the correct number of factors) is not satisfied in that DGP.

In our setting, we cannot impose this "strong factor assumption", because, as explained in the introduction, the data generating process $\Gamma_{it} = h(\alpha_i, \gamma_t)$ typically generates an infinite sequence of factors of decreasing strength. Demanding all those factors in equation (3) to be strong factors therefore makes no sense. Deriving a convergence rate for $\widehat{\beta}_{\text{LS}}$ faster than $(\min\{N,T\})^{-1/2}$ in our model therefore appears to very challenging, to say the least. This if of course, the key motivation for why we also consider the two-way grouped fixed effect estimator in this paper.

Remark 1. *If we change Assumption 4 to*

$$\sigma_r(\Gamma) \;\leq\; c\sqrt{NT}\, r^{-\rho}, \tag{15}$$

for all $r \in \{R_{NT}+1, \ldots \min\{N,T\}\}$*, wpa1, and some constant* $c > 0$*, then the result in equation* (14) *of Theorem 1 can be improved to*

$$\widehat{\beta}_{\text{LS}} - \beta^0 = O_P\left(R_{NT}^{1-\rho}\right) + O_P\left(R_{NT}\,(\min\{N,T\})^{-1/2}\right),$$

and we can then obtain consistency of $\widehat{\beta}$ *under the weaker condition* $\rho > 1$*. The condition* (15) *implies Assumption 4, but not vice versa, because Assumption 4 is a condition on the sum of the squared singular values, not on each of the singular values separately. It turns out to be technically much easier to verify Assumption 4 than to verify* (15) *for our main case of interest* $\Gamma_{it} = h(\alpha_i, \gamma_t)$,[5] *as we do in Lemma 1 below. This explains why we have chosen that formulation of the assumption and theorem in our baseline presentation.*

[5]This is because not only the decay of $\sigma_r(\Gamma)$ as $r \to \infty$ needs to be controlled, but also the convergence rate of the expressions as $N, T \to \infty$.

Despite the technical subtleties explained in the preceding remark, one should still interpret Assumption 4 as imposing a particular decay rate for the singular values Γ, as in display (15) of the remark. Thus, the leading few singular value can have a magnitude of $\sqrt{NT}$, as would be the case under the "strong factor assumption" in the usual interactive fixed effects model of Bai (2009). However, as N, T, r all converge to infinity we require the $\sigma_r(\Gamma)$ to converge at the polynomial rate $r^{-\rho}$ in order to satisfy the summability condition in Assumption 4.

The results in this section so far have not made any use of the structure $\Gamma_{it} = h(\alpha_i, \gamma_t)$. Theorem 1 is applicable to any other data generating process for Γ that satisfies Assumption 4. A full-rank matrix Γ satisfying that assumption could, for example, also be generated by a dynamic factor model (see e.g. Forni, Hallin, Lippi and Reichlin 2000, 2005, Stock and Watson 2002).[6]

In the following we now focus exclusively on the case $\Gamma_{it} = h(\alpha_i, \gamma_t)$. The following lemma provides conditions on the function $h(\cdot, \cdot)$ that guarantee that Assumption 4 is satisfied.

Lemma 1. *Assume $\alpha_i \in \Omega_\alpha$ and $\gamma_t \in \Omega_\gamma$, and that $h : \Omega_\alpha \times \Omega_\gamma \to \mathbb{R}$ is p times continuously differentiable in both arguments, and the domains $\Omega_\alpha \subset \mathbb{R}^{n_\alpha}$ and $\Omega_\gamma \subset \mathbb{R}^{n_\gamma}$ are smooth and bounded. Then for $\Gamma_{it} = h(\alpha_i, \gamma_t)$ Assumption 4 is satisfied for $R_{NT} \to \infty$ with $\rho = \frac{p}{\min\{n_\alpha, n_\gamma\}}$.*

Here, we measure the smoothness of the function $h(\cdot, \cdot)$ by p, the number of times it is continuously differentiable. The decay rate ρ of the singular values of Γ then depends on this measure of smoothness and the dimensions n_α and n_γ of the arguments α_i and γ_t. The smoother the function $h(\cdot, \cdot)$, for fixed dimensions n_α and n_γ, the faster to the eigenvalues of Γ converge to zero.

The proof of Lemma 1 crucially relies on the functional singular value decomposition in (2) and results on the decay rate of the corresponding singular values in Griebel and Harbrecht (2014). The only technical contribution of the proof is then to properly relate those known results on the functional singular value to the matrix singular values of Γ.

Notice that Lemma 1 requires no assumptions on the data generating process of α_i and γ_t, apart from boundedness of the domains Ω_α and Ω_γ, which can always be achieved by a reparameterization. Thus, those nuisance parameters can be arbitrarily correlated with each other (across i and over t) and with the regressors $X_{it,k}$. This result is analogous to the consistency Theorem 4.1 for $\widehat{\beta}_{\text{LS}}$ in Moon and Weidner (2015), where also no assumptions on the interactive fixed effects are imposed at all, apart from $\text{rank}(\lambda f') \leq R$.

From Theorem 1 and Lemma 1 we have the following corollary.

[6] One can generate an infinite number of "static factors", as in (3), via a dynamic factor model with a finite number of dynamic factors.

Corollary 1. *Let Assumptions 1 – 3 and the assumption on $h(.,.)$ in Lemma 1 be satisfied with $p > 3\min\{n_\alpha, n_\gamma\}/2$, and also let $R_{NT} \to \infty$ such that $R_{NT}/(\min\{N,T\})^{1/2} \to 0$. Then we have*

$$\widehat{\beta}_{\mathrm{LS}} - \beta^0 = o_P(1).$$

This is our final consistency result for the least-squares estimator of Bai (2009) in a data generating process given by (1). The convergence rate of the estimator was already discussed after Theorem 1 above, in particular, the difficulty in showing a convergence rate faster than $(\min\{N,T\})^{1/2}$ in our setting.

3.2 Further discussion

Here, we want to present some further intuition on the formal results on $\widehat{\beta}_{\mathrm{LS}}$ presented above. The discussion in this subsection is purely heuristic and does not aim to provide any formal derivations.

Remember the functional singular value decomposition in equation (2) of the introduction, which we now write as $h(\alpha_i, \gamma_t) = \sum_{r=1}^{\infty} \lambda_{ir}^0 f_{tr}^0$. For the sake of the following discussion, suppose that variation from $h(\alpha_i, \gamma_t)$ dominates the variation in $X_{it}'\beta$ and ε_{it} for the leading R_{NT} principal components of the residuals $Y_{it} - X_{it}'\beta - \sum_{r=1}^{R} \lambda_{ir} f_{tr} = \sum_{r=1}^{\infty} \lambda_{ir}^0 f_{tr}^0 - X_{it}'(\beta - \beta^0) + \varepsilon_{it}$. In this "best case scenario", the estimated factors $\sum_{r=1}^{R} \lambda_{ir} f_{tr}$ in the definition of $\widehat{\beta}_{\mathrm{LS}}$ in (5) will coincide with the leading R_{NT} components $\sum_{r=1}^{R} \lambda_{ir}^0 f_{tr}^0$ of $h(\alpha_i, \gamma_t)$, and we then have

$$\widehat{\beta}_{\mathrm{LS}} - \beta^0 = \zeta_{NT} + \xi_{NT},$$

where

$$\zeta_{NT} = \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it}' X_{it} \right)^{-1} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it}' \varepsilon_{it}$$

$$\xi_{NT} = \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it}' X_{it} \right)^{-1} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it}' \sum_{r=R+1}^{\infty} \lambda_{ir}^0 f_{tr}^0.$$

Under standard regularity conditions we have $\sqrt{NT}\zeta_{NT} \Rightarrow \mathcal{N}(0, \Sigma)$, and under the assumptions in the last subsection we have $\xi_{NT} = O_P\left(R_{NT}^{(3-2\rho)/2}\right)$. In this "best-case scenario" we can therefore have $R_{NT} \to \infty$ quick enough such that $\xi_{NT} = o_P(1/\sqrt{NT})$.

However, this is not a realistic scenario for $R_{NT} \to \infty$, because as R_{NT} grows, eventually the singular values of ε_{it} will dominate those of $\sum_{r=R+1}^{\infty} \lambda_{ir}^0 f_{tr}^0$, and the factor projection method will just project out idiosyncratic noise, or even contributions from $X_{it}'(\widehat{\beta}_{\mathrm{LS}} - \beta^0)$. This implies

that the problematic variation associated with $\lambda_{ir}^0 f_{tr}^0$ for most singular values r remains. This explains why it is so difficult to show anything better than the convergence rate results in Theorem 1 for the estimator $\widehat{\beta}_{\mathrm{LS}}$ in our setting.

4 Asymptotic results for the group fixed-effect estimator

The main goal of this section is to derive asymptotic results for the estimator $\widehat{\beta}_{\mathrm{GS}}$ defined in (9), which is the sample-splitting version of the group fixed-effect estimator. But we are first going to discuss the initial group fixed-effect estimator $\widehat{\beta}_{\mathrm{G}}$ defined in (7) without sample-splitting. We will not actually derive convergence rate results for $\widehat{\beta}_{\mathrm{G}}$ itself, but the discussion of the approximation bias of $\widehat{\beta}_{\mathrm{G}}$ will be a very useful precursor of the results for $\widehat{\beta}_{\mathrm{GS}}$.

4.1 Results for $\widehat{\beta}_{\mathrm{G}}$

We can rewrite our estimating equation for the group fixed-effect estimator in (6) as

$$Y = X \cdot \beta + \delta\, D_\delta' + D_\nu\, \nu' + \varepsilon, \tag{16}$$

where δ and ν are the $N \times C$ and $T \times G$ matrices of nuisance parameters, while D_δ and D_ν are $T \times C$ and $N \times G$ are binary matrices in which each row contains a single one, indicating the group membership of the corresponding unit or time period, respectively. By standard partitioned regression results we can then rewrite the group fixed-effect estimator in (7) as

$$\widehat{\beta}_{\mathrm{G}} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}_{it}' \widetilde{X}_{it} \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}_{it}' \widetilde{Y}_{it}, \qquad \widetilde{X}_k = M_N X_k M_T, \qquad \widetilde{Y} = M_N Y M_T, \tag{17}$$

where $\widetilde{X}_{it} = \left(\widetilde{X}_{it,1}, \ldots, \widetilde{X}_{it,K} \right)$, $\widetilde{Y}_{it}$ and $\widetilde{X}_{it,k}$ are the entries of the $N \times T$ matrices $\widetilde{X}_k$ and $\widetilde{Y}$, respectively, and $M_N = \mathbb{I}_N - D_\nu (D_\nu' D_\nu)^{-1} D_\nu'$ and $M_T = \mathbb{I}_T - D_\delta (D_\delta' D_\delta)^{-1} D_\delta'$ are projection matrices of dimesion $N \times N$ and $T \times T$, respectively.

Using this representation of the group fixed-effect estimator and the model in (13) we obtain that

$$\widehat{\beta}_{\mathrm{G}} - \beta^0 = \phi_{NT} + \kappa_{NT}, \tag{18}$$

where

$$\phi_{NT} := \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}_{it}' \widetilde{X}_{it} \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}_{it}' \, \varepsilon_{it}, \quad \kappa_{NT} := \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}_{it}' \widetilde{X}_{it} \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}_{it}' \, \widetilde{\Gamma}_{it}, \tag{19}$$

with $\widetilde{\Gamma}$ defined analogously to $\widetilde{X}_k$ and $\widetilde{Y}$ in (17). In the definition of ϕ_{NT} we can equivalently write $\widetilde{\varepsilon}_{it}$ instead of ε_{it}, but since M_N and M_T are idempotent matrices, and $\widetilde{X}_{it}$ is already the

projected regressor, this does not matter. The same is true, of course, for $\widetilde{\Gamma}_{it}$ vs Γ_{it} in the definition of κ_{NT}. However, the expressions in (19) turn out to be convenient as written.

Here, κ_{NT} is the approximation error of having replaced the nonlinear specification $\Gamma_{it} = h(\alpha_i, \gamma_t)$ in our model in (1) by the much simpler additive specification $\delta_{i,c_t} + \nu_{t,g_i}$ in the estimation equation (6). To see this, we can use standard matrix inequalities to bound the Euclidian norm of κ_{NT} by

$$\|\kappa_{NT}\| \leq \left\| \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}_{it}' \widetilde{X}_{it} \right)^{-1} \right\| \left(\max_k \left\| \widetilde{X}_k \right\|_F \right) \left\| \widetilde{\Gamma} \right\|_F, \tag{20}$$

where $\|\cdot\|_F$ refers to the Frobenius norm. Due to the definition of M_N and M_T we have

$$\left\| \widetilde{\Gamma} \right\|_F^2 = \min_{\delta \in \mathbb{R}^{N \times C}} \min_{\nu \in \mathbb{R}^{T \times G}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[h(\alpha_i, \gamma_t) - \delta_{i,c_t} - \nu_{t,g_i} \right]^2. \tag{21}$$

The last two displays show that κ_{NT} is small whenever $h(\alpha_i, \gamma_t)$ can be well approximated by $\delta_{i,c_t} + \nu_{t,g_i}$. In equation (8) we already informally discussed the magnitude of this approximation error, and found that it is of order $\|\alpha_i - \overline{\alpha}_{g_i}\|^2 + \|\gamma_t - \overline{\gamma}_{c_t}\|^2$. We now want to provide a more formal discussion of this and show that κ_{NT} is asymptotically small under appropriate regularity conditions.

