Inference after discretizing unobserved heterogeneity

Jad Beyhum, Martin Mugnier

WORKING PAPER N° 2024-57

Inference after discretizing unobserved heterogeneity

Jad Beyhum
Martin Mugnier

JEL Codes:
Keywords: Unobserved heterogeneity, k-means clustering, panel data, double machine learning, inference.

Inference after discretizing unobserved heterogeneity*

Jad Beyhum
Department of Economics, KU Leuven, Belgium
and
Martin Mugnier
Paris School of Economics, Paris, France

December 10, 2024

Abstract

We consider a linear panel data model with nonseparable two-way unobserved heterogeneity corresponding to a linear version of the model studied in Bonhomme et al. (2022). We show that inference is possible in this setting using a straightforward two-step estimation procedure inspired by existing discretization approaches. In the first step, we construct a discrete approximation of the unobserved heterogeneity by (k-means) clustering observations separately across the individual (i) and time (t) dimensions. In the second step, we estimate a linear model with two-way group fixed effects specific to each cluster. Our approach shares similarities with methods from the double machine learning literature, as the underlying moment conditions exhibit the same type of bias-reducing properties. We provide a theoretical analysis of a cross-fitted version of our estimator, establishing its asymptotic normality at parametric rate under the condition $\max(N,T) = o(\min(N,T)^3)$. Simulation studies demonstrate that our methodology achieves excellent finite-sample performance, even when T is negligible with respect to N.

Keywords: Unobserved heterogeneity, k-means clustering, panel data, double machine learning, inference

*The authors thank Hugo Freeman for sharing his code of the paper Freeman and Weidner (2023). Jad Beyhum thanks Juan Carlos Escanciano for pushing him to work in the direction of double machine learning applied to panel data, Geert Dhaene for useful comments, and Martin Weidner for funding a research visit in Oxford. Jad Beyhum gratefully acknowledges financial support from the Research Fund KU Leuven through the grant STG/23/014. Martin Mugnier gratefully acknowledges financial support from the French National Research Agency (ANR) "Investissements d'Avenir" grant ANR-17-EURE-0001.

1 Introduction

Appropriately accounting for unobserved heterogeneity is a recurrent theme in much empirical and structural work in economics. In this paper, we consider the following panel data model, for $i = 1, \dots, N$ units and $t = 1, \dots, T$ dates,

$$y_{it} = x_{it}^\top \beta + f(\alpha_i, \gamma_t) + v_{it}, \tag{1}$$

$$x_{itk} = h_k(\alpha_i, \gamma_t) + u_{itk}, \ k = 1, \dots, K, \tag{2}$$

where $y_{it} \in \mathbb{R}$ is an observable dependent variable, $x_{it} = (x_{it1}, \dots, x_{itK})^\top \in \mathbb{R}^K$ is a vector of observable covariates, $\alpha_i \in \mathbb{R}^{K_\alpha}$ and $\gamma_t \in \mathbb{R}^{K_\gamma}$ are unobservable fixed effects, f and $(h_k)_{k\in\{1,\dots,K\}}$ are unknown deterministic mappings from $\mathbb{R}^{K_\alpha} \times \mathbb{R}^{K_\gamma}$ to $\mathbb{R}$, and $v_{it} \in \mathbb{R}$ and $u_{it} = (u_{it1}, \dots, u_{itK})^\top \in \mathbb{R}^K$ are unobservable mean-zero error terms uncorrelated with $(x_{it}^\top, f(\alpha_i, \gamma_t))^\top$ and $(h_k(\alpha_i, \gamma_t))_{k\in\{1,\dots,K\}}$, respectively. The outcome contribution of the fixed effects, $f(\alpha_i, \gamma_t)$, may be flexibly correlated with the covariates through the contributions $(h_k(\alpha_i, \gamma_t))_{k\in\{1,\dots,K\}}$. This paper focuses on the estimation of and inference for the parameter $\beta \in \mathbb{R}^K$, a regression coefficient when appropriately controlling for the unobserved $f(\alpha_i, \gamma_t)$ in the pooled ordinary least-squares (OLS) regression of y_{it} on x_{it}. The numbers K, K_α, and K_γ of covariates, unit-specific and time-specific fixed effects do not vary with N and T and we consider an asymptotic regime where N and T grow to infinity. Some fixed effects may enter the first equation but not the second or vice versa.

This is a semiparametric linear panel data model with possibly continuous nonseparable two-way unobserved heterogeneity. Bonhomme et al. (2022) study a likelihood version of the model, replacing Equation (1) by a parametric specification of the density of y_{it} given x_{it} as a function of the parameters α_i, γ_t, and β. Such a model is also related to the literature on panel data models with interactive fixed effects. Bai (2009) posits Equation (1) but not (2) and imposes that f is known and interactive. Pesaran (2006), Greenaway-McGrevy et al. (2012), and Westerlund and Urbain (2015) consider our model and assume that both f and $(h_k)_{k\in\{1,\dots,K\}}$ are interactive and therefore known. Finally, Freeman and Weidner (2023) study Model (1) without specifying the link between x_{it} and the fixed effects as in Equation (2). With two-way panel data, such a nonlinear factor model is arguably one of the most flexible models one can think of for the unobserved heterogeneity. Since economic theory rarely provides foundations for the additive or interactive separability of unobserved

heterogeneity, it seems desirable to seek for the weakest possible form.[1]

Despite the growing interest in this type of model, to the best of our knowledge, no proven inference procedure has been proposed.[2] Bonhomme et al. (2022) and Freeman and Weidner (2023) derive rates of convergence for their estimators but remain short of proving asymptotic normality. In this paper, we show that inference at parametric rates is possible in our model. We consider the following two-step estimation procedure. In the first step, we follow Bonhomme et al. (2022) and construct a discrete approximation of the unobserved heterogeneity by (k-means) clustering observations separately across the individual (i) and time (t) dimensions. In the second step, we follow the approach of Freeman and Weidner (2023) and estimate a linear model with additively separable two-way group fixed effects specific to each cluster.

Our procedure is thus definitely inspired, though different, from those considered in Bonhomme et al. (2022) and Freeman and Weidner (2023). Similarly to Bonhomme et al. (2022), we cluster units and dates in the first step. However, our second step is an OLS regression with additively separable two-way group fixed effects while Bonhomme et al. (2022) use maximum likelihood with nonseparable one-way (or two-way) group fixed effects. Freeman and Weidner (2023) start by computing Bai (2009)'s interactive fixed effect estimates before clustering the estimated loadings and factors and finally estimating a linear model with additively separable two-way group fixed effects specific to each cluster.[3] In comparison with the approaches of Bonhomme et al. (2022) and Freeman and Weidner (2023), the main advantage of our procedure is that our second step linear regression relies on a bias-reducing moment as popularized by the literature on double machine learning (see Chernozhukov et al., 2018). This robust moment is obtained through the model (2) on x_{it}, which Freeman and Weidner (2023) do not impose, see Section 2.2 for more discussion.[4]

[1] While the additive separability between the covariates and the unobservable random variables is central to the results obtained in this paper, the additive separability between each fixed effect transformation and the corresponding error term is without loss of generality if α_i, γ_t are identically distributed across i and over t and the error term is independent from the fixed effects and the covariates (consider the L_2 projection on the space of fixed effects and its residual). See the introduction of Freeman and Weidner (2023) for a more formal discussion.

[2] When f and $(h_k)_{k\in\{1,\dots,K\}}$ are additive or multiplicative in their arguments, the literature on panel data models with two-way or interactive fixed effects does provide inference procedures. Such results, however, are not valid in our more general case where f and $(h_k)_{k\in\{1,\dots,K\}}$ are unknown.

[3] Freeman and Weidner (2023, page 5) note that improved rates of convergence can be obtained by using an additively separable two-way group fixed effects estimator rather than using a nonseparable one-way (or two-way) group fixed effects specific to the intersection of unit and time clusters as done in Bonhomme et al. (2022). We follow the route of Freeman and Weidner (2023) in that respect.

[4] Notably, the estimator in Freeman and Weidner (2023) does not rely on an orthogonal moment even when Model (2) holds.

Two important consequences are that it is possible to show asymptotic normality, as we discuss below, and the use of Bai (2009)'s estimator in the first step of the estimation is not necessary.

Deriving the asymptotic distribution of estimators based on first-step black box methods such as k-means clustering or other machine learning tools is challenging. The main difficulty arises because it is difficult in such cases to obtain a precise control of the stochastic relationship between the error terms and the first-step estimates. To overcome this issue, we borrow a popular approach from the double machine learning literature and consider a cross-fitted version of our estimator. We obtain asymptotic normality of the cross-fitted estimator under the condition $\max(N,T) = o(\min(N,T)^3)$. This is weaker than the typical requirement for asymptotic normality in interactive fixed effects models, that is, $\max(N,T) = o(\min(N,T)^2)$, see Bai (2009) and Westerlund and Urbain (2015). This relaxation is particularly noteworthy as it is achieved in a setting that also weakens the assumptions about g and $(h_k)_{k=1,\dots,K}$, since they do not need to be interactive and known.[5] Importantly, as in the double machine learning literature, cross-fitting only serves as a theoretical device, and our simulations indicate that cross-fitting may not be needed in practice.

Our paper contains a Monte Carlo study evaluating our estimator and its cross-fitted version. We find that our baseline estimator has excellent finite sample properties and over-performs its cross-fitted variant, which itself substantially improves over benchmark estimators. Notably, our estimators have almost nominal coverage even when T is very low compared to N, which is of practical importance since many real-world datasets exhibit this feature.

Finally, we would like to note that the approach studied in this paper is related but different from the literature on panel data models with grouped fixed effects (see Bonhomme and Manresa, 2015; Chetverikov and Manresa, 2022; Mugnier, 2024, among many others). In this literature, it is assumed that the grouped fixed effects structure is exact. Here, we do not assume that the fixed effects follow a group pattern but instead use clusters as an approximation device.

Outline. This paper is organized as follows. Section 2 introduces our estimator. Our asymptotic results are presented in Section 3. Section 4 reports Monte Carlo simulations.

[5] See the discussion in Section 3, following Assumption 4, for an explanation of the reasons enabling the relaxation of restrictions on N and T.

All proofs can be found in the appendix.

2 Estimator and link to double machine learning

In this section, we introduce the estimation approach afterward analyzed and used in the rest of the paper. We also discuss its connection to procedures developed in the double machine learning literature.

2.1 Estimator

Let $z_{it} := (x_{it}^\top, y_{it})^\top$. By plugging (2) into (1), we have

$$\begin{aligned} z_{itk} &= h_k(\alpha_i, \gamma_t) + u_{itk}, \ \ k = 1, \ldots, K, \\ z_{it(K+1)} &= h_{K+1}(\alpha_i, \gamma_t) + e_{it}, \end{aligned}$$

where $h_{K+1}(\alpha_i, \gamma_t) = f(\alpha_i, \gamma_t) + \sum_{k=1}^K \beta_k h_k(\alpha_i, \gamma_t)$ and $e_{it} = u_{it}^\top \beta + v_{it}$. Under the contemporaneous exogeneity assumption, $\mathbb{E}[u_{it} v_{it}] = 0$, which we maintain hereafter (see Assumption 3 below), a natural procedure is to first estimate e_{it} and u_{it} and then linearly regress the estimates of e_{it} on that of u_{it} to obtain an estimator of β. This procedure can be expected to deliver a reasonable estimator if an analogous orthogonality condition holds, at least asymptotically, for the estimated quantities (see Section 2.2 below for more details). Below, we describe the two main steps of the estimation procedure.

Step 1 (Two-way clustering). To estimate e_{it} and u_{it}, we start by constructing a discrete approximation of unobserved heterogeneity across units and dates using time-series or cross-sectional averages of the data.[6] This approach can be expected to perform well if such averages are informative about the underlying unobserved heterogeneity in a way that can be exploited by the discretization method (see Section 3.3 below). Below, we outline one possible approach based on the popular k-means clustering algorithm.[7] Let G

[6] An alternative approach would be to discretize solely across one dimension (either units or dates). However, as noted in Bonhomme et al. (2022) and Freeman and Weidner (2023), this leads to slower rates of convergence. See also Beyhum and Gautier (2023) for a similar argument in panel data models with interactive fixed effects.

