Discretizing Unobserved Heterogeneity[*]

Stéphane Bonhomme[†] Thibaut Lamadon[‡] Elena Manresa[§]

April 2017

Abstract

We study panel data estimators based on a discretization of unobserved heterogeneity when individual heterogeneity is not necessarily discrete in the population. We focus on *two-step grouped-fixed effects* estimators, where individuals are classified into groups in a first step using *kmeans* clustering, and the model is estimated in a second step allowing for group-specific heterogeneity. We analyze the asymptotic properties of these discrete estimators as the number of groups grows with the sample size, and we show that bias reduction techniques can improve their performance. In addition to reducing the number of parameters, grouped fixed-effects methods provide effective regularization. For instance, when allowing for the presence of time-varying unobserved heterogeneity we show they enjoy fast rates of convergence depending on the underlying dimension of heterogeneity. Finally, we document the finite sample properties of two-step grouped fixed-effects estimators in two applications: a structural dynamic discrete choice model of migration, and a model of wages with worker and firm heterogeneity.

JEL codes: C23, C38.

Keywords: Dimension reduction, panel data, structural models, kmeans clustering.

[*] We thank Anna Simoni, Manuel Arellano, Gary Chamberlain, Tim Christensen, Alfred Galichon, Chris Hansen, Joe Hotz, Arthur Lewbel, Anna Mikusheva, Roger Moon, Whitney Newey, Juan Pantano, Philippe Rigollet, Martin Weidner, and seminar audiences at the 2016 Summer Meetings of the Econometric Society, the 2016 Panel Data Conference in Perth, the 2016 IO/Econometrics Cornell/Penn State Conference, and various other places for comments. We also thank the IFAU for access to, and help with, the Swedish administrative data. The usual disclaimer applies.

[†] University of Chicago, sbonhomme@uchicago.edu

[‡] University of Chicago, lamadon@uchicago.edu

[§] Massachusetts Institute of Technology, the Sloan School of Management, emanresa@mit.edu

1 Introduction

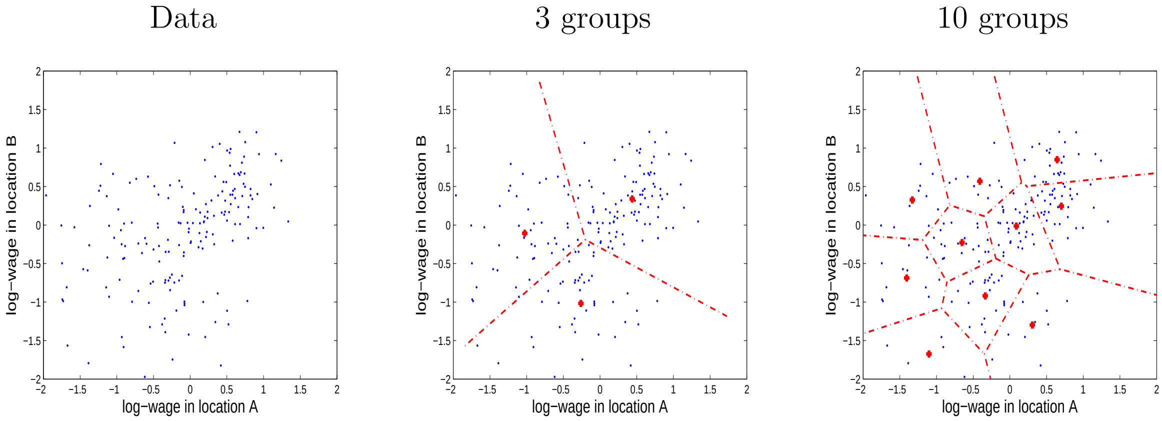
Unobserved heterogeneity is prevalent in modern economics, both in reduced-form and structural work, and accounting for it often makes large quantitative differences. In nonlinear panel data models fixed-effects approaches are conceptually attractive as they do not require restricting the form of unobserved heterogeneity. However, while these approaches are well understood from a theoretical perspective,[1] nonlinear fixed-effects estimators have not yet found wide applicability in empirical work. These methods raise computational difficulties due to the large number of parameters involved in estimation. Fixed-effects methods may also be infeasible in panels with insufficient variation, and they face challenges in the presence of multiple individual unobservables such as time-varying heterogeneity.

Discrete approaches to unobserved heterogeneity offer tractable alternatives. Consider as an example structural dynamic discrete choice models, which are popular in labor economics and industrial organizations. Starting with Keane and Wolpin (1997), numerous papers have modeled individual heterogeneity as a small number of unobserved types. In this context, discreteness is appealing for estimation as it leads to a finite number of unobserved state variables and reduces the number of parameters to estimate. However, the properties of discrete estimators have so far been studied under particular restrictions on the form of heterogeneity, typically under the assumption that heterogeneity is discrete in the population. In this paper we consider a class of easy-to-implement discrete estimators, and we study their properties in general nonlinear models while leaving the form of individual unobserved heterogeneity unspecified; that is, under "fixed-effects" assumptions.

We focus on *two-step grouped fixed-effects* estimation, which consists of a classification and an estimation steps. In a first step, individuals are classified based on a set of individual-specific moments using the *kmeans* clustering algorithm. Then, in a second step the model is estimated by allowing for group-specific heterogeneity. The aim of the kmeans classification is to group together individuals whose latent traits are most similar. The kmeans algorithm is a popular tool which has been extensively used and studied in machine learning and computer science, and fast and reliable implementations are available. Classifying individuals into types using kmeans is related to the grouped fixed-effects estimators recently introduced by Hahn and Moon (2010) and Bonhomme and Manresa (2015). However, unlike those methods, and

[1] Recent theoretical developments in the literature include general treatments of asymptotic properties of fixed-effects estimators as both dimensions of the panel increase, and methods for bias reduction and inference. This literature (as we do in this paper) focuses on models where all parameters, including the individual effects, are identified in a large-N, T sense. See among others Hahn and Newey (2004) and Arellano and Hahn (2007).

Figure 1: K-means clustering



Notes: Source NLSY79. The sample is described in Section 6. The kmeans partitions are indicated in dashed.

unlike random-effects methods such as finite mixtures, here the individual types and the model's parameters are estimated sequentially, as opposed to jointly.

When the number of groups is substantially smaller than the number of observations, two-step discrete estimators can improve computational tractability relative to existing methods. Figure 1 provides an illustration in a migration setting. According to the dynamic location choice model that we will describe in detail in Section 6, log-wages are informative about unobserved individual returns in locations A and B. Individuals are classified into groups based on location-specific means of log-wages. Depending on the number of groups K, the kmeans algorithm will deliver different partitions of individuals. Taking $K = 3$ will result in a drastic dimension reduction, however the approximation to the latent heterogeneity may be inaccurate. Taking a larger K, such as $K = 10$, may reduce approximation error while still substantially reducing the number of parameters relative to fixed-effects.

We characterize the statistical properties of two-step grouped fixed-effects estimators in settings where individual-specific unobservables are unrestricted. In other words, we use discrete heterogeneity as a dimension reduction device, instead of viewing discreteness as a substantive assumption about population unobservables. We show that grouped fixed-effects estimators generally suffer from an *approximation* bias that remains sizable unless the number of groups grows with the sample size. However, when the number of groups is relatively large estimating group membership becomes harder, and we show that this gives rise to an *incidental parameter* bias which has a similar order of magnitude as the one of conventional fixed-effects estimators.

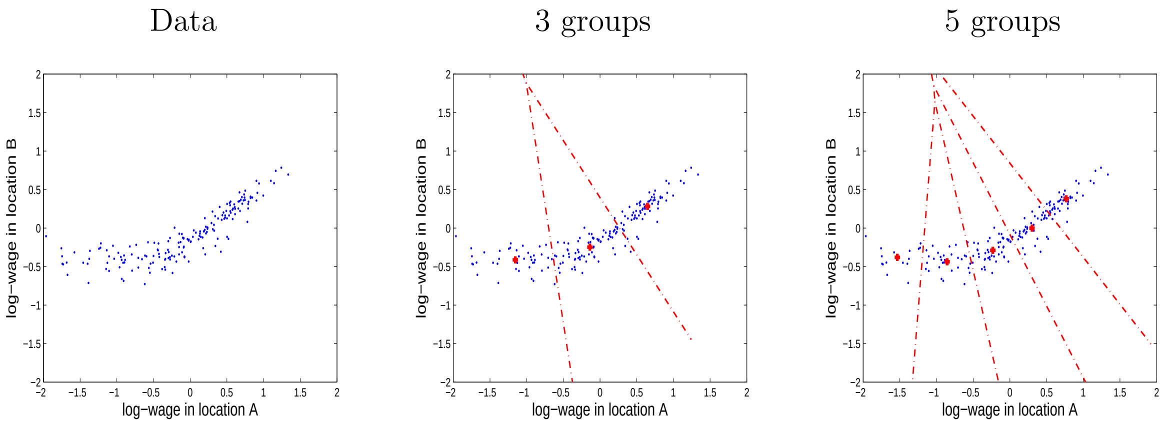
Importantly, our results show that estimation error in group membership has a non-negligible asymptotic impact on the performance of grouped fixed-effects estimators, which contrasts with existing results obtained under the assumption that heterogeneity is discrete in the population. Our asymptotic characterization motivates the use of bias reduction and inference methods from the literature on fixed-effects nonlinear panel data estimation. Specifically, we use the half-panel jackknife method of Dhaene and Jochmans (2015) to reduce bias.

Two-step grouped fixed-effects relies on two main inputs: the number of groups K, and the moments used for classification. We propose a simple data-driven choice of K which aims at controlling the approximation bias. We describe a generic approach to select moments based on individual-specific empirical distributions. Alternatively, moments such as individual means of outcomes or covariates can be used provided they are informative about unobserved heterogeneity (formally, an *injectivity* condition is needed). In addition, we propose a model-based iteration where individuals are re-classified based on the values of the group-specific objective function. We show in simulations that iterating may provide finite-sample improvements compared to the baseline two-step approach.

Implementation of our recommended two-step grouped fixed-effects procedure is straightforward. Given moments such as means or other characteristics of individual data, the kmeans algorithm is used to estimate the number of groups and the partition of individuals into groups. Given those, the model's parameters are estimated while allowing for group-specific fixed-effects. Bias-reduced estimates are then readily obtained by repeating the same procedure on two halves of the sample. Standard errors of bias-reduced estimators can be recovered using standard techniques. Finally, the model can be used to update the classification and compute an iterated estimator.

An appealing feature of grouped fixed-effects is its ability to exploit commonalities between different dimensions of heterogeneity. This can be seen in Figure 2, where in this example log-wages in the two locations are closely related to each other (that is, they approximately lie on a curve). Such a structure could arise from the presence of a one-dimensional ability factor, for example. The kmeans-based partition efficiently adapts to the data structure in a way that guarantees low approximation error. Consistently with this idea, we show that kmeans has fast rates of convergence even in cases where heterogeneity is high-dimensional, provided the *underlying* dimensionality of heterogeneity is low. In many economic models, agents' heterogeneity in preferences and technology is driven by low-dimensional economic types, which manifest themselves in potentially complex ways in the data. Through the use of kmeans, grouped fixed-effects provides a tool to exploit such underlying nonlinear factor

Figure 2: K-means in the presence of a low underlying dimension



Notes: Sample with the same conditional mean as in Figure 1, and one third of the conditional standard deviation. The kmeans partitions are indicated in dashed.

structures.[2]

We consider two extensions of the grouped fixed-effects approach where fixed-effects estimators are either infeasible or poorly behaved, and exploiting the presence of a low underlying dimension is key. In the first, the researcher's goal is to estimate a model on cross-sectional data or a short panel, while also having access to outside data (e.g., measurements of individual skills or firm productivity) which are informative about unobserved heterogeneity. We show that grouped fixed-effects estimators which use external measurements for classification have a similar asymptotic structure as in the baseline analysis, with the important difference that a statistical trade-off arises since setting K too large may worsen statistical performance. Hence, in this setting discretizing heterogeneity plays the role of a regularization scheme that reduces incidental parameter bias, in addition to alleviating the computational burden.

In the second extension we consider models where unobserved heterogeneity varies over time (that is, "time-varying fixed-effects"). Such models have applications in a variety of contexts, such as demand analysis in the presence of unobserved product attributes that vary across markets. We show that grouped fixed-effects estimators may enjoy fast rates of convergence depending on the underlying dimensionality of unobserved heterogeneity. For example, time-varying paths of unobservables have a low underlying dimension when they follow a low-

[2] Hence, though related to principal component analysis (PCA), kmeans differs from PCA as it allows the latent components to enter the model nonlinearly. See Hastie and Stuetzle (1989) and Chen, Hansen and Scheinkman (2009) for different approaches to nonlinear PCA.

dimensional linear or nonlinear factor structure, the interactive fixed-effects model of Bai (2009) being a special case. Our results provide a justification for using discrete estimators in settings where unobserved heterogeneity is high-dimensional, provided its underlying dimension is not too large.

We illustrate the properties of grouped fixed-effects methods in two different economic settings. First, we consider structural dynamic discrete choice models. Two-step methods provide alternatives to finite mixtures and related approaches, such as the ones developed in Arcidiacono and Jones (2003) and Arcidiacono and Miller (2011) for example. We set up a simulation exercise based on estimates from a simple dynamic model of location choice in the spirit of Kennan and Walker (2011), estimated on NLSY data. Using a data generating process with continuous heterogeneity, we assess the magnitude of the biases of grouped fixed-effects estimators and the performance of bias reduction.

Beyond traditional panel data applications, grouped fixed-effects methods are well-suited in "complex data" settings where several dimensions of heterogeneity interact with each other. As an illustration, we next revisit the estimation of workers' and firms' contributions to log-wage dispersion using matched employer-employee data. We focus on a short panel version of the model of Abowd, Kramarz and Margolis (1999), and report simulation results calibrated to Swedish administrative data. We compare the performance of two estimators: an additive version of the grouped fixed-effects estimator introduced in Bonhomme, Lamadon and Manresa (2015) which uses the wage distribution in the firm to classify firms into groups, and a fixed-effects estimator. We find that grouped fixed-effects alleviates the incidental parameter bias arising from low mobility rates of workers between firms.

Related literature and outline. The analysis of discrete estimators was initially done from a random-effects perspective, under functional form and/or independence assumptions on unobservables and how they relate to observed covariates. Heckman and Singer (1984)'s analysis of single-spell duration models provides a seminal example of this approach, in a setting where individual heterogeneity is independent of covariates and continuous. There is also a large literature on parametric and semi-parametric mixture models in statistics and econometrics; see McLachlan and Peel (2000), Frühwirth-Schnatter (2006), and Kasahara and Shimotsu (2009), among many others.

Previously to this paper, the properties of grouped fixed-effects estimators have been characterized under the assumption that unobserved heterogeneity is discrete in the population. Under suitable conditions, estimated type memberships converge to the true population types as both dimensions of the panel increase; see Hahn and Moon (2010), Lin and Ng (2012), Saggio (2012), Bonhomme and Manresa (2015), Bai and Ando (2015), Su, Shi and Phillips (2015), and Vogt and Linton (2015). In the context of structural dynamic discrete choice estimation, also under a discrete population framework, Buchinsky, Hahn and Hotz (2005) propose to classify types based on kmeans clustering and perform the Hotz and Miller (1993) estimation strategy using the estimated types. Pantano and Zheng (2013) use a related approach based on subjective expectations data.

There has been little work studying properties of discrete estimators as the sample size tends to infinity together with the number of groups. Important exceptions are Bester and Hansen (2016), who focus on a setup with known groups, and Gao, Lu and Zhou (2015) and Wolfe and Ohlede (2014), who derive results on stochastic blockmodels in networks.[3]

Finally, our analysis borrows from previous work on kmeans clustering and vector quantization; see among others Gersho and Gray (1992), Gray and Neuhoff (1998), Graf and Luschgy (2000, 2002), Linder (2002), and Levrard (2015), as well as the seminal analysis of kmeans by Pollard (1981, 1982a, 1982b).

The outline of the paper is as follows. We introduce the setup and two-step grouped fixed-effects estimators in Section 2. We study their asymptotic properties in Section 3. In Section 4 we focus on several practical aspects of the method: selection of the moments and the number of groups, and bias reduction and inference. In Section 5 we describe two extensions: grouped fixed-effects in short panels based on outside information for classification, and models with time-varying unobserved heterogeneity. We then present the two illustrations in Sections 6 and 7. Lastly we conclude in Section 8. A supplementary appendix contains additional results.[4]

2 Two-step grouped fixed-effects

We consider a panel data setup where outcome variables and covariates are denoted as $Y_i = (Y_{i1}, ..., Y_{iT})'$ and $X_i = (X_{i1}', ..., X_{iT}')'$, respectively, for $i = 1, ..., N$.[5] Following the literature (e.g., Hahn and Newey, 2004) the density of (Y_i, X_i), with respect to some measure, is denoted as $f(Y_i, X_i \,|\, \alpha_{i0}, \theta_0)$, where the α_{i0} are individual-specific vectors and θ_0 is a vector of common parameters. Throughout the analysis we leave the α_{i0} unrestricted, and we condition on them. In dynamic models the joint density is also conditioned on initial values (Y_{i0}, X_{i0}). We are

[3] Previous statistical analyses of stochastic blockmodels were done under discrete heterogeneity in the population; see for example Bickel and Chen (2009).

[4] Available at: https://sites.google.com/site/stephanebonhommeresearch/

[5] The focus on a balanced panel is for simplicity. One may allow T_i to differ across i's.

interested in estimating the parameter vector θ_0 as well as average effects depending on the individual effects α_{i0}, all of which are assumed to be identified. We defer formal assumptions until the next section.

In conditional models with strictly exogenous covariates we similarly denote the conditional density of Y_i given X_i as $f(Y_i \,|\, X_i, \alpha_{i0}, \theta_0)$. However in this case we do not specify the density of covariates parametrically. We allow the density of X_i to depend on an additional individual-specific vector μ_{i0} while leaving the relationship between X_i and (α_{i0}, μ_{i0}) unrestricted.

Hence the individual-specific distribution $f_i(Y_i, X_i)$ of (Y_i, X_i) depends on α_{i0}, or alternatively on (α_{i0}, μ_{i0}) in conditional models. We will see that the asymptotic properties of two-step grouped fixed-effects estimators will depend on the (underlying) dimension of α_{i0} or (α_{i0}, μ_{i0}); that is, on the dimensionality of individual heterogeneity. In conditional models, the dimension of the heterogeneity of the process of conditioning covariates will matter for performance in general. In the first part of the paper the dimension of α_{i0} or (α_{i0}, μ_{i0}) is kept fixed in the asymptotics. In this case fixed-effects is generally consistent as N, T tend to infinity, hence it is a natural benchmark to consider. In Section 5 we will instead consider settings where fixed-effects is *not* asymptotically well-behaved in general.