In Section 2.2.2 we described the clustering algorithms that delivers the group memberships g_i and c_t based on the initial estimates $\widehat{\lambda}_i^*$ and $\widehat{f}_t^*$. The goal of the clustering is to group units i with approximately the same value of α_i, and to group time periods t with approximately the same γ_t. It is therefore crucial that $\widehat{\lambda}_i^*$ and $\widehat{f}_t^*$ are good proxies for α_i and γ_t. Specifically, we require that there exist functions $\lambda^* : \mathcal{A} \to \mathbb{R}^{R_*}$ and $f^* : \mathcal{C} \to \mathbb{R}^{R_*}$ such that $\widehat{\lambda}_i^*$ and $\widehat{f}_t^*$ converge to the non-random limits $\lambda^*(\alpha_i)$ and $f^*(\gamma_t)$ as $N, T \to \infty$. The following assumption formalizes this and states all the regularity condition that we require on $h(\cdot,\cdot)$, $\lambda^*(\cdot)$, $f^*(\cdot)$, $\widehat{\lambda}_i^*$, $\widehat{f}_t^*$, and X_{it}.

Assumption 5. *There exists a sequence $\xi_{NT} > 0$ such that $\xi_{NT} \to 0$ as $N, T \to \infty$, and*

(i) The function $h(\cdot,\cdot)$ is at least twice continuously differentiable with uniformly bounded second derivatives.

(ii) Every unit i is a member of exactly one group $g_i \in \{1, \ldots, G\}$, and every time period t is a member of exactly one group $c_t \in \{1, \ldots, C\}$. The size of all G groups of units, and the size of all C groups of time periods is bounded uniformly by $Q_{\max}$.

(iii) There exists $B > 0$ such that $\|a - b\| \leq B \|\lambda^*(a) - \lambda^*(b)\|$ for all $a, b \in \mathcal{A}$, and $\|a - b\| \leq B \|f^*(a) - f^*(b)\|$ for all $a, b \in \mathcal{C}$, and the domains $\mathcal{A}$ and $\mathcal{C}$ are convex set.

(iv) $\frac{1}{N}\sum_{i=1}^{N}\left(\left\|\widehat{\lambda}_i^* - \lambda^*(\alpha_i)\right\|^2\right) = O_P\left(\xi_{NT}\right)$, $\frac{1}{T}\sum_{t=1}^{T}\left(\left\|\widehat{f}_t^* - f^*(\gamma_t)\right\|^2\right) = O_P\left(\xi_{NT}\right)$.

(v) $\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\lambda}_i^* - \widehat{\lambda}_{j(i)}^*\right\|^2 = O_P\left(\xi_{NT}\right)$ *for any matching function* $j(i) \in \{1,\ldots,N\}$ *such that* $g_i = g_{j(i)}$, *and* $\frac{1}{T}\sum_{t=1}^{T}\left\|\widehat{f}_t^* - \widehat{f}_{s(t)}^*\right\|^2 = O_P\left(\xi_{NT}\right)$ *for any matching function* $s(t) \in \{1,\ldots,T\}$ *such that* $c_t = c_{s(t)}$.

(vi) $\max_{k,i,t}\left|\widetilde{X}_{it,k}\right| = O_P(1)$, *and* $\operatorname{plim}_{N,T\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}_{it}'\widetilde{X}_{it} = \Omega$, *where* Ω *is a positive definite non-random matrix.*

Lemma 2. *Under Assumption 5 we have*

$$\kappa_{NT} = O_P(\xi_{NT})$$

The lemma shows that the approximation error κ_{NT} vanishes at the rate ξ_{NT} as $N, T \to \infty$. The assumptions and the lemma are formulated for arbitrary rates, but as will become clear from the following discussion, the best we can achieve in our setting is a rate of $\xi_{NT} = 1/\min(N,T)$, which coincides with $\xi_{NT} = 1/\sqrt{NT}$ in the special case that N and T grow at the same rate.

Part (i) of Assumption 5 requires the function $h(\cdot,\cdot)$ to be sufficiently smooth. This condition should not be surprising, because our informal discussion of the approximation error in equation (8) already relies on a second order Taylor expansion of $h(\cdot,\cdot)$, and the proof of Lemma 2 is based on exactly such an expansion.

Part (iii) and (iv) of the assumption are analogous to "Assumption 2 (injective moments)" in Bonhomme, Lamadon and Manresa (2021), except that they consider a one-way fixed effect setting while we consider a two-way fixed effect setting. Part (iii) requires the functions $\lambda^*(\cdot)$ and $f^*(\cdot)$ to be injective, that is, α_i and γ_t can be uniquely recovered from knowing $\lambda^*(\alpha_i)$ and $f^*(\gamma_t)$. A necessary condition for this is that

$$R^* \geq \max(d_\alpha, d_\gamma), \tag{22}$$

where d_α and d_γ are the dimensions of α_i and γ_t, respectively. Part (iv) requires the estimates $\widehat{\lambda}_i^*$ and $\widehat{f}_t^*$ to converge to $\lambda^*(\alpha_i)$ and $f^*(\gamma_t)$ at the average rate of $\xi_{NT}^{1/2}$. Since T observations are available for unit i we expect that $\widehat{\lambda}_i^*$ converges at a rate of $T^{1/2}$, and since N observations are available for time period t we expect that $\widehat{f}_t^*$ converges at a rate of $N^{1/2}$, see, for example, Theorem 1 and 2 in Bai (2003). This explains why $\xi_{NT} = 1/\min(N,T)$ is the best rate we can achieve here.

Part (v) of Assumption 5 is a high-level assumption on the clustering mechanism used to obtain the group memberships g_i and c_t. For units i and j in the same group, and for time periods t and s in the same group, we demand the average differences $\widehat{\lambda}_i^* - \widehat{\lambda}_j^*$ and $\widehat{f}_t^* - \widehat{f}_s^*$ to

be small as $N, T \to \infty$. In other words, we require that the clustering mechanism does what it is intended to do, namely forming groups such that the estimates $\widehat{\lambda}_i^*$ and $\widehat{f}_t^*$ for units i and time periods t in the same group are close to each other. For a given clutstering algorithms (e.g. the one describe in Section 2.2.2) one could prove that this assumption holds under further regularity conditions on the distribution of α_i and γ_t, see, for example, Lemma 1 in Bonhomme, Lamadon and Manresa 2021. In particular, a necessary condition for part (v) of Assumption 5 to hold is the following:

Regularity condition. $\frac{1}{N}\sum_{i=1}^{N}\left\|\alpha_i - \alpha_{j(i)}\right\|^2 = O_P\left(\xi_{NT}\right)$ *for any matching function* $j(i) \in \{1,\dots,N\}$ *such that* $g_i = g_{j(i)}$*, and* $\frac{1}{T}\sum_{t=1}^{T}\left\|\gamma_t - \gamma_{s(t)}\right\|^2 = O_P\left(\xi_{NT}\right)$ *for any matching function* $s(t) \in \{1,\dots,T\}$ *such that* $c_t = c_{s(t)}$.

This condition coincides with Assumption 5(v) in the unrealistic case that $\widehat{\lambda}_i^* = \alpha_i$ and $\widehat{f}_t^* = \gamma_t$. Starting from this unrealistic case and then applying the transformations $\lambda^* : \mathcal{A} \to \mathbb{R}^{R_*}$ and $f^* : \mathcal{C} \to \mathbb{R}^{R_*}$ and adding noise to the estimates then gives part (v) of Assumption 5. Crucially, for this regularity condition to hold, we need that $\xi_{NT} \gtrsim 1/\min(N^{2/d_\alpha}, T^{2/d_\gamma})$, see Lemma 2 in Bonhomme, Lamadon and Manresa (2021) for the analogous results in a one-way fixed effect model (also Graf and Luschgy 2002). Since our actual clustering method is not based on the unobserved α_i and γ_t, but on $\widehat{\lambda}_i^*$ and $\widehat{f}_t^*$ we require the stronger condition (in view of (22)) that

$$\xi_{NT} \gtrsim [\min(N,T)]^{-2/R^*}.$$

This is a necessary condition for Assumption 5(v) to be satisfied.[7] Therefore, if we want to achieve the best possible rate $\xi_{NT} = 1/\min(N,T)$, then we need $R^* \leq 2$, which according to (22) implies that $d_\alpha \leq 2$ and $d_\gamma \leq 2$. This discussion shows that our group fixed-effect estimator $\widehat{\beta}_{\mathrm{G}}$ suffers from a curse of dimensionality with regards to the dimensions of α_i and γ_t. However, this should be unsurprising, given the semi-parametric nature of the estimation problem – with non-parametric component $h(\alpha_i, \gamma_t)$. This also shows that there is a tradeoff between the LS estimator analyzed in Section 3 and the group fixed effects estimator discussed here – we will further compare those two estimators in our MC analysis below.

Finally, part (vi) of Assumption 5 requires some regularity conditions on the projected regressors $\widetilde{X}_k = M_N X_k M_T$ defined in (17).

This concludes our discussion of the approximation error κ_{NT}. We have argued that, under appropriate regularity conditions, including $\max(d_\alpha, d_\gamma) \leq 2$, we can use Lemma 2 to obtain

[7] Following the logic in Bonhomme, Lamadon and Manresa (2021) we believe that we actually only need $\xi_{NT} \gtrsim 1/\min(N^{2/d_\alpha}, T^{2/d_\gamma})$, that is, our group fixed effect estimator $\widehat{\beta}_{\mathrm{G}}$ truly cannot achieve a convergence rate faster than $1/\min(N^{2/d_\alpha}, T^{2/d_\gamma})$. Thus, if $R^* > \max(d_\alpha, d_\gamma)$, then $\xi_{NT} \gtrsim [\min(N,T)]^{-2/R^*}$ is probably not a necessary condition for the result of Lemma 2 itself, but only for our Assumption 5(v).

$\kappa_{NT} = 1/\sqrt{NT}$, for N and T growing to infinity at the same rate. Since $\widehat{\beta}_{\mathrm{G}} - \beta^0 = \phi_{NT} + \kappa_{NT}$ we could then conclude that $\widehat{\beta}_{\mathrm{G}} - \beta^0 = O_P(1/\sqrt{NT})$, if we could also show that $\phi_{NT} = O_P(1/\sqrt{NT})$.

From the definition of ϕ_{NT} in (19) one might think that it is easy to derive this results on ϕ_{NT} by imposing an approximate exogeneity condition on the regressors. However, the problem is that $\widetilde{X}_k$ depends on the group assignments of units i and time periods t, which were constructed based on $\widehat{\lambda}^*$ and $\widehat{f}^*$, which depend on the errors ε. Thus, $\widetilde{X}_k$ depends on ε in complicated ways through the group assignment, making a proof of $\phi_{NT} = O_P(1/\sqrt{NT})$ technically challenging. In principle, we expect that

$$\sqrt{NT}\,\phi_{NT} \Rightarrow \mathcal{N}(0, \Sigma_{\mathrm{G}}) \tag{23}$$

holds for an appropriate covariance matrix Σ_{G}, and our simulations evidence suggest that this is indeed the case. However, we are not aiming to prove this result in this paper. As explained already in Section 2, this technical difficulty in analyzing $\widehat{\beta}_{\mathrm{G}}$ is exactly why we introduced the split-sample version of the group fixed-effect estimator, for which we are going to derive results in the following.