[7] We could potentially use other clustering algorithms, but our theory and the reported simulations concern the case where k-means is used. In unreported simulations, we have found that using hierarchical clustering leads to similar results.

and C denote the number of unit and time groups, respectively (a rule to select them is outlined below). Let $\|\cdot\|$ denote the Euclidian norm.

Clustering algorithm for units. Let $a_i := \frac{1}{T}\sum_{t=1}^{T} z_{it},\ i \in \{1,\dots,N\}$. We use the algorithm

$$(\widehat{a}(1),\dots,\widehat{a}(G),g_1,\dots,g_N) \in \underset{\substack{a(1),\dots,a(G)\in\mathbb{R}^{K+1}\\ \tilde{g}_1,\dots,\tilde{g}_N\in\{1,\dots,G\}}}{\arg\min} \sum_{i=1}^{N} \|a_i - a(\tilde{g}_i)\|^2 .$$

Clustering algorithm for dates. Let $b_t := \frac{1}{N}\sum_{i=1}^{N} z_{it},\ t \in \{1,\dots,T\}$. We use the algorithm

$$\left(\widehat{b}(1),\dots,\widehat{b}(C),c_1,\dots,c_T\right) \in \underset{\substack{b(1),\dots,b(C)\in\mathbb{R}^{K+1}\\ \tilde{c}_1\dots,\tilde{c}_T\in\{1,\dots,C\}}}{\arg\min} \sum_{t=1}^{T} \|b_t - b(\tilde{c}_t)\|^2 .$$

Including both y_{it} and x_{it} as inputs of each clustering algorithm is crucial, as some fixed effects could enter Equation (1) but not Equation (2) or vice versa, and these fixed effects would not be accounted for by clustering solely on either x_{it} or y_{it}.

Fast computational routines exist to find exact solutions to both clustering problems for data sets of moderate sizes (e.g., du Merle et al., 1997; Aloise et al., 2009) and local minima for others (e.g., Hartigan–Wong's algorithm). If the quality of the local minima raises suspicion, we recommend using hierarchical clustering approaches (see, e.g., Section 14.3.12 in Hastie et al., 2009) as a sensitivity analysis, though we leave the verification of its approximation properties for further research.

Step 2 (Two-way group fixed effect estimator). The estimators of e_{it} and u_{it} are

$$\widehat{e}_{it} := y_{it} - \bar{y}_{g_i t} - \bar{y}_{i c_t} + \bar{y}_{g_i c_t},$$
$$\widehat{u}_{it} := x_{it} - \bar{x}_{g_i t} - \bar{x}_{i c_t} + \bar{x}_{g_i c_t},$$

where, for any variable w_{it}, we define

$$\bar{w}_{g_i t} := \frac{1}{N_{g_i}} \sum_{j=1}^{N} \mathbf{1}\{g_j = g_i\} w_{jt},$$
$$\bar{w}_{i c_t} := \frac{1}{T_{c_t}} \sum_{s=1}^{T} \mathbf{1}\{c_s = c_t\} w_{is},$$
$$\bar{w}_{g_i c_t} := \frac{1}{N_{g_i} T_{c_t}} \sum_{j=1}^{N} \sum_{s=1}^{T} \mathbf{1}\{g_j = g_i\} \mathbf{1}\{c_s = c_t\} w_{js},$$

with $N_{g_i} := \sum_{j=1}^{N} \mathbf{1}\{g_j = g_i\}$ and $T_{c_t} := \sum_{s=1}^{T} \mathbf{1}\{c_s = c_t\}$. These estimators correspond to within-group transformations applied to y_{it} and x_{it} in a similar fashion to the standard within transformations in standard linear panel data models with two-way fixed effects. The final estimator of β is the ordinary least squares estimator

$$\widehat{\beta} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{u}_{it} \widehat{u}_{it}^{\top} \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{u}_{it} \widehat{e}_{it}.$$

Note that $\widehat{\beta}$ can be reformulated as the following two-way group fixed effects estimator

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^K} \min_{\delta \in \mathbb{R}^{N \times C}} \min_{\nu \in \mathbb{R}^{T \times G}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - x_{it}^{\top} \beta - \delta_{i,c_t} - \nu_{g_i,t} \right)^2.$$

An appealing feature of the two-way group fixed effects estimator is that, since it relies on linear regression, usual standard errors (with a degree of freedom correction) can be used.[8]

Choice of the number of clusters. In practice, we use the data-driven choice of the number of clusters developed by Bonhomme et al. (2022) to select G and C. Let us outline the procedure. First, let the k-means objective functions be

$$Q_g(G) := \frac{1}{N} \sum_{i=1}^{N} \left\| a_i - \widehat{a}(g_i) \right\|^2,$$
$$Q_c(C) := \frac{1}{T} \sum_{t=1}^{T} \left\| b_t - \widehat{b}(c_t) \right\|^2.$$

[8] To be precise, our contribution here is to combine the clustering technique of Bonhomme et al. (2022) with the second estimation step of Freeman and Weidner (2023). Given the output of any given clustering procedure $(g_1, \ldots, g_N, c_1, \ldots, c_T)$, the two-way group fixed effects estimator was first proposed and analyzed, to the best of our knowledge, in Freeman and Weidner (2023).

The quantities $Q_g(G)$ and $Q_c(C)$ measure the approximation errors made through the clustering. Let us also define the following empirical variances which measure the degree of noise stemming from the inputs of the clustering procedures:

$$\widehat{V}_g := \frac{1}{NT^2} \sum_{i=1}^{N} \sum_{t=1}^{T} \|z_{it} - a_i\|^2 ,$$
$$\widehat{V}_c := \frac{1}{N^2T} \sum_{t=1}^{T} \sum_{i=1}^{N} \|z_{it} - b_t\|^2 .$$

The data-driven choice of the number of clusters balances the approximation error and the input noise in the following manner:

$$\widehat{G} := \min_{G \geq 1} \{G : \ Q_g(G) \leq \widehat{V}_g\},$$
$$\widehat{C} := \min_{C \geq 1} \{C : \ Q_c(C) \leq \widehat{V}_c\},$$

where $\widehat{G}$ and $\widehat{C}$ are the chosen number of unit and time clusters, respectively. We refer to Bonhomme et al. (2022) for more details on this selection procedure, including some off-the-shelf theoretical guarantees that are valid without modification in our setting.

2.2 Link to double machine learning

The double machine learning literature relies on a two-step estimation procedure where the second-step is based on a Neyman-orthogonal moment (Chernozhukov et al., 2018). Such moments are bias-reducing because they limit the influence of the error in estimating the nuisance parameters in the first step and, therefore, make inference possible. This robustness property arises because the difference between the empirical version of the Neyman-orthogonal moment and the empirical moment from an oracle estimator knowing the nuisance parameters is composed of terms which are the sums either of products of estimation errors or products of an estimation error and an error term, see in particular the discussion in Section 1 of Chernozhukov et al. (2018). It turns out that the moment on which our second-step estimator is based exhibits the same type of robustness properties. To see this, note that our second-step estimator solves the following empirical moment

$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{u}_{it}(\widehat{e}_{it} - \widehat{u}_{it}^{\top}\beta) = 0. \tag{3}$$

Moment (3) approximates the empirical moment

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} u_{it}(e_{it} - u_{it}^{\top}\beta) = 0, \tag{4}$$

solved by an infeasible "oracle" OLS estimator knowing u_{it} and e_{it}. Notice that

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \widehat{u}_{it}(\widehat{e}_{it} - \widehat{u}_{it}^{\top}\beta) - \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} u_{it}(e_{it} - u_{it}^{\top}\beta) = a^* + b^* + c^*,$$

where

$$\begin{aligned}
a^* &:= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\widehat{u}_{it} - u_{it})(\widehat{e}_{it} - e_{it} - (\widehat{u}_{it} - u_{it})^{\top}\beta),\\
b^* &:= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\widehat{u}_{it} - u_{it})v_{it},\\
c^* &:= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} u_{it}(\widehat{e}_{it} - e_{it} - (\widehat{u}_{it} - u_{it})^{\top}\beta).
\end{aligned}$$

Hence, the difference between the moments (3) and (4) is the sum of a term a^*, corresponding to the sum of the products of two estimation errors, and two terms b^* and c^* which are sums of products of an estimation error and an error term. All of these terms are, therefore, sums of products of "small terms" and will thus be asymptotically negligible. This explains why our estimator can be asymptotically normal.

In contrast, Freeman and Weidner (2023) solve the empirical moment

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}(y_{it} - \check{f}_{it} - x_{it}^{\top}\beta) = 0, \tag{5}$$

where $\check{f}_{it}$ is some estimator of $f_{it} := f(\alpha_i, \gamma_t)$. Their rationale is that (5) approximates the moment

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}(y_{it} - f_{it} - x_{it}^{\top}\beta) = 0, \tag{6}$$

solved by an infeasible "oracle" OLS estimator knowing f_{it}. However, the difference between

(5) and (6) is

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}(y_{it} - \check{f}_{it} - x_{it}^{\top}\beta) - \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}(y_{it} - f_{it} - x_{it}^{\top}\beta)$$
$$= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}(\check{f}_{it} - f_{it}) =: d^*,$$

which is the sum of the product of covariates and an estimation error $\check{f}_{it} - f_{it}$ which is likely to be correlated with x_{it}. Hence, it will not be possible in general to show that $d^* = o_P(1/\sqrt{NT})$, which explains why inference does not work with the moments from Freeman and Weidner (2023). We note, however, that the moment (3) that we use is only valid because of the model (2) on x_{it}, which is not imposed by Freeman and Weidner (2023).

3 Asymptotic theory

This section provides theoretical guarantees for a cross-fitted variant of our estimator. In Section 3.1, we motivate and discuss the use of cross-fitting. Section 3.2 introduces the cross-fitted version of the two-step estimator. Section 3.3 provides sufficient conditions for its asymptotic normality. Section 3.4 formally presents our large sample results.

3.1 On the use of cross-fitting

Deriving the limiting distribution of the least-squares estimator $\widehat{\beta}$ is challenging, as it requires controlling the dependence between the clusters estimated in the first step and the error terms of the data used in the second step. This difficulty is a common feature of many two-step estimators based on some highly nonlinear black-box first-step estimators.[9]

This type of issue has also been encountered in the literature on double machine learning (Chernozhukov et al., 2018). The solution taken in this research area is to use cross-fitting. The data is split into different folds, and the first-step and second-step estimations are performed on different folds. The role of the folds is then reversed, and the second-step estimators over the different folds are averaged to improve efficiency. Under independent observations, this mechanically eliminates the dependence between the first-step estimator and the data used in the second step, therefore solving the aforementioned problem.

[9] In particular, without a control of the dependence between the two steps, one cannot use concentration arguments on u_{it} and v_{it} to bound the terms b^* and c^* introduced in Section 2.2.

In this section, we follow this strategy to establish the asymptotic normality at the parametric $\sqrt{NT}$-rate of a cross-fitted version of our estimator that learns clusters and estimates the slope coefficient from separate batches of the data. Thus, we improve on the asymptotic expansions derived in Bonhomme et al. (2022) in a semiparametric linear version of their model, and in Freeman and Weidner (2023) by adding a similar nonseparable structure on the covariates. We note that Freeman and Weidner (2023) also study a cross-fitted version of their estimator, for which they are able to derive better asymptotic properties (but no inference procedure is proven to be valid).

The cross-fitting exercise is more a proof device than a recommendation. Monte Carlo simulations in Section 4 demonstrate improved performance for the original estimator $\widehat{\beta}$ over its cross-fitted version which already performs reasonably well. We note that the cross-fitted estimator in Freeman and Weidner (2023) exhibits the same type of behavior: while the authors derive theoretical properties for the cross-fitted estimator, the latter exhibits poorer performance relative to the estimator without cross-fitting in simulations.

Cross-fitting has been shown to reduce estimator performance in simulations across various settings (Dukes and Vansteelandt, 2021; Chen et al., 2022; Vansteelandt et al., 2024; Wang et al., 2024; Shi et al., 2024). Moreover, it has been demonstrated that cross-fitting is not always essential for achieving asymptotic results in double machine learning when the learners adhere to a natural leave-one-out stability property (Chen et al., 2022) or the lasso is used (Chernozhukov et al., 2015). These findings suggest that in certain contexts, cross-fitting is not only unnecessary but may even be counterproductive. Our simulation results indicate that k-means clustering is one such learner where cross-fitting can be omitted without compromising performance.