The two-step grouped fixed-effects method consists of a *classification* step and an *estimation* step. In the classification step we rely on a set of individual-specific moments $h_i = \frac{1}{T}\sum_{t=1}^{T} h(Y_{it}, X_{it})$ to learn about individual heterogeneity α_{i0}. Classification consists in partitioning individual units into K groups based on the moments h_i, where K is chosen by the researcher. In our asymptotic analysis we will require h_i to be informative about α_{i0} in a precise sense, and we will let K grow with the sample size. In Section 4 we will discuss the important questions of how to choose the moments and the number of groups. The partition of individual units, corresponding to group indicators $\widehat{k}_i$, is obtained by finding the best grouped approximation to the moments h_i based on K groups; that is, we solve:

$$\left(\widehat{h}, \widehat{k}_1, ..., \widehat{k}_N\right) = \underset{\left(\widetilde{h}, k_1, ..., k_N\right)}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| h_i - \widetilde{h}(k_i) \right\|^2, \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean norm, $\{k_i\} \in \{1, ..., K\}^N$ are partitions of $\{1, ..., N\}$ into at most K groups, and $\widetilde{h} = \left(\widetilde{h}(1)', ..., \widetilde{h}(K)'\right)'$ are vectors. Note that $\widehat{h}(k)$ is simply the mean of h_i in group $\widehat{k}_i = k$.

In the estimation step we maximize the log-likelihood function with respect to common parameters and group-specific effects, where the groups are given by the $\widehat{k}_i$ estimated in the first step. Letting $\ell_i(\alpha_i, \theta) = \ln f(Y_i \,|\, X_i, \alpha_i, \theta)/T$ denote the scaled individual log-likelihood, we define the estimator as:

$$\left(\widehat{\theta},\widehat{\alpha}\right)=\underset{(\theta,\alpha)}{\mathrm{argmax}}\ \sum_{i=1}^{N}\ell_i\left(\alpha\left(\widehat{k}_i\right),\theta\right),\tag{2}$$

where the maximization is with respect to θ and $\alpha=(\alpha(1)',...,\alpha(K)')'$.

The optimization problem in (1) is referred to as *kmeans* in machine learning and computer science. In (1) the minimum is taken with respect to all possible partitions $\{k_i\}$, in addition to values $\widetilde{h}(1),...,\widetilde{h}(K)$. Computing a global minimum may be challenging, yet fast and stable heuristic algorithms have been developed, such as iterative descent, genetic algorithms or variable neighborhood search. Lloyd's algorithm is often considered to be a simple and reliable benchmark.[6] In the asymptotic analysis, consistently with most of the statistical literature on classification estimators dating back to Pollard (1981, 1982a), we will focus on the properties of the global minimum in (1). Note that, while we focus on an unweighted version of kmeans, the quadratic loss function in (1) could accommodate different weights on different components of h_i (e.g., based on inverse variances).

The optimization problem in (2) involves estimating substantially fewer parameters than fixed-effects maximum likelihood. Indeed, the latter would require maximizing $\sum_{i=1}^{N}\ell_i\left(\alpha_i,\theta\right)$ with respect to θ and $\alpha_1,...,\alpha_N$ (this would correspond to taking $K=N$ in (2)). In contrast, in the estimation step in our approach one only needs to estimate K values $\alpha(1),...,\alpha(K)$. This dimension reduction can result in a substantial simplification of the computational task when K is small relative to N.

Let us now briefly introduce two illustrative examples to which we shall return several times.

Example 1: dynamic discrete choice model. A prototypical structural dynamic discrete choice model features the following elements (see for example Aguirregabiria and Mira, 2010): choices $j_{it}\in\{1,...,J\}$, payoff variables Y_{it}, and observed and unobserved state variables X_{it} and α_i, respectively. As an example, in the location choice model of Section 6, j_{it} is location at time t, and log-wages Y_{it} depend on latent location-specific returns $\alpha_i(j_{it})$. The individual log-likelihood function conditional on initial choices and state variables typically takes the form:

$$\ell_i(\alpha_i,\theta)=\frac{1}{T}\sum_{t=1}^{T}\underbrace{\ln f\left(j_{it}\,|\,X_{it},\alpha_i,\theta\right)}_{\text{choices}}+\underbrace{\ln f\left(X_{it}\,|\,j_{i,t-1},X_{i,t-1},\alpha_i,\theta\right)}_{\text{state variables}}+\underbrace{\ln f\left(Y_{it}\,|\,j_{it},X_{it},\alpha_i,\theta\right)}_{\text{payoff variables}}.\tag{3}$$

[6] See Steinley (2006) and Bonhomme and Manresa (2015) for algorithms and references. Different implementations of kmeans are available in standard software such as R, Matlab or Stata.

Note that in such models the law of motion of observed covariates (state variables X_{it}) is fully specified given α_i.

Computing choice probabilities $f\left(j_{it}\,|\,X_{it},\alpha_i,\theta\right)$ in (3) requires solving the dynamic optimization problem, which can be demanding. In two-step grouped fixed-effects, one estimates a partition $\{\widehat{k}_i\}$ in a first step that does not require solving the model. In the second step, the partition $\{\widehat{k}_i\}$ is taken as given and the log-likelihood in (3) is maximized with respect to θ and type-specific parameters $\alpha(k)$. This may reduce the computational burden compared both to fixed-effects maximum likelihood, and to random-effects mixture approaches which are commonly based on iterative algorithms. As moment vectors h_i to be used in the classification step one may take moments of payoff variables, observed state variables, and choices. One may also use individual-specific conditional choice probabilities, possibly based on a coarsened version of X_{it}. Moments will be required to satisfy an *injectivity* condition, to be defined in the next section. In the application in Section 6 we will use means of log-wages in a first step, while also relying on a likelihood-based iteration which exploits the full model's structure, hence using information on choices.

Example 2: linear regression. We will use a simple regression example to illustrate our assumptions and results. Consider the following model for a scalar outcome:

$$Y_{it} = \rho_0 Y_{i,t-1} + X_{it}'\beta_0 + \alpha_{i0} + U_{it}, \tag{4}$$

where $|\rho_0| < 1$. A two-step grouped fixed-effects estimator in this model can be based on the moment vector $h_i = (\overline{Y}_i, \overline{X}_i')'$. The estimation step then consists in regressing Y_{it} on $Y_{i,t-1}$, X_{it}, and group indicators. In this model with conditioning covariates the properties of two-step grouped fixed-effects will depend on the dimension (more precisely, the *underlying* dimension) of $(\alpha_{i0}, \mu_{i0}')'$, where $\mu_{i0} = \text{plim}_{T\rightarrow\infty} \frac{1}{T}\sum_{t=1}^{T} X_{it}$. In particular, performance will also depend on the dimensionality of the heterogeneity affecting the X's.

3 Asymptotic properties of two-step grouped fixed-effects

In this section we study asymptotic properties of the two steps in turn, classification and estimation, in an environment without any restriction on individual effects. At the end of the section we compare our results with previous results obtained under discrete population heterogeneity.

3.1 Classification step

Our first result is to derive a rate of convergence for the kmeans estimator $\widehat{h}(\widehat{k}_i)$ in (1). Let q and $r \geq q$ denote the dimensions of α_{i0} and h_i, respectively. The dimensions q and r are kept fixed as N, T, K tend jointly to infinity.[7] We make the following assumption.

Assumption 1. *(moments, first step) There is a Lipschitz continuous function φ such that, as N, T tend to infinity: $\frac{1}{N}\sum_{i=1}^{N} \|h_i - \varphi(\alpha_{i0})\|^2 = O_p\left(1/T\right)$.*

The probability limit of h_i is a function of α_{i0}, which indexes the joint distribution of (Y_i, X_i). The function φ depends on population parameter values, and need not be known to the econometrician. The rate in Assumption 1 will hold under weak conditions on the serial dependence of $\varepsilon_{it} = h(Y_{it}, X_{it}) - \varphi(\alpha_{i0})$, such as suitable mixing conditions, which are commonly made when studying asymptotic properties of fixed-effects panel data estimators.

Example 2 (continued). Consider classifying individuals based on the moment vector $h_i = (\overline{Y}_i, \overline{X}_i')'$ in Example 2. We have, under standard conditions: $\text{plim}_{T\to\infty}\, h_i = \left(\frac{\alpha_{i0}+\mu_{i0}'\beta_0}{1-\rho_0}, \mu_{i0}'\right)' = \varphi(\alpha_{i0}, \mu_{i0})$. In this example, as in conditional models more generally, there are thus two types of individual effects: those that enter the outcome distribution conditional on covariates (that is, α_{i0}), and those that only enter the distribution of covariates (that is, μ_{i0}). The full vector of individual effects to be approximated in the classification step is then (α_{i0}, μ_{i0}).[8]

Let us define the following quantity, which we refer to as the *approximation bias* of α_{i0}:

$$B_\alpha(K) = \min_{(\alpha, \{k_i\})} \frac{1}{N}\sum_{i=1}^{N} \|\alpha_{i0} - \alpha(k_i)\|^2,$$

where, similarly as in (1), the minimum is taken with respect to all $\{k_i\}$ and $\alpha(k)$. The term $B_\alpha(K)$ represents the approximation error one would make if one were to discretize the population unobservables α_{i0} directly. It is a non-increasing function of K. In conditional models such as Example 2 where the distribution of covariates depends on μ_{i0}, the relevant

[7] In Subsection 5.2 we will consider settings with time-varying unobserved heterogeneity where the dimensions of α_{i0} and h_i increase with the sample size.

[8] Note that there could be additional heterogeneity in the variance of h_i, for example, which need not be included in (α_{i0}, μ_{i0}). Correct specification of a Gaussian likelihood is not needed in this example. Moreover, given that $|\rho_0| < 1$ the impact of the initial condition Y_{i0} vanishes as T tends to infinity, so the marginal distribution of Y_{i0} can be left fully unrestricted.

approximation bias is $B_{(\alpha,\mu)}(K)$. Later we will review existing results about the convergence rate of $B_\alpha(K)$ (or alternatively $B_{(\alpha,\mu)}(K)$) in various settings.

We have the following characterization of the rate of convergence of $\widehat{h}(\widehat{k}_i)$. In the asymptotic we let $T = T_N$ and $K = K_N$ tend to infinity jointly with N. All proofs are in Appendix A.

Lemma 1. *Let Assumption 1 hold. Then, as N, T, K tend to infinity:*

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{h}(\widehat{k}_i) - \varphi(\alpha_{i0})\right\|^2 = O_p\left(\frac{1}{T}\right) + O_p\left(B_\alpha(K)\right).$$

Lemma 1 provides an upper bound on the rate of convergence of the discrete estimator $\widehat{h}(\widehat{k}_i)$ of $\varphi(\alpha_{i0})$. The bound has two terms: an $O_p(1/T)$ term which has a similar order of magnitude as the convergence rate of the fixed-effects estimator $h_i = \frac{1}{T}\sum_{t=1}^{T} h(Y_{it}, X_{it})$, and an $O_p\left(B_\alpha(K)\right)$ term which reflects the presence of an approximation error. Lemma 1 will be instrumental in deriving the asymptotic properties of estimators of common parameters and average effects in the next subsections. Nonetheless, using an alternative machine learning classifier in the first step will deliver second-step estimators with analogous properties, provided the classifier satisfies the convergence rate of Lemma 1.

Approximation bias: convergence rates. $B_\alpha(K)$ is closely related to the dimension of unobserved heterogeneity. This quantity has been extensively studied in the literature on vector quantization, where it is referred to as the "empirical quantization error".[9] Graf and Luschgy (2002) provide explicit characterizations in the case where α_{i0} has compact support with a nonsingular probability distribution.[10] As N, K tend to infinity, their Theorem 5.3 establishes that $B_\alpha(K) = O_p(K^{-\frac{2}{q}})$. This implies that $B_\alpha(K) = O_p(K^{-2})$ when α_{i0} is one-dimensional, and $B_\alpha(K) = O_p(K^{-1})$ when α_{i0} is two-dimensional, for example.

The quality of approximation of the discretization depends on the *underlying dimensionality* of the heterogeneity, not on its number of components. For example, when φ is Lipschitz we have: $B_{\varphi(\alpha)}(K) = O_p(B_\alpha(K))$.[11] This is precisely the reason why $B_\alpha(K)$ shows up in Lemma 1.

[9] Empirical quantization errors can be mapped to covering numbers commonly used in empirical process theory. Specifically, it can be shown that if the ϵ-covering number, for the Euclidean norm, of the set $\{\alpha_{10}, ..., \alpha_{N0}\}$ is such that $\mathcal{N}(\epsilon, \{\alpha_{i0}\}, \|\cdot\|) \geq K$, then $B_\alpha(K) \leq \epsilon^2$.

[10] While results on empirical quantization errors have been derived in the large-N limit under general conditions, see for example Theorem 6.2 in Graf and Luschgy (2000), rates as N and K tend to infinity jointly are so far limited to distributions with compact support; see p.875 in Graf and Luschgy (2002).

[11] This is a direct consequence of the fact that, if $\varphi(\alpha_{i0}) = a(\xi_{i0})$ and $\|a(\xi') - a(\xi)\| \leq \tau\|\xi' - \xi\|$ for all (ξ, ξ'), then: $\min_{(b,\{k_i\})} \frac{1}{N}\sum_{i=1}^{N}\|\varphi(\alpha_{i0}) - b(k_i)\|^2 \leq \tau^2 \min_{(\xi,\{k_i\})} \frac{1}{N}\sum_{i=1}^{N}\|\xi_{i0} - \xi(k_i)\|^2$.

More generally, if the dimensions of $\varphi(\alpha_{i0})$ are linked to each other in some way so its underlying dimension is low, the approximation bias may still be relatively small for moderate K. In those cases, discretizing the data jointly using kmeans may allow exploiting the presence of such a low dimension in the data as opposed to discretizing each component of h_i separately.[12] Note that the convergence rate in Lemma 1 does not depend on the actual dimension of h_i, only on the underlying dimension of α_{i0} (through the approximation bias $B_\alpha(K)$).

Convergence rate with many groups. We end this subsection by establishing a tighter bound on the rate of convergence of the kmeans estimator $\widehat{h}(\widehat{k}_i)$, when the number of groups is relatively large compared to T (though still possibly small relative to N).

Corollary 1. *Let $\varepsilon_i = h_i - \varphi(\alpha_{i0})$, and let $C = \text{plim}_{N,T\to\infty} \frac{1}{N}\sum_{i=1}^N T\left\|\varepsilon_i\right\|^2$. Suppose that there is an $\eta > 0$ such that $T\,B_{\varphi(\alpha)}(K^{1-\eta}) \xrightarrow{p} 0$ as N, T, K tend to infinity. Suppose also that, for any diverging $K_{N,T}$ sequence, $T\,B_\varepsilon(K_{N,T}) \xrightarrow{p} 0$ as N, T tend to infinity. Then, as N, T, K tend to infinity:*

$$\frac{1}{N}\sum_{i=1}^N \left\|\widehat{h}(\widehat{k}_i) - \varphi(\alpha_{i0})\right\|^2 = \frac{C}{T} + o_p\left(\frac{1}{T}\right).$$

Under the conditions of Corollary 1, the kmeans objective is: $\frac{1}{N}\sum_{i=1}^N \|h_i - \widehat{h}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right)$, hence in this regime grouped fixed-effects and fixed-effects are first-order equivalent. This happens when K grows sufficiently fast relative to T. As an example, when α_{i0} is scalar and $B_{\varphi(\alpha)}(K) = O_p(K^{-2})$ the condition requires TK^{-2} to tend to zero.[13] The condition on B_ε should be satisfied quite generally. As a simple example, it is satisfied when ε_i is normal with zero mean and variance Σ/T for some $\Sigma > 0$.

3.2 Estimation step

We now turn to the second step. In the following $\mathbb{E}_{\alpha_{i0}}$ denotes an expectation taken with respect to the joint distribution $f_i(Y_i, X_i)$, which depends on α_{i0}. For conciseness we simply write $\mathbb{E} = \mathbb{E}_{\alpha_{i0}}$. Similarly, $\mathbb{E}_\alpha$ is indexed by a generic α. In conditional models such as Example 2 the expectations are indexed by (α_{i0}, μ_{i0}) or a generic (α, μ).

[12] Another possibility would be to bypass the first step and optimize: $\sum_{i=1}^N \ell_i(\alpha(h_i), \theta)$ with respect to parameters θ and functions $\alpha(\cdot) : \mathbb{R}^r \to \mathbb{R}^q$ (belonging to some nonparametric class). By comparison, an attractive feature of the two-step approach we study is its ability to exploit low underlying dimensionality in h_i.

[13] In fact, if for some $d > 0$ the quantity $K^{\frac{2}{d}} B_{\varphi(\alpha)}(K)$ tends to a positive constant the first condition in Corollary 1 can be replaced by: $T\,B_{\varphi(\alpha)}(K) \xrightarrow{p} 0$.