4.2 Results for $\widehat{\beta}_{\mathrm{GS}}$

The split-sample version of the group fixed effect estimator was introduced in Section 2.2.3 above. Using the Frisch-Waugh-Lovell theorem we can rewrite $\widehat{\beta}_{\mathrm{GS}}$ in equation (9) as follows:

$$\widehat{\beta}_{\mathrm{GS}} = \left(\sum_{s=1}^{4} \sum_{(i,t)\in\mathcal{O}_s} \widetilde{X}_{it}^{(s)\prime} \widetilde{X}_{it}^{(s)} \right)^{-1} \sum_{s=1}^{4} \sum_{(i,t)\in\mathcal{O}_s} \widetilde{X}_{it}^{(s)\prime} Y_{it},$$

where the projected regressors $\widetilde{X}_{it}^{(s)} = \left(\widetilde{X}_{it,1}^{(s)}, \ldots, \widetilde{X}_{it,K}^{(s)} \right)'$ for each subpanel $s \in \{1,2,3,4\}$, each regressor $k = 1, \ldots, K$, and observations $(i,t) \in \mathcal{O}_s$ within that subpanel, are the residuals of the least-squares problem

$$\min_{\delta} \min_{\nu} \sum_{(i,t)\in\mathcal{O}_s} \left(X_{it,k} - \delta_{i,c_t^{(s)}} - \nu_{t,g_i^{(s)}} \right)^2. \tag{24}$$

Following the decomposition of $\widehat{\beta}_{\mathrm{G}}$ in (18), we can now introduce the analogous decomposition for $\widehat{\beta}_{\mathrm{GS}}$ by

$$\widehat{\beta}_{\mathrm{GS}} - \beta^0 = \phi_{NT}^{(\mathrm{GS})} + \kappa_{NT}^{(\mathrm{GS})}, \tag{25}$$

where

$$\phi_{NT}^{(\mathrm{GS})} := \left(\sum_{s=1}^{4} \sum_{(i,t)\in\mathcal{O}_s} \widetilde{X}_{it}^{(s)\prime} \widetilde{X}_{it}^{(s)} \right)^{-1} \sum_{s=1}^{4} \sum_{(i,t)\in\mathcal{O}_s} \widetilde{X}_{it}^{(s)\prime} \varepsilon_{it},$$

$$\kappa_{NT}^{(\mathrm{GS})} := \left(\sum_{s=1}^{4} \sum_{(i,t)\in\mathcal{O}_s} \widetilde{X}_{it}^{(s)\prime} \widetilde{X}_{it}^{(s)} \right)^{-1} \sum_{s=1}^{4} \sum_{(i,t)\in\mathcal{O}_s} \widetilde{X}_{it}^{(s)\prime} \widetilde{\Gamma}_{it}^{(s)},$$

Here, $\phi_{NT}^{(\mathrm{GS})}$ is a variance term that we will show to be unbiased and asymptotically normal, and $\kappa_{NT}^{(\mathrm{GS})}$ is the approximation error from having replaced $h(\alpha_i, \gamma_t)$ by the linear grouped fixed effect in the estimation for $\widehat{\beta}_{\mathrm{GS}}$ in (9). The $\widetilde{\Gamma}_{it}^{(s)}$ are the residuals of the least-squares problem (24) when $X_{it,k}$ is replaced by $\Gamma_{it} = h(\alpha_i, \gamma_t)$.

For each of the four subpaneles $s \in \{1, 2, 3, 4\}$, the discussion of the approximation error $\kappa_{NT}^{(\mathrm{GS})}$ is identical to the discussion of the approximation error κ_{NT} of $\widehat{\beta}_{\mathrm{G}}$, see, in particular, the bounds (20) and (21) above. It is therefore straightforward to obtain the analogue of Lemma 2 for the approximation error of the split-sample estimator.

Lemma 3. *Under Assumption A.1 (in appendix) we have*

$$\kappa_{NT}^{(\mathrm{GS})} = O_P(\xi_{NT})$$

Assumption A.1 is stated in the appendix, but it is simply a restatement of Assumption 5 for each subpanel $s \in \{1, 2, 3, 4\}$. Those assumptions were discussed after Lemma 2 above. In particular, the best possible convergence rate we can hope for here is $\xi_{NT} = 1/\min(N, T)$, but that rate is only attainable for $d_\alpha \leq 2$ and $d_\gamma \leq 2$.

The key difference between $\widehat{\beta}_{\mathrm{G}}$ and $\widehat{\beta}_{\mathrm{GS}}$ is that for the split-sample estimator we can derive the asymptotic behavior of the variance term very easily $\phi_{NT}^{(\mathrm{GS})}$. For this purpose, we impose the following assumptions.

Assumption 6.

(i) Conditional on X, α, γ, we assume that ε_{it} is independently distributed across i and over t, such that $\sigma_{it}^2 := \mathbb{E}\left(\varepsilon_{it}^2 \,\middle|\, X, \alpha, \gamma\right) \leq B < \infty$, for some constant B that is independent of i, t, N, T.

(ii) We have $\operatorname{plim}_{N,T\to\infty} \frac{1}{NT} \sum_{s=1}^{4} \sum_{(i,t)\in\mathcal{O}_s} \widetilde{X}_{it}^{(s)\prime} \widetilde{X}_{it}^{(s)} = \Omega > 0$, and for each $s \in \{1, \ldots, S\}$ we have $\operatorname{plim}_{N,T\to\infty} \frac{1}{NT} \sum_{(i,t)\in\mathcal{O}_s} \sigma_{it}^2 \widetilde{X}_{it}^{(s)\prime} \widetilde{X}_{it}^{(s)} = \Sigma^{(s)}$. Furthermore, we assume that, for $s \in \{1, 2, 3, 4\}$, all the third-order sample moments of $\widetilde{X}_{it}^{(s)\prime} \varepsilon_{it}$ across $(i, t) \in \mathcal{O}_s$ are bounded as $N, T \to \infty$.

Assumption 6 together with the sample splitting method used to construct $\widehat{\beta}_{\mathrm{GS}}$ guarantees that, within each subpabel $s \in \{1,2,3,4\}$, the $\widetilde{X}_{it}^{(s)\prime}\varepsilon_{it}$ are zero mean and independently distributed across $\sum_{(i,t)\in\mathcal{O}_s}$. Here, the split-panel construction is crucial, since it guarantees that $\widetilde{X}_{it}^{(s)}$ is independent of ε_{it}. The remaining conditions in Assumption 6 are regularity conditions to allow us to apply the Lyapunov central limit theorem for each subpanel and to guarantee that $\phi_{NT}^{(\mathrm{GS})}$ has a finite asymptotic variance. We therefore obtain the following lemma.

Lemma 4. *Under Assumption 6 we have, as $N,T \to \infty$,*

$$\sqrt{NT}\,\phi_{NT}^{(\mathrm{GS})} \Rightarrow \mathcal{N}(0,\Sigma_{\mathrm{GS}}), \qquad \Sigma_{\mathrm{GS}} = \Omega^{-1}\left(\sum_{s=1}^{4}\Sigma^{(s)}\right)\Omega^{-1}.$$

Combining equation (25) with Lemma 3 and Lemma 4 then gives the following theorem.

Theorem 2. *Under Assumption 6 and Assumption A.1 we have*

$$\widehat{\beta}_{\mathrm{GS}} - \beta^0 = O_P\left(\frac{1}{\sqrt{NT}} + \xi_{NT}\right) = o_P(1).$$

Analogous to Corollary 1 for the least-squared estimator of Bai (2009), we have this obtained a consistency result for $\widehat{\beta}_{\mathrm{GS}}$ as well. We have not derived asymptotic inference results using either of these estimators, but in the following section we explain how we use those estimators to construct confidence intervals in our simulations and empirical application.

5 Implementation

The asymptotic results derived for $\widehat{\beta}_{\mathrm{LS}}$, $\widehat{\beta}_{\mathrm{G}}$, and $\widehat{\beta}_{\mathrm{SG}}$ in the last two sections are insightful for how those estimates should be used in practice. In particular, our discussions and derivations are helpful to appreciate the limitations and assumptions needed for the estimation approaches, and we will summarize those again in our conclusion section below.

In the following Monte Carlo simulations and empirical application we will employ the estimates $\widehat{\beta}_{\mathrm{LS}}$, $\widehat{\beta}_{\mathrm{G}}$, and $\widehat{\beta}_{\mathrm{SG}}$ in a way that goes beyond our formal asymptotic results. In particular, we will use all those estimators to construct confidence intervals and we will also apply Jackknife methods for bias correction. In this section, we want to briefly explain how those confidence intervals and bias corrected estimates are constructed.

To formulate standard errors for each estimator we initially ignore the approximation error discussed in our formal results and assume the residual terms follow the asymptotic behaviour of independent and identically distributed zero mean disturbance. For example, in section 4.1 where we split the residual term into ϕ and κ, we will ignore the κ term and estimate standard errors as if we are left with only ϕ. We use the jackknife corrections to address the residual terms related to approximation error in both the factor and grouped fixed-effects estimation models.

For factor model standard errors we construct the heteroscedasticity-consistent estimator from White (1980) as follows. Take $\Omega = \sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}'_{it}\widetilde{X}_{it}$ and $\widehat{\Sigma} = \sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{u}^2_{it}\widetilde{X}'_{it}\widetilde{X}_{it}$ where $\widehat{u}_{it} = \widetilde{Y}_{it} - \sum_k \widehat{\beta}_{LS,k}\widetilde{X}_{it,k}$ and for a matrix A, in this context, $\widetilde{A}$ represents the matrix with factors projected. We must make a degrees of freedom correction for the factor projection by the ratio $dfc = \sqrt{\frac{NT}{(N-R)(T-R)}}$. Then the vector of standard errors are,

$$\text{se}(\widehat{\beta}_{\text{LS}}) = dfc \cdot \sqrt{\text{diag}\left(\Omega^{-1}\widehat{\Sigma}\Omega^{-1}\right)}.$$

As above, we use this same standard error estimator for jackknife corrected estimates.

For the grouped fixed-effects models we use clustered standard errors where clusters are taken as the combination of i and t clusters. That is, for the matrices of clusters D_α and D_γ for i and t respectively we take clusters as the Kronecker product between these two matrices, $D_\alpha \otimes D_\gamma$. Remember here that the columns of D_α, resp. D_γ, are the cluster assignments of i, resp. t with a 1 entry if that observation is in the cluster and a 0 otherwise. Take m as the index for cluster assignment with $M = GC$ the total number of clusters. Hence, $D_\alpha \otimes D_\gamma := \mathcal{D}$ is an NT by M matrix with $\mathcal{D}_m$ representing a column of this matrix and $\mathcal{D}_{n,m}$ representing an entry. A combination (i,t) can be can be identified by the row, n, of the matrix $\mathcal{D}$ as $t = \lceil n/N \rceil$ and $i = n - (\lceil n/N \rceil - 1)\, N$, which is similar to the usual matrix flattening procedure. Then, the column-vector $\mathcal{D}_m$ consists of a 1 if the (i,t) combination implied by that row, n, is in that column's cluster and 0 otherwise.

Define as above $\Omega = \sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}'_{it}\widetilde{X}_{it}$ and $\widehat{u}_{it} = \widetilde{Y}_{it} - \sum_k \widehat{\beta}_{G,k}\widetilde{X}_{it,k}$ where in this context for matrix A, the matrix $\widetilde{A}$ represents the matrix with group fixed-effects projected out. Call the index function $n(i,t) = i + (t-1)N$, such that $\mathcal{D}_{n(i,t),m}$ returns the binary indicator of whether (i,t) is in the m^{th} combination cluster. Now define $\widehat{\Sigma} = \sum_{m=1}^{M}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathcal{D}_{n(i,t),m}\widehat{u}^2_{it}\widetilde{X}'_{it}\widetilde{X}_{it}$. This collapses the familiar block-diagonal matrix where values within each block corresponds to a combination cluster and are unrestricted but zero outside each block. The clustered standard errors can thus be defined as

$$\text{se}(\widehat{\beta}_G) = dfc \cdot \sqrt{\text{diag}\left(\Omega^{-1}\widehat{\Sigma}\Omega^{-1}\right)}$$

where in this context $dfc = \sqrt{\frac{NT}{(N-G)(T-C)}}$. The standard error estimator is identical for the split sample version except there are many more combination clusters by the nature of this split sample estimators clustering method.

Finally, we also apply Jackknife bias corrections to the estimators to reduce both the approximation bias and the incidental parameter bias of the various estimates. We follow Fernández-Val and Weidner (2016) to estimate the jackknife bias corrected analog to each estimator as follows. This procedure is closely related to Dhaene and Jochmans (2015). First, split the sample along

the i dimension into two $N/2$ by T samples. For each of these samples run and call the related estimates from estimator E, $\widehat{\beta}_E^{1,1}$ and $\widehat{\beta}_E^{1,2}$, respectively. Repeat this process along the t dimension to return $\widehat{\beta}_E^{2,1}$ and $\widehat{\beta}_E^{2,2}$. Then the final jackknife bias corrected analog for estimator E is

$$\widehat{\beta}_{E,JK} = 3\widehat{\beta}_E - \frac{1}{2}\left(\widehat{\beta}_E^{1,1} + \widehat{\beta}_E^{1,2}\right) - \frac{1}{2}\left(\widehat{\beta}_E^{2,1} + \widehat{\beta}_E^{2,2}\right),$$

where $\widehat{\beta}_E$ is simply the estimate without any sample split. We maintain the assumption that standard errors are the same across split samples so we can simply take the standard error estimate from the whole sample.

6 Monte Carlo simulations

For our Monte Carlo simulations, we choose a data generating process with a single regressor ($K = 1$), and we generate outcomes and regressor as follows:

$$\begin{aligned} Y_{it} &= X_{it}\beta + h(\alpha_i, \gamma_t) + \varepsilon_{it}, \\ X_{it} &= g(\alpha_i, \gamma_t) + \mu_{it}, \end{aligned} \tag{26}$$

with

$$\varepsilon_{it},\ \alpha_i,\ \gamma_t,\ \mu_{it} \sim \text{all mutually independent and i.i.d.} \mathcal{N}(0,1) \tag{27}$$

This setting assumes that the endogeneity in X_{it} depends on the specification of $g(.,.)$ vis–à–vis $h(.,.)$. The decay in singular values for either the unobserved term in Y_{it} or for X_{it} can be directly manipulated through the specification of $h(.,.)$ and $g(.,.)$, which will dictate the number of significant factors in each decomposition.