3.2 Alternative estimator with cross-fitting

Let us outline our alternative estimator based on cross-fitting.

Cross-fitting. To simplify, we consider only four folds:

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

We also use the notation $N_d := |\mathcal{N}_d|$ and $T_d := |\mathcal{T}_d|$. This type of division in four folds is appropriate for panel data and also appears in Freeman and Weidner (2023).[10]

Estimation algorithm. As for $\widehat{\beta}$, we estimate the group memberships in the first step via a clustering method and then compute an OLS estimator. The main difference is that the data used in each of these two steps do not intersect but the final estimator still uses variation across the full dataset. The estimation procedure to obtain the resulting cross-fitted two-way group fixed effect estimator is as follows.

For each fold, $d \in \{1, \ldots, 4\}$:

1. Apply k-means clustering to the data in $\{\mathcal{O}_{\tilde{d}}\}$, where

$$\tilde{d} = \begin{cases} 2 & \text{if } d = 1, \\ 1 & \text{if } d = 2, \\ 4 & \text{if } d = 3, \\ 3 & \text{if } d = 4, \end{cases}$$

   to obtain the unit cluster indicators $g_i^d \in \{1, \ldots, G_d\}$.

2. Apply k-means clustering to the data in $\{\mathcal{O}_{\tilde{d}}\}$, where

$$\tilde{d} = \begin{cases} 3 & \text{if } d = 1, \\ 4 & \text{if } d = 2, \\ 1 & \text{if } d = 3, \\ 2 & \text{if } d = 4, \end{cases}$$

   to obtain the time cluster indicators $c_t^d \in \{1, \ldots, C_d\}$.

[10] In unreported simulations, we have not found any improvement resulting from increasing the number of folds.

3. Estimate e_{it} and u_{it} on fold d by $\widehat{e}_{it}^d := y_{it} - \bar{y}_{g_i^d t} - \bar{y}_{ic_t^d} + \bar{y}_{g_i^d c_t^d}$ and $\widehat{u}_{it}^d := x_{it} - \bar{x}_{g_i^d t} - \bar{x}_{ic_t^d} + \bar{x}_{g_i^d c_t^d}$, and where, for any variable w_{it}, we define

$$\bar{w}_{g_i^d t} := \frac{1}{N_{g_i^d}^d} \sum_{j \in \mathcal{N}_d} \mathbf{1}\{g_j^d = g_i^d\} w_{jt},$$
$$\bar{w}_{ic_t^d} := \frac{1}{T_{c_t^d}^d} \sum_{s \in \mathcal{T}_d} \mathbf{1}\{c_s^d = c_t^d\} w_{is},$$
$$\bar{w}_{g_i^d c_t^d} := \frac{1}{N_{g_i^d}^d T_{c_t^d}^d} \sum_{(j,s) \in \mathcal{O}_d} \mathbf{1}\{g_j^d = g_i^d\} \mathbf{1}\{c_s^d = c_t^d\} w_{js},$$

with $N_{g_i^d}^d := \sum_{j \in \mathcal{N}_d} \mathbf{1}\{g_j^d = g_i^d\}$ and $T_{c_t^d}^d := \sum_{s \in \mathcal{T}_d} \mathbf{1}\{c_s^d = c_t^d\}$.

The final estimator is

$$\widehat{\beta}^{\mathrm{CF}} := \left(\sum_{d=1}^4 \sum_{(i,t) \in \mathcal{O}_d} \widehat{u}_{it}^d (\widehat{u}_{it}^d)^\top \right)^{-1} \sum_{d=1}^4 \sum_{(i,t) \in \mathcal{O}_d} \widehat{u}_{it}^d \widehat{e}_{it}^d,$$

that is, the linear regression of the $\widehat{e}_{it}^d$ on the $\widehat{u}_{it}^d$. This estimator can equivalently be formulated as

$$\widetilde{\beta}^{\mathrm{CF}} = \arg\min_{\beta \in \mathbb{R}^K} \min_{\delta} \min_{\nu} \sum_{i=1}^N \sum_{t=1}^T \left[y_{it} - x_{it}^\top \beta - \sum_{d=1}^4 \mathbf{1}\{(i,t) \in \mathcal{O}_d\} \left(\delta_{i,c_t^d}^d + \nu_{t,g_i^d}^d \right) \right]^2. \tag{7}$$

In summary, to obtain the unit cluster indicators $g_i^d \in \{1, \ldots, G_d\}$ (resp. the time cluster indicators $c_t^d \in \{1, \ldots, C_d\}$), we use the fold that contains the same units as $\mathcal{O}_d$ but different dates (resp. the same dates as $\mathcal{O}_d$ but different units). A similar trick is used for clustering time periods. We then use these clusters to estimate e_{it} and u_{it}, before running a linear regression on such estimates.[11]

The clustering steps are carried out using straightforward adapted versions of the algorithm introduced in Section 2, which we display below for completeness. For all

[11] As for our baseline estimator, the second step (7) of our cross-fitted estimator corresponds to the second step of the cross-fitted estimator in Freeman and Weidner (2023). The clustering steps differ between the two papers.

$d \in \{1, \ldots, 4\}$, the clustering algorithm is applied to the two empirical averages:

$$a_i^d := \frac{1}{T_{\tilde{d}}} \sum_{t \in \mathcal{T}_{\tilde{d}}} z_{it}, \quad i \in \mathcal{N}_d,$$
$$b_t^d := \frac{1}{N_{\tilde{d}}} \sum_{i \in \mathcal{N}_{\tilde{d}}} z_{it}, \quad t \in \mathcal{T}_d.$$

Clustering algorithm for units. We use the algorithm

$$\left(\widehat{a}^d(1), \ldots, \widehat{a}^d(G_d), \{g_i^d,\ i \in \mathcal{N}_d\}\right) \in \underset{\substack{a(1), \ldots, a(G_d) \in \mathbb{R}^{K+1} \\ g_i \in \{1, \ldots, G_d\},\ i \in \mathcal{N}_d}}{\arg\min} \sum_{i \in \mathcal{N}_d} \left\| a_i^d - a(g_i) \right\|^2.$$

Clustering algorithm for dates. We use the algorithm

$$\left(\widehat{b}^d(1), \ldots, \widehat{b}^d(C_d), \{c_t^d,\ t \in \mathcal{T}_d\}\right) \in \underset{\substack{b(1), \ldots, b(C_d) \in \mathbb{R}^{K+1} \\ c_t \in \{1, \ldots, C_d\},\ t \in \mathcal{T}_d}}{\arg\min} \sum_{t \in \mathcal{T}_d} \left\| b_t^d - b(c_t) \right\|^2.$$

In practice, we use the data-driven rule outlined in Section 2 to select the number of clusters G^d and C^d in the different folds $d \in \{1, \ldots, 4\}$.

3.3 Assumptions

Consider the following assumptions.

Assumption 1 (Heterogeneity) *The functions $(h_k)_{k \in \{1, \ldots, K+1\}}$ are bounded and twice differentiable with second-order derivatives bounded uniformly in the support of (α_i, γ_t).*

Assumption 2 (Injectivity) *For all $d \in \{1, \ldots, 4\}$:*

(*i*) *There exists a Lipschitz-continuous function φ_d^α such that $\frac{1}{N_d} \sum_{i \in \mathcal{N}_d} \left\| a_i^d - \varphi_d^\alpha(\alpha_i) \right\|^2 = O_P(1/T)$ as N, T tend to infinity. Moreover, there exists a Lipschitz-continuous function ψ_d^α such that, for all $i \in \mathcal{N}_d$, $\alpha_i = \psi_d^\alpha(\varphi_d^\alpha(\alpha_i))$.*

(*ii*) *There exists a Lipschitz-continuous function φ_d^γ such that $\frac{1}{T_d} \sum_{t \in \mathcal{T}_d} \left\| b_t^d - \varphi_d^\gamma(\gamma_t) \right\|^2 = O_P(1/N)$ as N, T tend to infinity. Moreover, there exists a Lipschitz-continuous function ψ_d^γ such that, for all $t \in \mathcal{T}_d$, $\gamma_t = \psi_d^\gamma(\varphi_d^\gamma(\gamma_t))$.*

Assumption 1 is a mild regularity condition on $(h_k)_{k\in\{1,\dots,K+1\}}$. Assumption 2 is similar to Assumption 2 in Bonhomme et al. (2022). It is best understood in the case of pointwise limits, where $\text{plim}_{T\to\infty} a_i^d = \varphi_d^\alpha(\alpha_i)$ and $\text{plim}_{N\to\infty} b_t^d = \varphi_d^\gamma(\gamma_t)$, which can be justified by laws of large numbers. In this case, Assumption 2 requires that individuals (resp. time periods) with similar values of time-series (resp. cross-sectional) averages of z_{it} have similar values of unit-specific (resp. time-specific) fixed effects and vice versa, with equality in the limit. Intuitively, such an injectivity property suggests that matching on observed panel data averages may be sufficient to control for unobserved heterogeneity (i.e., matching on the fixed effects). Such ideas have been exploited in Zeleneev (2020), Bonhomme et al. (2022), and Freeman and Weidner (2023), among others.

The following assumption rules out any cross-section or time-series dependence of the error terms.

Assumption 3 (Errors)

(*i*) *The error terms* u_{it} *and* v_{it} *are i.i.d. across* i *and* t *with finite variance.*

(*ii*) $\mathbb{E}[u_{11}] = \mathbb{E}[v_{11}] = 0$.

(*iii*) *The covariance matrix* $\Sigma_U := \mathbb{E}[u_{11}u_{11}^\top]$ *is positive definite.*

(*iv*) *The error terms* $\{(u_{it}, v_{it}),\ i \in \{1,\dots,N\},\ t \in \{1,\dots,T\}\}$ *are independent of* $\{(\alpha_i, \gamma_t),\ i \in \{1,\dots,N\},\ t \in \{1,\dots,T\}\}$.

(*v*) *The error terms* $\{v_{it},\ i \in \{1,\dots,N\},\ t \in \{1,\dots,T\}\}$ *are independent of* $\{(u_{it}, \alpha_i, \gamma_t),\ i \in \{1,\dots,N\},\ t \in \{1,\dots,T\}\}$.

Assumption 3(i) implies that the data from the different folds are independent. It is arguably strong, but relaxing it would require obtaining a precise control of the dependence between the outcome of the clustering algorithm and the error terms, which, as mentioned earlier, is quite challenging with black-box methods such as k-means. In the simulations reported in Section 4, we find that our estimator still performs very well under time series correlation. We note that the assumption of i.i.d. errors is often made in papers studying sophisticated panel data models; see, for instance, Moon and Weidner (2015), Chen et al. (2021), Assumption S2(i) in Bonhomme et al. (2022), and Freeman and Weidner (2023). Similarly to us, these papers derive their main theoretical results with this assumption but present simulation evidence showing that the restriction is not necessary.

Assumption 3(ii) requires that errors have zero mean. Assumption 3(iii) is a standard non-collinearity condition on the covariates in the second-step regression. Assumption 3(iv) requires that the error terms are independent of the fixed effects, a standard assumption in the panel data literature. Assumption 3(v) stipulates that the error term v_{it} of the outcome equation is jointly independent of the error terms u_{it} of (2) and the fixed effects.

The following assumption specifies the relative rates at which N, T, and the numbers of clusters G_d and C_d can grow.

Assumption 4 (Asymptotics) *For all $d \in \{1, \ldots, 4\}$, as N, T, G_d, C_d tend to infinity,*

(i) $\max(N,T) = o(\min(N,T)^3)$.

(ii) $G_d = o(N)$, $C_d = o(T)$.

Assumption 4(i) is weaker than the rate conditions on N and T typically found in the literature on panel data models with interactive fixed effects, that is $\max(N,T) = o(\min(N,T)^2)$, see Bai (2009) and Westerlund and Urbain (2015). This is an important result because the improvement is obtained while relaxing the modeling assumption that g and h_k are interactive and known. We relax the condition in Bai (2009) thanks to the use of an orthogonal moment coming from (2), while we improve on Westerlund and Urbain (2015) because we estimate both unit and time-specific fixed effects in the first step, while Westerlund and Urbain (2015) only estimates the factors (corresponding to the time-specific fixed effects in an interactive fixed effects model), see also Footnote 6 for a related discussion. In contrast, the rate condition (i) is stronger than that for grouped fixed effects models such as in Bonhomme and Manresa (2015), where T can grow at an arbitrary polynomial rate with respect to N. This is because we do not assume that the data has a group structure, and instead, we only use clustering as an approximation device.