Assumption 2. *(regularity)*

(i) *Observations are i.i.d. across individuals conditional on the α_{i0}'s. Moreover, $\ell_i(\alpha_i, \theta)$ is three times differentiable in both its arguments. In addition, the parameter spaces Θ for θ_0 and $\mathcal{A}$ for α_{i0} are compact, and θ_0 belongs to the interior of Θ.*

(ii) *For all $\eta > 0$, $\inf_{\alpha_{i0}} \inf_{\|(\alpha_i,\theta)-(\alpha_{i0},\theta_0)\|>\eta} \mathbb{E}[\ell_i(\alpha_{i0}, \theta_0)] - \mathbb{E}[\ell_i(\alpha_i, \theta)]$ is bounded away from zero for large enough T. For all $\theta \in \Theta$, let $\overline{\alpha}_i(\theta) = \text{argmax}_{\alpha_i} \lim_{T\to\infty} \mathbb{E}(\ell_i(\alpha_i, \theta))$. $\inf_{\alpha_{i0}} \inf_\theta \lim_{T\to\infty} \mathbb{E}(-\frac{\partial^2 \ell_i(\overline{\alpha}_i(\theta),\theta)}{\partial\alpha_i\partial\alpha_i'})$ is positive definite. $\lim_{N,T\to\infty} \frac{1}{N}\sum_{i=1}^N \mathbb{E}(\ell_i(\overline{\alpha}_i(\theta), \theta))$ has a unique maximum at θ_0 on Θ and its second derivative $-H$ is negative definite.*

(iii) *$\sup_{\alpha_{i0}} \sup_{(\alpha_i,\theta)} |\mathbb{E}(\ell_i(\alpha_i, \theta))| = O(1)$, $\max_{i=1,\dots,N} \sup_{(\alpha_i,\theta)} |\ell_i(\alpha_i, \theta) - \mathbb{E}(\ell_i(\alpha_i, \theta))| = o_p(1)$, and $\frac{1}{N}\sum_{i=1}^N (\ell_i(\alpha_{i0}, \theta_0) - \mathbb{E}(\ell_i(\alpha_{i0}, \theta_0)))^2 = O_p(T^{-1})$, and similarly for the first three derivatives of ℓ_i. $\sup_{\alpha_{i0}} \sup_\theta \|\frac{\partial}{\partial\alpha'}\big|_{\alpha_{i0}} \mathbb{E}_\alpha(\frac{\partial \ell_i(\overline{\alpha}_i(\theta),\theta)}{\partial\alpha_i})\| = O(1)$, $\sup_{\alpha_{i0}} \|\frac{\partial}{\partial\alpha'}\big|_{\alpha_{i0}} \mathbb{E}_\alpha(\text{vec}\, \frac{\partial^2 \ell_i(\alpha_{i0},\theta_0)}{\partial\theta\partial\alpha_i'})\| = O(1)$, and $\sup_{\alpha_{i0}} \|\frac{\partial}{\partial\alpha'}\big|_{\alpha_{i0}} \mathbb{E}_\alpha(\text{vec}\, \frac{\partial^2 \ell_i(\alpha_{i0},\theta_0)}{\partial\alpha_i\partial\alpha_i'})\| = O(1)$.*[14]

(iv) *The function $\widehat{\ell}_i(\theta) = \ell_i(\widehat{\alpha}(\widehat{k}_i, \theta), \theta)$ is three times differentiable on a neighborhood of θ_0, where $\widehat{\alpha}(k, \theta)$ for all k is the solution of (2) given any $\theta \in \Theta$. Moreover, $\frac{1}{N}\sum_{i=1}^N \|\frac{\partial^2 \widehat{\ell}_i(\theta)}{\partial\theta\partial\theta'}\|^2 = O_p(1)$ uniformly in a neighborhood of θ_0, and similarly for the third derivative of $\widehat{\ell}_i$.*

Most conditions in Assumption 2 are commonly assumed in nonlinear panel data models. The uniqueness of θ_0 and α_{i0} in (ii) is an identification condition. Hahn and Kuersteiner (2011) show that the convergence rates in (iii) are satisfied in stationary dynamic models under suitable mixing conditions on time-series dependence, existence of certain moments, and relative rates of N and T (specifically, $N = O(T)$, see their Lemma 4). The differentiability condition on the sample objective function in (iv) is not needed in order to characterize the first-order properties of fixed-effects estimators.[15] Theorem 1 can be established absent this condition when $\mathbb{E}(-\frac{\partial^2 \ell_i(\alpha_i,\theta_0)}{\partial\alpha_i\partial\alpha_i'})$ is uniformly bounded away from zero at all α_i, not only at the true α_{i0}.

Assumption 3. *(injective mapping) There exists a Lipschitz continuous function ψ such that $\alpha_{i0} = \psi(\varphi(\alpha_{i0}))$.*

[14] When A is a matrix, $\|A\|$ denotes the spectral norm of A.

[15] This is due to the fact that, under suitable conditions, fixed-effects estimators of individual effects are uniformly consistent in the sense that: $\max_{i=1,\dots,N} \|\widehat{\alpha}_i - \alpha_{i0}\| = o_p(1)$; see, e.g., Hahn and Kuersteiner (2011). In contrast, our characterization of grouped fixed-effects is based on establishing a rate of convergence for the average $\frac{1}{N}\sum_{i=1}^N \|\widehat{\alpha}(\widehat{k}_i) - \alpha_{i0}\|^2$.

Assumption 3 requires the individual moment $h_i = \frac{1}{T}\sum_{t=1}^{T} h(Y_{it}, X_{it})$ to be informative about α_{i0}, in the sense that $\text{plim}_{T\rightarrow\infty}\, h_i = \varphi(\alpha_{i0})$ and $\alpha_{i0} = \psi(\varphi(\alpha_{i0}))$, hence φ is *injective*. The injectivity of the mapping between the heterogeneity α_{i0} and the limiting moment $\varphi(\alpha_{i0})$ is a key requirement for consistency of two-step grouped fixed-effects estimators. In Section 4 we will describe a distribution-based moment choice which guarantees that φ is injective when the α_{i0}'s are identified. Finally, note that neither φ nor ψ need to be known to the econometrician.

Example 2 (continued). In Example 2, when using a grouped fixed-effects estimator based on a Gaussian quasi-likelihood, Assumptions 2 and 3 can be verified under standard conditions on error terms and covariates and a stationary initial condition, as done in Supplementary Appendix S1. In particular, the expectations in Assumption 2 are indexed by (α_{i0}, μ_{i0}) or a generic (α, μ). In addition, letting $\psi(h_{i1}, h_{i2}) = ((1-\rho_0)h_{i1} - h_{i2}'\beta_0, h_{i2}')'$, we have $(\alpha_{i0}, \mu_{i0}')' = \psi(\varphi(\alpha_{i0}, \mu_{i0}))$, so φ is injective. Note that both φ and ψ depend on true parameter values.

We now characterize asymptotic properties of the two-step grouped fixed-effects estimators of θ_0 and α_{i0}. For this, let us denote:

$$s_i = \frac{\partial \ell_i(\alpha_{i0}, \theta_0)}{\partial \theta} + \mathbb{E}\left(\frac{\partial^2 \ell_i(\alpha_{i0}, \theta_0)}{\partial \theta \partial \alpha_i'}\right)\left[\mathbb{E}\left(-\frac{\partial^2 \ell_i(\alpha_{i0}, \theta_0)}{\partial \alpha_i \partial \alpha_i'}\right)\right]^{-1}\frac{\partial \ell_i(\alpha_{i0}, \theta_0)}{\partial \alpha_i},$$

and:

$$\begin{aligned} H = \lim_{N,T\rightarrow\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left(-\frac{\partial^2 \ell_i(\alpha_{i0}, \theta_0)}{\partial \theta \partial \theta'}\right) \\ - \mathbb{E}\left(\frac{\partial^2 \ell_i(\alpha_{i0}, \theta_0)}{\partial \theta \partial \alpha_i'}\right)\left[\mathbb{E}\left(-\frac{\partial^2 \ell_i(\alpha_{i0}, \theta_0)}{\partial \alpha_i \partial \alpha_i'}\right)\right]^{-1}\mathbb{E}\left(\frac{\partial^2 \ell_i(\alpha_{i0}, \theta_0)}{\partial \alpha_i \partial \theta'}\right). \end{aligned}$$

The individual-specific efficient score for θ_0, s_i, coincides with the score of the *target* log-likelihood $\ell_i(\overline{\alpha}_i(\theta), \theta)$ (e.g., Arellano and Hahn, 2007, 2016). H is the corresponding Hessian matrix (it is non-singular by Assumption 2 (iii)). That is:

$$s_i = \left.\frac{\partial}{\partial \theta}\right|_{\theta_0} \ell_i\left(\overline{\alpha}_i(\theta), \theta\right), \quad H = \underset{N,T\rightarrow\infty}{\text{plim}} \; -\frac{1}{N}\sum_{i=1}^{N} \left.\frac{\partial^2}{\partial \theta \partial \theta'}\right|_{\theta_0} \ell_i\left(\overline{\alpha}_i(\theta), \theta\right). \qquad (5)$$

We have the following result.

Theorem 1. *Let Assumptions 1, 2 and 3 hold. Then, as N, T, K tend to infinity:*

$$\widehat{\theta} \;=\; \theta_0 + H^{-1}\frac{1}{N}\sum_{i=1}^{N} s_i + O_p\left(\frac{1}{T}\right) + O_p\left(B_\alpha(K)\right) + o_p\left(\frac{1}{\sqrt{NT}}\right), \qquad (6)$$

and:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\right\|^2 = O_p\left(\frac{1}{T}\right)+O_p\left(B_\alpha(K)\right). \tag{7}$$

Theorem 1 holds irrespective of the relative rates of N and K, so in particular K may be small relative to N. The result shows the presence of two types of bias for $\widehat{\theta}$: the approximation bias $B_\alpha(K)$ that vanishes as K increases, and a contribution akin to a form of incidental parameter bias that decreases at the rate $1/T$.[16] In conditional models such as Example 2 the relevant approximation bias is $B_{(\alpha,\mu)}(K)$.

The next corollary characterizes the properties of the grouped fixed-effects estimator of θ_0 as K grows relatively fast compared to T, but still slowly compared to N.

Corollary 2. *Let Assumptions 1, 2 and 3 hold, and suppose that the conditions of Corollary 1 are satisfied. Let $\widehat{\alpha}_i(\theta)=\text{argmax}_{\alpha_i}\,\ell_i(\alpha_i,\theta)$, $\widehat{g}_i(\theta)=\frac{\partial^2\ell_i(\widehat{\alpha}_i(\theta),\theta)}{\partial\theta\partial\alpha_i'}(\frac{\partial^2\ell_i(\widehat{\alpha}_i(\theta),\theta)}{\partial\alpha_i\partial\alpha_i'})^{-1}$, and let $\overline{\mathbb{E}}(\cdot\,|\,\cdot)$ be a conditional expectation* across *individuals (see the proof for details). Suppose in addition:*

(i) ℓ_i is four times differentiable, and its fourth derivatives satisfy similar uniform boundedness properties as the first three.

(ii) $\inf_{\alpha_{i0}}\inf_{\alpha_i}\lim_{T\to\infty}\mathbb{E}(-\frac{\partial^2\ell_i(\alpha_i,\theta_0)}{\partial\alpha_i\partial\alpha_i'})$ is positive definite.

(iii) $\gamma(h)=\overline{\mathbb{E}}[\widehat{\alpha}_i(\theta_0)\,|\,h_i=h]$ and $\lambda(h)=\overline{\mathbb{E}}[\widehat{g}_i(\theta_0)\,|\,h_i=h]$ are differentiable with respect to h, uniformly bounded with uniformly bounded first derivatives. Moreover, uniformly in h, $\overline{\mathbb{E}}[\|\widehat{\alpha}_i(\theta_0)-\gamma(h_i)\|^2\,|\,h_i=h]=O(T^{-1})$, $\overline{\mathbb{E}}[\|\widehat{g}_i(\theta_0)-\lambda(h_i)\|^2\,|\,h_i=h]=O(T^{-1})$, $\overline{\mathbb{E}}[\|\widehat{\alpha}_i(\theta_0)-\gamma(h_i)\|^3\,|\,h_i=h]=o(T^{-1})$, and $\overline{\mathbb{E}}[\|\widehat{g}_i(\theta_0)-\lambda(h_i)\|^3\,|\,h_i=h]=o(T^{-1})$.

Then, as N,T,K tend to infinity such that K/N tends to zero we have:

$$\widehat{\theta} \;=\; \theta_0+H^{-1}\frac{1}{N}\sum_{i=1}^{N}s_i+\frac{B}{T}+o_p\left(\frac{1}{T}\right)+o_p\left(\frac{1}{\sqrt{NT}}\right), \tag{8}$$

where the expression of the constant B is given in the proof.

[16] Although Theorem 1 is formulated in a likelihood setup, it holds more generally for M-estimators, interpreting $\sum_{i=1}^{N}\ell_i(\alpha_i,\theta)$ as the objective function in the M-estimation. In addition, a similar result holds for partial likelihood estimators where the objective function $\sum_{i=1}^{N}\ell_{i1}(\alpha_{i1},\theta_1)+\ell_{i2}(\alpha_{i1},\alpha_{i2},\theta_1,\theta_2)$ is maximized sequentially, first estimating (α_1,θ_1) based on ℓ_1, and then estimating (α_2,θ_2) given (α_1,θ_1) based on ℓ_2; see Supplementary Appendix S1. Such sequential estimators are commonly used in empirical applications, and we use this approach in our illustrations.

Condition (ii) requires the expected log-likelihood to be strictly concave with respect to α_i at all parameter values, not only at α_{i0}. This condition, which plays a technical role in the proof, was not used to establish Theorem 1.

Corollary 2 shows that, when K is sufficiently large so that the approximation bias $B_\alpha(K)$ is small relative to $1/T$, and when in addition K/N tends to zero, the grouped fixed-effects estimator of θ_0 satisfies a similar expansion as the fixed-effects estimator, with a different first-order bias term; see, e.g., Hahn and Newey (2004, p.1302) for an expression of the bias of fixed-effects. More generally, Theorem 1 and Corollary 2 imply that, when $B_\alpha(K)$ is of a similar or lower order of magnitude compared to $1/T$, the asymptotic distribution of two-step grouped fixed-effects estimators has a similar structure as that of conventional fixed-effects estimators. Like fixed-effects, grouped fixed-effects estimators suffer in general from an $O_p(1/T)$ bias term. In the next section we will show how a bias reduction technique can be used to improve the performance of grouped fixed-effects estimators, and discuss how to construct asymptotically valid confidence intervals as N/T tends to a constant.

3.3 Average effects

Average effects are of interest in many economic settings. For example, effects of counterfactual policies can often be written as averages over the cross-sectional agent heterogeneity. Here we characterize the asymptotic behavior of grouped fixed-effects estimators of such quantities.

Let $m_i\left(\alpha_i, \theta\right)$ be a shorthand for $\frac{1}{T}\sum_{t=1}^T m\left(X_{it}, \alpha_i, \theta\right)$. A grouped fixed-effects estimator of the population average $M_0 = \frac{1}{N}\sum_{i=1}^N m_i\left(\alpha_{i0}, \theta_0\right)$ is:

$$\widehat{M} = \frac{1}{N}\sum_{i=1}^N m_i\left(\widehat{\alpha}(\widehat{k}_i), \widehat{\theta}\right).$$

We make the following assumption.

Assumption 4. *(average effects)*

(i) $m_i(\alpha_i, \theta)$ *is twice differentiable with respect to* α_i *and* θ*.*

(ii) $\sup_{\alpha_{i0}} \mathbb{E}(\|m_i(\alpha_{i0}, \theta_0)\|) = O(1)$*,* $\max_{i=1,\ldots,N} \sup_{(\alpha_i,\theta)} \|m_i(\alpha_i, \theta)\| = O_p(1)$*, and similarly for the first two derivatives. In addition,* $\frac{1}{N}\sum_{i=1}^N \|\frac{\partial m_i(\alpha_{i0},\theta_0)}{\partial \theta'} - \mathbb{E}(\frac{\partial m_i(\alpha_{i0},\theta_0)}{\partial \theta'})\|^2 = O_p(T^{-1})$*,* $\frac{1}{N}\sum_{i=1}^N \|\frac{\partial m_i(\alpha_{i0},\theta_0)}{\partial \alpha_i'} - \mathbb{E}(\frac{\partial m_i(\alpha_{i0},\theta_0)}{\partial \alpha_i'})\|^2 = O_p(T^{-1})$*, and* $\sup_{\alpha_{i0}} \|\frac{\partial}{\partial \alpha'}\Big|_{\alpha_{i0}} \mathbb{E}_\alpha(\text{vec}\, \frac{\partial m_i(\alpha_{i0},\theta_0)}{\partial \alpha_i'})\| = O(1)$*.*

Given the quantities s_i and H introduced in the previous subsection, let us define:

$$s_i^m = \mathbb{E}\left(\frac{\partial m_i\left(\alpha_{i0},\theta_0\right)}{\partial \alpha_i'}\right)\left[\mathbb{E}\left(-\frac{\partial^2 \ell_i\left(\alpha_{i0},\theta_0\right)}{\partial \alpha_i \partial \alpha_i'}\right)\right]^{-1}\frac{\partial \ell_i(\alpha_{i0},\theta_0)}{\partial \alpha_i} + \mathbb{E}\left(\frac{\partial m_i\left(\alpha_{i0},\theta_0\right)}{\partial \theta'}\right) H^{-1}\frac{1}{N}\sum_{j=1}^N s_j$$
$$+\,\mathbb{E}\left(\frac{\partial m_i\left(\alpha_{i0},\theta_0\right)}{\partial \alpha_i'}\right)\left[\mathbb{E}\left(-\frac{\partial^2 \ell_i\left(\alpha_{i0},\theta_0\right)}{\partial \alpha_i \partial \alpha_i'}\right)\right]^{-1}\mathbb{E}\left(\frac{\partial^2 \ell_i\left(\alpha_{i0},\theta_0\right)}{\partial \alpha_i \partial \theta'}\right) H^{-1}\frac{1}{N}\sum_{j=1}^N s_j.$$

We have the following corollary to Theorem 1.

Corollary 3. *Let the assumptions of Theorem 1 hold. Let Assumption 4 hold. Then, as N, T, K tend to infinity:*

$$\widehat{M} \;=\; M_0 + \frac{1}{N}\sum_{i=1}^N s_i^m + O_p\left(\frac{1}{T}\right) + O_p\left(B_\alpha(K)\right) + o_p\left(\frac{1}{\sqrt{NT}}\right).$$

3.4 Comparison with results under discrete heterogeneity

It is useful to compare the results of this section, obtained in an environment where population heterogeneity is unrestricted and a growing number of groups K is used in estimation, to existing results on the performance of grouped fixed-effects estimators in discrete population settings. When the population consists of K^* groups, where K^* is a known fixed number, Hahn and Moon (2010) and Bonhomme and Manresa (2015) provide conditions under which estimated group membership $\widehat{k}_i$ tends in probability to the population group membership k_i^* for every individual i, up to arbitrary labeling of the groups.[17] Their conditions imply that the probability of misclassifying at least one individual unit tends to zero as N, T tend to infinity and N/T^η tends to zero for any $\eta > 0$. In this asymptotic the grouped fixed-effects estimators are not affected by incidental parameter bias. In other words, the asymptotic distribution of $\widehat{\theta}$ is not affected by the fact that group membership has been estimated.[18]

The results derived in this section contrast sharply with this previous literature. Under discrete population heterogeneity, according to perfect classification (or "oracle") results the grouped fixed-effects estimator $\widehat{h}(\widehat{k}_i)$ would have a convergence rate $O_p(1/NT)$. In contrast, here the convergence rate of $\widehat{h}(\widehat{k}_i)$ *cannot* be $o_p(1/T)$. Indeed, by definition we have:

[17] Assumptions include groups having positive probability and being separated in the population. Under suitable conditions K^* can be consistently estimated using information criteria or sequential tests.