We set $\beta = 1$ and,

$$h(a,b) = g(a,b) = \frac{1}{\sqrt{2\pi}\theta}\exp\left(\frac{(a-b)^2}{\theta^2}\right), \quad \theta = (1/2)^3. \tag{28}$$

The θ value here dictates the speed of decay in singular values for $h(.,.)$ and $g(.,.)$, holding fixed the variation in their arguments, where a lower value implies a slower decay. This particular value for θ was chosen as it implies a slow decay in singular values such that the endogenous component of the unobserved term and X persists even as many factors are included. The value for θ carries no fundamental economic meaning. Note, the nature of bias in this simulation is by design monotonic and positive for illustrative purposes.

Table 2 below shows the results from 10,000 Monte Carlo simulations. These results display our theoretical result on bias reduction succinctly. We see that as we increase the number of

factors the average bias reduces and the standard deviation of estimates increases. We also see a significant improvement in bias using the grouped fixed-effects estimator, without a large increase in standard deviation. The GFE split sample estimator performs much worse in terms of bias, which is expected given the significantly smaller candidate pool for clustering in this estimator. The jackknife analog to each estimator reduces bias in all cases except the factor model with 5 factors, but significantly increases standard deviation in all cases.

If we compare the mean standard error estimates to standard deviation across simulations we see evidence that the standard error calculation may underestimate the true standard error of the estimator. In light of discussion in Section 5, we explicitly ignore fixed-effects approximation error and assumed only a noise term remains when estimating standard errors, which may explain this discrepancy. The divergence between estimated standard errors and standard deviation across simulations is particularly noticeable for the factor model with a large number of factors and for jackknife bias corrected estimators. For large factor models it is likely our inference approach misses out dependence structures introduced by the factor projection. This divergence is less pronounced for the group fixed-effects estimator without bias correction. Since we do not expressly advocate a particular inference approach for any estimator used in this paper we do not discuss this issue any further and leave it for future research.

Table 2: Monte Carlo simulations

	Mean Bias	Stand. Dev.	Mean $\widehat{se}$	% Coverage
LS factor model (5 factors)	0.0843	0.0122	0.0105	0.00
LS factor model (20 factors)	0.0162	0.0151	0.0111	0.65
LS factor model (50 factors)	0.0124	0.0306	0.0137	0.56
LS factor model Jackknife (5 factors)	-0.5307	0.0373	0.0105	0.00
LS factor model Jackknife (20 factors)	-0.0100	0.0286	0.0111	0.51
LS factor model Jackknife (50 factors)	-0.0066	0.0752	0.0137	0.27
Group fixed-effects	0.0019	0.0189	0.0179	0.93
GFE jackknife	-0.0005	0.0343	0.0179	0.70
GFE splits	0.0205	0.0188	0.0126	0.57

N = T = 100 with 10,000 repetitions.

Mean bias is simply the mean of the bias across simulations. Standard deviation is the standard deviation of the estimates, again across simulations. Mean $\widehat{se}$ is the mean across simulations of the standard error estimate. Coverage is defined here as the percentage of estimates that contain the true β in its 95% confidence intervals implies by its $\widehat{se}$.

7 Empirical application

We apply our estimation procedure to an analysis of the UK housing market, following Giglio, Maggiori and Stroebel (2016) (GMS16). Specifically, we study the effects of extremely long lease agreements on the price of housing, when compared to freehold agreements. In the UK housing market it is common for real estate property to be sold under each agreement. GMS16 posit that any change in price due to exogenous variation in whether the property was sold under extremely long lease or freehold must be attributed to so–called "housing bubbles associated with a failure of the transversality condition". The empirical challenge in making this comparison, and much discussed in GMS16, is to sufficiently control for observable and unobservable covariates such that variation in the variable of interest can be reasonably described as exogenous.

In the following, we compare estimates using our method with the more flexible approach taken in their paper. We note first that given differences in data, these results should not be directly compared with GMS16. Rather, this should be seen as an internal validity check across estimation models, i.e., to check if the aggregated setting produce similar estimates to the granular setting from GMS16 within the same set of data.

Consider the granular model from GMS16

$$Y_{iprt} = ExtremelyLongLease_i\beta + controls'_{it}\delta + \phi_{prt} + \varepsilon_{iprt} \tag{29}$$

where i are individual *transactions* (i.e. not necessarily properties), p is property type, r are regions and t is the month of transaction. Controls include hedonic variables, e.g. number of bedrooms, bathrooms and floorspace. ϕ_{prt} is a scalar fixed effect particular to the region, property type and month, and is identified via variation across transactions i. Compare this to an aggregated setting,

$$Y_{rt} = ExtremelyLongLease_{rt}\beta + controls'_{rt}\delta + h(\alpha_r, \gamma_t) + \varepsilon_{rt} \tag{30}$$

where Y_{rt}, $ExtremelyLongLease_{rt}$ and $controls_{rt}$ are the sample means aggregated to the region and transaction month. The multidimensional array with entries ϕ_{prt} varies with higher rank than the matrix with entries $h(\alpha_r, \gamma_t)$ because the latter is constant across p if extended to the equivalent multidimensional array with dimensions across (p, r, t). This is why we believe the model in (29) will better capture fixed-effects.

For purposes of this exercise we take the granular model with fixed-effects below as being, in theory, the better model to approximate unobserved heterogeneity. Hence we refer to this as the benchmark model. We use this benchmark approach to understand how well each estimator performs in practical instances where granular levels of aggregation are not always available, for example when data is aggregated for privacy reasons or for other feasibility reason. Hence,

estimates close to the granular model estimates should be seen as performing "well" in this setting.

Table 3 shows that when we control for fixed effects in the granular model there is a 0.3% reduction in price when a long leasehold transaction is made compared to a freehold. Whilst this is statistically significant, it translates to a decrease in the median house price of less than £1,000 so is arguably a small reduction economically. The OLS estimates do not change much across the different aggregation schemes and perhaps unsurprisingly the panel aggregated OLS has a much higher standard deviation due to the lower effective sample size. In the panel setting the factor model shows a convergence to the granular model with fixed effects as factors are increased and, interestingly, also to the grouped fixed-effects estimate, which is the closest to the benchmark estimates.[8] These results show a similar pattern to the simulation exercise where, according to the benchmark model, we see a bias reduction as the number of factors increases and when using the group fixed-effects estimator.

Table 3: Empirical Results

	Model	Estimate	Standard Errors
Granular Model	Ordinary Least Squares	0.203	0.0054
(29)	with Fixed Effects	-0.003	0.0006
Panel Model	Ordinary Least Squares	0.229	0.106
(30)	LS factor model (5 factors)	0.024	0.012
	LS factor model (15 factors)	0.007	0.007
	LS factor model (30 factors)	0.007	0.008
	Group fixed-effects	0.006	0.020

UK housing market results for N = 2088 and T = 48.

8 Conclusions

Panel regressions are very popular estimation tools, because they allow to control for omitted variables that are unobserved and potentially correlated with the observed covariates. Both Pesaran (2006) and Bai (2009), and most of the literature following those seminal papers, assume that those unobserved omitted variables take the form of a low-rank matrix, which can be interpreted as a static factor model or interactive fixed effects. In this paper, we deviate from this interactive fixed effect model by assuming that the unobserved omitted variables enter the

[8] In Table 3, our usual computation for the clustered standard errors of the group fixed-effect estimator was infeasible here due to the sample size. These standard error estimates are generated by resampling region clusters with replacement over 10,000 resamples.

model in the more general form $h(\alpha_i, \gamma_t)$, where $h(\cdot, \cdot)$ is an unknown smooth function, and α_i and γ_t are (multidimensional) fixed effects that can be arbitrarily correlated across i, over t, and with the observed covariates.

We first explore the behavior of Bai's least-squares esimator in this new setting. We show that this LS esimator estimator is still consistent, as long as the number of factors used in the estimation is allowed to grow asymptotically. However, as explained in detail in Section 3, it seems impossible to derive convergence rates faster than $(\min\{N, T\})^{1/2}$ for this estimator in our setting.

We therefore develop a new estimation approach called the two-way grouped fixed effects approach, which generalize ideas in Bonhomme, Lamadon and Manresa (2021) to our two-way setting. We derive convergence rate results for the resulting new estimators and show that, depending on the dimension of α_i and γ_t, and the relative size of N and T, convergence rates up to $\sqrt{NT}$ can be achieved with our new estimation approach.

We also explore the performance of those various estimators in simulations and in an empirical application. We find that both Bai's least-squares esimator and our grouped fixed effect estimators tend to perform well in practice. Interestingly, the theoretical convergence rate of $(\min\{N, T\})^{1/2}$ for the LS esimator may often understate the performance of this esimator in practice.

We also find that Jackknife bias correction helps to further reduce the bias of the various estimators, but at the cost of increasing the variance. Overall, the (Jackknife corrected) group fixed-effects estimator tends to have the smallest bias, but not necessarily the smallest variance. The empirical application shows that, according to our benchmark estimation, the LS estimation approach improves with more factors and that the group fixed-effects estimator does indeed provide a bias reduction compared to the LS estimator.

In the simulation exercise and empirical application we implemented standard error calculations for each estimator, but we leave formal inference results in the setting of our paper as an open question for future research.

References

Ahn, S. C., Y. H. Lee, and P. Schmidt (2001, April). GMM estimation of linear panel data models with time-varying individual effects. *Journal of Econometrics 101*(2), 219–255.

Ahn, S. C., Y. H. Lee, and P. Schmidt (2013). Panel data models with multiple time-varying individual effects. *Journal of Econometrics 174*(1), 1–14.

Amjad, M., D. Shah, and D. Shen (2018). Robust synthetic control. *Journal of Machine Learning Research 19*(22), 1–51.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2017). Matrix completion methods for causal panel data models.

Auerbach, E. (2019). Identification and estimation of a partially linear regression model using network data. *arXiv preprint arXiv:1903.09679*.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*(1), 135–171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*(4), 1229–1279.

Bai, J. and P. Wang (2016). Econometric analysis of large factor models. *Annual Review of Economics 8*, 53–80.

Bodelet, J. and J. Shan (2020). Nonparametric additive factor models. *arXiv preprint arXiv:2003.13119*.

Bonhomme, S., T. Lamadon, and E. Manresa (2021). Discretizing unobserved heterogeneity. *Econometrica (Forthcoming)*.

Chen, M., I. Fernández-Val, and M. Weidner (2020). Nonlinear factor models for network and panel data. *Journal of Econometrics*.

Chernozhukov, V., C. Hansen, Y. Liao, and Y. Zhu (2021). Inference for low-rank models. *arXiv preprint arXiv:2107.02602*.

Chudik, A. and M. H. Pesaran (2013). Large panel data models with cross-sectional dependence: a survey. *CAFE Research Paper* (13.15).

Chudik, A., M. H. Pesaran, and E. Tosetti (2011a). Weak and strong cross-section dependence and estimation of large panels. *The Econometrics Journal 14*(1), C45–C90.

Chudik, A., M. H. Pesaran, and E. Tosetti (2011b). Weak and strong cross-section dependence and estimation of large panels. *Econometrics Journal 14*, 45–90.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica 78*(3), 883–931.

Dhaene, G. and K. Jochmans (2015). Split-panel jackknife estimation of fixed-effect models. *The Review of Economic Studies 82*(3), 991–1030.

Dzemski, A. (2019). An empirical model of dyadic link formation in a network with unobserved heterogeneity. *Review of Economics and Statistics 101*(5), 763–776.

Fernández-Val, I., H. Freeman, and M. Weidner (2021). Low-rank approximations of nonseparable panel models. *The Econometrics Journal 24*(2), C40–C77.

Fernández-Val, I. and M. Weidner (2016). Individual and time effects in nonlinear panel models with large n, t. *Journal of Econometrics 192*(1), 291–312.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized dynamic-factor model:

Identification and estimation. *Review of Economics and statistics 82*(4), 540–554.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2005). The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association 100*(471), 830–840.

Freyberger, J. (2017, 09). Non-parametric Panel Data Models with Interactive Fixed Effects. *The Review of Economic Studies 85*(3), 1824–1851.

Galvao, A. F. and K. Kato (2014). Estimation and inference for linear panel data models under misspecification when both n and t are large. *Journal of Business & Economic Statistics 32*(2), 285–309.

Gao, C., Y. Lu, H. H. Zhou, et al. (2015). Rate-optimal graphon estimation. *The Annals of Statistics 43*(6), 2624–2652.

Giglio, S., M. Maggiori, and J. Stroebel (2016). No-bubble condition: Model-free tests in housing markets. *Econometrica 84*(3), 1047–1091.

Giglio, S. and D. Xiu (2021). Asset pricing with omitted factors. *Journal of Political Economy 129*(7), 000–000.

Gobillon, L. and T. Magnac (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics 98*(3), 535–551.