Assumption 4 (ii) stipulates that the number of unit clusters (resp. time clusters) must be negligible with respect to N (resp. T). Intuitively, this is necessary because, otherwise, the within transformations applied to the data to estimate e_{it} and u_{it} would create non-negligible time series and cross-section dependence in the generated regressors of the second step, precluding the estimator from being $\sqrt{NT}$-consistent.

The last assumption concerns the approximation error of an infeasible "oracle" approximation procedure that would directly cluster the unobserved unit and time fixed effects. We follow Bonhomme et al. (2022) and define such approximation errors as, for all

$d \in \{1, \ldots, 4\}$,

$$B_\alpha^d(G_d) = \min_{\substack{\alpha(1), \ldots, \alpha(G_d) \in \mathbb{R}^{K_\alpha} \\ \tilde{g}_i \in \{1, \ldots, G_d\},\ i \in \mathcal{N}_d}} \sum_{i \in \mathcal{N}_d} \|\alpha_i - \alpha(\tilde{g}_i)\|^2$$

and

$$B_\alpha^d(G_d) \min_{\substack{\gamma(1), \ldots, \gamma(C_d) \in \mathbb{R}^{K_\gamma} \\ \tilde{c}_t \in \{1, \ldots, C_d\},\ t \in \mathcal{T}_d}} \sum_{t \in \mathcal{T}_d} \|\gamma_t - \gamma(\tilde{c}_t)\|^2 .$$

Lemma 2 in Section 3.4 below suggests that, due to the injectivity condition (Assumption 2), the k-means clustering algorithm used in the first step achieves an approximation error close to the infeasible oracle k-means algorithm (that is $B_\alpha^d(G_d)$, $B_\gamma^d(C_d)$). Next, we require this approximation error of our clustering algorithm to be small enough for our estimator to be asymptotically normal. This is subsumed in the next assumption below.

Assumption 5 (Approximation error) *For all $d \in \{1, \ldots, 4\}$, as N, T, G_d, C_d tend to infinity,*

$$B_\alpha^d(G_d) = o_P\left((NT)^{-1/4}\right) \text{ and } B_\gamma^d(C_d) = o_P\left((NT)^{-1/4}\right).$$

Assumption 5 requires the oracle approximation error resulting from discretizing the unobserved heterogeneity to decrease sufficiently fast as the sample size increases. Intuitively, this condition requires the number of clusters to increase at a rate governed by the difficulty of the approximation problem, which itself depends on the the dimensions of the fixed effects K_α and K_γ. As discussed in Freeman and Weidner (2023) and Bonhomme et al. (2022), a precise dependence of the approximation error on K_α and K_γ can be obtained under further regularity conditions on the distribution of α_i and γ_t.

Lemma 1 (Graf and Luschgy (2002)) *Suppose that α_i and γ_t are i.i.d. with compact supports. Then, for all $d \in \{1, \ldots, 4\}$, as N, T, G_d, C_d tend to infinity we have*

$$B_\alpha^d(G_d) = O_P\left((G_d)^{-\frac{2}{K_\alpha}}\right) \text{ and } B_\gamma^d(C_d) = O_P\left((C_d)^{-\frac{2}{K_\gamma}}\right).$$

Lemma 1 shows that the approximation error decreases at a rate inversely proportional to the dimension of the underlying fixed effects. The assumption that α_i and γ_t are i.i.d with compact support is only a sufficient condition that may not be necessary. While it may

be restrictive for some applications (if $N = T$, it implies that α_i and γ_t are independent for all $i \neq \sigma_{NT}(t)$ for some permutation σ_{NT} of $\{1, \ldots, N\}$) and the result might hold under departures from this assumption, proving the validity of such an extension is beyond the scope of this paper. In our Monte Carlo study, our estimator continues to perform well when the time-specific fixed effects exhibit autocorrelation and have an unbounded support. We note that the assumption of i.i.d. fixed effects with compact support is invoked in Assumption S2(i) in Bonhomme et al. (2022). Using Lemma 1, we obtain the following corollary which gives sufficient conditions for Assumption 5.

Corollary 1 *Suppose that α_i and γ_t are i.i.d. with compact supports. Then, Assumption 5 holds if for all $d \in \{1, \ldots, 4\}$, as N, T, G_d, C_d tend to infinity, we have*

$$(NT)^{K_\alpha/8} = o(G_d) \text{ and } (NT)^{K_\gamma/8} = o(C_d).$$

Note that, when N and T grow at the same rate, the rate conditions of Corollary 1 and Assumption 4(ii) can only hold together if $K_\alpha \leq 3$ and $K_\gamma \leq 3$, so that we are imposing a restriction on the dimensions of the fixed effect spaces.

3.4 Asymptotic results

Our first asymptotic result is Lemma 2 below. It states that our clustering algorithm put together units (resp. time periods) with similar unit (resp. time) fixed effect up to the oracle approximation error. A similar type of result is Lemma 1 in Bonhomme et al. (2022).

Lemma 2 *Let Assumption 2 hold. Then, for all $d \in \{1, \ldots, 4\}$, as N, T, G_d, C_d tend to infinity we have*

$$(i) \quad \frac{1}{N_d} \sum_{i \in \mathcal{N}_d} \left\| \alpha_i - \frac{1}{N^d_{g_i^d}} \sum_{j \in \mathcal{N}_d} \mathbf{1}\{g_j^d = g_i^d\} \alpha_j \right\|^2 = O_P\left(\frac{1}{T} + B_\alpha^d(G_d)\right),$$

$$(ii) \quad \frac{1}{T_d} \sum_{t \in \mathcal{T}_d} \left\| \gamma_t - \frac{1}{T^d_{c_t^d}} \sum_{s \in \mathcal{T}_d} \mathbf{1}\{c_s^d = c_t^d\} \gamma_s \right\|^2 = O_P\left(\frac{1}{N} + B_\gamma^d(C_d)\right).$$

Lemma 2 suggests that injectivity ensures that if the approximation error resulting from discretizing the unobserved heterogeneity based on the unobserved heterogeneity itself, $B_\alpha(G_d)$ and $B_f(C_d)$, is small, then the approximation error resulting from discretizing the unobserved heterogeneity based on discretizing time-series or cross-sectional averages of the data is small as N, T tend to infinity.

Next, we state the main result of the paper, that is, the asymptotic normality of the cross-fitted version of our two-step estimator.

Theorem 1 *Under Assumptions 1, 2, 3, 4, and 5, as N, T, G_d, C_d tend to infinity, we have*

$$\sqrt{NT}(\widehat{\beta}^{\mathrm{CF}} - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma_U^{-1}\sigma^2),$$

where Σ_U is defined in Assumption 3(iii) and $\sigma^2 := \mathbb{E}[v_{11}^2]$.

Theorem 1 justifies inference on β based on Gaussian approximations of the asymptotic distribution. This contrasts with the properties of GFE estimators in nonlinear likelihood models (Bonhomme et al., 2022). Indeed, when both the outcome and covariate models are not restricted to be linear, classification noise affects the properties of second-step estimators in general through an incidental parameter bias. Theorem 1 shows that a linear structure is free of such bias and thus allows the researcher to avoid using (potentially computationally difficult and not proven valid) bias reduction or bootstrap techniques for inference.

4 Simulations

Let us now consider Monte Carlo simulations to evaluate the finite sample performance of our estimator $\widehat{\beta}$ and its cross-fitted version $\widehat{\beta}^{\mathrm{CF}}$. All results in this section are averages over 10,000 replications. In all simulations, we always use 30 random starting values and the Hartigan-Wong algorithm to optimize the k-means objective functions.[12]

First, we describe the data-generating process. We consider the sample sizes $N \in \{50, 100\}$ and $T \in \{10, 20, 30, 40, 50\}$. There is a single regressor, that is, $K = 1$, and we set $\beta = 1$. The error terms u_{it1} and v_{it} are i.i.d. $\mathcal{N}(0, 1)$ random variables. The fixed effects α_i and γ_t are i.i.d. Gamma$(1, 1)$ random variables (so that $K_\alpha = K_\gamma = 1$). The functions f and h_1 are

$$f(\alpha_i, \gamma_t) = \left(0.5 \times \alpha_i^{10} + 0.5 \times \gamma_t^{10} + 2\right)^{1/10},$$
$$h_1(\alpha_i, \gamma_t) = \left(0.5 \times \alpha_i^{10} + 0.5 \times \gamma_t^{10} + 5\right)^{1/5}.$$

This data-generating process is inspired by the constant elasticity of substitution (CES)

[12] The results are not sensitive to the implementation of k-means.

specification for time-varying unobserved heterogeneity proposed in page 631 of Bonhomme et al. (2022).

Baseline results. We start by evaluating our baseline estimator $\widehat{\beta}$, where the number of clusters G and C are chosen according to the rule outlined in Section 2. For inference, we use heteroskedasticity-consistent standard errors, that is, the standard error for $\widehat{\beta}$ is

$$\text{se}(\widehat{\beta}) := \sqrt{\frac{NT}{(N-G)(T-C)}} \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{u}_{it1}^2 \right)^{-1} \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\widehat{e}_{it} - \widehat{\beta} \widehat{u}_{it1} \right)^2 \widehat{u}_{it1}^2 \right)^{1/2},$$

where the factor $\sqrt{\frac{NT}{(N-G)(T-C)}}$ is a degree of freedom correction. We compute 95% confidence intervals based on the Gaussian approximation using the standard error $\text{se}(\widehat{\beta})$.

Then, in the same designs, we study the cross-fitted estimator $\widehat{\beta}^{\text{CF}}$. For all $d \in \{1, \ldots, 4\}$, we set G_d and C_d in each fold according to the data-driven rule described in Section 2. The standard errors are heteroskedasticity-consistent and computed as

$$\text{se}(\widehat{\beta}^{\text{CF}}) := \sqrt{\frac{NT}{df}} \left(\frac{1}{NT} \sum_{d=1}^{4} \sum_{(i,t) \in \mathcal{O}_d} \left(\widehat{u}_{it1}^d \right)^2 \right)^{-1} \left(\frac{1}{NT} \sum_{d=1}^{4} \sum_{(i,t) \in \mathcal{O}_d} \left(\widehat{e}_{it}^d - \widehat{\beta}^{\text{CF}} \widehat{u}_{it1}^d \right)^2 \left(\widehat{u}_{it1}^d \right)^2 \right)^{1/2},$$

where $df := \sum_{d=1}^{D} \left(\frac{N}{2} - G^d \right) \left(\frac{T}{2} - C^d \right)$ is the number of degrees of freedom. We continue to use a Gaussian approximation to build the confidence intervals.

The results for both $\widehat{\beta}$ and $\widehat{\beta}^{\text{CF}}$ are reported in Table 1. The columns "Bias" and "Variance" report the estimators' bias and variance. The columns "Coverage" and "Width" present the coverage and width of the 95% confidence intervals. We will use these names to refer to the same quantities (albeit sometimes for different estimators) in other tables of this section.

We find that our estimators, $\widehat{\beta}$ and $\widehat{\beta}^{\text{CF}}$, exhibit small bias and variance. The baseline estimator, $\widehat{\beta}$, achieves coverage levels close to the nominal 95% across nearly all sample sizes. In comparison, the cross-fitted estimator has slightly lower coverage, though it remains reasonably close to the 95% benchmark. Based on these findings, we recommend that practitioners primarily use the baseline estimator, $\widehat{\beta}$. These results confirm that cross-fitting serves primarily as a theoretical tool to facilitate asymptotic proofs rather than offering practical advantages in finite samples. Intuitively, the slightly weaker performance of $\widehat{\beta}^{\text{CF}}$ arises from its reliance on only half the observations for clustering. Interestingly,

N	T	Bias	Variance	Coverage	Width
Results for $\widehat{\beta}$					
50	10	0.001	0.003	0.957	0.227
50	20	0.001	0.001	0.958	0.149
50	30	0.001	0.001	0.961	0.119
50	40	0.001	0.001	0.957	0.102
50	50	0.001	0.000	0.958	0.090
100	10	0.001	0.001	0.967	0.161
100	20	0.000	0.001	0.960	0.105
100	30	0.000	0.000	0.958	0.083
100	40	0.000	0.000	0.958	0.071
100	50	0.000	0.000	0.957	0.063
Results for $\widehat{\beta}^{\mathrm{CF}}$					
50	10	0.001	0.004	0.875	0.195
50	20	0.002	0.002	0.914	0.133
50	30	0.002	0.001	0.920	0.108
50	40	0.002	0.001	0.922	0.093
50	50	0.001	0.001	0.928	0.083
100	10	0.002	0.002	0.883	0.136
100	20	0.002	0.001	0.914	0.093
100	30	0.002	0.000	0.923	0.075
100	40	0.002	0.000	0.929	0.065
100	50	0.002	0.000	0.925	0.058

Table 1: Baseline results.