[18] Although Hahn and Moon (2010) and Bonhomme and Manresa (2015) study joint estimation of parameters and groups, similar results to the ones they derive hold for two-step grouped fixed-effects estimators under discrete population heterogeneity.

$\frac{1}{N}\sum_{i=1}^{N}\|\widehat{h}(\widehat{k}_i)-\varphi(\alpha_{i0})\|^2 \geq B_{\varphi(\alpha)}(K)$ almost surely. Now suppose that the rate of $\widehat{h}(\widehat{k}_i)$ were $o_p(1/T)$. In that case $T\,B_{\varphi(\alpha)}(K)$ would tend to zero (corresponding to K growing sufficiently fast). However, from Corollary 1 the convergence rate of $\widehat{h}(\widehat{k}_i)$ would then be proportional to $1/T$ as in fixed-effects, leading to a contradiction.[19] "Oracle" asymptotic results thus appear fragile to departures from exact discreteness in the population.

Theorem 1 and Corollaries 2 and 3 show that, in an environment with possibly non-discrete heterogeneity, classification noise does affect the properties of second-step estimators. As the number of groups increases in order to control approximation bias, group membership estimates $\widehat{k}_i$ become increasingly noisy as groups become harder to distinguish. The order of magnitude of the resulting bias is $1/T$, as in conventional fixed-effects estimators. This framework, which leads to very different properties compared to previous results obtained under discrete heterogeneity, motivates combining discrete estimators with bias reduction techniques as we will describe in the next section.

4 Applying grouped fixed-effects

In this section we focus on practical aspects of grouped fixed-effects estimation. We first discuss the choice of moments for classification, and a model-based iteration. We then propose a method to select the number of groups. Finally, we show how to perform bias reduction and inference.

4.1 Choice of moments for the classification

When applying two-step grouped fixed-effects the choice of moments is a key input, since it determines the quality of the approximation to the unobserved heterogeneity. Specific models may suggest particular individual summary statistics to be used in the classification step. In linear models such as Example 2, individual averages of outcomes and covariates are natural choices. A general approach which does not rely on specific features of the model is to make use of the entire empirical distribution of the data, thereby capturing all the relevant heterogeneity in the classification step.

To outline the distribution-based approach, consider a static model with outcomes and exogenous covariates. Let $W_{it}=(Y_{it},X_{it}')'$, and denote $\widehat{F}_i(w)=\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}\{W_{it}\leq w\}$ the empirical cumulative distribution function of W_{it}.[20] We propose to classify individuals based

[19] This argument requires taking $\eta=0$ in the conditions of Corollary 1; see footnote 13 for a sufficient condition.

[20] In a dynamic setting such as Example 1, one could consider classifying individuals based on joint individual

on $h_i = \widehat{F}_i$, using the norm $\|g\|_\omega^2 = \int g(w)^2\omega(w)dw$, where ω is an integrable function. The classification step then is: $\min_{(\{k_i\},\widetilde{h})} \sum_{i=1}^N \|h_i - \widetilde{h}(k_i)\|_\omega^2$, where the $\widetilde{h}(k)$'s are functions. In practice we discretize the integral, leading to a standard (weighted) kmeans objective function. We discuss asymptotic properties in Supplementary Appendix S1, and show that for this choice of moments the injectivity condition of Assumption 3 is automatically satisfied when the α_{i0}'s are identified. We will use a distribution-based classification in the illustration on matched employer-employee data in Section 7.

Model-based iteration. Given two-step estimates $\widehat{\theta}$ and $\widehat{\alpha}$ from (2), a new partition of individual units can be computed according to the following model-based classification rule:

$$\widehat{k}_i^{(2)} = \underset{k\in\{1,...,K\}}{\operatorname{argmax}} \ \ell_i\left(\widehat{\alpha}\left(k\right), \widehat{\theta}\right), \qquad \text{for all } i = 1, ..., N. \tag{9}$$

This classification exploits the full structure of the likelihood model. Second-step estimates can then be updated as:

$$\left(\widehat{\theta}^{(2)}, \widehat{\alpha}^{(2)}\right) = \underset{(\theta,\alpha)}{\operatorname{argmax}} \sum_{i=1}^N \ell_i\left(\alpha\left(\widehat{k}_i^{(2)}\right), \theta\right). \tag{10}$$

The method may be iterated further. In Supplementary Appendix S1 we derive an asymptotic expansion for the iterated estimator $\widehat{\theta}^{(2)}$, similar to the one in Theorem 1, in fully specified likelihood models. We will see in the application in Section 6 that this likelihood-based iteration can provide improvements in finite samples.[21]

One-step estimation. A related estimator is the *one-step* grouped fixed-effects estimator, which is defined as follows:

$$\left(\widehat{\theta}^{1step}, \widehat{\alpha}^{1step}, \{\widehat{k}_i^{1step}\}\right) = \underset{(\theta,\alpha,\{k_i\})}{\operatorname{argmax}} \sum_{i=1}^N \ell_i\left(\alpha\left(k_i\right), \theta\right), \tag{11}$$

where the maximum is taken with respect to all possible parameter values (θ, α) and all possible partitions $\{k_i\}$ of $\{1, ..., N\}$ into at most K groups. This corresponds to the classification

frequencies such as: $h_i(j, j', x, x', y) = \frac{1}{T}\sum_{t=1}^T \mathbf{1}\{j_{i,t-1} \leq j, j_{it} \leq j', X_{i,t-1} \leq x, X_{it} \leq x', Y_{it} \leq y\}$. Such an approach could be combined with the model-based iteration described below.

[21] In addition, as an alternative to this likelihood-based approach the iteration can be based on modifying the moments h_i. Specifically, one can use $\widehat{\psi}(h_i)$ as moments in the classification step, where $\widehat{\psi}$ is a consistent estimate of any generalized inverse ψ appearing in Assumption 3. In the application to firm and worker heterogeneity in Section 7 we will show results using such a moment-based iteration.

maximum likelihood estimator of Bryant and Williamson (1978); see also Hahn and Moon (2010) and Bonhomme and Manresa (2015). Unlike in two-step grouped fixed-effects, (11) requires optimizing the likelihood function with respect to every partition and parameter value. This poses two difficulties. First, the estimator may be substantially more computationally intensive than two-step methods. Second, this complicates the statistical analysis since the discrete classification depends on parameter values and the objective function of the one-step estimator is therefore not smooth.[22] We now return to Example 2 and characterize properties of two-step and one-step grouped fixed-effects estimators in more detail.

Example 2 (continued). By Theorem 1, under conditions formally spelled out in Supplementary Appendix S1 the two-step estimators of ρ_0 and β_0 based on $h_i = (\overline{Y}_i, \overline{X}_i')'$ in model (4) have bias $O_p(1/T) + O_p(B_{(\alpha,\mu)}(K))$. Note that, as the dimension of X_{it} increases, $B_{(\alpha,\mu)}(K)$ decreases at a slower rate as a function of K. In Supplementary Appendix S1 we derive the first-order bias term for the one-step estimator (11) in model (4). Under normality, the bias takes a simple form that combines the bias of the within estimator with a "between" component which tends to zero as the number of groups increases. The rate of convergence of the approximation bias is $1/K^2$ in this case, irrespective of the dimension of the vector of covariates. This reflects the fact that one-step estimation delivers model-based moments which can improve the performance of grouped fixed-effects.[23]

4.2 Choice of the number of groups

The other key input for the method is the number of groups K. Here we propose a simple data-driven selection rule which aims at controlling approximation bias as the sample size increases. For simplicity the rule is based on the classification step alone. Let:

$$\widehat{Q}(K) = \min_{(h^K, \{k_i^K\})} \frac{1}{N} \sum_{i=1}^{N} \left\| h_i - h^K(k_i^K) \right\|^2$$

[22] In the case of the kmeans estimator, Pollard (1981, 1982a) derived asymptotic properties for fixed K and T as N tends to infinity. Deriving the properties of one-step estimators in (11) as N, T, K tend jointly to infinity is an interesting avenue for future work.

[23] In this example one can consider other possibilities for estimation that exploit features of the model in the classification step. In Supplementary Appendix S3 we present a "double grouped fixed-effects" estimator where we discretize all components of h_i *separately*, and include the indicators of estimated groups additively in the second-step regression. We report simulation results for a probit model. This strategy can be used in linear or linear-index models.

be the value of the kmeans objective function corresponding to K groups. For a given constant $\xi > 0$, we suggest taking:

$$\widehat{K} = \min_{K \in \mathbb{N}} \left\{ K, \, \widehat{Q}(K) \leq \xi \, \frac{\widehat{V}_h}{T} \right\}, \tag{12}$$

where $\widehat{V}_h$ is a consistent estimator of $V_h = \text{plim}_{N,T\to\infty} \frac{1}{N} \sum_{i=1}^{N} T \left\| h_i - \varphi(\alpha_{i0}) \right\|^2$.[24]

A default choice is $\xi = 1$. However, a more aggressive choice $\xi < 1$ may be preferable in situations where h_i is only weakly informative about α_{i0}. In practice we recommend taking $K = \widehat{K}$ with $\xi = 1$, and checking how the results of the grouped fixed-effects estimator and its bias-corrected version vary with ξ, as a check of whether the number of groups is sufficiently large to ensure a small approximation bias. We will illustrate the impact of ξ in our illustration on firm and worker heterogeneity.

We have the following result.

Corollary 4. *Let the conditions of Theorem 1 hold. Take $K \geq \widehat{K}$, where $\widehat{K}$ is given by (12). Then, as N, T tend to infinity:*

$$\widehat{\theta} \;\; = \;\; \theta_0 + H^{-1} \frac{1}{N} \sum_{i=1}^{N} s_i + O_p\left(\frac{1}{T}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right). \tag{13}$$

Expansion (13) holds for any $K \geq \widehat{K}$. In this environment (unlike the ones we consider in Section 5 below) taking $K = N$ as in fixed-effects also leads to (13). However, when the underlying dimensionality of unobserved heterogeneity is not too large Corollary 4 offers a justification for using a smaller K. Indeed, the data-driven rule to select K depends on this underlying dimensionality through the rate of decay of $\widehat{Q}(K)$. In particular, if $K^{\frac{2}{d}} \widehat{Q}(K)$ tends to a constant, where d is the underlying dimension of h_i, then $\widehat{K}$ in (12) is of the order of $T^{d/2}$.[25] As an example, when $d = 1$ $\widehat{K}$ will be of a similar order of magnitude as $\sqrt{T}$. In situations where $\sqrt{T}$ is small relative to N and the likelihood function is hard to evaluate or optimize, two-step grouped fixed-effects based on $\widehat{K}$ can thus represent a substantial decrease in computational cost compared to fixed-effects estimation.

[24] In the case where $\varepsilon_{it} = h(Y_{it}, X_{it}) - \varphi(\alpha_{i0})$ are independent over time, a consistent estimator of V_h is: $\widehat{V}_h = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \|h(Y_{it}, X_{it}) - h_i\|^2$. More generally, with dependent data, trimming or bootstrap strategies may be used for consistent estimation of V_h; see Hahn and Kuersteiner (2011) and Arellano and Hahn (2016).

[25] Under suitable conditions it can be shown that $\widehat{Q}(K) = O_p(B_{\varphi(\alpha)}(K)) + o_p(T^{-1})$, where the first term depends on the underlying dimensionality of $\varphi(\alpha_{i0})$.

4.3 Bias reduction and inference

The asymptotic analysis shows that grouped fixed-effects estimators and fixed-effects estimators of common parameters and average effects have a similar asymptotic structure, including when the number of groups is estimated (by Corollary 4). This similarity motivates adapting existing bias reduction techniques to grouped fixed-effects estimation. A variety of methods have been developed in the nonlinear panel data literature to perform bias reduction; see Arellano and Hahn (2007) for a review.

We consider the half-panel jackknife method of Dhaene and Jochmans (2015). Specifically, when estimating θ_0 half-panel jackknife works as follows:[26] We first compute the two-step grouped fixed-effects estimator $\widehat{\theta}$ on the full sample, using our data-driven selection of K. Then, we compute $\widehat{\theta}_1$ and $\widehat{\theta}_2$ on the first $T/2$ periods and the last $T/2$ periods, respectively, re-selecting K in each sample (considering T even for simplicity). The bias-reduced estimator is then:

$$\widehat{\theta}^{BR} = 2\widehat{\theta} - \frac{\widehat{\theta}_1 + \widehat{\theta}_2}{2}.$$

The half-panel jackknife method requires stationary panel data, however it can allow for serial correlation and dynamics.

To derive the asymptotic distribution of $\widehat{\theta}^{BR}$, let us suppose that, as N, T tend to infinity, the $O_p(1/T)$ term on the right-hand side of (13) takes the form $C/T + o_p(1/T)$, for some constant $C > 0$. For example, this will be the case when K is taken such that it grows sufficiently fast relative to T, under the conditions of Corollary 1. From Theorem 1, under standard conditions on the asymptotic behavior of the score $\frac{1}{N}\sum_{i=1}^N s_i$ the bias-reduced grouped fixed-effects estimator then has the following distribution as N, T tend to infinity such that N/T tends to a non-zero constant:

$$\sqrt{NT}\left(\widehat{\theta}^{BR} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \Omega\right). \tag{14}$$

In (14), Ω coincides with the asymptotic variance of the two-step grouped fixed-effects estimator, which in turn coincides with that of the fixed-effects estimator; that is:

$$\Omega = H^{-1}\left(\lim_{N,T\to\infty} \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[s_i s_i'\right]\right) H^{-1}.$$

This asymptotic variance can be consistently estimated using several methods, for example using a HAC formula clustered at the individual level, replacing α_{i0} and θ_0 by their (possibly bias-corrected) grouped fixed-effects estimates $\widehat{\alpha}(\widehat{k}_i)$ and $\widehat{\theta}$.

[26] From Corollary 3 the same bias-reduction and inference techniques can be used when estimating average effects M_0.

5 Grouped fixed-effects in other settings

We now consider two settings where, in contrast with the analysis so far, fixed-effects estimators are poorly behaved or infeasible and grouped fixed-effects provides a consistent alternative. In the first setting, in order to estimate the model on a short panel or a cross-section the researcher uses additional information ("measurements") about the unobserved heterogeneity. In the second case unobserved heterogeneity is time-varying. When the underlying dimension of unobserved heterogeneity is not too large, two-step grouped fixed effects provides accurate estimates of parameters of interest.

5.1 Classification based on outside information

Consider a setting where the time dimension available to estimate the model is short. We denote the number of periods as S. A special case is $S=1$, where only cross-sectional data is available. Outcomes $Y_i=(Y_{i1},...,Y_{iS})$ and covariates $X_i=(X_{i1}',...,X_{iS}')'$ are drawn from the individual-specific distribution $f_i(Y_i,X_i)$, which depends on α_{i0}. As in the previous sections, $f(Y_i\,|\,X_i,\alpha_{i0},\theta_0)$ is indexed by a parameter vector θ_0, while the conditional distribution of X_i given α_{i0} is unrestricted.[27]

Suppose the researcher has access to T *measurements* $W_i=(W_{i1}',...,W_{iT}')'$ drawn from an individual-specific distribution $f_i(W_i)$ indexed by the same individual heterogeneity α_{i0}. Individual summary statistics $h_i=\frac{1}{T}\sum_{t=1}^{T}h(W_{it})$ are assumed to be informative about α_{i0} according to Assumptions 1 and 3.[28] We assume that, while S may be very small, T is relatively large. Unlike S, the number of measurements T will be required to tend to infinity in the asymptotic analysis. Moreover, another important difference with the setup considered in the previous sections is that the measurements of α_{i0} are assumed independent of the outcome variables and covariates of interest.

Assumption 5. *(measurements) h_i and (Y_i,X_i) are conditionally independent given α_{i0}.*

Classifying individuals according to outside measurements may be natural in a number of situations in economics. For example, in structural models of the labor market the researcher may have access to measures of academic ability or some dimensions of skills (cognitive or non-cognitive, as in Cunha *at al.*, 2010), such as test scores or psychometric measures taken before

[27] In conditional models $f_i(Y_i,X_i)$ is indexed by (α_{i0},μ_{i0}), where the conditional distribution of X_i given (α_{i0},μ_{i0}) is unrestricted.

[28] In conditional models where the distribution of (Y_i,X_i) depends on (α_{i0},μ_{i0}), the moments h_i need to be informative about (α_{i0},μ_{i0}).

the individual entered the labor market. Consistency of two-step grouped fixed-effects in these settings will rely on measurements W_i and outcomes and covariates (Y_i, X_i) depending on the *same* vector of unobserved traits α_{i0}.

Another example is the decomposition of log-wage dispersion into terms reflecting worker and firm heterogeneity (as in Abowd *et al.*, 1999). In Section 7 we will show that the grouped fixed-effects estimator of Bonhomme *et al.* (2015), where the distribution of wages in the firm is used for classification, fits the setup analyzed here. We will study its performance in simulation experiments, and show that it can alleviate the finite-sample bias which arises from low worker mobility rates.

We now turn to the asymptotic properties of grouped fixed-effects in this context. We make the following assumptions, where $\ell_i(\alpha_i, \theta) = \ln f(Y_i \,|\, X_i, \alpha_{i0}, \theta_0)/S$.

Assumption 6. *(regularity) Part (i) in Assumption 2 holds. In addition:*

(i) For each i $(Y_{i1}, ..., Y_{iS})$ and $(X'_{i1}, ..., X'_{iS})'$ are stationary. $\overline{\alpha}_i(\theta)$ and θ_0 uniquely maximize $\mathbb{E}(\ell_i(\alpha_i, \theta))$ and $\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N \mathbb{E}(\ell_i(\overline{\alpha}_i(\theta), \theta))$, respectively. The minimum eigenvalue of $(-\frac{\partial^2 \ell_i(\alpha_i,\theta)}{\partial\alpha_i \partial\alpha'_i})$ is bounded away from zero almost surely, uniformly in i and (α_i, θ).