Graf, S. and H. Luschgy (2002). Rates of convergence for the empirical quantization error. *The Annals of Probability 30*(2), 874–897.

Graham, B. S. (2017). An econometric model of network formation with degree heterogeneity. *Econometrica 85*(4), 1033–1063.

Griebel, M. and H. Harbrecht (2014). Approximation of bi-variate functions: singular value decomposition versus sparse grids. *IMA journal of numerical analysis 34*(1), 28–54.

Gunsilius, F. and S. M. Schennach (2019). Independent nonlinear component analysis. Technical report, cemmap working paper.

Harding, M. and C. Lamarche (2011). Least squares estimation of a panel data model with multifactor error structure and endogenous covariates. *Economics Letters 111*(3), 197–199.

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983). Stochastic blockmodels: First steps. *Social networks 5*(2), 109–137.

Holtz-Eakin, D., W. Newey, and H. S. Rosen (1988, November). Estimating vector autoregressions with panel data. *Econometrica 56*(6), 1371–95.

Juodis, A. (2020). This shock is different: Estimation and inference in misspecified two-way fixed effects panel regressions. *Working Paper*.

Juodis, A. and V. Sarafidis (2018). Fixed t dynamic panel data estimators with multifactor

errors. *Econometric Reviews 37*(8), 893–929.

Juodis, A. and V. Sarafidis (2020). A linear estimator for factor-augmented fixed-t panels with endogenous regressors. *Journal of Business & Economic Statistics*, 1–15.

Karabiyik, H., F. C. Palm, and J.-P. Urbain (2019). Econometric analysis of panel data models with multifactor error structures. *Annual Review of Economics 11*, 495–522.

Kim, D. and T. Oka (2014). Divorce law reforms and divorce rates in the usa: an interactive fixed-effects approach. *Journal of Applied Econometrics 29*(2), 231–245.

Latala, R. (2005). Some estimates of norms of random matrices. *Proc. Amer. Math. Soc. 133*, 1273–1282.

Lee, N., H. R. Moon, and M. Weidner (2012). Analysis of interactive fixed effects dynamic linear panel regression with measurement error. *Economics Letters 117*(1), 239–242.

Li, H., Q. Li, and Y. Shi (2017). Determining the number of factors when the number of factors can increase with sample size. *Journal of Econometrics 197*(1), 76–86.

Lu, X. and L. Su (2016). Shrinkage estimation of dynamic panel data models with interactive fixed effects. *Journal of Econometrics 190*(1), 148–175.

Menzel, K. (2021). Bootstrap with cluster-dependence in two or more dimensions. *Econometrica 89*(5), 2143–2188.

Moon, H. R., M. Shum, and M. Weidner (2018). Estimation of random coefficients logit demand models with interactive fixed effects. *Journal of Econometrics 206*(2), 613–644.

Moon, H. R. and M. Weidner (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica 83*(4), 1543–1579.

Moon, H. R. and M. Weidner (2017). Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory 33*(1), 158–195.

Norkutė, M., V. Sarafidis, T. Yamagata, and G. Cui (2021). Instrumental variable estimation of dynamic linear panel data models with defactored regressors and a multifactor error structure. *Journal of Econometrics 220*(2), 416–446.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics 92*(4), 1004–1016.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics 168*(2), 244–258.

Pesaran, M. H. (2006, 07). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica 74*(4), 967–1012.

Pesaran, M. H. and E. Tosetti (2011). Large panels with common factors and spatial correlation. *Journal of Econometrics 161*(2), 182–202.

Robertson, D. and V. Sarafidis (2015). Iv estimation of panels with factor residuals. *Journal of Econometrics 185*(2), 526–541.

Sarafidis, V. and D. Robertson (2009). On the impact of error cross-sectional dependence in short dynamic panel estimation. *The Econometrics Journal 12*(1), 62–81.

Stock, J. H. and M. W. Watson (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics 20*(2), 147–162.

Su, L. and Q. Chen (2013). Testing homogeneity in panel data models with interactive fixed effects. *Econometric Theory 29*(6), 1079–1135.

Su, L. and X. Wang (2017). On time-varying factor models: Estimation and testing. *Journal of Econometrics 198*(1), 84–101.

Totty, E. (2017). The effect of minimum wages on employment: A factor model approach. *Economic Inquiry 55*(4), 1712–1737.

Westerlund, J., Y. Petrova, and M. Norkute (2019). Cce in fixed-t panels. *Journal of Applied Econometrics 34*(5), 746–761.

Westerlund, J. and J.-P. Urbain (2013). On the estimation and inference in factor-augmented panel regressions with correlated loadings. *Economics Letters 119*(3), 247–250.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, 817–838.

Wolfe, P. J. and S. C. Olhede (2013). Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*.

Zeleneev, A. (2020). Identification and estimation of network models with nonparametric unobserved heterogeneity.

A Appendix

A.1 Proofs for Section 3

We first establish a technical lemma, which is afterwards used to proof the main text theorem. Remember that we write $\|\cdot\|$ for the spectral norm of a matrix. Define the projection matrix $P_A = A(A'A)^\dagger A'$ for any matrix A and remember we write the annihilation matrix $M_A = \mathbb{I} - P_A$. Here, $\dagger$ refers to the Moore-Penrose inverse.

Lemma A.1. *Let Assumption 3 hold and consider $N, T \to \infty$. Furthermore, assume that*

$$Y = \sum_{k=1}^{K} X_k \beta_k^0 + e^* + e, \tag{A.1}$$

with $\text{rank}(e^*) = R_{NT} \leq \min(N,T)/2$, $\|e\| = \mathcal{O}_P(\eta_{NT})$, $\|X_k\| = \mathcal{O}_P(\sqrt{NT})$, *and* $\frac{1}{\sqrt{NT}}\text{Tr}(X_k e') = \mathcal{O}_P(\xi_{NT})$, *for* $k = 1, \ldots, K$. *Then, the LS estimator in* (5) *calculated with* $R = R_{NT}$ *factors in the estimation procedure, satisfies* $\widehat{\beta}_{\text{LS}} - \beta^0 = \mathcal{O}_P\left((\xi_{NT} + R_{NT}\eta_{NT})/\sqrt{NT}\right)$.

Proof of Lemma A.1. This proof is relatively minor modification of the consistency proof for the LS estimator in Moon and Weidner (2015), and more technical details can be found there. For simplicity we just write R, η, ξ instead of R_{NT}, η_{NT}, ξ_{NT} in this proof. We rewrite the definition of $\widehat{\beta}_{\text{LS}}$ as

$$\widehat{\beta}_{\text{LS}} = \underset{\beta}{\text{argmin}}\, \mathcal{L}_{NT}(\beta),$$

$$\mathcal{L}_{NT}(\beta) := \min_{\{\lambda \in \mathbb{R}^{N\times R},\, f \in \mathbb{R}^{T\times R}\}} \frac{1}{NT}\text{Tr}\left[\left(Y - X\cdot\beta - \lambda f'\right)\left(Y - X\cdot\beta - \lambda f'\right)'\right]. \tag{A.2}$$

Since $\text{rank}(e^*) = R$ we can write $e^* = \lambda^* f^{*\prime}$ for some $N \times R$ matrix λ^* and $T \times R$ matrix $f^{*\prime}$.

We now first establish a lower bound on $\mathcal{L}_{NT}(\beta)$. Let $\Delta\beta = \beta - \beta^0$. Consider the definition of $\mathcal{L}_{NT}(\beta)$ in equation (A.2) and plug in the model $Y = \beta\cdot X + \lambda^* f^{*\prime} + e$. We then have

$$\begin{aligned}
\mathcal{L}_{NT}(\beta) &= \min_{\{\lambda \in \mathbb{R}^{N\times R},\, f \in \mathbb{R}^{T\times R}\}} \frac{1}{NT}\text{Tr}\left[\left(\Delta\beta\cdot X + e + \lambda^* f^{*\prime} - \lambda f'\right)\left(\Delta\beta\cdot X + e + \lambda^* f^{*\prime} - \lambda f'\right)'\right] \\
&\geq \min_{\left\{\tilde{\lambda} \in \mathbb{R}^{N\times(2R)},\, \tilde{f} \in \mathbb{R}^{T\times(2R)}\right\}} \frac{1}{NT}\text{Tr}\left[\left(\Delta\beta\cdot X + e - \tilde{\lambda}\tilde{f}'\right)\left(\Delta\beta\cdot X + e - \tilde{\lambda}\tilde{f}'\right)'\right] \\
&= \frac{1}{NT}\min_{\tilde{f} \in \mathbb{R}^{T\times(2R)}} \text{Tr}\left[\left(\Delta\beta\cdot X + e\right) M_{\tilde{f}} \left(\Delta\beta\cdot X + e\right)'\right] \\
&= \frac{1}{NT}\min_{\tilde{f} \in \mathbb{R}^{T\times(2R)}} \Bigg\{ \text{Tr}\left[\left(\Delta\beta\cdot X\right) M_{\tilde{f}} \left(\Delta\beta\cdot X\right)'\right] + \text{Tr}\left(ee'\right) - \text{Tr}\left(eP_{\tilde{f}}e'\right) \\
&\qquad\qquad + 2\text{Tr}\left[\left(\Delta\beta\cdot X\right) e'\right] - 2\text{Tr}\left[\left(\Delta\beta\cdot X\right) P_{\tilde{f}} e'\right] \Bigg\} \\
&\geq \frac{1}{NT}\Bigg\{ \sum_{r=2R+1}^{T} \mu_r\left[(\Delta\beta\cdot X)'(\Delta\beta\cdot X)\right] + \text{Tr}\left(ee'\right) - 2R\|e\|^2 \\
&\qquad\qquad + 2\text{Tr}\left[\left(\Delta\beta\cdot X\right) e'\right] - 4R\|e\|\|\Delta\beta\cdot X\| \Bigg\} \\
&\geq b\,\|\Delta\beta\|^2 + \frac{1}{NT}\,\text{Tr}\left(ee'\right) + \mathcal{O}_P\left(\frac{R\,\eta^2}{NT}\right) + \mathcal{O}_P\left(\frac{(\xi + R\,\eta)\,\|\Delta\beta\|}{\sqrt{NT}}\right).
\end{aligned} \tag{A.3}$$

Here, we applied the inequality $|\text{Tr}(A)| \leq \text{rank}(A)\|A\|$ with $A = (\Delta\beta\cdot X) P_{\tilde{f}} e'$ and also with $A = eP_{\tilde{f}}e'$. We also used that $\min_{\tilde{f}} \text{Tr}\left[\left(\Delta\beta\cdot X\right) M_{\tilde{f}} \left(\Delta\beta\cdot X\right)'\right] = \sum_{r=2R+1}^{T} \mu_r\left[(\Delta\beta\cdot X)'(\Delta\beta\cdot X)\right]$. In the last step of (A.3) we applied the various assumptions in the lemma.

Next, we establish an upper bound on $\mathcal{L}_{NT}(\beta^0)$. We can choose $\lambda = \lambda^*$ and $f = f^*$ in the minimization problem in (A.2), and therefore

$$\mathcal{L}_{NT}(\beta^0) \leq \frac{1}{NT}\,\text{Tr}\left(ee'\right). \tag{A.4}$$

Since we could choose $\beta = \beta^0$ in the minimization of β, the optimal $\widehat{\beta}_{\text{LS}}$ needs to satisfy $\mathcal{L}_{NT}(\widehat{\beta}_{\text{LS}}) \leq \mathcal{L}_{NT}(\beta^0)$. Together with (A.3) and (A.4) this gives

$$b\,\|\widehat{\beta}_{\text{LS}} - \beta^0\|^2 + \mathcal{O}_P\left(\frac{(\xi + R\,\eta)\|\widehat{\beta}_{\text{LS}} - \beta^0\|}{\sqrt{NT}}\right) + \mathcal{O}_P\left(\frac{R\,\eta^2}{NT}\right) \leq 0\,. \tag{A.5}$$

Since $\sqrt{R} = O(R)$ it follows that $\|\widehat{\beta}_{\text{LS}} - \beta^0\| = \mathcal{O}_P\left((\xi + R\,\eta)/\sqrt{NT}\right)$, which is what we wanted to show. ■

Using Lemma A.1 we are now ready to prove Theorem 1.

Proof of Theorem 1. To apply Lemma A.1 we first need to define e and e^* such that (A.1) is an implication of our model (13). Decompose $\Gamma = \sum_{r=1}^{\min\{N,T\}} \lambda_r^* f_r^{*\prime}$, which is a reformulation of the singular value decomposition of a matrix. Define $e^* = \sum_{r=1}^{R_{NT}} \lambda_r^* f_r^{*\prime}$ such that $\text{rank}(e^*) = R_{NT}$. Also define $e = S + \varepsilon$ where $S = \Gamma - \sum_{r=1}^{R_{NT}} \lambda_r^* f_r^{*\prime}$. With these definitions model (13) can be rewritten as (A.1) and it remains to show Assumptions 1-3 are sufficient for Lemma A.1 and to characterise the sequences η_{NT} and ξ_{NT}.