N	T	Bias	Variance	Coverage	Width
Two-way fixed effects estimator					
50	10	0.056	0.003	0.645	0.161
50	20	0.059	0.002	0.495	0.116
50	30	0.060	0.001	0.376	0.095
50	40	0.062	0.001	0.283	0.083
50	50	0.062	0.001	0.227	0.074
100	10	0.056	0.002	0.507	0.115
100	20	0.060	0.001	0.319	0.082
100	30	0.061	0.001	0.196	0.068
100	40	0.062	0.001	0.125	0.059
100	50	0.063	0.001	0.070	0.053
Estimator of Bai (2009)					
50	10	0.069	0.010	0.500	0.172
50	20	0.061	0.008	0.439	0.121
50	30	0.053	0.007	0.425	0.102
50	40	0.046	0.007	0.424	0.093
50	50	0.040	0.007	0.409	0.088
100	10	0.057	0.006	0.535	0.131
100	20	0.046	0.004	0.519	0.093
100	30	0.035	0.003	0.546	0.077
100	40	0.027	0.002	0.555	0.068
100	50	0.024	0.002	0.551	0.062
GFE estimator of Freeman and Weidner (2023)					
50	10	0.059	0.009	0.649	0.225
50	20	0.041	0.005	0.722	0.182
50	30	0.033	0.003	0.745	0.156
50	40	0.029	0.003	0.757	0.138
50	50	0.025	0.002	0.760	0.126
100	10	0.044	0.006	0.645	0.164
100	20	0.026	0.003	0.761	0.135
100	30	0.017	0.002	0.821	0.116
100	40	0.012	0.001	0.867	0.103
100	50	0.009	0.001	0.888	0.093

Table 2: Results for alternative estimators.

our estimators continue to perform remarkably well even when T is much smaller than N, a scenario frequently encountered in real-world datasets. This robustness makes our approach particularly valuable for practical applications in such settings.

Alternative estimators. We compare our estimators with three benchmarks. The first benchmark is the classical two-way fixed effects estimator, employing heteroskedasticity-robust standard errors. The second is the estimator proposed by Bai (2009), with $T/2$ factors. As demonstrated by Freeman and Weidner (2023), this estimator is consistent within our model. We construct 95% confidence intervals for this estimator using a Gaussian approximation, with heteroskedasticity-robust standard errors as described in Bai (2009),

incorporating a degrees-of-freedom correction following Freeman and Weidner (2023). Finally, we evaluate the performance of the two-step group fixed effects (GFE) estimator introduced by Freeman and Weidner (2023). Our implementation adheres to their methodology, clustering only the first five loadings and factors, employing their hierarchical clustering approach with a minimum single linkage algorithm, and using clustered heteroskedasticity-robust standard errors with a degrees-of-freedom correction.

It is important to note that, to the best of our knowledge, no theoretical results have established the asymptotic normality of either the estimator of Bai (2009) or the GFE estimator of Freeman and Weidner (2023) in our specific context. Confidence intervals are computed solely for exploratory purposes to assess whether asymptotic normality might plausibly hold.

The results presented in Table 2 indicate that all alternative estimators exhibit substantially higher bias compared to our proposed estimators. The estimator of Bai (2009) demonstrates coverage levels that deviate significantly from nominal values, suggesting that it is not asymptotically normal in this context. Notably, for small sample sizes, Bai (2009)'s estimator is the most biased, while the two-way fixed effects estimator shows increasing bias as the sample size grows. Among the alternatives, the GFE estimator proposed by Freeman and Weidner (2023) systematically exhibits lower bias than Bai (2009)'s estimator. Furthermore, its coverage improves as the sample size increases, which suggests potential asymptotic normality in this setting. This aligns with the simulations reported by Freeman and Weidner (2023), although no formal proof of asymptotic normality is provided. Despite its advantages over other alternative estimators, the GFE estimator still falls short when compared to our proposed estimators. It exhibits greater bias, higher variance, wider confidence intervals, and lower coverage levels. Notably, its performance deteriorates significantly when the time dimension is negligible compared to the cross-sectional dimension.

Sensitivity to time series dependence. Next, we study the behavior of our estimators $\widehat{\beta}$ and $\widehat{\beta}^{\mathrm{CF}}$ under time series dependence. To this end, we vary the distributions of γ_t, u_{it1}, and v_{it}. Specifically, γ_t now follows an AR(1) process with parameter $\rho = 0.7$ and disturbances from a Gamma distribution with shape parameter $(1-\rho)^2/(1-\rho^2)$ and scale parameter $(1-\rho)/(1-\rho^2)$. This specification ensures that the mean and variance of γ_t remain consistent with those in our main design. When $\rho = 0$, the process reverts to the main design. The process is initialized with a Gamma$(1,1)$ distribution, and we discard

the first $10,000$ observations as a burn-in period.

The error terms u_{it1} and v_{it} also follow AR(1) processes. Specifically, we set $u_{i11} \sim \mathcal{N}(0,1)$ and $v_{i1} \sim \mathcal{N}(0,1)$, and for all $i \in \{1,\ldots,N\}$ and $t \in \{2,\ldots,T\}$,

$$u_{it1} = \kappa u_{i(t-1)1} + \mathcal{N}(0,(1-\kappa^2)) \quad \text{and} \quad v_{it} = \kappa v_{i(t-1)} + \mathcal{N}(0,(1-\kappa^2)),$$

where κ is set to either 0 or 0.7.

The case $\kappa = 0$ isolates the effect of time series dependence in the time fixed effects alone. In this scenario, we continue to use heteroskedasticity-robust standard errors, as there is no reason to expect the asymptotic distribution to differ.

When $\kappa = 0.7$, autocorrelation is introduced into the error terms. In this case, we estimate the long-run variance of $u_{it1}v_{it}$ using the standard errors proposed by Arellano (1987), applying a degrees-of-freedom correction. These standard errors are robust to autocorrelation in the error terms.

The results are presented in Table 3. We observe that the performance of our estimators remains largely unaffected by autocorrelation in the time-fixed effects. This indicates that even with strong serial correlation in γ_t, the clustering procedure effectively discretizes the time-fixed effects. While time series dependence in the errors slightly increases the bias and variance of the estimators, the coverage actually slightly improves due to the adjusted standard error estimation (and the confidence intervals become wider as expected). These results confirm that our estimators exhibit strong performance even under serial correlation.

		$\widehat{\beta}$				$\widehat{\beta}^{\text{CF}}$			
N	T	Bias	Variance	Coverage	Width	Bias	Variance	Coverage	Width
		$\rho = 0.7$ and $\kappa = 0$							
50	10	0.002	0.003	0.961	0.232	0.003	0.004	0.865	0.194
50	20	0.002	0.001	0.956	0.151	0.002	0.002	0.905	0.133
50	30	0.002	0.001	0.958	0.121	0.002	0.001	0.916	0.108
50	40	0.001	0.001	0.960	0.103	0.001	0.001	0.921	0.093
50	50	0.001	0.001	0.959	0.092	0.001	0.001	0.927	0.083
100	10	0.000	0.001	0.963	0.166	0.001	0.002	0.878	0.136
100	20	0.001	0.001	0.958	0.107	0.002	0.001	0.910	0.093
100	30	0.001	0.000	0.958	0.085	0.002	0.000	0.923	0.075
100	40	0.000	0.000	0.957	0.072	0.001	0.000	0.926	0.065
100	50	0.000	0.000	0.958	0.064	0.002	0.000	0.923	0.058
		$\rho = 0.7$ and $\kappa = 0.7$							
50	10	0.002	0.006	0.963	0.345	-0.001	0.006	0.900	0.272
50	20	0.001	0.003	0.958	0.243	0.001	0.003	0.909	0.199
50	30	0.002	0.002	0.955	0.199	0.002	0.002	0.914	0.169
50	40	0.001	0.002	0.957	0.173	0.002	0.002	0.916	0.150
50	50	0.001	0.001	0.956	0.155	0.001	0.001	0.921	0.136
100	10	0.001	0.003	0.973	0.247	0.000	0.003	0.914	0.191
100	20	0.000	0.002	0.963	0.172	0.002	0.002	0.920	0.140
100	30	0.001	0.001	0.961	0.140	0.002	0.001	0.928	0.119
100	40	0.000	0.001	0.960	0.122	0.001	0.001	0.929	0.105
100	50	0.001	0.001	0.963	0.109	0.002	0.001	0.929	0.095

Table 3: Results for $\widehat{\beta}$ and $\widehat{\beta}^{\text{CF}}$ under time series dependence.

References

Aloise, D., Hansen, P., and Liberti, L. (2009). An improved column generation algorithm for minimum sum-of-squares clustering. *Mathematical Programming*, 131:195 – 220.

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, 49(4):431–434.

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

Beyhum, J. and Gautier, E. (2023). Factor and factor loading augmented estimators for panel regression with possibly nonstrong factors. *Journal of Business & Economic Statistics*, 41(1):270–281.

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

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

Chen, L., Dolado, J. J., and Gonzalo, J. (2021). Quantile factor models. *Econometrica*, 89(2):875–910.

Chen, Q., Syrgkanis, V., and Austern, M. (2022). Debiased machine learning without sample-splitting for stable estimators. *Advances in Neural Information Processing Systems*, 35:3096–3109.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review*, 105(5):486–490.

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

du Merle, O., Hansen, P., Jaumard, B., and Mladenović, N. (1997). An interior point algorithm for minimum sum-of-squares clustering. *SIAM J. Sci. Comput.*, 21:1485–1505.

Dukes, O. and Vansteelandt, S. (2021). Inference for treatment effect parameters in potentially misspecified high-dimensional models. *Biometrika*, 108(2):321–334.

Freeman, H. and Weidner, M. (2023). Linear panel regressions with two-way unobserved heterogeneity. *Journal of Econometrics*, 237(1):105498.

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

Greenaway-McGrevy, R., Han, C., and Sul, D. (2012). Asymptotic distribution of factor augmented estimators for panel regression. *Journal of Econometrics*, 169(1):48–53.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition.

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

Mugnier, M. (2024). A simple and computationally trivial estimator for grouped fixed effects models.

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

Shi, C., Zhu, J., Shen, Y., Luo, S., Zhu, H., and Song, R. (2024). Off-policy confidence interval estimation with confounded markov decision process. *Journal of the American Statistical Association*, 119(545):273–284.

Vansteelandt, S., Dukes, O., Van Lancker, K., and Martinussen, T. (2024). Assumption-lean cox regression. *Journal of the American Statistical Association*, 119(545):475–484.

Wang, Y., Ying, A., and Xu, R. (2024). Doubly robust estimation under covariate-induced dependent left truncation. *Biometrika*, page asae005.

Westerlund, J. and Urbain, J.-P. (2015). Cross-sectional averages versus principal components. *Journal of Econometrics*, 185(2):372–377.

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

A Proof of Lemma 2

We only prove the first statement, as the proof of the second one is similar. We proceed in two steps.