(ii) $\sup_{\alpha_{i0}} \sup_{(\alpha_i,\theta)} |\mathbb{E}(\ell_i(\alpha_i, \theta))| = O(1)$, and similarly for the first three derivatives of ℓ_i. Second and third derivatives of $\ell_i(\alpha_i, \theta)$ are uniformly $O_p(1)$ in (α_i, θ) and i. In addition, $\sup_{\alpha_{i0}} \sup_{\theta} \|\frac{\partial}{\partial\alpha'}\big|_{\alpha_{i0}} \mathbb{E}_{\alpha}(\frac{\partial \ell_i(\overline{\alpha}_i(\theta),\theta)}{\partial\alpha_i})\| = O(1)$, $\sup_{\alpha_{i0}} \|\frac{\partial}{\partial\alpha'}\big|_{\alpha_{i0}} \mathbb{E}_{\alpha}(\text{vec}\, \frac{\partial^2 \ell_i(\alpha_{i0},\theta_0)}{\partial\theta\partial\alpha'_i})\| = O(1)$, and $\sup_{\alpha_{i0}} \|\frac{\partial}{\partial\alpha'}\big|_{\alpha_{i0}} \mathbb{E}_{\alpha}(\text{vec}\, \frac{\partial^2 \ell_i(\alpha_{i0},\theta_0)}{\partial\alpha_i\partial\alpha'_i})\| = O(1)$.

(iii) $\sup_{\alpha_{i0}} \sup_{\theta} \text{Var}(\frac{\partial \ell_i(\overline{\alpha}_i(\theta),\theta)}{\partial\alpha_i}) = O(1/S)$, and $\sup_{\alpha_{i0}} \text{Var}(\text{vec}\, \frac{\partial}{\partial\theta'}|_{\theta_0} \frac{\partial \ell_i(\overline{\alpha}_i(\theta),\theta)}{\partial\alpha_i}) = O(1/S)$.

Strict concavity of the log-likelihood in (i) was not required in Assumption 2. This limits the scope of the theorem to strictly concave likelihood models. Examples of strictly concave panel data likelihood models are the Logit, Probit, ordered Probit, Multinomial Logit, Poisson, or Tobit regression models; see Chen *et al.* (2014) and Fernández-Val and Weidner (2015). In regression models part (ii) requires covariates X_{is} to have bounded support. Note that here we do not require the log-likelihood function to be concave in all parameters, only in individual effects. The other conditions in Assumption 6 are similar to those in Assumption 2, with the difference that here there are S available periods on every individual in the second step, where S may or may not tend to infinity.

We have the following result.[29]

[29] An analogous result to Theorem 2 also holds for partial likelihood estimation under similar conditions, although for conciseness we do not formally spell it out.

Theorem 2. *Let Assumptions 1, 3, 5 and 6 hold. Then, as N, T, K tend to infinity such that K/NS tends to zero:*

$$\widehat{\theta} = \theta_0 + H^{-1}\frac{1}{N}\sum_{i=1}^{N} s_i + O_p\left(\frac{1}{T}\right) + O_p(B_\alpha(K)) + O_p\left(\frac{K}{NS}\right) + o_p\left(\frac{1}{\sqrt{NS}}\right).$$

Under the conditions of Theorem 2 a fixed-effects estimator only based on (Y_i, X_i), which maximizes the likelihood $\sum_{i=1}^N \ell_i(\alpha_i, \theta)$, satisfies: $\widehat{\theta}^{FE} = \theta_0 + H^{-1}\frac{1}{N}\sum_{i=1}^N s_i + O_p(S^{-1}) + o_p((NS)^{-\frac{1}{2}})$. In particular, fixed-effects may be severely biased when S is small, and it is generally inconsistent for S fixed. In contrast, since it takes advantage of the measurements data, the two-step grouped fixed-effects estimator is still consistent even when $S = 1$, as N, T, K tend to infinity such that K/NS tends to zero.

The expansion in Theorem 2 is similar to that in Theorem 1, with one important difference: here, unlike in the setting analyzed in the previous sections, increasing K comes at a cost that is reflected in the term $O_p(K/NS)$. Intuitively, when choosing K too large the grouped fixed-effects estimator gets close to fixed-effects, which is generally not well-behaved asymptotically in this setting.[30] Hence, in this environment discretizing unobserved heterogeneity has a second advantage in addition to lowering the computational burden, as the discrete regularization leads to a reduction of the incidental parameter bias.

5.2 Time-varying unobserved heterogeneity

We now return to the setup of Section 2, with the difference that unobserved heterogeneity $\alpha_{i0} = (\alpha_{i0}(1)', ..., \alpha_{i0}(T)')'$ is time-varying, where $\alpha_{i0}(t)$ has fixed dimension q. We focus on static models where the likelihood function takes the form:

$$\ln f(Y_i \,|\, X_i, \alpha_i, \theta) = \sum_{t=1}^{T} \ln f(Y_{it} \,|\, X_{it}, \alpha_i(t), \theta),$$

and denote $\ell_{it}(\alpha_i(t), \theta) = \ln f(Y_{it} \,|\, X_{it}, \alpha_i(t), \theta)$ and $\ell_i(\alpha_i, \theta) = \ln f(Y_i \,|\, X_i, \alpha_i, \theta)/T$. As before we leave the relationship between X_i and α_{i0} unrestricted.

Allowing for time-varying unobserved heterogeneity is of interest in many economic settings. For example, in demand models for differentiated products unobserved product characteristics

[30] This feature of the problem should be kept in mind when using the data-driven selection of K proposed in Subsection 4.2. Indeed, in order to obtain an analogous result to Corollary 4 in the setting of Theorem 2, one needs to take K such that K/NS is $O(T^{-1})$. This requires: $\widehat{K}/NS = O_p(T^{-1})$. Developing a data-driven method to select K that is justified more generally is left to future work.

may vary across markets t as well as products i (as in Berry *et al.*, 1995). Fixed-effects methods are popular alternatives to instrumental variables strategies. As an example, Moon, Shum and Weidner (2014) model unobserved product characteristics through a factor-analytic "interactive fixed-effects" specification in the spirit of Bai (2009). In comparison, here we show that grouped fixed-effects methods are able to approximate general unobservables with low underlying dimensionality, through delivering a data-based classification of products in terms of their unobserved attributes.

Let $r \geq qT$. Let $h_i = h(Y_i, X_i)$ be an r-dimensional vector with $h_i = \varphi(\alpha_{i0}) + \varepsilon_i$, where here the function φ maps $\mathbb{R}^{qT}$ to $\mathbb{R}^r$. Here the h_i's may contain entire *sequences* of outcomes or covariates, for example. Let us start with a definition and an assumption.

Definition 1. *(sub-Gaussianity) Let Z be a random vector of dimension m. We say that Z is sub-Gaussian if there exists a scalar constant $\lambda > 0$ such that $\mathbb{E}\left[\exp(\tau' Z)\right] \leq \exp(\lambda \|\tau\|^2)$ for all $\tau \in \mathbb{R}^m$.*

Assumption 7. *(moments, first step) $\varepsilon = (\varepsilon_1', ..., \varepsilon_N')'$ satisfies Definition 1 for a constant λ independent of the sample size. In addition, the ratio r/T tends to a positive constant as T tends to infinity, φ is Lipschitz continuous, and there exists a Lipschitz continuous ψ such that $\alpha_{i0} = \psi(\varphi(\alpha_{i0}))$.*

Assumption 7 requires the $\varepsilon = (\varepsilon_1', ..., \varepsilon_N')'$ to be sub-Gaussian (e.g., Vershynin, 2010). This is stronger than Assumption 1. For example, i.i.d. Gaussian random variables and i.i.d. bounded random variables are sub-Gaussian. More generally, this assumption allows for dependence across observations. As an example, in the case where $\varepsilon \sim \mathcal{N}(0, \Sigma)$ Assumption 7 holds provided the maximal eigenvalue of Σ is bounded from above by 2λ. This allows for weak forms of dependence in ε, across both individual units and time periods.[31] This condition is only needed for the models with time-varying heterogeneity of this subsection. Lastly, φ is required to be injective similarly as in Assumption 3.

Assumption 8. *(regularity) Part (i) in Assumption 2 holds (with $\mathcal{A}$ denoting the parameter space for $\alpha_{i0}(t)$). In addition:*

(i) For all i, t, θ, $\mathbb{E}(\ell_{it}(\alpha_i(t), \theta))$ has a unique maximum on $\mathcal{A}$, denoted as $\overline{\alpha}_i(\theta, t)$. θ_0 uniquely maximizes $\lim_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \mathbb{E}(\ell_{it}(\overline{\alpha}_i(\theta, t), \theta))$. In addition, the minimum eigenvalue of $(-\frac{\partial^2 \ell_{it}(\alpha_i(t), \theta)}{\partial \alpha_i(t) \partial \alpha_i(t)'})$ is bounded away from zero almost surely, uniformly in i, t, and $(\alpha_i(t), \theta)$.

[31] Related conditions have been used in the literature on large approximate factor models (Chamberlain and Rothschild, 1983, Bai and Ng, 2002).

(ii) $\sup_{\alpha_{i0}(t)} \sup_{(\alpha_i(t),\theta)} |\mathbb{E}(\ell_{it}(\alpha_i(t),\theta))| = O(1)$, *and similarly for the first three derivatives of* ℓ_{it}. *Moreover, second and third derivatives of* $\ell_{it}(\alpha_i(t),\theta)$ *are uniformly* $O_p(1)$ *in* $(\alpha_i(t),\theta)$ *and* i,t. *Further,* $\sup_{\alpha_{i0}(t)} \sup_{\theta} \|\frac{\partial}{\partial \alpha(t)'}|_{\alpha_{i0}(t)} \mathbb{E}_{\alpha(t)}(\frac{\partial \ell_{it}(\overline{\alpha}_i(\theta,t),\theta)}{\partial \alpha_i(t)})\| = O(1)$, $\sup_{\alpha_{i0}(t)} \|\frac{\partial}{\partial \alpha(t)'}|_{\alpha_{i0}(t)} \mathbb{E}_{\alpha(t)}(\text{vec}\, \frac{\partial^2 \ell_{it}(\alpha_{i0}(t),\theta_0)}{\partial \theta \partial \alpha_i(t)'})\| = O(1)$, *and we have in addition* $\sup_{\alpha_{i0}(t)} \|\frac{\partial}{\partial \alpha(t)'}|_{\alpha_{i0}(t)} \mathbb{E}_{\alpha(t)}(\text{vec}\, \frac{\partial^2 \ell_{it}(\alpha_{i0}(t),\theta_0)}{\partial \alpha_i(t) \partial \alpha_i(t)'})\| = O(1)$.

(iii) *For each* $\theta \in \Theta$, $(T \frac{\partial \ell_i(\overline{\alpha}_i(\theta),\theta)}{\partial \alpha_i})_{i=1,...,N}$ *satisfies Definition 1 for a common constant* λ.[32] *Moreover,* $(T \text{ vec}\, \frac{\partial}{\partial \theta'}|_{\theta_0} \frac{\partial \ell_i(\overline{\alpha}_i(\theta),\theta)}{\partial \alpha_i})_{i=1,...,N}$ *satisfies Definition 1.*

Similarly as Assumption 5, Assumption 8 restricts the scope of the next theorem to likelihood models that are strictly concave in α's. In particular, concavity is used to establish consistency of the grouped fixed-effects estimator. The tail condition on scores in part (iii) is also instrumental in order to deal with the presence of time-varying unobserved heterogeneity. In Supplementary Appendix S1 we provide sufficient conditions for Assumption 8 in a regression example (Example 3 below).

Theorem 3. *Let Assumptions 7 and 8 hold. Then, as* N, T, K *tend to infinity such that* $(\ln K)/T$, K/N, *and* $B_\alpha(K)/T$ *tend to zero:*

$$\widehat{\theta} = \theta_0 + H^{-1} \frac{1}{N} \sum_{i=1}^{N} s_i + O_p\left(\frac{\ln K}{T}\right) + O_p\left(\frac{K}{N}\right) + O_p\left(\frac{B_\alpha(K)}{T}\right), \tag{15}$$

and:

$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left\| \widehat{\alpha}(\widehat{k}_i, t) - \alpha_{i0}(t) \right\|^2 = O_p\left(\frac{\ln K}{T}\right) + O_p\left(\frac{K}{N}\right) + O_p\left(\frac{B_\alpha(K)}{T}\right). \tag{16}$$

In Theorem 3 the expansion of $\widehat{\theta}$ and the convergence rate of $\widehat{\alpha}(\widehat{k}_i, t)$ have three components. The K/N part reflects the fact that we are estimating KT parameters (that is, the $\alpha(k,t)$) using NT observations. Hence, as in Theorem 2, and unlike the setup studied in the first part of the paper, here increasing K may worsen the convergence rate. The $(\ln K)/T$ term is equal to the logarithm of the number of possible partitions of N individual units into K groups (that is, K^N) divided by the number of observations. This term reflects the presence of an incidental parameter bias due to noisy group classification, similarly as the $1/T$ term in Theorem 1. It arises from the application of a union bound argument and the use of the sub-Gaussianity assumptions.[33]

[32] Here we denote: $\overline{\alpha}_i(\theta) = (\overline{\alpha}_i(\theta,1)', ..., \overline{\alpha}_i(\theta,T)')'$.

[33] A related term arises as a component of the convergence rate in the study of network stochastic blockmodels in Gao, Lu and Zhou (2015).

The third component of the rate in Theorem 3 is the scaled approximation bias:[34]

$$\frac{B_\alpha(K)}{T} = \min_{(\alpha,\{k_i\})} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \|\alpha_{i0}(t) - \alpha(k_i, t)\|^2 .$$

As in the case studied in Section 3 this term depends on the underlying dimensionality of $\alpha_{i0}(t)$. When no restrictions are made on $\alpha_{i0}(t)$ except bounded support, one can only bound $B_\alpha(K)/T$ by $O_p(K^{-\frac{2}{qT}})$, which does not tend to zero unless K is extremely large relative to T. Restrictions on the underlying dimension of $\alpha_{i0}(t)$ allow one to separate its contribution from that of time-varying errors. Examples of latent processes with low underlying dimensionality are linear or nonlinear factor models of the form $\alpha_{i0}(t) = \alpha(\xi_{i0}, t)$, where α is Lipschitz continuous in its first argument and the factor loading ξ_{i0} has fixed dimension $d > 0$. In that case the approximation bias is $B_\alpha(K)/T = O_p(K^{-\frac{2}{d}})$. In Supplementary Appendix S3 we show the results of a small simulation exercise that illustrates the convergence rate in (16).

It is interesting to interpret the present setup with time-varying heterogeneity as a semi-parametric model. In this perspective the K/N term in Theorem 3 arises due to the $\alpha_{i0}(t)$ being fully unrestricted functions of time. A faster rate could be achieved under smoothness restrictions, when using kernel or sieve estimators in the second step to estimate the group-specific nonparametric functions $\widehat{\alpha}(k, t)$. More generally, when combining grouped fixed-effects with nonparametric methods grouping can lead to asymptotic gains relative to a pure fixed-effects approach, even when population heterogeneity is not discrete.[35] A key condition is that the underlying dimensionality of the (functional) individual-specific parameters be sufficiently low.

Lastly, unlike Theorems 1 and 2, Theorem 3 cannot be directly used to motivate the use of standard bias-reduction techniques, since $(\ln K)/T$ dominates $1/T$ asymptotically. In models with time-varying heterogeneity, the development of bias reduction and inference methods for common parameters and average effects, and of methods to select K, are important questions that we leave for future work.

Example 3: regression with time-varying unobservables. In this example we allow for time-varying unobservables possibly correlated with covariates (that is, "time-varying fixed-

[34] Similarly as in Theorems 1 and 2, in conditional models where the distribution of X_{it} also depends on $\mu_{i0}(t)$ the relevant approximation bias is $B_{(\alpha,\mu)}(K)/T$.

[35] An example is a nonparametric regression of the form: $Y_{it} = m(X_{it}, \alpha_{i0}) + U_{it}$, where m is a nonparametric function. Vogt and Linton (2015) study such a model under discrete population heterogeneity.

effects") in a regression model, and consider:

$$Y_{it} = X_{it}'\beta_0 + \alpha_{i0}(t) + U_{it}. \tag{17}$$

As an example, in a logit demand model Y_{it} could be the log market share of product i in market t, and X_{it} could include the price and other observed product attributes. Let $X_{it} = \mu_{i0}(t) + V_{it}$. Grouped fixed-effects may be based on $h_i = (Y_i', X_i')'$, in which case ε_i is a linear transformation of $(U_i', V_i')'$. Assumption 7 then requires $(U_i', V_i')'$ to be sub-Gaussian, as shown in Supplementary Appendix S1. Theorem 3 implies that $\widehat{\beta}$ is consistent as N, T, K tend to infinity such that $(\ln K)/T$, K/N, and $B_\alpha(K)/T$ tend to zero. The result is quite general as it allows for unspecified form of heterogeneity, although the performance of the estimator may not be as good when the underlying dimension of $(\alpha_{i0}(t), \mu_{i0}(t))$ is large. A special case of Example 3 is when $\alpha_{i0}(t)$ has an exact grouped structure, as in Bonhomme and Manresa (2015). Another special case with low underlying dimension is the interactive fixed-effects model with $\alpha_{i0}(t) = \lambda_{0i}' f_{0t}$, as in Bai (2009) and Pesaran (2006). Compared to interactive fixed-effects, an advantage of grouped fixed-effects is that the researcher need not specify the functional form of time-varying unobservables. This may be of particular interest in nonlinear models where explicitly allowing for linear or nonlinear factor-analytic specifications can be computationally challenging (see Chen *et al.*, 2014).[36]

6 A dynamic model of location choice

In this section we study a structural dynamic model of location choice. In such environments discrete estimation improves tractability since it leads to unobserved state variables having a small number of points of support and the number of parameters being relatively small. Here we report simulation results for two-step estimators and their bias-corrected and iterated versions, which show the ability of discrete approximations to deliver accurate estimates of structural parameters and counterfactual effects.

[36] Similarly as in Example 2, an alternative estimator is "double grouped fixed-effects", where Y_i and all components of X_i are used separately to form sets of groups, which are then all included as controls additively in the regression and interacted with time indicators. Under similar conditions it can be shown that the convergence rate of the estimator of β_0 is the same as in Theorem 3, except that the relevant approximation bias is the maximum among the *unidimensional* approximation biases corresponding to $\alpha_{i0}(t)$ and all components of $\mu_{i0}(t)$. Hence, as in Example 2, specific features of the model (here its additive structure) can be exploited in order to improve statistical performance.

6.1 Model and estimation

We consider a model of location choices over J possible alternatives. There is a continuum of agents i who differ in their permanent type $\alpha_i \in \mathbb{R}^J$ which defines their wage in each location. Log-wages in location j, net of age effects and other demographics, are given by: $\ln W_{it}(j) = \alpha_i(j) + \varepsilon_{it}(j)$, where $\varepsilon_{it}(j)$ are assumed to be i.i.d over time, agents, and locations, distributed as normal $(0, \sigma^2)$. The flow utility of being in location j at time t is given by: $U_{it}(j) = \rho W_{it}(j) + \xi_{it}(j)$, where $\xi_{it}(j)$ are unobserved shocks i.i.d across agents, time and locations, and distributed as type-I extreme value. When moving between two locations j and j' the agent faces a cost $c(j,j') = c\,\mathbf{1}\{j' \neq j\}$.