First, use the norm inequality $\|S + \varepsilon\| \leq \|S\| + \|\varepsilon\|$ with $\|\varepsilon\| = O_P(\sqrt{\max\{N,T\}})$ from Assumption 1 (ii) to show $\|e\| \leq \|S\| + O_P(\sqrt{\max\{N,T\}})$. To bound $\|S\|$ use the fact that the spectral norm is bounded by the Frobenius norm and Assumption 4 to show

$$\begin{aligned}\|S\|^2 \leq \|S\|_F^2 &= \sum_{r=R_{NT}+1}^{\infty} \sigma_r^2(\Gamma) \\ &\leq O_P\big(NTR_{NT}^{1-2\rho}\big).\end{aligned}$$

This shows that $\|e\|$ is asymptotically bounded in probability by the sequence η_{NT} with

$$\eta_{NT} = \sqrt{\max\{N,T\}} + \sqrt{NT}R_{NT}^{(1-2\rho)/2}.$$

That is, $\|e\| = O_P(\eta_{NT})$.

Secondly, the bound on $\|X_k\|$ is direct from Assumption 1.(i) again because the spectral norm is bounded by the Frobenius norm. That is, $\|X_k\|^2 \leq \|X_k\|_F^2 = \sum_{i=1}^{N}\sum_{t=1}^{T} X_{it,k}^2 = O_P(NT)$.

Lastly, we need to show that $\frac{1}{\sqrt{NT}}\text{Tr}(X_k e') = O_P(\xi_{NT})$ and to find ξ_{NT}. To do this we decompose e and use the Cauchy-Schwarz inequality, the triangle inequality and linearity of the trace operator in the following,

$$\begin{aligned}\left|\frac{1}{\sqrt{NT}}\text{Tr}(X_k e')\right| &= \left|\frac{1}{\sqrt{NT}}\text{Tr}(X_k(S+\varepsilon)')\right| \\ &\leq \frac{1}{\sqrt{NT}}\,\|X_k\|_F\,\|S\|_F + \frac{1}{\sqrt{NT}}|\text{Tr}(X_k\varepsilon')| \\ &= O_P(1)\,\|S\|_F + O_P(1).\end{aligned} \tag{A.6}$$

The third line follows from Assumption 1.(i) and Assumption 2. From above we know $\|S\|_F = O_P\big(\sqrt{NT}R_{NT}^{(1-2\rho)/2}\big)$, hence we have found $\xi_{NT} = \sqrt{NT}R_{NT}^{(1-2\rho)/2} + 1$.

Thus, we have shown that all conditions for Lemma A.1 are satisfied and found the rates η_{NT} and ξ_{NT}. This shows that LS estimation in (5) on the model (13) with $R = R_{NT}$ factors satisfies $\widehat{\beta}_{\text{LS}} - \beta^0 = O_P\left((\xi_{NT} + R_{NT}\eta_{NT})/\sqrt{NT}\right)$, with

$$\begin{aligned} O_P\left(\frac{(\xi_{NT} + R_{NT}\eta_{NT})}{\sqrt{NT}}\right) &= O_P\Big(R_{NT}^{(1-2\rho)/2}\Big) + O_P\Big(\frac{1}{\sqrt{NT}}\Big) + O_P\Big(R_{NT}^{(3-\rho)/2}\Big) \\ &\quad + \mathcal{O}_P\left(R_{NT}\sqrt{\frac{\max\{N,T\}}{NT}}\right) \\ &= O_P\Big(R_{NT}^{(3-\rho)/2}\Big) + O_P\Big(R_{NT}\min\{N,T\}^{-1/2}\Big). \end{aligned}$$

■

Proof of Remark 1. Note that if we weaken the singular value decay to that supposed in Remark 1, i.e. $\sigma_r(\Gamma) = c\sqrt{NT}r^{-\rho}$, and otherwise maintain Assumptions 1-3 we can further bound the bias in LS estimation found in Theorem 1 as follows. For $\|S\|_F$, note,

$$\begin{aligned} \|S\|_F^2 &= \sum_{r=R_{NT}+1}^{\infty} \sigma_r^2(\Gamma) \\ &\leq \sum_{r=R_{NT}+1}^{\infty} cNTr^{-2\rho} \qquad wpa.1 \qquad \text{(Assumption 4)} \\ &\leq cNT\int_{R_{NT}}^{\infty} r^{-2\rho}dr \qquad wpa.1 \qquad \text{(integral bound)} \\ &= \frac{c}{2\rho-1}NTR_{NT}^{1-2\rho} \qquad wpa.1 \end{aligned}$$

In the third line we use an integral bound and the fourth line simply evaluates this integral. From line two all arguments are $wpa.1$, hence $\|S\|_F = \mathcal{O}_P(\sqrt{NT}R_{NT}^{(1-2\rho)/2})$, where $(c/2\rho - 1)$ is the bounding constant. We can then directly bound

$$\begin{aligned} \|S\| &= \max_{r\in\{R_{NT}+1,\ldots,\min\{N,T\}\}} \sigma_r(\Gamma) \\ &= \mathcal{O}_P\big(\sqrt{NT}(R_{NT}+1)^{-\rho}\big), \end{aligned}$$

where we use the convention that singular values are indexed in descending order. We then simplify the last bound to $\|S\| = \mathcal{O}_P\big(\sqrt{NT}R_{NT}^{-\rho}\big)$, replacing $R_{NT}+1$ with R_{NT} as $R_{NT}\to\infty$. We can then rely on the same working in the proof of Theorem 1 to show that the conditions in Lemma A.1 are satisfied with $\xi_{NT} = \sqrt{NT}R_{NT}^{(1-2\rho)/2} + 1$ and $\eta_{NT} = \sqrt{\max\{N,T\}} +$

$\sqrt{NT}R_{NT}^{-\rho}$, where the second term in η_{NT} is slightly different to Theorem 1. Hence, $\widehat{\beta}_{\text{LS}} - \beta^0 = O_P\left((\xi_{NT} + R_{NT}\eta_{NT})/\sqrt{NT}\right)$, with

$$\begin{aligned} O_P\left(\frac{(\xi_{NT} + R_{NT}\eta_{NT})}{\sqrt{NT}}\right) &= O_P\Big(R_{NT}^{(1-2\rho)/2}\Big) + O_P\Big(\frac{1}{\sqrt{NT}}\Big) + O_P\Big(R_{NT}^{1-\rho}\Big) \\ &\quad + O_P\left(R_{NT}\sqrt{\frac{\max\{N,T\}}{NT}}\right) \\ &= O_P\Big(R_{NT}^{1-\rho}\Big) + O_P\Big(R_{NT}\min\{N,T\}^{-1/2}\Big). \end{aligned}$$

■

To prove Lemma 1 we rely on the following result from (Griebel and Harbrecht, 2014), which we state without proof.

Lemma A.2 (Theorem 3.5 in Griebel and Harbrecht 2014)**.** *Let $g \in L^p(\Omega_\alpha \times \Omega_\gamma)$ and $p > \min\{n_\alpha, n_\gamma\}/2$, then*

$$\left\| g - \sum_{l=1}^{R} \sigma_l(\varphi_l \otimes \psi_l) \right\|_{L^2(\Omega_1 \times \Omega_2)} = O\left(R^{\frac{1}{2} - \frac{p}{\min\{n_\alpha, n_\gamma\}}}\right). \tag{A.7}$$

In the following proof we use the Frobenius norm, which as a reminder is defined as $\|A\|_F^2 = \sum_{i=1}^{N}\sum_{t=1}^{T}|A_{it}|^2$ for any $N \times T$ matrix A.

Proof of Lemma 1. From Lemma A.2 we have,

$$\begin{aligned} \mathbb{E}\Big[\Big(h(\alpha_i,\gamma_t) - \sum_{s=1}^{R}\sigma_r\varphi_r(\alpha_i)\psi_r(\gamma_t)\Big)^2\Big] &= \int_{\Omega_\alpha}\int_{\Omega_\gamma}\big(h(a,c) - \sum_{s=1}^{R}\sigma_r\varphi_r(a)\psi_r(c)\big)^2 f_\alpha(a)da f_\gamma(c)dc \\ &\leq \int_{\Omega_\alpha}\int_{\Omega_\gamma}\big(h(a,c) - \sum_{s=1}^{R}\sigma_r\varphi_r(a)\psi_r(c)\big)^2 dadc \sup_a(f_\alpha(a)) \sup_c(f_\gamma(c)) \\ &= \left\| g - \sum_{l=1}^{R}\sigma_l\varphi_l \otimes \psi_l \right\|^2_{L^2(\Omega_\alpha \times \Omega_\gamma)} O(1) \\ &= O\left(R^{1 - \frac{2p}{\min\{n_\alpha, n_\gamma\}}}\right) \end{aligned} \tag{A.8}$$

where in the second line we use a supremum bound on the probabilities, in the third line we use the definition of the $L^2(\Omega_\alpha \times \Omega_\gamma)$-norm and in the final line we use Lemma A.2. This shows that, in expectations, the entry-wise functional representation decays at polynomial rate $r^{1-2\rho}$, with $\rho = p/\min\{n_\alpha, n_\gamma\}$.

Using the Markov inequality gives

$$\left(\Gamma_{it} - \sum_{\ell=1}^{r} \sigma_\ell \varphi_\ell(\alpha)\psi_\ell(\gamma)'\right)^2 = \mathcal{O}_P\left(r^{1-\frac{2p}{\min\{n_\alpha, n_\gamma\}}}\right),$$

which we use to bound singular values of the matrix Γ as follows.

We know

$$\Gamma_{it} = h(\alpha_i, \gamma_t) = \sum_{r=1}^{\infty} \sigma_r \varphi_r(\alpha_i)\psi_r(\gamma_t) = \sum_{r=1}^{\infty} \sigma_r u_{ir} v_{tr}$$

and in matrix form,

$$\Gamma = h(\alpha, \gamma) = \sum_{r=1}^{\infty} \sigma_r \varphi_r(\alpha)\psi_r(\gamma)' = \sum_{r=1}^{\infty} \sigma_r u_r v_r'.$$

Hence, we have

$$\begin{aligned}
\sum_{\ell=r+1}^{\min\{N,T\}} \sigma_\ell^2(\Gamma) &= \min_{\lambda \in \mathbb{R}^{N\times r}} \min_{f \in \mathbb{R}^{T\times r}} \left\|\Gamma - \lambda\, f'\right\|_F^2 \\
&\leq \left\|\Gamma - \sum_{\ell=1}^{r} \sigma_\ell \varphi_\ell(\alpha)\psi_\ell(\gamma)'\right\|_F^2 \\
&= \sum_i \sum_t \left(\sum_{\ell=r+1}^{\infty} \sigma_\ell \varphi_\ell(\alpha_i)\psi_\ell(\gamma_t)\right)^2 \\
&= \sum_i \sum_t \mathcal{O}_P\left(r^{1-\frac{2p}{\min\{n_\alpha, n_\gamma\}}}\right) \\
&= NT\mathcal{O}_P\left(r^{1-\frac{2p}{\min\{n_\alpha, n_\gamma\}}}\right).
\end{aligned}$$

Hence, we have $\frac{1}{NT}\sum_{\ell=r+1}^{\min\{N,T\}} \sigma_\ell^2(\Gamma) = \mathcal{O}_P\left(r^{1-2\rho}\right)$ with $\rho = p/\min\{n_\alpha, n_\gamma\}$, and Assumption 4 is satisfied. ■

A.2 Proofs for Section 4

Proof of Lemma 2. From Section 4 we have

$$\kappa_{NT} := \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \widetilde{X}_{it}'\widetilde{X}_{it}\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} \widetilde{X}_{it}'\,\widetilde{\Gamma}_{it},$$

with $\widetilde{\Gamma}$ defined analogously to $\widetilde{X}_k$ and $\widetilde{Y}$.

Take

$$\|\kappa_{NT}\| := \left\|\left(\sum_{i=1}^{N}\sum_{t=1}^{T} \widetilde{X}_{it}'\widetilde{X}_{it}\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} \widetilde{X}_{it}'\,\widetilde{\Gamma}_{it}\right\|.$$

Using the inequality $\|Az\| \leq \|A\| \|z\|$ for general matrices A and vectors z we find

$$\|\kappa_{NT}\| \leq \left\| \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}'_{it} \widetilde{X}_{it} \right)^{-1} \right\| \left\| \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}'_{it} \widetilde{\Gamma}_{it} \right\|.$$

Use $\left| \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}_{it,k} \widetilde{\Gamma}_{it} \right| \leq \sum_{i=1}^{N} \sum_{t=1}^{T} \left| \widetilde{X}_{it,k} \widetilde{\Gamma}_{it} \right|$ and Hölder's inequality such that

$$\begin{bmatrix} \left| \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}_{it,1} \widetilde{\Gamma}_{it} \right| \\ \vdots \\ \left| \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}_{it,1} \widetilde{\Gamma}_{it} \right| \end{bmatrix} \leq \begin{bmatrix} \sum_{i=1}^{N} \sum_{t=1}^{T} \left| \widetilde{X}_{it,1} \widetilde{\Gamma}_{it} \right| \\ \vdots \\ \sum_{i=1}^{N} \sum_{t=1}^{T} \left| \widetilde{X}_{it,K} \widetilde{\Gamma}_{it} \right| \end{bmatrix} \leq \begin{bmatrix} \|\mathrm{vec}(X_1)\|_\infty \\ \vdots \\ \|\mathrm{vec}(X_K)\|_\infty \end{bmatrix} \left\| \mathrm{vec}(\widetilde{\Gamma}) \right\|_1,$$

where $\mathrm{vec}(A)$ vectorises a matrix A such that $\|\mathrm{vec}(A)\|_\infty = \max_{i,t} |A_{it}|$ yields the maximum norm and $\|\mathrm{vec}(A)\|_1 = \sum_{i=1}^{N} \sum_{t=1}^{T} |A_{it}|$ yields the entry-wise 1-norm of such a matrix.