Step 1. In this step, we show that

$$\frac{1}{N_d}\sum_{i\in\mathcal{N}_d}\left\|\varphi_d^\alpha(\alpha_i)-\frac{1}{N_{g_i^d}^d}\sum_{j=1}^N \mathbf{1}\{g_j^d=g_i^d\}\varphi_d^\alpha(\alpha_j)\right\|^2 = O_P\left(\frac{1}{T}\right)+O_P(B_\alpha^d(G_d)). \tag{8}$$

By the triangle inequality and the classical inequality $ab \leq (a^2+b^2)/2$, we have

$$\begin{aligned}
&\frac{1}{N_d}\sum_{i\in\mathcal{N}_d}\left\|\varphi_d^\alpha(\alpha_i)-\frac{1}{N_{g_i^d}^d}\sum_{j=1}^{N}\mathbf{1}\{g_j^d=g_i^d\}\varphi_d^\alpha(\alpha_j)\right\|^2 \\
&\leq \frac{2}{N_d}\sum_{i\in\mathcal{N}_d}\left\|\varphi_d^\alpha(\alpha_i)-\widehat{a}_i^d(g_i^d))\right\|^2+\frac{2}{N_d}\sum_{i\in\mathcal{N}_d}\left\|\widehat{a}_i^d(g_i^d)-\frac{1}{N_{g_i^d}^d}\sum_{j=1}^{N}\mathbf{1}\{g_j^d=g_i^d\}\varphi_d^\alpha(\alpha_j)\right\|^2.
\end{aligned} \tag{9}$$

Under Assumption 2(i), Lemma 1 in Bonhomme et al. (2022) yields

$$\frac{1}{N_d}\sum_{i\in\mathcal{N}_d}\left\|\varphi_d^\alpha(\alpha_i)-\widehat{a}_i^d(g_i^d))\right\|^2=O_P\left(\frac{1}{T}\right)+O_P(B_\alpha^d(G_d)). \tag{10}$$

Next, using that $\widehat{a}_i^d(g_i^d)=\frac{1}{N_{g_i^d}^d}\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d=g_i^d\}a_j^d$, we have

$$\begin{aligned}
&\frac{1}{N_d}\sum_{i\in\mathcal{N}_d}\left\|\widehat{a}_i^d(g_i^d)-\frac{1}{N_{g_i^d}^d}\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d=g_i^d\}\varphi_d^\alpha(\alpha_j)\right\|^2 \\
&=\frac{1}{N_d}\sum_{i\in\mathcal{N}_d}\left\|\frac{1}{N_{g_i^d}^d}\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d=g_i^d\}(a_j^d-\varphi_d^\alpha(\alpha_j))\right\|^2 \\
&\leq\frac{1}{N_d}\sum_{i\in\mathcal{N}_d}\frac{1}{\left(N_{g_i^d}^d\right)^2}\left(\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d=g_i^d\}\left\|a_j^d-\varphi_d^\alpha(\alpha_j)\right\|\right)^2 \\
&\leq\frac{1}{N_d}\sum_{i\in\mathcal{N}_d}\frac{1}{N_{g_i^d}^d}\left(\sum_{j\in\mathcal{N}_d}\left\|a_j^d-\varphi_d^\alpha(\alpha_j)\right\|^2\right) \\
&=\frac{1}{N_d}\sum_{g=1}^{G_d}\sum_{i\in\mathcal{N}_d}\mathbf{1}\{g_i^d=g\}\frac{1}{N_g^d}\left(\sum_{j\in\mathcal{N}_d}\left\|a_j^d-\varphi_d^\alpha(\alpha_j)\right\|^2\right) \\
&=\frac{1}{N_d}\sum_{j\in\mathcal{N}_d}\left\|a_j^d-\varphi_d^\alpha(\alpha_j)\right\|^2 \\
&=O_P\left(\frac{1}{T}\right),
\end{aligned} \tag{11}$$

where the first inequality follows from the triangle inequality, the second inequality is a consequence of the Cauchy–Schwarz inequality, and the last equality follows from Assumption 2. Combining (9), (10), and (11), we obtain (8).

Step 2. In this second step, we prove the result of the lemma. We have

$$
\begin{aligned}
&\frac{1}{N_d}\sum_{i\in\mathcal{N}_d}\left\|\alpha_i-\frac{1}{N^d_{g_i^d}}\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d=g_i^d\}\alpha_j\right\|^2\\
&=\frac{1}{N_d}\sum_{i\in\mathcal{N}_d}\left\|\frac{1}{N^d_{g_i^d}}\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d=g_i^d\}(\alpha_i-\alpha_j)\right\|^2\\
&\le\frac{1}{N_d}\sum_{i\in\mathcal{N}_d}\frac{1}{\left(N^d_{g_i^d}\right)^2}\left(\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d=g_i^d\}\left\|\alpha_i-\alpha_j\right\|\right)^2\\
&\le\frac{1}{N_d}\sum_{i\in\mathcal{N}_d}\frac{1}{N^d_{g_i^d}}\left(\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d=g_i^d\}\left\|\alpha_i-\alpha_j\right\|^2\right)
\end{aligned}
\tag{12}
$$

where the first inequality follows from the triangle inequality and the second inequality is a consequence of the Cauchy–Schwarz inequality. Next, by Assumption 2(i), there exists a constant $L>0$ such that

$$
\begin{aligned}
&\frac{1}{N_d}\sum_{i\in\mathcal{N}_d}\frac{1}{N^d_{g_i^d}}\left(\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d=g_i^d\}\left\|\alpha_i-\alpha_j\right\|^2\right)\\
&=\frac{1}{N_d}\sum_{i\in\mathcal{N}_d}\frac{1}{N^d_{g_i^d}}\left(\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d=g_i^d\}\left\|\psi_d^\alpha(\varphi_d^\alpha(\alpha_i)-\psi_d^\alpha(\varphi_d^\alpha(\alpha_j)\right\|^2\right)\\
&\le\frac{L}{N_d}\sum_{i\in\mathcal{N}_d}\frac{1}{N^d_{g_i^d}}\left(\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d=g_i^d\}\left\|\varphi_d^\alpha(\alpha_i)-\varphi_d^\alpha(\alpha_j)\right\|^2\right).
\end{aligned}
\tag{13}
$$

Moreover, we have

$$
\begin{aligned}
&\frac{1}{N_d}\sum_{i\in\mathcal{N}_d}\frac{1}{N^d_{g_i^d}}\left(\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d=g_i^d\}\left\|\varphi_\alpha^d(\alpha_i)-\varphi_\alpha^d(\alpha_j)\right\|^2\right)\\
&=\frac{1}{N_d}\sum_{i\in\mathcal{N}_d}\frac{1}{N^d_{g_i^d}}\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d=g_i^d\}\left(\varphi_d^\alpha(\alpha_i)^\top(\varphi_d^\alpha(\alpha_i)-\varphi_d^\alpha(\alpha_j))-\varphi_d^\alpha(\alpha_j)^\top(\varphi_d^\alpha(\alpha_i)-\varphi_d^\alpha(\alpha_j))\right)\\
&=\frac{2}{N_d}\sum_{i\in\mathcal{N}_d}\varphi_d^\alpha(\alpha_i)^\top\varphi_d^\alpha(\alpha_i)-\frac{2}{N_d}\sum_{i\in\mathcal{N}_d}\frac{1}{N^d_{g_i^d}}\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d=g_i^d\}\varphi_d^\alpha(\alpha_i)^\top\varphi_d^\alpha(\alpha_j)\\
&=\frac{2}{N_d}\sum_{i\in\mathcal{N}_d}\varphi_d^\alpha(\alpha_i)^\top\left(\varphi_d^\alpha(\alpha_i)-\frac{1}{N^d_{g_i^d}}\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d=g_i^d\}\varphi_d^\alpha(\alpha_j)\right)\\
&=\frac{2}{N_d}\sum_{i\in\mathcal{N}_d}\left\|\varphi_d^\alpha(\alpha_i)-\frac{1}{N^d_{g_i^d}}\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d=g_i^d\}\varphi_d^\alpha(\alpha_j)\right\|^2\\
&=O_P\left(\frac{1}{T}+B_\alpha^d(G_d)\right),
\end{aligned}
$$

where in the last equality we used (8). Combining the last result with (12) and (13), we obtain the result of the lemma.

B On Theorem 1

This section concerns the proof of Theorem 1. It is divided as follows. In Section B.1, we introduce some notation used in the proof. Section B.2 contains the body of the proof of Theorem 1. This proof relies on auxiliary lemmas stated and proved in Section B.3. The proofs of the auxiliary lemmas themselves depend on technical lemmas stated and proved in Section B.4.

B.1 Notation

For all $i \in \{1, \dots, N\}$, $t \in \{1, \dots, T\}$, and $k \in \{1, \dots, K\}$, we let $h_{itk} := h_k(\alpha_i, \gamma_t)$ and $f_{it} := f(\alpha_i, \gamma_t)$. For all $k \in \{1, \dots, K\}$ and $d \in \{1, \dots, 4\}$, we use the notation

$$\begin{aligned}
\widetilde{h}^d_{itk} &:= h_{itk} - \left(\bar{h}_{g^d_i t}\right)_k - \left(\bar{h}_{ic^d_t}\right)_k + \left(\bar{h}_{g^d_i c^d_t}\right)_k, \\
\widetilde{f}^d_{it} &:= f_{it} - \bar{f}_{g^d_i t} - \bar{f}_{ic^d_t} + \bar{f}_{g^d_i c^d_t}, \\
\widetilde{u}^d_{itk} &:= u_{itk} - \left(\bar{u}_{g^d_i t}\right)_k - \left(\bar{u}_{ic^d_t}\right)_k + \left(\bar{u}_{g^d_i c^d_t}\right)_k.
\end{aligned}$$

B.2 Proof of Theorem 1

We have

$$\begin{aligned}
\widehat{\beta}^{\mathrm{CF}} &= \left(\sum_{d=1}^{4} \sum_{(i,t)\in\mathcal{O}_d} \widehat{u}^d_{it} (\widehat{u}^d_{it})^\top \right)^{-1} \sum_{d=1}^{4} \sum_{(i,t)\in\mathcal{O}_d} \widehat{u}^d_{it} \widehat{e}^d_{it} \\
&= \left(\sum_{(i,t)\in\mathcal{O}_d} \widehat{u}^d_{it} (\sum_{d=1}^{4} \widehat{u}^d_{it})^\top \right)^{-1} \sum_{(i,t)\in\mathcal{O}_d} \sum_{d=1}^{4} \widehat{u}^d_{it} y_{it}.
\end{aligned}$$

Since $y_{it} = x_{it}^\top \beta + f_{it} + v_{it}$, this yields

$$\widehat{\beta}^{\mathrm{CF}} = \beta + \left(\sum_{d=1}^{4} \sum_{(i,t)\in\mathcal{O}_d} \widehat{u}^d_{it} (\widehat{u}^d_{it})^\top \right)^{-1} \sum_{d=1}^{4} \sum_{(i,t)\in\mathcal{O}_d} \widehat{u}^d_{it} f_{it} + \left(\sum_{d=1}^{4} \sum_{(i,t)\in\mathcal{O}_d} \widehat{u}^d_{it} (\widehat{u}^d_{it})^\top \right)^{-1} \sum_{d=1}^{4} \sum_{(i,t)\in\mathcal{O}_d} \widehat{u}^d_{it} v_{it}.$$

We obtain

$$\sqrt{NT}(\widehat{\beta}^{\mathrm{CF}}-\beta)=\left(\frac{1}{NT}\sum_{d=1}^{4}\sum_{(i,t)\in\mathcal{O}_d}\widehat{u}_{it}^{d}(\widehat{u}_{it}^{d})^{\top}\right)^{-1}\frac{1}{\sqrt{NT}}\sum_{d=1}^{4}\sum_{(i,t)\in\mathcal{O}_d}\widehat{u}_{it}^{d}f_{it}$$
$$+\left(\frac{1}{NT}\sum_{d=1}^{4}\sum_{(i,t)\in\mathcal{O}_d}\widehat{u}_{it}^{d}(\widehat{u}_{it}^{d})^{\top}\right)^{-1}\frac{1}{\sqrt{NT}}\sum_{d=1}^{4}\sum_{(i,t)\in\mathcal{O}_d}\widehat{u}_{it}^{d}v_{it}.$$

By Lemmas 3, 4, and 5 and the continuous mapping theorem, we obtain

$$\sqrt{NT}(\widehat{\beta}^{\mathrm{CF}}-\beta)=\Sigma_U^{-1}\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}u_{it}v_{it}+o_P(1).$$

The result follows from the central limit theorem and Slutsky's lemma.