Agent i faces uncertainty about her own type α_i. While we assume she knows the distribution from which the components of α_i are drawn, she only observes $\alpha_i(j)$ in the locations j she has visited, and she forms expectations about the value she might get in locations she has not visited yet. At time t, let $\mathcal{J}_{it}$ denote the set of locations that agent i has visited. Let $\alpha_i(\mathcal{J}_{it})$ denote the set of corresponding realized location-specific types. The information set of the agent is: $S_{it} = (j_{it}, \mathcal{J}_{it}, \alpha_i(\mathcal{J}_{it}))$. Note that we assume that wage shocks $\varepsilon_{it}(j)$ do not affect the decision to move to another location. This assumption is useful for tractability, though not essential to our approach.

We consider an infinite horizon environment, where agents discount time at a common β. At time t, let $V_t(j, S_{i,t-1})$ denote the expected value function associated with choosing location j given state $S_{i,t-1}$ and behaving optimally in the future. Value functions are derived in Supplementary Appendix S2. The conditional choice probabilities are then:

$$\Pr(j_{it} = j \mid S_{i,t-1}) = \frac{\exp V_t(j, S_{i,t-1})}{\sum_{j'=1}^{J} \exp V_t(j', S_{i,t-1})}. \tag{18}$$

Given an i.i.d sample of wages and locations $(W_{i1}, ..., W_{iT}, j_{i1}, ..., j_{iT})$ we first estimate the location-specific returns $\alpha_i(j_{it})$ as follows:

$$(\widehat{\alpha}, \{\widehat{k}_i\}) = \underset{(\alpha, \{k_i\})}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\ln W_{it} - \alpha(k_i, j_{it})\right)^2, \tag{19}$$

which amounts to classifying individuals according to location-specific means of log-wages.

In the second step, we maximize the log-likelihood of choices; that is:

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{j=1}^{J} \mathbf{1}\{j_{it} = j\} \ln \Pr\left(j_{it} = j \mid j_{i,t-1}, \mathcal{J}_{i,t-1}, \widehat{\alpha}(\widehat{k}_i, \mathcal{J}_{i,t-1}), \theta\right), \tag{20}$$

where θ contains the structural parameters (including utility and cost parameters ρ and c), and $\mathcal{J}_{i,t-1}$ denotes the set of locations visited by i until period $t-1$. The likelihood is conditional on the initial location and location-specific return of the agent in period 0. We use a steepest ascent algorithm to maximize the objective in (20), analogous to the nested fixed point method of Rust (1994).[37] Lastly, given parameter estimates $\widehat{\alpha}(k,j)$, $\widehat{\sigma}^2$, and $\widehat{\theta}$, one can update the estimated partition of individuals using the full model's structure, as in (9) and (10). Details are given in Supplementary Appendix S2.

Note that agents in the model face uncertainty about future values of realized types $\alpha_i(j)$. The decision problem is only discretized for estimation purposes. This approach contrasts with a model with discrete types in the population, where after observing her type in one location the agent would in many cases be able to infer her type in all other possible locations. In the present setup, uncertainty about future types diminishes over time as the agent visits more locations, but it does not disappear (until all locations have been visited).

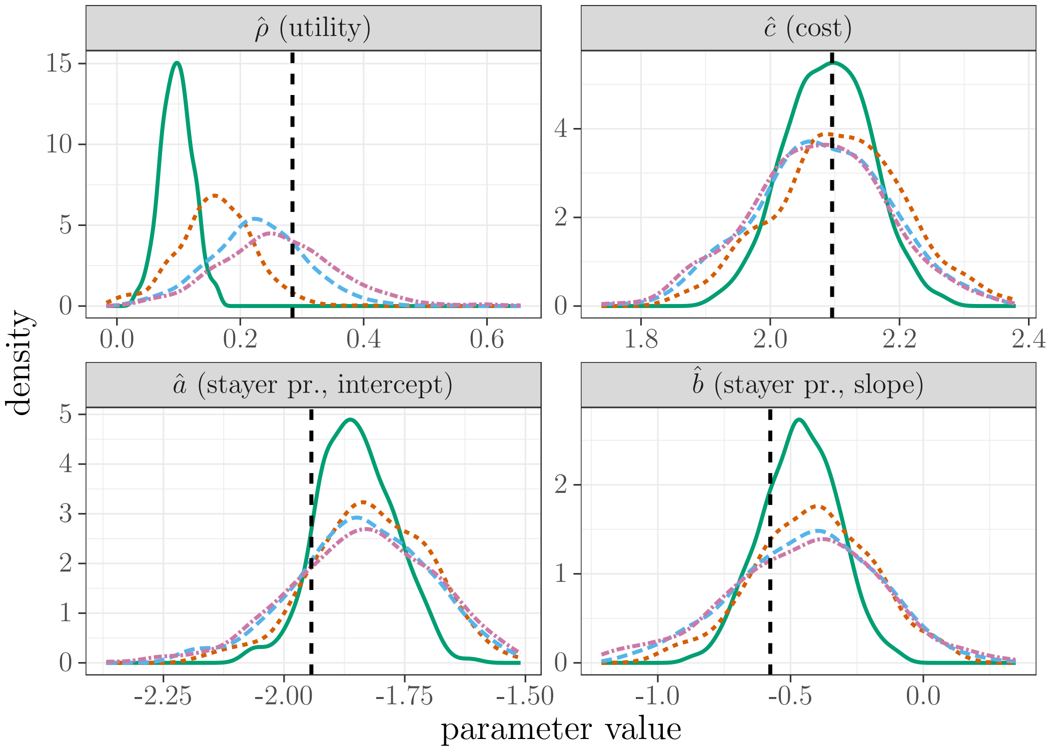
6.2 Simulation exercise

We calibrate the model to the NLSY79, using observations on males who were at least 22 years old in 1979. We keep observations until 1994. Log-wages are regressed on indicators of years of education and race and a full set of age indicators. We then compute log-wage residuals $\ln W_{it}$. We focus on a stylized setup with $J = 2$ large regions: North-East and South (region A), and North-Central and West (B). There are $N = 1889$ workers, who are observed on average for 12.3 years with a maximum of $T = 16$ years. The probability of moving between the two regions is low in the data: 1.5% per year, and only 10.5% of workers move at all during the observation period. Mean log-wage residuals are .09 higher in region A compared to B.

To construct the data generating process (DGP) we first estimate the model using grouped fixed-effects with $K = 10$ groups, with an iteration starting from a first step based on location-specific mean log-wages as in (19). Here $\alpha_i(j)$ is the expected log-wage in location j, so location-specific means of log-wages satisfy the injectivity condition in Assumption 3. Following Kennan and Walker (2011), each agent is a "stayer type" with some probability (which depends on the initial $\alpha_i(j_{i1})$ though a logistic specification), in which case his mobility cost is infinite; "mover types" have mobility cost c. Hence, while the main parameters of interest are ρ and c, the model also features the intercept and slope coefficients in the probability of being a stayer type (a and b). The estimates we obtain are $\widehat{\rho} = .28$, $\widehat{c} = 2.10$, $\widehat{a} = -1.94$, and $\widehat{b} = -.58$. According to the DGP the probability of being a stayer type is high and depends negatively on the initial

[37] Alternatively, in this model the CCP method of Hotz and Miller (1993) or the iterative method of Aguirregabiria and Mira (2002) could be used. A computational alternative to maximize the objective in (20) is the MPEC method of Su and Judd (2012).

Figure 3: Parameter estimates across simulations



Notes: Solid is two-step grouped fixed-effects, dotted is bias-corrected, dashed is iterated once and bias-corrected, dashed-dotted is iterated three times and bias-corrected. The vertical line indicates the true parameter value. $N = 1889$, $T = 16$. *Unobserved heterogeneity is continuously distributed in the DGP. The number of groups K is estimated in every replication.* 500 *replications.*

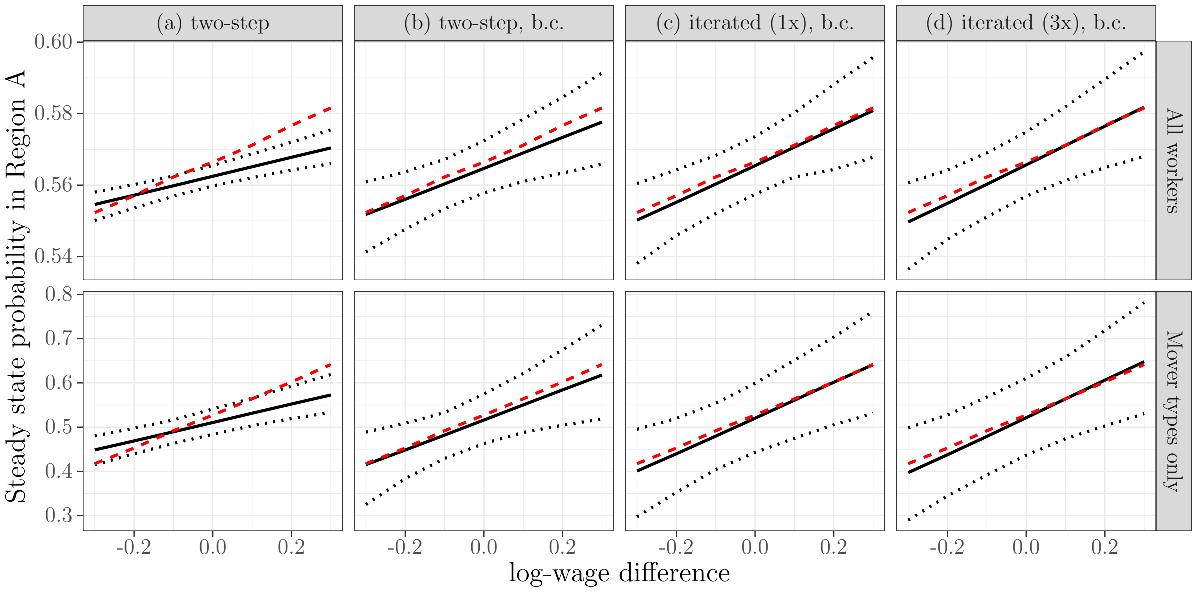
location-specific return $\alpha_i(j_{i1})$. Costs are also high for non-stayer types. The effect of wages on utility is positive, although we will see below that it is quantitatively small.[38]

We next solve and simulate the model (as described in Supplementary Appendix S2) based on these parameter values, together with i.i.d. normal specifications of shocks to log wages and $(\alpha_i(A), \alpha_i(B))$, with means and variances calibrated to our estimates. The model is simulated for $T = 16$ periods, the α's being drawn independently of the initial location. Note that the α's are not discrete in the DGP, although we use a discrete approach in estimation. In Figure 3

[38] The model reproduces well the probability of moving, both unconditionally and conditional on past wages; in particular it reproduces the negative relationship between past wages and mobility. It also reproduces means and variances of log-wages by location. However, the model does not fit well average wages after mobility, as it tends to predict mean wage increases upon job move while the data do not show such a pattern.

we report the results of 500 Monte Carlo simulations for the four parameters of the model. We use a kmeans routine with 100 randomly generated starting values, and checked that varying this number had no noticeable impact on the results. We estimate the number of groups based on (12), with $\xi = 1$, in every simulation (and in every subsample when using half-panel jackknife for bias correction). We show four types of estimates: two-step grouped fixed-effects (solid curve), bias-corrected two-step grouped fixed-effects (dotted), a single iteration and bias-corrected (dashed), and iterated three times and bias-corrected (dashed-dotted).

Figure 4: Long-run effects of wages



Notes: Difference in log-wage between the two regions (x-axis), and steady-state probability of working in region A (y-axis). (a) is two-step grouped fixed-effects, (b) is bias-corrected, (c) is iterated once and bias-corrected, (d) is iterated three times and bias-corrected. The dashed curve indicates the true value. Solid curves are means, and dotted curves are 97.5% and 2.5% percentiles, across simulations. 500 *replications.*

The results in Figure 3 show that grouped-fixed-effects estimators have moderate bias for all parameters except for the wage coefficient in utility ρ.[39] Using both bias reduction and an iteration improves the performance of the estimator of ρ substantially. Note that, when combined with half-panel jackknife, a single iteration seems sufficient to reduce bias. At the same time, bias reduction tends to be associated with a variance increase. In Supplementary

[39] The estimated number of groups $\widehat{K}$ is around 7 on average.

Appendix S2 we show the results of various alternative estimators: two-step grouped fixed-effects for fixed values of K, a fixed-effects estimator and its bias-corrected counterpart, and a random-effects estimator with a fixed number of types computed using the EM algorithm. We find that setting K too low relative to our suggested procedure may be associated with less accuracy, and that the statistical performance of two-step grouped fixed-effects estimators (which enjoy computational advantages) is competitive with fixed-effects and random-effects alternatives.

Counterfactual: long-run effects of wages. As an example of a counterfactual experiment that the model can be used for, we next compute the steady-state probability of working in region A when varying the log-wage differential between A and B (that is, $\mathbb{E}(\alpha_i(A) - \alpha_i(B))$). In Figure 4 we show the log-wage differential on the x-axis, and the probability of working in A on the y-axis. The dashed curves on the graphs show the estimates from the NLSY, while the solid and dotted curves are means and 95% pointwise bands across simulations for the two-step estimator, two-step bias-corrected, iterated once and bias-corrected, and iterated three times and bias-corrected, respectively.[40]

The results in Figure 4 show that the model predicts small effects of wages on mobility on average. When increasing the wage in A relative to B by 30% the probability of working in A increases by less than 2 percentage points, from 56.7% to 58.2%. When focusing on workers who are not of a "stayer type" (bottom panel), whose mobility may be affected by the change in wages, we see a more substantial effect, as increasing the wage in region A by 30% increases the long-run probability of working in A from 52.8% to 64.0%. In both cases the two-step estimators are biased downward. In contrast the bias-corrected and bias-corrected iterated estimators are close to unbiased. However they are less precisely estimated, reflecting the estimates of the wage coefficient ρ in Figure 3.

This application illustrates the potential usefulness of discrete grouped fixed-effects estimators in the presence of continuous unobserved heterogeneity. Computation of two-step estimators is particularly easy, and the jackknife bias reduction and the iteration (only once, or three times) provide finite-sample improvements at moderate computational cost. This stylized illustration thus suggests that the methods we propose could be useful in structural models.

[40] In the counterfactual we keep the probability of being a "stayer type" constant. Hence we abstract from the fact that wage increases could affect the distribution of mobility costs, in addition to the effect on utility that we focus on.

7 Firm and worker heterogeneity

In the second illustration we consider the question of assessing the sources of wage dispersion across workers and firms. We consider an additive model in worker and firm heterogeneity:

$$Y_{it} = \eta_i + \psi_{j(i,t)} + \varepsilon_{it}, \tag{21}$$

where Y_{it} denote log-wages, worker i works in firm $j(i,t)$ at time t, and η_i and ψ_j denote unobserved attributes of worker i and firm j, respectively. Equation (21) corresponds to the model of Abowd *et al.* (1999) for matched employer-employee data, where we abstract from observed covariates for simplicity. Our interest centers on the decomposition of the variance of log-wages into a worker component, a firm component, a component reflecting the sorting of workers into heterogeneous firms, and an idiosyncratic match component:

$$\operatorname{Var}(y_{i1}) = \operatorname{Var}(\eta_i) + \operatorname{Var}\left(\psi_{j(i,1)}\right) + 2\operatorname{Cov}\left(\eta_i, \psi_{j(i,1)}\right) + \operatorname{Var}(\varepsilon_{i1}). \tag{22}$$

Identification of firm effects ψ_j comes from job movements. As an example, with two time periods the fixed-effects estimators of the ψ_j's are obtained from:

$$Y_{i2} - Y_{i1} = \psi_{j(i,2)} - \psi_{j(i,1)} + \varepsilon_{i2} - \varepsilon_{i1},$$

which is uninformative for workers who remain in the same firm in the two periods. When the number of job movers into and out of firm j is low, ψ_j may be poorly estimated; see Abowd *et al.* (2004), Andrews *et al.* (2008), and Jochmans and Weidner (2016). This source of incidental parameter bias may be particularly severe in short panels.

To alleviate this "low-mobility bias", Bonhomme *et al.* (2015, BLM) propose to reduce the number of firm-specific parameters by grouping firms based on firm-level observables in a first step. Then, in a second step, the ψ_j's are recovered at the group level, thus pooling information across job movers within firm groups. Specifically, given a kmeans-based classification $\{\widehat{k}_j\}$ of firms, the $\psi(k)$'s are estimated based on the following criterion:

$$\min_{(\psi(1),\ldots,\psi(K))} \sum_{i=1}^{n} \left\| \widehat{\mathbb{E}}\left(Y_{i2} - Y_{i1} \,|\, \widehat{k}_{j(i,1)}, \widehat{k}_{j(i,2)}\right) - \psi\left(\widehat{k}_{j(i,2)}\right) + \psi\left(\widehat{k}_{j(i,1)}\right) \right\|^2,$$

where $\widehat{\mathbb{E}}$ denotes a group-pair average and n denotes the number of workers, subject to a single normalization (e.g., $\psi(K) = 0$).

This estimator is a two-step grouped fixed-effects estimator based on outside information, as analyzed in Subsection 5.1. Here N is the number of firms, S is the number of available observations to estimate the firm-specific parameters ψ_j (that is, S is the number of job movers per firm), and T is the number of measurements on firm heterogeneity. In BLM firms are classified based on their empirical wage distribution functions. Using only job stayers in the classification is consistent with conditional independence in Assumption 5, provided wage observations are independent within firms. In this case T is the number of job stayers per firm, which is typically much larger than the number of job movers in short panels.[41]

Simulation exercise. We focus on a two-period model, where ε_{it} are independent of $j(i,1)$, $j(i,2)$, η's, and ψ's, have zero means, and are i.i.d. across workers and time. Following BLM we adopt a correlated random-effects approach to model worker heterogeneity within firms. The parameters of the model are the firm fixed-effects ψ_j, the means and variances of worker effects in each firm $\mu_j = \mathbb{E}\left(\eta_i \,|\, j(i,1)=j\right)$ and $\sigma_j^2 = \text{Var}\left(\eta_i \,|\, j(i,1)=j\right)$, and the variance of idiosyncratic errors $s^2 = \text{Var}\left(\varepsilon_{i1}\right)$. We will be estimating the components of the variance decomposition in (22). In addition we will report estimates of the correlation between worker and firm effects, $\text{Corr}\left(\eta_i, \psi_{j(i,1)}\right)$, which is commonly interpreted as a measure of sorting.