Take the $\|\cdot\|$ to show

$$\begin{aligned} \left\| \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}'_{it} \widetilde{\Gamma}_{it} \right\| &= \left(\sum_k \left| \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}_{it,k} \widetilde{\Gamma}_{it} \right|^2 \right)^{1/2} \\ &\leq \left(\sum_k \left(\left\| \mathrm{vec}(\widetilde{X}_k) \right\|_\infty \left\| \mathrm{vec}(\widetilde{\Gamma}) \right\|_1 \right)^2 \right)^{1/2} \\ &= \left(\sum_k \left(\left\| \mathrm{vec}(\widetilde{X}_k) \right\|_\infty \right)^2 \right)^{1/2} \left\| \mathrm{vec}(\widetilde{\Gamma}) \right\|_1 \leq \left(\sum_k \left\| \mathrm{vec}(\widetilde{X}_k) \right\|_\infty \right) \left\| \mathrm{vec}(\widetilde{\Gamma}) \right\|_1, \end{aligned}$$

where in the last line we use that $\left\| \mathrm{vec}(\widetilde{\Gamma}) \right\|_1$ is a scalar and that $\left\| \mathrm{vec}(X_k) \right\|_\infty > 0 \ \forall k$. Thus we can bound the norm of κ_{NT} by

$$\|\kappa_{NT}\| \leq \left\| \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{X}'_{it} \widetilde{X}_{it} \right)^{-1} \right\| \left(\sum_{k=1}^{K} \left\| \mathrm{vec}(\widetilde{X}_k) \right\|_\infty \right) \left\| \mathrm{vec}(\widetilde{\Gamma}) \right\|_1.$$

Concentrate on $\left\| \mathrm{vec}(\widetilde{\Gamma}) \right\|_1$. Let n_i^N be the size of each i's cluster and n_t^T be the size of each t's cluster, then

$$\widetilde{\Gamma}_{it} = h(\alpha_i, \gamma_t) - \frac{1}{n_i^N} \sum_{j \in g_i} h(\alpha_j, \gamma_t) - \frac{1}{n_t^T} \sum_{s \in c_t} h(\alpha_i, \gamma_s) + \frac{1}{n_i^N} \frac{1}{n_t^T} \sum_{j \in g_i} \sum_{s \in c_t} h(\alpha_j, \gamma_s).$$

Take the following Taylor expansions,

$$\begin{aligned} h(\alpha_j, \gamma_s) &= h(\alpha_i, \gamma_t) + \frac{\partial h(\alpha_i, \gamma_t)}{\partial \alpha'} (\alpha_j - \alpha_i) + \frac{\partial h(\alpha_i, \gamma_t)}{\partial \gamma'} (\gamma_s - \gamma_t) + r(i, j, t, s) \\ h(\alpha_j, \gamma_t) &= h(\alpha_i, \gamma_t) + \frac{\partial h(\alpha_i, \gamma_t)}{\partial \alpha_i'} (\alpha_j - \alpha_i) + r'(i, j, t) \\ h(\alpha_i, \gamma_s) &= h(\alpha_i, \gamma_t) + \frac{\partial h(\alpha_i, \gamma_t)}{\partial \gamma'} (\gamma_s - \gamma_t) + r''(t, s, i), \end{aligned}$$

where r, r' and r'' are remainder terms from the Taylor expansion.

From these expansions we have

$$\frac{1}{n_i^N}\sum_{j\in g_i} h(\alpha_j,\gamma_t) = h(\alpha_i,\gamma_t) + \frac{1}{n_i^N}\sum_{\substack{j\in g_i,\\ j\neq i}}\left(\frac{\partial h(\alpha_i,\gamma_t)}{\partial \alpha'}(\alpha_j-\alpha_i) + r'(i,j,t)\right),$$

$$\frac{1}{n_t^T}\sum_{s\in c_t} h(\alpha_i,\gamma_s) = h(\alpha_i,\gamma_t) + \frac{1}{n_t^T}\sum_{\substack{s\in c_t,\\ s\neq t}}\left(\frac{\partial h(\alpha_i,\gamma_t)}{\partial \gamma'}(\gamma_s-\gamma_t) + r''(t,s,i)\right),$$

and

$$\begin{aligned}
\frac{1}{n_i^N}\frac{1}{n_t^T}\sum_{j\in g_i}\sum_{s\in c_t} h(\alpha_j,\gamma_s) &= \frac{1}{n_i^N n_t^T}h(\alpha_i,\gamma_t) + \frac{1}{n_i^N n_t^T}\left(\sum_{\substack{j\in g_i,\\ j\neq i}}\sum_{\substack{s\in c_t,\\ s\neq t}} h(\alpha_j,\gamma_s) + \sum_{\substack{j\in g_i,\\ j\neq i}} h(\alpha_j,\gamma_t) + \sum_{\substack{s\in c_t,\\ s\neq t}} h(\alpha_i,\gamma_s)\right)\\
&= h(\alpha_i,\gamma_t) + \frac{1}{n_i^N n_t^T}\sum_{\substack{j\in g_i,\\ j\neq i}}\sum_{\substack{s\in c_t,\\ s\neq t}}\left(\frac{\partial h(\alpha_i,\gamma_t)}{\partial \alpha'}(\alpha_j-\alpha_i) + \frac{\partial h(\alpha_i,\gamma_t)}{\partial \gamma'}(\gamma_s-\gamma_t) + r(i,j,t,s)\right)\\
&+ \frac{1}{n_i^N n_t^T}\sum_{\substack{j\in g_i,\\ j\neq i}}\left(\frac{\partial h(\alpha_i,\gamma_t)}{\partial \alpha'}(\alpha_j-\alpha_i) + r'(i,j,t)\right) + \frac{1}{n_i^N n_t^T}\sum_{\substack{s\in c_t,\\ s\neq t}}\left(\frac{\partial h(\alpha_i,\gamma_t)}{\partial \gamma'}(\gamma_s-\gamma_t) + r''(t,s,i)\right)\\
&= h(\alpha_i,\gamma_t) + \frac{1}{n_i^N}\sum_{\substack{j\in g_i,\\ j\neq i}}\left(\frac{\partial h(\alpha_i,\gamma_t)}{\partial \alpha'}(\alpha_j-\alpha_i) + r'(i,j,t)\right) + \frac{1}{n_t^T}\sum_{\substack{s\in c_t,\\ s\neq t}}\left(\frac{\partial h(\alpha_i,\gamma_t)}{\partial \gamma'}(\gamma_s-\gamma_t) + r''(t,s,i)\right)\\
&+ \frac{1}{n_i^N n_t^T}\sum_{\substack{j\in g_i,\\ j\neq i}}\sum_{\substack{s\in c_t,\\ s\neq t}} r(i,j,t,s).
\end{aligned}$$

We explicitly split the sum in the second line to make clearer the fact that almost all terms cancel out once we difference these identities. From the last line it should be clear that,

$$\widetilde{\Gamma}_{it} = \frac{1}{n_i^N n_t^T}\sum_{\substack{j\in g_i,\\ j\neq i}}\sum_{\substack{s\in c_t,\\ s\neq t}} r(i,j,t,s).$$

From $h(.,.)$ being twice continuously differentiable and a uniformly bounded second derivative, we have $r(i,j,t,s) = O\left(\|\alpha_i-\alpha_j\|\,\|\gamma_t-\gamma_s\| + \|\alpha_i-\alpha_j\|^2 + \|\gamma_t-\gamma_s\|^2\right)$.

For the entry-wise 1-norm, we have,

$$\begin{aligned}\left\|\text{vec}(\widetilde{\Gamma})\right\|_1 &= \sum_i \sum_t \left| \frac{1}{n_i^N n_t^T} \sum_{\substack{j\in g_i,\\ j\neq i}} \sum_{\substack{s\in c_t,\\ s\neq t}} r(i,j,t,s) \right| \\ &\leq \sum_i \sum_t \left| \frac{1}{n_i^N n_t^T} \sum_{\substack{j\in g_i,\\ j\neq i}} \sum_{\substack{s\in c_t,\\ s\neq t}} O\left(\|\alpha_i-\alpha_j\|\,\|\gamma_t-\gamma_s\|\right) \right| \\ &+ \sum_i \sum_t \left| \frac{1}{n_i^N n_t^T} \sum_{\substack{j\in g_i,\\ j\neq i}} \sum_{\substack{s\in c_t,\\ s\neq t}} O\left(\|\alpha_i-\alpha_j\|^2\right) \right| + \sum_i \sum_t \left| \frac{1}{n_i^N n_t^T} \sum_{\substack{j\in g_i,\\ j\neq i}} \sum_{\substack{s\in c_t,\\ s\neq t}} O\left(\|\gamma_t-\gamma_s\|^2\right) \right|.\end{aligned}$$

Now, concentrate on the second term,

$$\begin{aligned}\sum_i \sum_t \left| \frac{1}{n_i^N n_t^T} \sum_{\substack{j\in g_i,\\ j\neq i}} \sum_{\substack{s\in c_t,\\ s\neq t}} O\left(\|\alpha_i-\alpha_j\|^2\right) \right| &\leq \sum_i \sum_t \left| \frac{(n_i^N-1)(n_t^T-1)}{n_i^N n_t^T} \max_{\substack{j\in g_i,\\ j\neq i}} O\left(\|\alpha_i-\alpha_j\|^2\right) \right| \\ &= O(T) \sum_i \max_{\substack{j\in g_i,\\ j\neq i}} \|\alpha_i-\alpha_j\|^2\end{aligned}$$

Use Assumption 5(iii) to show for $j \in g_i$,

$$\begin{aligned}\|\alpha_i-\alpha_j\|^2 &\leq B^2 \|\lambda(\alpha_i)-\lambda(\alpha_j)\|^2 \\ &= B^2 \left\|\lambda(\alpha_i)-\widehat{\lambda}_i-(\lambda(\alpha_j)-\widehat{\lambda}_j)+\widehat{\lambda}_i-\widehat{\lambda}_j\right\|^2 \\ &\leq B^2 \left(\left\|\lambda(\alpha_i)-\widehat{\lambda}_i\right\|+\left\|\lambda(\alpha_j)-\widehat{\lambda}_j\right\|+\left\|\widehat{\lambda}_i-\widehat{\lambda}_j\right\|\right)^2.\end{aligned}$$

An application of Cauchy-Schwarz and Assumption 5(iv) gives

$$\begin{aligned}\sum_{i=1}^N \max_{\substack{j\in g_i,\\ j\neq i}} \|\alpha_i-\alpha_j\|^2 &\leq B^2 \sum_{i=1}^N \max_{\substack{j\in g_i,\\ j\neq i}} \Bigg(\left\|\lambda(\alpha_i)-\widehat{\lambda}_i\right\|^2 + \left\|\lambda(\alpha_j)-\widehat{\lambda}_j\right\|^2 + \left\|\widehat{\lambda}_i-\widehat{\lambda}_j\right\|^2 \\ &+2\left\|\lambda(\alpha_i)-\widehat{\lambda}_i\right\|\left\|\lambda(\alpha_j)-\widehat{\lambda}_j\right\| + 2\left\|\lambda(\alpha_i)-\widehat{\lambda}_i\right\|\left\|\widehat{\lambda}_i-\widehat{\lambda}_j\right\| + 2\left\|\lambda(\alpha_j)-\widehat{\lambda}_j\right\|\left\|\widehat{\lambda}_i-\widehat{\lambda}_j\right\| \Bigg) \\ &\leq NO_P(\xi_{NT}) + 2B^2 \sqrt{\sum_{i=1}^N \left\|\lambda(\alpha_i)-\widehat{\lambda}_i\right\|^2} \sqrt{\sum_{i=1}^N \max_{\substack{j\in g_i,\\ j\neq i}} \left\|\lambda(\alpha_j)-\widehat{\lambda}_j\right\|^2} \\ &+2B^2 \sqrt{\sum_{i=1}^N \left\|\lambda(\alpha_i)-\widehat{\lambda}_i\right\|^2} \sqrt{\sum_{i=1}^N \max_{\substack{j\in g_i,\\ j\neq i}} \left\|\widehat{\lambda}_i-\widehat{\lambda}_j\right\|^2} \\ &+2B^2 \sqrt{\sum_{i=1}^N \max_{\substack{j\in g_i,\\ j\neq i}} \left\|\lambda(\alpha_j)-\widehat{\lambda}_j\right\|^2} \sqrt{\sum_{i=1}^N \max_{\substack{j\in g_i,\\ j\neq i}} \left\|\widehat{\lambda}_i-\widehat{\lambda}_j\right\|^2}.\end{aligned}$$

Note that

$$\sum_{i=1}^{N} \max_{\substack{j\in g_i,\\ j\neq i}} \left\|\lambda(\alpha_j) - \widehat{\lambda}_j\right\|^2 \leq \sum_{i=1}^{N} \sum_{\substack{j\in g_i,\\ j\neq i}} \left\|\lambda(\alpha_j) - \widehat{\lambda}_j\right\|^2$$
$$\leq (Q_{\max} - 1) \sum_{i=1}^{N} \left\|\lambda(\alpha_i) - \widehat{\lambda}_i\right\|^2,$$

hence we have

$$\sum_{i=1}^{N}\sum_{t=1}^{T} \left| \frac{1}{n_i^N n_t^T} \sum_{\substack{j\in g_i,\\ j\neq i}} \sum_{\substack{s\in c_t,\\ s\neq t}} O\left(\|\alpha_i - \alpha_j\|^2\right) \right| = NTO_P(\xi_{NT})$$

The t-dimension analogy is direct such that $\sum_{i=1}^{N}\sum_{t=1}^{T} \left| \frac{1}{n_i^N n_t^T} \sum_{\substack{j\in g_i,\\ j\neq i}} \sum_{\substack{s\in c_t,\\ s\neq t}} O\left(\|\gamma_t - \gamma_s\|^2\right) \right| = NTO_P(\xi_{NT})$.