B.3 Auxiliary lemmas

Lemma 3 *Under Assumptions 1, 2, 3, 4, and 5, for every fold $d\in\{1,\dots,4\}$, we have*

$$\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}\widehat{u}_{it}^{d}(\widehat{u}_{it}^{d})^{\top}=\Sigma_U+o_P(1).$$

Proof. Fix $d\in\{1,\dots,4\}$ and $k,\ell\in\{1,\dots,K\}$. We have

$$\begin{aligned}
&\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}\widehat{u}_{itk}^{d}\widehat{u}_{it\ell}^{d}\\
&=\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}\widehat{u}_{itk}^{d}u_{it\ell}\\
&=\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}\left(\widetilde{u}_{itk}^{d}+\widetilde{h}_{itk}^{d}\right)(u_{it\ell}+h_{it\ell})\\
&=\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}u_{itk}u_{it\ell}+\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}\widetilde{h}_{itk}^{d}\widetilde{h}_{it\ell}^{d}+\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}\widetilde{h}_{itk}^{d}u_{it\ell}\\
&\quad+\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}u_{itk}\widetilde{h}_{it\ell}^{d}+\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}(\widetilde{u}_{itk}^{d}-u_{itk})u_{it\ell}\\
&=\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}u_{itk}u_{it\ell}+o_P(1),
\end{aligned}$$

where we used Lemmas 6, 7, and 9 and Assumptions 4 and 5 in the last equality. The

result follows from the law of large numbers and the continuous mapping theorem. □

Lemma 4 *Under Assumptions 1, 2, 3, 4, and 5, for every $d \in \{1, \ldots, 4\}$, we have*

$$\frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} \widehat{u}_{it}^d f_{it} = o_P\left(\frac{1}{\sqrt{NT}}\right).$$

Proof. For any $k \in \{1, \ldots, K\}$, it holds that

$$\begin{aligned}
&\frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} \widehat{u}_{itk}^d f_{it} \\
&= \frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} \left(\widetilde{u}_{itk}^d + \widetilde{h}_{itk}^d\right) f_{it} \\
&= \frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} \widetilde{h}_{itk}^d \widetilde{f}_{it}^d + \frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} u_{itk} \widetilde{f}_{it}^d \\
&= o_P(1),
\end{aligned}$$

by Lemmas 6 and 9 and Assumptions 4 and 5. □

Lemma 5 *Under Assumptions 1, 2, 3, 4, and 5, for every $d \in \{1, \ldots, 4\}$, we have*

$$\frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} \widehat{u}_{it} v_{it} = \frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} u_{it} v_{it} + o_P(1).$$

Proof. For every $k \in \{1, \ldots, K\}$, it holds that

$$\begin{aligned}
&\frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} \widehat{u}_{itk} v_{it} \\
&= \frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} \left(\widetilde{u}_{itk}^d + \widetilde{h}_{itk}^d\right) v_{it} \\
&= \frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} u_{itk} v_{it} + \frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} \left(\widetilde{u}_{itk}^d - u_{itk}\right) v_{it} + \frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} \widetilde{h}_{itk}^d v_{it} \\
&= \frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} u_{itk} v_{it} + o_P(1),
\end{aligned}$$

by Lemmas 8 and 9 and Assumptions 4 and 5. □

B.4 Technical lemmas

Lemma 6 *Under Assumptions 1, 2, and 3, for all $k \in \{1, \dots, K\}$ and $d \in \{1, \dots, 4\}$, we have*

$$\frac{1}{N_d T_d} \sum_{(i,t)\in\mathcal{O}_d} \left(\widetilde{h}^d_{itk}\right)^2 = O_P\left(\frac{1}{T^2} + \frac{1}{N^2} + B^d_\alpha(G_d)^2 + B^d_\gamma(C_d)^2\right)$$

and

$$\frac{1}{N_d T_d} \sum_{(i,t)\in\mathcal{O}_d} \left(\widetilde{f}^d_{it}\right)^2 = O_P\left(\frac{1}{T^2} + \frac{1}{N^2} + B^d_\alpha(G_d)^2 + B^d_\gamma(C_d)^2\right).$$

Proof. By Assumption 1 and relying on analogous Taylor expansions as in the proof of Lemma 2 in Freeman and Weidner (2023), we have

$$\widetilde{h}^d_{itk} = O\left(\frac{1}{N^d_{g^d_i}} \sum_{j\in\mathcal{N}_d} \mathbf{1}\{g^d_j = g^d_i\} \left\|\alpha_i - \alpha_j\right\|^2 + \frac{1}{T^d_{c^d_t}} \sum_{s\in\mathcal{T}_d} \mathbf{1}\{c^d_s = c^d_t\} \left\|\gamma_t - \gamma_s\right\|^2\right),$$

uniformly in i, t. By the triangle inequality, this yields

$$\begin{aligned}
&\frac{1}{N_d T_d} \sum_{(i,t)\in\mathcal{O}_d} \left|\widetilde{h}^d_{itk}\right| \\
&= O\left(\frac{1}{N_d} \sum_{i\in\mathcal{N}_d} \frac{1}{N^d_{g^d_i}} \sum_{j\in\mathcal{N}_d} \mathbf{1}\{g^d_j = g^d_i\} \left\|\alpha_i - \alpha_j\right\|^2 + \frac{1}{T_d} \sum_{t\in\mathcal{T}_d} \frac{1}{T^d_{c^d_t}} \sum_{s\in\mathcal{T}_d} \mathbf{1}\{c^d_s = c^d_t\} \left\|\gamma_t - \gamma_s\right\|^2\right).
\end{aligned}$$

Now, notice that

$$\begin{aligned}
&\frac{1}{N_d} \sum_{i\in\mathcal{N}_d} \frac{1}{N^d_{g^d_i}} \sum_{j\in\mathcal{N}_d} \mathbf{1}\{g^d_j = g^d_i\} \left\|\alpha_i - \alpha_j\right\|^2 \\
&= \frac{1}{N_d} \sum_{i\in\mathcal{N}_d} \frac{1}{N^d_{g^d_i}} \sum_{j\in\mathcal{N}_d} \mathbf{1}\{g^d_j = g^d_i\} \left(\alpha_i^\top(\alpha_i - \alpha_j) - \alpha_j^\top(\alpha_i - \alpha_j)\right) \\
&= \frac{2}{N_d} \sum_{i\in\mathcal{N}_d} \alpha_i^\top \alpha_i - \frac{2}{N_d} \sum_{i\in\mathcal{N}_d} \frac{1}{N^d_{g^d_i}} \sum_{j\in\mathcal{N}_d} \mathbf{1}\{g^d_j = g^d_i\} \alpha_i^\top \alpha_j \\
&= \frac{2}{N_d} \sum_{i\in\mathcal{N}_d} \alpha_i^\top \left(\alpha_i - \frac{1}{N^d_{g^d_i}} \sum_{j\in\mathcal{N}_d} \mathbf{1}\{g^d_j = g^d_i\} \alpha_j\right) \\
&= \frac{2}{N_d} \sum_{i\in\mathcal{N}_d} \left\|\alpha_i - \frac{1}{N^d_{g^d_i}} \sum_{j\in\mathcal{N}_d} \mathbf{1}\{g^d_j = g^d_i\} \alpha_j\right\|^2 \\
&= O_P\left(\frac{1}{T} + B^d_\alpha(G_d)\right),
\end{aligned}$$

where we used Lemma 2 to obtain the last equality. Similarly, we have

$$\frac{1}{T_d}\sum_{t\in\mathcal{T}_d}\frac{1}{T^d_{c^d_t}}\sum_{s\in\mathcal{T}_d}\mathbf{1}\{c^d_s=c^d_t\}\left\|\gamma_t-\gamma_s\right\|^2=O_P\left(\frac{1}{N}+B^d_\gamma(C_d)\right).$$

This yields

$$\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}\left|\widetilde{h}^d_{itk}\right|=O_P\left(\frac{1}{T}+B^d_\alpha(G_d)+\frac{1}{N}+B^d_\gamma(C_d)\right).$$

We obtain the result using that

$$\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}\left(\widetilde{h}^d_{itk}\right)^2\leq\left(\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}\left|\widetilde{h}^d_{itk}\right|\right)^2.$$

The proof of the second statement is similar and, therefore, omitted. □

Lemma 7 *Under Assumptions 1, 2, and 3, for all $k,\ell\in\{1,\dots,K\}$ and $d\in\{1,\dots,4\}$, we have*

$$\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}(\widetilde{u}^d_{itk}-u_{itk})u_{it\ell}=O_P\left(\frac{G_d}{N}+\frac{C_d}{T}+\frac{G_dC_d}{NT}\right).$$

Proof. We have

$$\begin{aligned}\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}(\widetilde{u}^d_{itk}-u_{itk})u_{it\ell}&=\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}\left(\bar{u}^d_{g^d_it}+\bar{u}^d_{ic^d_t}-\bar{u}^d_{g^d_ic^d_t}\right)_k u_{it\ell}\\&=J_1+J_2+J_3,\end{aligned}$$

where

$$\begin{aligned}J_1&:=\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}\left(\bar{u}^d_{g^d_it}\right)_k u_{it\ell},\\J_2&:=\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}\left(\bar{u}^d_{ic^d_t}\right)_k u_{it\ell},\\J_3&:=\frac{1}{N_dT_d}\sum_{(i,t)\in\mathcal{O}_d}\left(\bar{u}^d_{g^d_ic^d_t}\right)_k u_{it\ell}.\end{aligned}$$

Let us bound J_1. It holds that

$$\begin{aligned}
J_1 &= \frac{1}{N_d T_d} \sum_{(i,t)\in\mathcal{O}_d} \left(\frac{1}{N^d_{g_i^d}} \sum_{j\in\mathcal{N}_d} \mathbf{1}\{g_j^d = g_i^d\} u_{jtk} \right) u_{it\ell} \\
&= \frac{1}{N_d T_d} \sum_{g=1}^{G_d} \sum_{(i,t)\in\mathcal{O}_d} \left(\frac{1}{N_g^d} \sum_{j\in\mathcal{N}_d} \mathbf{1}\{g_j^d = g\} u_{jtk} \right) \mathbf{1}\{g_i^d = g\} u_{it\ell} \\
&= \frac{1}{N_d T_d} \sum_{g=1}^{G_d} \sum_{t\in\mathcal{T}_d} \left(\frac{1}{\sqrt{N_g^d}} \sum_{j\in\mathcal{N}_d} \mathbf{1}\{g_j^d = g\} u_{jtk} \right) \left(\frac{1}{\sqrt{N_g^d}} \sum_{j\in\mathcal{N}_d} \mathbf{1}\{g_j^d = g\} u_{jt\ell} \right).
\end{aligned}$$

By the triangle inequality,

$$\begin{aligned}
|J_1| &\leq \frac{1}{N_d T_d} \sum_{g=1}^{G_d} \sum_{t\in\mathcal{T}_d} \left| \frac{1}{\sqrt{N_g^d}} \sum_{j\in\mathcal{N}_d} \mathbf{1}\{g_j^d = g\} u_{jtk} \right| \left| \frac{1}{\sqrt{N_g^d}} \sum_{j\in\mathcal{N}_d} \mathbf{1}\{g_j^d = g\} u_{jt\ell} \right| \\
&\leq \frac{1}{N_d T_d} \sum_{k=1}^{K} \sum_{g=1}^{G_d} \sum_{t\in\mathcal{T}_d} \left(\frac{1}{\sqrt{N_g^d}} \sum_{j\in\mathcal{N}_d} \mathbf{1}\{g_j^d = g\} u_{jtk} \right)^2.
\end{aligned}$$

Next, since $(u_{jtk})_{j\in\mathcal{N}_d}$ are mean-zero independent random variables, independent of $(g_j^d)_{j\in\mathcal{N}_d}$, and each with the same distribution as u_{11k}, we have

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{\sqrt{N_g^d}} \sum_{j\in\mathcal{N}_d} \mathbf{1}\{g_j^d = g\} u_{jtk} \right)^2 \right] &= \mathbb{E}\left[\frac{1}{N_g^d} \sum_{j\in\mathcal{N}_d} \mathbf{1}\{g_j^d = g\} u_{jtk}^2 \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{1}{N_g^d} \sum_{j\in\mathcal{N}_d} \mathbf{1}\{g_j^d = g\} u_{jtk}^2 \,\middle|\, (g_j^d)_{j\in\mathcal{N}_j} \right] \right] \\
&= E\left[\frac{1}{N_g^d} \sum_{j\in\mathcal{N}_d} \mathbf{1}\{g_j^d = g\} \mathbb{E}[u_{11k}^2] \right] \\
&= \mathbb{E}[u_{11k}^2] \\
&\leq \max_{k\in\{1,\ldots,K\}} \mathbb{E}[u_{11k}^2].
\end{aligned}$$