In the baseline DGP firm heterogeneity is continuous and three-dimensional, and its underlying dimension equals one. Specifically, the vector of firm-specific parameters is:

$$\alpha_j = \left(\psi_j\,,\, \mu_j\,,\, \sigma_j^2\right) = \left(\psi_j,\, \mathbb{E}\left(\eta_i | \psi_{j(i,1)} = \psi_j\right)\,,\, \text{Var}\left(\eta_i | \psi_{j(i,1)} = \psi_j\right)\right),$$

so all firm-specific parameters are (nonlinear) functions of the scalar firm effects ψ_j. This specification is consistent with theoretical models of worker-firm sorting where firms are characterized by their scalar productivity level.[42] In Supplementary Appendix S2 we report simulations using several alternative designs. We study cases where the underlying dimension of firm heterogeneity is equal to two, which allows for a second dimension of latent firm heterogeneity in addition to the wage effects ψ_j (and we provide evidence suggesting that the underlying dimension of firm heterogeneity is low in the data). We also consider a DGP where firm heterogeneity is discrete in the population.

We start by estimating model (21) on Swedish register data, following BLM. We select male workers full-year employed in 2002 and 2004, and define as job movers the workers whose firm

[41] A difference with the setting of Theorem 2 is that here the likelihood function is not separable across firms, due to the fact that two firms are linked by the workers who move between them. We conjecture that Theorem 2 could be extended to such network settings, although formally developing this extension exceeds the scope of this paper.

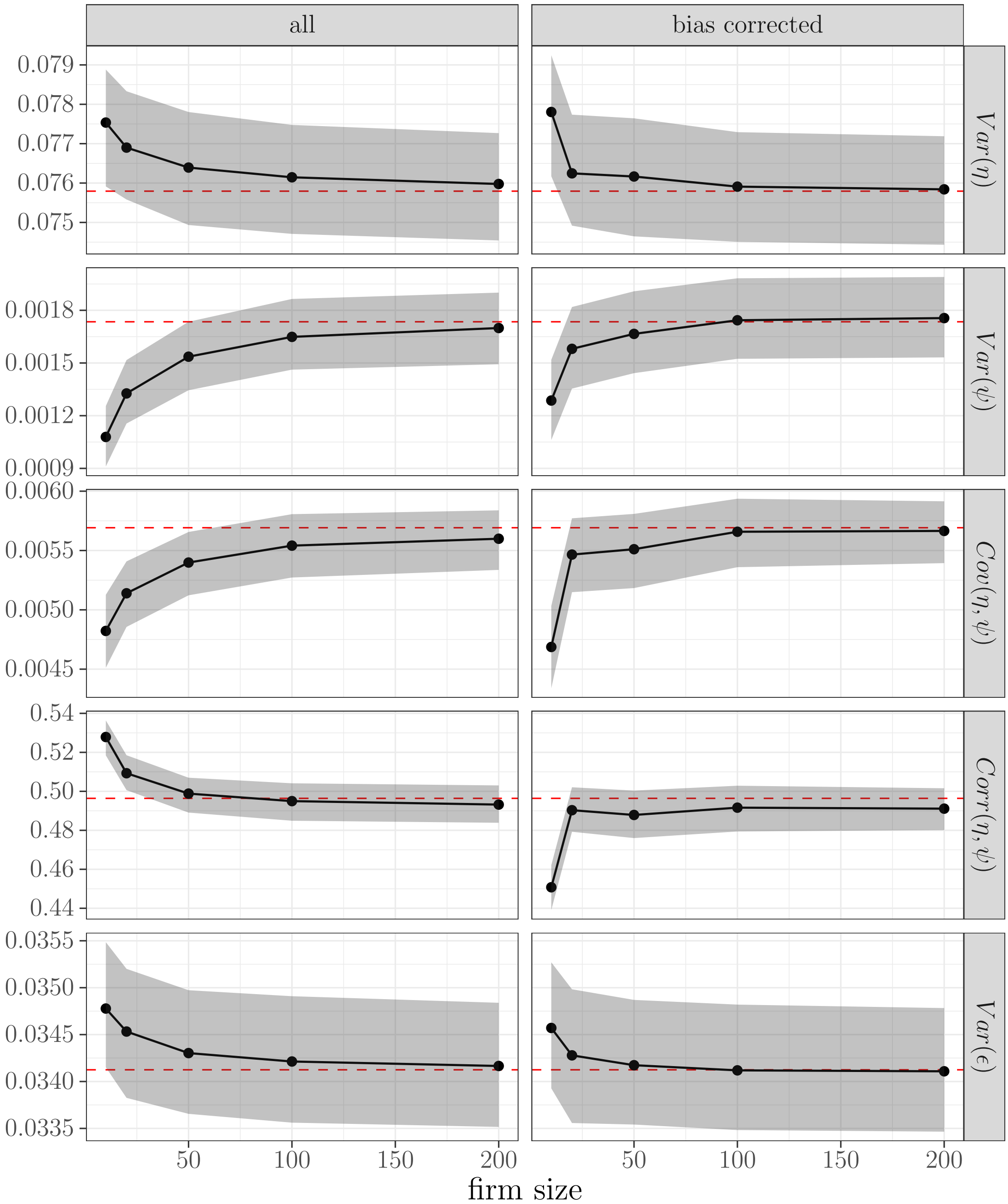
[42] Here Assumption 3 requires the mapping $\psi_j \mapsto \left(\psi_j + \mathbb{E}\left(\eta_i \,|\, \psi_j\right), \text{Var}\left(\eta_i \,|\, \psi_j\right)\right)$ to be injective. Firm-specific means of log-wages being monotone in ψ_j is sufficient, though not necessary, for this to hold.

Table 1: Estimates of firm and worker heterogeneity across simulations

Firm size	Var (η_i)	Var (ψ_j)	Cov (η_i, ψ_j)	Corr (η_i, ψ_j)	Var (ε_{i1})	$\hat{K}$
			true values			
-	0.0758	0.0017	0.0057	0.4963	0.0341	
			two-step estimator			
10	0.0775 [0.076,0.079]	0.0011 [0.001,0.001]	0.0048 [0.005,0.005]	0.5281 [0.519,0.537]	0.0348 [0.034,0.035]	3.0 [3,3]
20	0.0769 [0.076,0.078]	0.0013 [0.001,0.002]	0.0051 [0.005,0.005]	0.5091 [0.500,0.518]	0.0345 [0.034,0.035]	4.0 [4,4]
50	0.0764 [0.075,0.078]	0.0015 [0.001,0.002]	0.0054 [0.005,0.006]	0.4986 [0.490,0.507]	0.0343 [0.034,0.035]	6.0 [6,6]
100	0.0761 [0.075,0.077]	0.0016 [0.001,0.002]	0.0055 [0.005,0.006]	0.4955 [0.487,0.504]	0.0342 [0.034,0.035]	8.4 [8,9]
200	0.0760 [0.075,0.077]	0.0017 [0.001,0.002]	0.0056 [0.005,0.006]	0.4930 [0.483,0.503]	0.0342 [0.034,0.035]	11.3 [11,12]
			two-step estimator, bias-corrected			
10	0.0778 [0.076,0.079]	0.0013 [0.001,0.002]	0.0047 [0.004,0.005]	0.4511 [0.439,0.463]	0.0346 [0.034,0.035]	
20	0.0763 [0.075,0.078]	0.0016 [0.001,0.002]	0.0055 [0.005,0.006]	0.4902 [0.479,0.501]	0.0343 [0.034,0.035]	
50	0.0762 [0.075,0.078]	0.0017 [0.001,0.002]	0.0055 [0.005,0.006]	0.4876 [0.476,0.499]	0.0342 [0.034,0.035]	
100	0.0759 [0.075,0.077]	0.0017 [0.002,0.002]	0.0056 [0.005,0.006]	0.4923 [0.481,0.502]	0.0341 [0.034,0.035]	
200	0.0759 [0.074,0.077]	0.0017 [0.002,0.002]	0.0056 [0.005,0.006]	0.4909 [0.480,0.503]	0.0341 [0.033,0.035]	
			fixed-effects estimator			
10	0.1342 [0.132,0.136]	0.0342 [0.033,0.036]	-0.0267 [-0.028,-0.025]	-0.3949 [-0.409,-0.382]	0.0173 [0.017,0.018]	
20	0.1002 [0.099,0.102]	0.0130 [0.012,0.014]	-0.0056 [-0.006,-0.005]	-0.1548 [-0.169,-0.139]	0.0256 [0.025,0.026]	
50	0.0848 [0.083,0.086]	0.0055 [0.005,0.006]	0.0019 [0.002,0.002]	0.0895 [0.072,0.107]	0.0307 [0.030,0.031]	
100	0.0802 [0.079,0.082]	0.0035 [0.003,0.004]	0.0039 [0.004,0.004]	0.2311 [0.212,0.250]	0.0324 [0.032,0.033]	
200	0.0780 [0.077,0.079]	0.0026 [0.002,0.003]	0.0048 [0.004,0.005]	0.3359 [0.319,0.352]	0.0333 [0.033,0.034]	

Notes: Means and 95% confidence intervals. See Supplementary Appendix S2 for a description of the estimators. Unobserved heterogeneity is continuously distributed in the DGP. The number of groups $\widehat{K}$ *is estimated in every replication, using (12) with* $\xi = 1$, *and it is reported in the last column of the first panel. We use the kmeans routine from R, with* 100 *starting values.* 500 *simulations.*

Figure 5: Estimates of firm and worker heterogeneity across simulations



Notes: Means (solid line) and 95% confidence intervals. The dashed line indicates the true parameter value. Unobserved heterogeneity is continuously distributed in the DGP. The number of groups K is estimated in every replication. 500 *replications.*

IDs change between the two years. We focus on firms that are present throughout the period. There are about 20,000 job movers in the sample. We use two-step grouped fixed-effects with a classification based on the firms' empirical distributions of log-wages in 2002, evaluated at 20 percentiles of the overall log-wage distribution, with $K = 10$ groups. In the second step, we estimate the model's parameters $\widehat{\psi}(\widehat{k}_j)$, $\widehat{\mu}(\widehat{k}_j)$, $\widehat{\sigma}^2(\widehat{k}_j)$, and $\widehat{s}^2$. This step relies on simple mean and covariance restrictions, as we describe in Supplementary Appendix S2.

Given parameter estimates, we then simulate a two-period model where firm heterogeneity is continuously distributed. Specifically, the ψ_j's are drawn from a normal distribution, calibrated to match the mean and variance of the $\widehat{\psi}(\widehat{k}_j)$'s. We draw 120,000 workers in the cross-section, including 20,000 job movers. We run simulations for different firm sizes, from 10 workers per firm to 200 workers per firm. The total number of job movers is kept constant, so the number of movers per firm increases with firm size.

In Table 1 and Figure 5 we report the mean and 95% confidence intervals of grouped fixed-effects and fixed-effects estimators of the components of the variance decomposition (22), across 500 simulations. The number of groups is estimated in every simulation. We see that biases of two-step grouped fixed-effects estimators decrease quite rapidly when firm size grows, although biases are not negligible when firms are small. As an example, the variance of firm effects is two thirds of the true value on average when firm size equals 10, and 75% of the true value for a firm size of 20. Moreover, bias correction tends to provide performance improvements: for example, biases for the variance of firm effects become 25% and 5% for firm sizes of 10 and 20, respectively. Note that bias correction is not associated with large increases in dispersion.[43] In addition, the last column in the table shows that the estimated number of groups is rather small, and close to proportional to the square root of firm size (which is to be expected in this DGP with one-dimensional underlying heterogeneity).

Lastly, in the bottom panel of Table 1 we report the results for a fixed-effects estimator, which is computationally feasible in this linear setting. We see that fixed-effects is substantially biased. This shows that incidental parameter bias due to low mobility is particularly acute in this DGP. The contrast between fixed-effects and grouped fixed-effects is in line with Theorem 2, since here the number S of movers per firm is small relative to the total number T of workers in the firm which we use to group firms.[44] Hence grouped fixed-effects, possibly combined with

[43] To implement the bias-correction method of Dhaene and Jochmans (2015) we select two halves within each firm at random. We re-estimate the number of groups in each half-sample. Note that in this particular setting one could alternatively average across multiple random permutations within firm.

[44] In Supplementary Appendix S2 we report the results for a bias-corrected version of the fixed-effects estimator. We find that, although the correction helps, the modified estimator is still substantially biased.

bias reduction, provides an effective regularization in this context. In Supplementary Appendix S2 we illustrate this point further by contrasting the performance of fixed-effects and grouped fixed-effects estimators as the number of job movers per firm varies.

8 Conclusion

Two-step grouped fixed-effects method based on an initial data-driven classification are effective dimension reduction devices. In this paper we have analyzed some of their properties under general assumptions on the form of individual unobserved heterogeneity. We have seen that grouped fixed-effects estimators are subject to an approximation bias, when the population is not discrete, and an incidental parameter bias, since groups are estimated with noise. We have shown in two illustrations that bias reduction methods can improve the performance of discrete estimators.

Grouped fixed-effects methods may be particularly well-suited when unobservables are multi-dimensional, provided their underlying dimension is low. A case in point is a model with time-varying unobservables with an underlying low-dimensional nonlinear factor structure. In such settings important questions for future work are the choice of the number of groups and the characterization of asymptotic distributions.

Finally, grouped fixed-effects methods could be of interest beyond the two empirical illustrations we have considered here. Other settings include models with multi-sided heterogeneity, nonlinear factor models, nonparametric or semi-parametric panel data models such as quantile regression with individual effects, and network models, for example. We also envision grouped fixed-effects methods, and more generally classification and clustering methods, to be useful in structural analysis. For example, in dynamic migration models with a large number of locations, a curse of dimensionality arises when workers keep track of their full history of locations (as in Kennan and Walker, 2011). Extending clustering methods to address this dimensionality challenge would be interesting.

References

[1] Abowd, J., F. Kramarz, and D. Margolis (1999): “High Wage Workers and High Wage Firms”, *Econometrica*, 67(2), 251–333.

[2] Aguirregabiria, V., and P. Mira (2002): “Swapping the Nested Fixed-Point Algorithm: A Class of Estimators for Discrete Markov Decision Models,” *Econometrica*, 70(4), 1519–1543.

[3] Aguirregabiria, V., and P. Mira (2010): “Dynamic discrete choice structural models: A survey,” *Journal of Econometrics*, 156, 38–67.

[4] Arcidiacono, P., and J. B. Jones (2003): ‘Finite Mixture Distributions, Sequential Likelihood and the EM Algorithm”, *Econometrica*, 71(3), 933–946.

[5] Arcidiacono, P., and R. Miller (2011): ‘Conditional Choice Probability Estimation of Dynamic Discrete Choice Models With Unobserved Heterogeneity”, *Econometrica*, 79(6), 1823–1867.

[6] Arellano, M., and J. Hahn (2007): “Understanding Bias in Nonlinear Panel Models: Some Recent Developments,”. In: R. Blundell, W. Newey, and T. Persson (eds.): *Advances in Economics and Econometrics, Ninth World Congress*, Cambridge University Press.

[7] Arellano, M., and J. Hahn (2016): “A likelihood-Based Approximate Solution to the Incidental Parameter Problem in Dynamic Nonlinear Models with Multiple Effects,” *Global Economic Review*, 45(3), 251–274.

[8] Bai, J. (2009), “Panel Data Models with Interactive Fixed Effects,” *Econometrica*, 77, 1229–1279.

[9] Bai, J., and T. Ando (2015):“Panel Data Models with Grouped Factor Structure Under Unknown Group Membership,” *Journal of Applied Econometrics.*

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

[11] Berry, S., J. Levinsohn, and A. Pakes (1995): “Automobile Prices in Market Equilibrium,” *Econometrica*, 841–890.

[12] Bester, A., and C. Hansen (2016): "Grouped Effects Estimators in Fixed Effects Models", *Journal of Econometrics*, 190(1), 197–208.

[13] Bickel, P.J., and A. Chen (2009): "A Nonparametric View of Network Models and Newman-Girvan and Other Modularities," *Proc. Natl. Acad. Sci. USA*, 106, 21068–21073.

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

[15] Bonhomme, S., T. Lamadon, and E. Manresa (2015): "A Distributional Framework for Matched Employer-Employee Data," unpublished manuscript.

[16] Bryant, P. and Williamson, J. A. (1978): "Asymptotic Behaviour of Classification Maximum Likelihood Estimates," *Biometrika*, 65, 273–281.

[17] Buchinsky, M., J. Hahn, and J. Hotz (2005): "Cluster Analysis: A tool for Preliminary Structural Analysis," unpublished manuscript.

[18] Card, D., J. Heining, and P. Kline (2013): "Workplace Heterogeneity and the Rise of West German Wage Inequality," *Quarterly Journal of Economics*, 128(3), 967–1015.

[19] Chamberlain, G., and M. Rotschild (1983): "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets," *Econometrica*, 51(5), 1281–1304.

[20] Chen, M., I. Fernández-Val, and M. Weidner (2014): "Nonlinear Panel Models with Interactive Effects," unpublished manuscript.

[21] Chen, X., L. P. Hansen, and J. Scheinkman (2009): "Nonlinear principal components and long-run implications of multivariate diffusions," *Annals of Statistics*, 4279–4312.

[22] Cunha, F., J. Heckman, and S. Schennach (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation", *Econometrica*, 78(3), 883–931.

[23] Dhaene, G. and K. Jochmans (2015): "Split Panel Jackknife Estimation," *Review of Economic Studies*, 82(3), 991–1030.

[24] Fernández-Val, I., and M. Weidner (2016): "Individual and Time Effects in Nonlinear Panel Data Models with Large N, T," *Journal of Econometrics*, 196, 291–312.

[25] Frühwirth-Schnatter, S. (2006): *Finite Mixture and Markov Switching Models*, Springer.

[26] Gao, C., Y. Lu, and H. H. Zhou (2015): "Rate-Optimal Graphon Estimation," *Annals of Statistics*, 43(6), 2624–2652.

[27] Gersho, A., and R.M. Gray (1992): *Vector Quantization and Signal Compression.* Kluwer Academic Press.

[28] Graf, S., and H. Luschgy (2000): *Foundations of Quantization for Probability Distributions.* Springer Verlag, Berlin, Heidelberg.