Finally,

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T} \left| \frac{1}{n_i^N n_t^T} \sum_{\substack{j\in g_i,\\ j\neq i}} \sum_{\substack{s\in c_t,\\ s\neq t}} O\left(\|\alpha_i - \alpha_j\| \, \|\gamma_t - \gamma_s\|\right) \right| \\
&\leq \sum_{i=1}^{N}\sum_{t=1}^{T} \frac{(n_i^N - 1)(n_t^T - 1)}{n_i^N n_t^T} \max_{\substack{j\in g_i,\\ j\neq i}} \max_{\substack{s\in c_t,\\ s\neq t}} \|\alpha_i - \alpha_j\| \, \|\gamma_t - \gamma_s\| \, |O(1)| \\
&= O(NT) \left(\frac{1}{N} \sum_{i=1}^{N} \max_{\substack{j\in g_i,\\ j\neq i}} \|\alpha_i - \alpha_j\| \right) \left(\frac{1}{T} \sum_{t=1}^{T} \max_{\substack{s\in c_t,\\ s\neq t}} \|\gamma_t - \gamma_s\| \right) \\
&\leq O(NT) \sqrt{\frac{1}{N} \sum_{i=1}^{N} \max_{\substack{j\in g_i,\\ j\neq i}} \|\alpha_i - \alpha_j\|^2} \sqrt{\frac{1}{T} \sum_{t=1}^{T} \max_{\substack{s\in c_t,\\ s\neq t}} \|\gamma_t - \gamma_s\|^2} \\
&= NTO_P(\xi_{NT}),
\end{aligned}$$

where we use Jensen's inequality in the third step.

Lastly, use Assumption 5.(v), which implies $\left(\sum_{i=1}^{N}\sum_{t=1}^{T} \widetilde{X}_{it}'\widetilde{X}_{it}\right)^{-1} = O_p(1/NT)$, to show

$$\|\kappa_{NT}\| = O_p(\xi_{NT})$$
$$\Rightarrow \kappa_{NT} = O_p(\xi_{NT})$$

■

For each partition $\mathcal{O}_q$, with $q \in \{1, 2, 3, 4\}$, the $N_q \times G^{(q)}$ matrix $D_\nu^{(q)}$, respectively $T_q \times C^{(q)}$ matrix $D_\delta^{(q)}$ represent the i, respectively t, cluster assignment matrices for $(i, t) \in \mathcal{O}_q$ where the

columns of each matrix are binary indicators of cluster assignment. That is, any given column of $D_\nu^{(q)}$ represents a cluster equal to 1 of that row is a member of the cluster and 0 otherwise, and likewise for $D_\delta^{(q)}$. Here $G^{(q)}$ are the number of i clusters and $C^{(q)}$ are the number of t clusters in $\mathcal{O}_q$. For each partition define the annihilation matrix $M_\nu^{(q)} = D_\nu^{(q)} \left([D_\nu^{(q)}]' D_\nu^{(q)} \right)^{-1} [D_\nu^{(q)}]'$ and $M_\delta^{(q)} = D_\delta^{(q)} \left([D_\delta^{(q)}]' D_\delta^{(q)} \right)^{-1} [D_\delta^{(q)}]'$. To perform within-cluster mean-differences we can then take, for matrix $A^{(q)}$ being the partition $\mathcal{O}_q$ of matrix A, $\check{A}^{(q)} = M_\nu^{(q)} A^{(q)} M_\delta^{(q)}$.[9] Take $\check{A}$ as the block matrix with blocks $\check{A}^{(q)}$. Further, for each regressor, k, let $\check{X}_k$ be defined similarly for each k separately such that $\check{X}_{it}$ a K dimensional column vector.

Assumption A.1. *Let $\mathcal{O}_q$ denote partitions for cluster formation and $\mathcal{O}_q^*$ denote partitions for proxy sampling. Across each partition, $\alpha_i^{(q)}$ has common support $\mathcal{A}$ for each q, $\gamma^{(q)}$ has common support $\mathcal{C}$ for each q, and both of these are bounded and convex sets. Also, assume each partition is of equal size, up to rounding error, such that they all grow proportionally with N, T. There exists a sequence $\xi_{NT} > 0$ common to all partitions such that $\xi_{NT} \to 0$ as $N, T \to \infty$, and*

(i) The function $h(\cdot, \cdot)$ is at least twice continuously differentiable with uniformly bounded second derivatives.

(ii) For each q, every unit $i \in \mathcal{O}_q$ is a member of exactly one group $g_i^{(q)} \in \{1, \ldots, G^{(q)}\}$, and every time period t is a member of exactly one group $c_t^{(q)} \in \{1, \ldots, C^{(q)}\}$. The size of all $G^{(q)}$ groups of units, and the size of all $C^{(q)}$ groups of time periods is bounded uniformly by $Q_{\max}$ for all q.

(iii) There exists $B > 0$ such that for all q there is, $\|a - b\| \leq B \left\| \lambda^{(q)}(a) - \lambda^{(q)}(b) \right\|$ for all $a, b \in \mathcal{A}$, and $\|a - b\| \leq B \left\| f^{(q)}(a) - f^{(q)}(b) \right\|$ for all $a, b \in \mathcal{C}$.

(iv) For each q there is,
$\frac{1}{N_q^* T_q^*} \sum_{i=1}^N \sum_{t=1}^T \mathbb{1}\{(i,t) \in \mathcal{O}_q^*\} \left(\left\| \widehat{\lambda}_i^{(q)} - \lambda^{(q)}(\alpha_i) \right\|^2 \right) = O_P\left(\xi_{NT}\right)$,
$\frac{1}{N_q^* T_q^*} \sum_{i=1}^N \sum_{t=1}^T \mathbb{1}\{(i,t) \in \mathcal{O}_q^*\} \left(\left\| \widehat{f}_t^{(q)} - f^{(q)}(\gamma_t) \right\|^2 \right) = O_P\left(\xi_{NT}\right)$.

(v) For each q there is,
$\frac{1}{N_q^* T_q^*} \sum_{i=1}^N \sum_{t=1}^T \mathbb{1}\{(i,t) \in \mathcal{O}_q^*\} \left\| \widehat{\lambda}_i^{(q)} - \widehat{\lambda}_{j(i)}^{(q)} \right\|^2 = O_P\left(\xi_{NT}\right)$ *for any matching function* $(j(i), t) \in \mathcal{O}_q$ *such that* $g_i^{(q)} = g_{j(i)}^{(q)}$, *and*
$\frac{1}{N_q^* T_q^*} \sum_{i=1}^N \sum_{t=1}^T \mathbb{1}\{(i,t) \in \mathcal{O}_q^*\} \left\| \widehat{f}_t^{(o)} - \widehat{f}_{s(t)}^{(o)} \right\|^2 = O_P\left(\xi_{NT}\right)$ *for any matching function* $(i, s(t)) \in \mathcal{O}_q$ *such that* $c_t^{(q)} = c_{s(t)}^{(q)}$.

[9] Note these are very similar to the $\tilde{A}$ variables in the main text, but here we make the distinction that projection is done at the partition level.

(vi) $\max_{k,i,t} \left|\check{X}_{it,k}\right| = O_P(1)$, *and* $\operatorname{plim}_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \check{X}_{it}' \check{X}_{it} = \Omega$, *where* Ω *is a positive definite non-random matrix.*

Proof of Lemma 3. Recall from the proof of Lemma 2 the definition of κ_{NT}. Take the split sample version as follows,

$$\begin{aligned}\kappa_{NT}^{(GS)} &:= \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \check{X}_{it}'\check{X}_{it}\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} \check{X}_{it}'\,\check{\Gamma}_{it} \\ &= \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \check{X}_{it}'\check{X}_{it}\right)^{-1} \sum_{o=1}^{4} \sum_{(i,t)\in\mathcal{O}_q} [\check{X}_{it}^{(q)}]'\,\check{\Gamma}_{it}^{(q)}.\end{aligned}$$

By Assumption A.1 and the proof steps of Lemma 2 we have that for each partition $\sum_{(i,t)\in\mathcal{O}_q} [\check{X}_{it}^{(q)}]'\,\check{\Gamma}_{it}^{(q)} = O_P(N_q T_q \xi_{NT})$, where N_q and T_q are the number of i and t, respectively, in partition q. Thus we have $\sum_{o=1}^{4} \sum_{(i,t)\in\mathcal{O}_q} [\check{X}_{it}^{(q)}]'\,\check{\Gamma}_{it}^{(q)} = \sum_{o=1}^{4} O_P(N_q T_q \xi_{NT}) \leq O_P(NT\xi_{NT})$. The statement of the lemma then follows from $\sum_{i=1}^{N}\sum_{t=1}^{T} \check{X}_{it}'\check{X}_{it} = O_P(NT)$. ■

Proof of Lemma 4. Using the definition of $\phi_{NT}^{(\mathrm{GS})}$ in the main text we have

$$\sqrt{NT}\,\phi_{NT}^{(\mathrm{GS})} := \widehat{\Omega}^{-1} \sum_{s=1}^{4} \phi_{NT}^{(s)}$$

where

$$\widehat{\Omega} := \frac{1}{NT} \sum_{s=1}^{4} \sum_{(i,t)\in\mathcal{O}_s} \widetilde{X}_{it}^{(s)\prime} \widetilde{X}_{it}^{(s)}, \qquad \phi_{NT}^{(s)} := \frac{1}{\sqrt{NT}} \sum_{(i,t)\in\mathcal{O}_s} \widetilde{X}_{it}^{(s)\prime}\,\varepsilon_{it}.$$

By construction, the projected regressors $\widetilde{X}_{it}^{(s)}$ for subpanel $s \in \{1,2,3,4\}$ only depend on $X = (X_{it})$, and on outcomes Y_{it} (and thus error terms ε_{it}) that are not in that subpanel, i.e. $(i,t) \notin \mathcal{O}_s$. Therefore, under Assumption 6(i), we have that for $s \in \{1,2,3,4\}$, conditional on $\{\widetilde{X}_{it}^{(s)} : (i,t) \in \mathcal{O}_s\}$, the $\widetilde{X}_{it}^{(s)\prime}\varepsilon_{it}$ are mean zero and independently distributed across all the observations $(i,t) \in \mathcal{O}_s$ in that subpanel. Using the regularity conditions in Assumption 6(ii), for each $s \in \{1,2,3,4\}$, we can therefore apply Lyapunov's CLT to find

$$\left(\widehat{\Sigma}^{(s)}\right)^{-1} \phi_{NT}^{(s)} \Rightarrow \mathcal{N}(0, \mathbb{1}_K), \qquad \widehat{\Sigma}^{(s)} := \sum_{(i,t)\in\mathcal{O}_s} \sigma_{it}^2 \widetilde{X}_{it}^{(s)\prime} \widetilde{X}_{it}^{(s)},$$

and the limiting distributions of $\left(\widehat{\Sigma}^{(s)}\right)^{-1} \phi_{NT}^{(s)}$ are independent across s. Using that $\widehat{\Sigma}^{(s)}$ converges to the constant $\Sigma^{(s)}$ we thus find that

$$\sum_{s=1}^{4} \phi_{NT}^{(s)} \Rightarrow \mathcal{N}\left(0, \sum_{s=1}^{4} \Sigma^{(s)}\right).$$

Since $\widehat{\Omega}$ converges to $\Omega > 0$, the continuous mapping theorem then gives the statement of the lemma. ■