As a result, we get

$$\mathbb{E}[|J_1|] \leq \frac{G_d}{N_d} \max_{k\in\{1,\ldots,K\}} \mathbb{E}[u_{11k}^2].$$

This yields

$$J_1 = O_P\left(\frac{G_d}{N} \right).$$

Similarly, we have

$$J_2 = O_P\left(\frac{C_d}{T}\right).$$

Moreover, it holds that

$$\begin{aligned} J_3 &= \frac{1}{N_d T_d} \sum_{(i,t)\in\mathcal{O}_d} \left(\frac{1}{N^d_{g_i^d} T^d_{c_t^d}} \sum_{(j,s)\in\mathcal{O}_d} \mathbf{1}\{g_j^d = g_i^d\}\mathbf{1}\{c_s^d = c_t^d\} u_{jsk} \right) u_{it\ell} \\ &= \frac{1}{N_d T_d} \sum_{g=1}^{G_d} \sum_{c=1}^{C_d} \Bigg[\left(\frac{1}{\sqrt{N_g^d T_c^d}} \sum_{(j,s)\in\mathcal{O}_d} \mathbf{1}\{g_j^d = g\}\mathbf{1}\{c_s^d = c\} u_{jsk} \right) \\ &\quad \times \left(\frac{1}{\sqrt{N_g^d T_c^d}} \sum_{(j,s)\in\mathcal{O}_d} \mathbf{1}\{g_j^d = g\}\mathbf{1}\{c_s^d = c\} u_{js\ell} \right) \Bigg]. \end{aligned}$$

Then, by arguments similar to the ones allowing to bound J_1, we obtain

$$J_3 = O_P\left(\frac{G_d C_d}{NT}\right).$$

The result follows from combining the bounds on J_1, J_2, and J_3. □

Lemma 8 *Under Assumptions 1, 2, and 3, for all $k, \ell \in \{1, \dots, K\}$ and $d \in \{1, \dots, 4\}$, we have*

$$\frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} (\widetilde{u}^d_{itk} - u_{itk}) v_{it} = O_P\left(\sqrt{\frac{G_d}{N}} + \sqrt{\frac{C_d}{T}} + \sqrt{\frac{G_d C_d}{NT}} \right).$$

Proof.

$$\begin{aligned} \frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} (\widetilde{u}^d_{itk} - u_{itk}) v_{it} &= \frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} \left(\bar{u}^d_{g_i^d t} + \bar{u}^d_{ic_t^d} - \bar{u}^d_{g_i^d c_t^d} \right)_k v_{it} \\ &= J_1 + J_2 + J_3, \end{aligned}$$

where

$$J_1 := \frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} \left(\bar{u}^d_{g_i^d t}\right)_k v_{it},$$
$$J_2 := \frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} \left(\bar{u}^d_{i c_t^d}\right)_k v_{it},$$
$$J_3 := \frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} \left(\bar{u}^d_{g_i^d c_t^d}\right)_k v_{it}.$$

Let us bound J_1. First, notice that by Assumption 3, $(v_{it})_{(i,t)\in\mathcal{O}_d}$ is a sequence of mean-zero independent random variables mutually independent of $(\bar{u}^d_{g_i^d t})_{(i,t)\in\mathcal{O}_d}$ and each with the same distribution as v_{11}. Hence, we have $\mathbb{E}[J_1] = 0$ and

$$\begin{aligned}
\mathbb{E}[J_1^2] &= \mathbb{E}\left[\left(\frac{1}{\sqrt{N_d T_d}} \sum_{(i,t)\in\mathcal{O}_d} \left(\bar{u}^d_{g_i^d t}\right)_k v_{it}\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{N_d T_d} \sum_{(i,t)\in\mathcal{O}_d} \left(\bar{u}^d_{g_i^d t}\right)_k^2 v_{it}^2\right] \\
&= \mathbb{E}[v_{11}^2]\mathbb{E}\left[\frac{1}{N_d T_d} \sum_{(i,t)\in\mathcal{O}_d} \left(\bar{u}^d_{g_i^d t}\right)_k^2\right].
\end{aligned}$$

Second, it holds that

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{N_d T_d}\sum_{(i,t)\in\mathcal{O}_d}\left(\bar{u}^d_{g_i^d t}\right)_k^2\right] \\
&= \mathbb{E}\left[\frac{1}{N_d T_d}\sum_{(i,t)\in\mathcal{O}_d}\left(\frac{1}{N^d_{g_i^d}}\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d = g_i^d\}u_{jtk}\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{N_d}\sum_{i\in\mathcal{N}_d}\left(\frac{1}{N^d_{g_i^d}}\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d = g_i^d\}u_{j1k}\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{N_d}\sum_{i\in\mathcal{N}_d}\left(\frac{1}{N^d_{g_i^d}}\right)^2\sum_{(j,m)\in\mathcal{N}_d^2}\mathbf{1}\{g_j^d = g_i^d = g_m^d\}u_{j1k}u_{m1k}\right] \\
&= \mathbb{E}\left[\frac{1}{N_d}\sum_{i\in\mathcal{N}_d}\left(\frac{1}{N^d_{g_i^d}}\right)^2\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d = g_i^d\}u_{j1k}^2\right] \\
&= \mathbb{E}\left[\frac{1}{N_d}\sum_{g=1}^{G_d}\sum_{i\in\mathcal{N}_d}\left(\frac{1}{N^d_{g}}\right)^2\sum_{j\in\mathcal{N}_d}\mathbf{1}\{g_j^d = g_i^d = g\}u_{j1k}^2\right] \\
&= \mathbb{E}\left[\frac{1}{N_d}\sum_{g=1}^{G_d}1\right]\mathbb{E}\left[u_{11k}^2\right] \\
&\leq \frac{G_d}{N_d}\max_{k=1,\ldots,K}\mathbb{E}\left[u_{11k}^2\right].
\end{aligned}
$$

This yields

$$
J_1 = O_P\left(\sqrt{\frac{G_d}{N}}\right).
$$

Similarly, we have

$$
J_2 = O_P\left(\sqrt{\frac{C_d}{T}}\right).
$$

Finally, following the arguments used to bound J_1, we have $\mathbb{E}[J_3] = 0$ and

$$
\mathbb{E}[J_3^2] = \mathbb{E}[v_{11}^2]\mathbb{E}\left[\frac{1}{N_d T_d}\sum_{(i,t)\in\mathcal{O}_d}\left(\bar{u}^d_{g_i^d c_t^d}\right)_k^2\right].
$$

Next, notice that

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{N_d T_d}\sum_{(i,t)\in\mathcal{O}_d}\left(\bar{u}^d_{g_i^d c_t^d}\right)_k^2\right] \\
&= \mathbb{E}\left[\frac{1}{N_d T_d}\sum_{(i,t)\in\mathcal{O}_d}\left(\frac{1}{N^d_{g_i^d}T^d_{c_t^d}}\sum_{(j,s)\in\mathcal{O}_d}\mathbf{1}\{g_j^d=g_i^d\}\mathbf{1}\{c_s^d=c_t^d\}u_{jsk}\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{N_d T_d}\sum_{(i,t)\in\mathcal{O}_d}\left(\frac{1}{N^d_{g_i^d}T^d_{c_t^d}}\right)^2\sum_{(j,s),(m,h)\in\mathcal{O}_d}\mathbf{1}\{g_j^d=g_m^d=g_i^d\}\mathbf{1}\{c_s^d=c_h^d=c_t^d\}u_{jsk}u_{mhk}\right] \\
&= \mathbb{E}\left[\frac{1}{N_d T_d}\sum_{(i,t)\in\mathcal{O}_d}\left(\frac{1}{N^d_{g_i^d}T^d_{c_t^d}}\right)^2\sum_{(j,s)\in\mathcal{O}_d}\mathbf{1}\{g_j^d=g_i^d\}\mathbf{1}\{c_s^d=c_t^d\}u_{jsk}^2\right] \\
&= \mathbb{E}\left[\frac{1}{N_d T_d}\sum_{g=1}^{G_d}\sum_{c=1}^{C_d}\sum_{(i,t),(j,s)\in\mathcal{O}_d}\left(\frac{1}{N_g^d T_c^d}\right)^2\mathbf{1}\{g_j^d=g_i^d=g\}\mathbf{1}\{c_s^d=c_t^d=c\}u_{jsk}^2\right] \\
&= \mathbb{E}\left[\frac{1}{N_d T_d}\sum_{g=1}^{G_d}\sum_{c=1}^{C_d}\right]\mathbb{E}\left[u_{11k}^2\right] \\
&\leq \frac{G_d C_d}{N_d T_d}\max_{k=1,\ldots,K}\mathbb{E}\left[u_{11k}^2\right].
\end{aligned}
$$

This yields

$$
J_3 = O_P\left(\sqrt{\frac{G_d C_d}{NT}}\right).
$$

We obtain the result by combining the bounds on J_1, J_2, and J_3. □

Lemma 9 *Under Assumptions 1, 2, 3, 4, and 5, for all $k,\ell \in \{1,\ldots,K\}$ and $d \in \{1,\ldots,4\}$, we have*

$$
\begin{aligned}
\frac{1}{\sqrt{N_d T_d}}\sum_{(i,t)\in\mathcal{O}_d}\widetilde{h}^d_{itk}v_{it} &= o_P(1), \\
\frac{1}{\sqrt{N_d T_d}}\sum_{(i,t)\in\mathcal{O}_d}u_{itk}\widetilde{f}^d_{it} &= o_P(1), \\
\frac{1}{\sqrt{N_d T_d}}\sum_{(i,t)\in\mathcal{O}_d}\widetilde{h}^d_{itk}u_{it\ell} &= o_P(1).
\end{aligned}
$$

Proof. We only prove the first statement, as the proofs of the other two are similar. First,

notice that, by Assumption 3, $(v_{it})_{(i,t)\in\mathcal{O}_d}$ is independent of $(\widetilde{h}^d_{itk})_{(i,t)\in\mathcal{O}_d}$. Hence, we have

$$\mathbb{E}\left[\frac{1}{\sqrt{N_d T_d}}\sum_{(i,t)\in\mathcal{O}_d}\widetilde{h}^d_{itk}v_{it}\right]=0.$$

Moreover, because the v_{it} are i.i.d., it holds that

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{\sqrt{N_d T_d}}\sum_{(i,t)\in\mathcal{O}_d}\widetilde{h}^d_{itk}v_{it}\right)^2\right]\\
&=\mathbb{E}\left[\frac{1}{N_d T_d}\sum_{(i,t)\in\mathcal{O}_d}\left(\widetilde{h}^d_{itk}\right)^2 v_{it}^2\right]\\
&=\mathbb{E}\left[\frac{1}{N_d T_d}\sum_{(i,t)\in\mathcal{O}_d}\left(\widetilde{h}^d_{itk}\right)^2\right]\mathbb{E}[v_{11}^2].
\end{aligned}$$

By Lemma 6 and Assumptions 4 and 5, we have

$$\frac{1}{N_d T_d}\sum_{(i,t)\in\mathcal{O}_d}\left(\widetilde{h}^d_{itk}\right)^2=O_P\left(\frac{1}{T^2}+\frac{1}{N^2}+B^d_\alpha(G_d)^2+B^d_\gamma(C_d)^2\right)=o_P(1).$$

Since $\left(\widetilde{h}^d_{itk}\right)^2$ is bounded (because h_k is bounded itself by Assumption 1), this yields

$$\mathbb{E}\left[\frac{1}{N_d T_d}\sum_{(i,t)\in\mathcal{O}_d}\left(\widetilde{h}^d_{itk}\right)^2\right]=o(1).$$

We obtain the result since this implies

$$\mathbb{E}\left[\left(\frac{1}{\sqrt{N_d T_d}}\sum_{(i,t)\in\mathcal{O}_d}\widetilde{h}^d_{itk}v_{it}\right)^2\right]=o(1).$$

□