[29] Graf, S., and H. Luschgy (2002): "Rates of Convergence for the Empirical Quantization Error", *Annals of Probability.*

[30] Gray, R. M., and D. L. Neuhoff (1998): "Quantization", *IEEE Trans. Inform. Theory*, (Special Commemorative Issue), 44(6), 2325–2383.

[31] Hahn, J. and W.K. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models", *Econometrica*, 72, 1295–1319.

[32] Hahn, J., and H. Moon (2010): "Panel Data Models with Finite Number of Multiple Equilibria," *Econometric Theory*, 26(3), 863–881.

[33] Hastie, T., and W. Stuetzle (1989): "Principal Curves," *Journal of the American Statistical Association*, 84, 502–516.

[34] Heckman, J.J., and B. Singer (1984): "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52(2), 271–320.

[35] Hotz, J., and R. Miller (1993): "Conditional Choice Probabilities and the Estimation of Dynamic Models", *Review of Economic Studies*, 60(3), 497–529.

[36] Hsu, D., S.M. Kakade, and T. Zhang (2012): "A Tail Inequality for Quadratic Forms of Subgaussian Random Vectors," *Electron. Commun. Probab.*, 17, 52, 1–6.

[37] Jochmans, K., and M. Weidner (2016): "Fixed-Effect Regressions on Network Data, arXiv preprint arXiv:1608.01532.

[38] Kasahara, H., and K. Shimotsu (2009): "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," *Econometrica*, 77(1), 135–175.

[39] Keane, M., and K. Wolpin (1997): "The Career Decisions of Young Men," *Journal of Political Economy*, 105(3), 473–522.

[40] Kennan, J., and J. Walker (2011): "The Effect of Expected Income on Individual Migration Decisions", *Econometrica*, 79(1), 211–251.

[41] Levina, E., and P. J. Bickel (2004): "Maximum Likelihood Estimation of Intrinsic Dimension," *Advances in neural information processing systems*, 777–784.

[42] Lin, C. C., and S. Ng (2012): "Estimation of Panel Data Models with Parameter Heterogeneity when Group Membership is Unknown", *Journal of Econometric Methods*, 1(1), 42–55.

[43] Linder, T. (2002): *Learning-Theoretic Methods in Vector Quantization*, Principles of Nonparametric Learning. L. Gyorfi, editor, CISM Lecture Notes, Wien, New York.

[44] McLachlan, G., and D. Peel (2000): *Finite Mixture Models*, Wiley Series in Probabilities and Statistics.

[45] Moon, H. R., M. Shum, and M. Weidner (2014): "Estimation of Random Coefficients Logit Demand Models with Interactive Fixed Effects," unpublished manuscript.

[46] Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, 4, 2111–2245.

[47] Pantano, J., and Y. Zheng (2013): "Using Subjective Expectations Data to Allow for Unobserved Heterogeneity in Hotz-Miller Estimation Strategies," unpublished working paper.

[48] Pesaran, M. H. (2006): "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure," *Econometrica*, 74(4), 967–1012.

[49] Pollard, D. (1981): "Strong Consistency of K-means Clustering," *Annals of Statistics*, 9, 135–140.

[50] Pollard, D. (1982a): "A Central Limit Theorem for K-Means Clustering," *Annals of Probability*, 10, 919–926.

[51] Pollard, D. (1982b): "Quantization and the Method of K-Means," *IEEE Transactions on Information Theory*, vol. IT-28 (2), 199–205.

[52] Raginsky, M., and Lazebnik, S. (2005): "Estimation of Intrinsic Dimensionality Using High-Rate Vector Quantization," *In Advances in neural information processing systems*, 1105–1112.

[53] Rust, J. (1994): "Structural Estimation of Markov Decision Processes," *Handbook of econometrics*, 4(4), 3081–3143.

[54] Saggio, R. (2012): "Discrete Unobserved Heterogeneity in Discrete Choice Panel Data Models," CEMFI Master Thesis.

[55] Sorkin, I. (2016): "Ranking Firms Using Revealed Preference," unpublished manuscript.

[56] Steinley, D. (2006): "K-means Clustering: A Half-Century Synthesis," *Br. J. Math. Stat. Psychol.*, 59, 1–34.

[57] Su, C. and K. Judd (2012): "Constrained Optimization Approaches to Estimation of Structural Models", *Econometrica*, 80(5), 2213–2230.

[58] Su, L., Z. Shi, and P. C. B. Phillips (2015): "Identifying Latent Structures in Panel Data," to appear in *Econometrica.*

[59] Vershynin, R. (2010): *Introduction to the Non-Asymptotic Analysis of Random Matrices*, in Y. C. Eldar and G. Kutyniok, ed., Compressed Sensing: Theory and Applications. Cambridge University Press.

[60] Vogt and O. Linton (2015): "Classification of Nonparametric Regression Functions in Longitudinal Data Models," to appear in the *Journal of the Royal Statistical Society: Series B.*

[61] Wolfe, P. J., and S. C. Ohlede (2014): "Nonparametric Graphon Estimation", Arkiv.

APPENDIX

A Proofs

A.1 Proof of Lemma 1

Let us define:[45]

$$(h^*, \{k_i^*\}) = \underset{(\widetilde{h}, \{k_i\})}{\text{argmin}} \sum_{i=1}^{N} \left\| \varphi(\alpha_{i0}) - \widetilde{h}(k_i) \right\|^2 . \tag{A1}$$

By definition of $(\widehat{h}, \{\widehat{k}_i\})$, we have (almost surely):

$$\sum_{i=1}^{N} \left\| h_i - \widehat{h}(\widehat{k}_i) \right\|^2 \leq \sum_{i=1}^{N} \| h_i - h^*(k_i^*) \|^2 .$$

Letting $\varepsilon_i = \sum_{t=1}^{T} \varepsilon_{it}/T$, with $\varepsilon_{it} = h(Y_{it}, X_{it}) - \varphi(\alpha_{i0})$, we thus have, by the triangular inequality:

$$\frac{1}{N} \sum_{i=1}^{N} \left\| \varphi(\alpha_{i0}) - \widehat{h}(\widehat{k}_i) \right\|^2 = O_p \underbrace{\left(\frac{1}{N} \sum_{i=1}^{N} \| \varphi(\alpha_{i0}) - h^*(k_i^*) \|^2 \right)}_{= B_{\varphi(\alpha)}(K)} + O_p \left(\tfrac{1}{N} \textstyle\sum_{i=1}^{N} \| \varepsilon_i \|^2 \right) .$$

Lemma 1 thus follows from the fact that, by Assumption 1, $\frac{1}{N} \sum_{i=1}^{N} \| \varepsilon_i \|^2 = O_p(1/T)$ and $B_{\varphi(\alpha)}(K) = O_p(B_\alpha(K))$.

A.2 Proof of Corollary 1

Let $\{k_i\} = \{k_{i1}\} \cap \{k_{i2}\}$ be the intersection of two partitions of $\{1, ..., N\}$: a first partition with (the integer part of) $K_1 = K^{1-\eta}$ groups, and a second partition with (the integer part of) $K_2 = K^\eta$ groups. Since $\left(\widehat{h}, \{\widehat{k}_i\} \right)$ solves (1), we have:

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^{N} \left\| h_i - \widehat{h}(\widehat{k}_i) \right\|^2 &= \frac{1}{N} \sum_{i=1}^{N} \left\| \varphi(\alpha_{i0}) + \varepsilon_i - \widehat{h}(\widehat{k}_i) \right\|^2 \\ &\leq \min_{(\widetilde{h}_1, \widetilde{h}_2, \{k_{i1}\}, \{k_{i2}\})} \frac{1}{N} \sum_{i=1}^{N} \left\| \varphi(\alpha_{i0}) - \widetilde{h}_1(k_{i1}) + \varepsilon_i - \widetilde{h}_2(k_{i2}) \right\|^2 \\ &= O_p \left(B_{\varphi(\alpha)}(K_1) \right) + O_p \left(B_\varepsilon \left(K_2 \right) \right) = o_p \left(\frac{1}{T} \right) . \end{aligned}$$

Hence, Corollary 1 follows from the fact that: $\frac{1}{N} \sum_{i=1}^{N} \| h_i - \varphi(\alpha_{i0}) \|^2 = C/T + o_p(1/T)$.

[45] The literature on vector quantization provides general results on existence of optimal empirical quantizers; that is, solutions to (A1). See for example Chapter 1 in Graf and Luschgy (2000).

A.3 Proof of Theorem 1

In this proof and the rest of the appendix we will use the following notation. $v_i = \frac{\partial \ell_i}{\partial \alpha_i}$, $v_i^{\alpha} = \frac{\partial^2 \ell_i}{\partial \alpha_i \partial \alpha_i'}$, $v_i^{\theta} = \frac{\partial^2 \ell_i}{\partial \theta \partial \alpha_i'}$, and $v_i^{\alpha\alpha} = \frac{\partial^3 \ell_i}{\partial \alpha_i \partial \alpha_i' \otimes \partial \alpha_i'}$ (which is a $q \times q^2$ matrix). When there is no ambiguity we will omit the dependence on true parameter values from the notation. Let, for all θ and $k \in \{1, ..., K\}$:

$$\widehat{\alpha}(k,\theta) = \underset{\alpha \in \mathcal{A}}{\text{argmax}} \sum_{i=1}^{N} \mathbf{1}\{\widehat{k}_i = k\} \ell_i(\alpha, \theta). \tag{A2}$$

Let $\overline{\alpha}_i(\theta) = \text{argmax}_{\alpha_i \in \mathcal{A}} \lim_{T \to \infty} \mathbb{E}(\ell_i(\alpha_i, \theta))$. Let also $\delta = \frac{1}{T} + B_{\alpha}(K)$ (or more generally $\delta = \frac{1}{T} + B_{(\alpha,\mu)}(K)$ in conditional models).

The proof consists of three steps. We will first establish that $\widehat{\theta}$ is consistent for θ_0. Then, we will expand the score equation around θ_0:

$$0 = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \ell_i(\widehat{\alpha}(\widehat{k}_i, \widehat{\theta}), \widehat{\theta})}{\partial \theta} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \ell_i(\widehat{\alpha}(\widehat{k}_i, \theta_0), \theta_0)}{\partial \theta} + \left(\frac{\partial}{\partial \theta'}\Big|_{\widetilde{\theta}} \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \ell_i(\widehat{\alpha}(\widehat{k}_i, \theta), \theta)}{\partial \theta} \right) \left(\widehat{\theta} - \theta_0\right),$$

where $\widetilde{\theta}$ lies between θ_0 and $\widehat{\theta}$. We will then establish the following main intermediate results:

$$\frac{1}{N} \sum_{i=1}^{N} \frac{\partial \ell_i(\widehat{\alpha}(\widehat{k}_i, \theta_0), \theta_0)}{\partial \theta} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \theta}\Big|_{\theta_0} \ell_i\left(\overline{\alpha}_i(\theta), \theta\right) + O_p(\delta), \tag{A3}$$

$$\frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2}{\partial \theta \partial \theta'}\Big|_{\theta_0} \left(\ell_i\left(\widehat{\alpha}(\widehat{k}_i, \theta), \theta\right) - \ell_i\left(\overline{\alpha}_i(\theta), \theta\right) \right) = o_p(1). \tag{A4}$$

By (5), the first part of Theorem 1 will then come from approximating $\frac{\partial}{\partial \theta'}\big|_{\widetilde{\theta}} \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \ell_i(\widehat{\alpha}(\widehat{k}_i, \theta), \theta)}{\partial \theta}$ by its value at θ_0, using that $\widetilde{\theta}$ is consistent and part (*iv*) in Assumption 2. The second part of the theorem will then follow.

We will focus on the case where $\mathbb{E}(\ell_i(\alpha_i, \theta))$ does not depend on T, and likewise for its first three derivatives. Allowing those population moments to depend on T would be needed in order to deal with the presence of non-stationary initial conditions in dynamic models. This could be done at the cost of making the notation in the proof more involved.

Consistency of $\widehat{\theta}$. We will establish that:

$$\sup_{\theta \in \Theta} \left| \frac{1}{N} \sum_{i=1}^{N} \ell_i\left(\widehat{\alpha}(\widehat{k}_i, \theta), \theta\right) - \frac{1}{N} \sum_{i=1}^{N} \ell_i\left(\overline{\alpha}_i(\theta), \theta\right) \right| = o_p(1). \tag{A5}$$

Compactness of the parameter space, continuity of the target likelihood, and identification of θ_0 (from part (*ii*) in Assumption 2), will then imply that $\widehat{\theta}$ is consistent for θ_0 (e.g., Theorem 2.1 in Newey and McFadden, 1994).

To show (A5) we first note that, for every $\theta \in \Theta$: $\mathbb{E}\left(v_i\left(\overline{\alpha}_i(\theta), \theta\right)\right) = 0$, where the expectation is taken with respect to $f_i(Y_i, X_i)$, which depends on α_{i0}. This shows that $\overline{\alpha}_i(\theta)$, which is unique by

Assumption 2 (*ii*), is a function of θ and α_{i0}. We will denote this function as $\overline{\alpha}_i(\theta) = \overline{\alpha}\left(\theta, \alpha_{i0}\right)$.[46] From Assumption 2 (*ii*) and (*iii*) we have that both:

$$\frac{\partial \overline{\alpha}\left(\theta, \alpha_{i0}\right)}{\partial \theta'} = \mathbb{E}\left[-v_i^{\alpha}\left(\overline{\alpha}_i(\theta), \theta\right)\right]^{-1} \mathbb{E}\left[v_i^{\theta}\left(\overline{\alpha}_i(\theta), \theta\right)\right]'$$

and:

$$\frac{\partial \overline{\alpha}\left(\theta, \alpha_{i0}\right)}{\partial \alpha_i'} = \mathbb{E}\left[-v_i^{\alpha}\left(\overline{\alpha}_i(\theta), \theta\right)\right]^{-1} \left.\frac{\partial}{\partial \alpha'}\right|_{\alpha_{i0}} \mathbb{E}_{\alpha}\left[v_i\left(\overline{\alpha}_i(\theta), \theta\right)\right] \tag{A6}$$

are uniformly bounded. This shows that $\overline{\alpha}\left(\theta, \alpha_{i0}\right)$ is Lipschitz continuous with uniformly bounded Lipschitz coefficients.

Let now $a(k, \theta) = \overline{\alpha}\left(\theta, \psi\left(\widehat{h}\left(k\right)\right)\right)$, where ψ is defined in Assumption 3. Define the fixed-effects estimator of α_i, for given θ, as $\widehat{\alpha}_i(\theta) = \operatorname{argmax}_{\alpha_i \in \mathcal{A}} \ \ell_i(\alpha_i, \theta)$. We have, for all θ (that is, pointwise):

$$\frac{1}{N}\sum_{i=1}^{N} \ell_i\left(a(\widehat{k}_i, \theta), \theta\right) \leq \frac{1}{N}\sum_{i=1}^{N} \ell_i\left(\widehat{\alpha}(\widehat{k}_i, \theta), \theta\right) \leq \frac{1}{N}\sum_{i=1}^{N} \ell_i\left(\widehat{\alpha}_i(\theta), \theta\right) = \frac{1}{N}\sum_{i=1}^{N} \ell_i\left(\overline{\alpha}_i(\theta), \theta\right) + O_p\left(\frac{1}{T}\right),$$

where the last equality follows from expanding the log-likelihood around $\overline{\alpha}_i(\theta)$ (as in Arellano and Hahn, 2007, for example).

Now, for some $a_i(\theta)$ between $\widehat{\alpha}_i\left(\theta\right)$ and $a(\widehat{k}_i, \theta)$:

$$\frac{1}{N}\sum_{i=1}^{N} \ell_i\left(a(\widehat{k}_i, \theta), \theta\right) - \frac{1}{N}\sum_{i=1}^{N} \ell_i\left(\widehat{\alpha}_i\left(\theta\right), \theta\right) = \frac{1}{2N}\sum_{i=1}^{N} \left(a(\widehat{k}_i, \theta) - \widehat{\alpha}_i\left(\theta\right)\right)' v_i^{\alpha}\left(a_i\left(\theta\right), \theta\right) \left(a(\widehat{k}_i, \theta) - \widehat{\alpha}_i\left(\theta\right)\right).$$

By Assumption 2 (*iii*), $\max_{i=1,\dots,N} \sup_{(\alpha_i, \theta)} \|v_i^{\alpha}(\alpha_i, \theta)\| = O_p(1)$. Moreover:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} \left\|a(\widehat{k}_i, \theta) - \widehat{\alpha}_i\left(\theta\right)\right\|^2 &= \frac{1}{N}\sum_{i=1}^{N} \left\|a(\widehat{k}_i, \theta) - \overline{\alpha}_i\left(\theta\right)\right\|^2 + O_p\left(\frac{1}{T}\right) \\
&= \frac{1}{N}\sum_{i=1}^{N} \left\|\overline{\alpha}\left(\theta, \psi\left(\widehat{h}\left(\widehat{k}_i\right)\right)\right) - \overline{\alpha}\left(\theta, \psi(\varphi(\alpha_{i0}))\right)\right\|^2 + O_p\left(\frac{1}{T}\right) \\
&= O_p\left(\frac{1}{N}\sum_{i=1}^{N} \left\|\widehat{h}\left(\widehat{k}_i\right) - \varphi(\alpha_{i0})\right\|^2\right) + O_p\left(\frac{1}{T}\right) = O_p(\delta),
\end{aligned} \tag{A7}$$

where we have used Lemma 1 and Assumption 3, and the fact that $\overline{\alpha}$ is Lipschitz with respect to its second argument. This implies that, pointwise in $\theta \in \Theta$:

$$\left|\frac{1}{N}\sum_{i=1}^{N} \ell_i\left(\widehat{\alpha}(\widehat{k}_i, \theta), \theta\right) - \frac{1}{N}\sum_{i=1}^{N} \ell_i\left(\widehat{\alpha}_i\left(\theta\right), \theta\right)\right| = O_p(\delta). \tag{A8}$$

[46] In conditional models where the distribution of covariates depends also on μ_{i0}, $\overline{\alpha}_i(\theta)$ will be a function of both types of individual effects; that is: $\overline{\alpha}\left(\theta, \alpha_{i0}, \mu_{i0}\right)$.

In order to establish uniform convergence of the grouped fixed-effects log-likelihood, let us first recall the uniform convergence of the fixed-effects log-likelihood (from Assumption 2 (i)-(ii)-(iii)):

$$\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}_i(\theta),\theta\right)-\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\overline{\alpha}_i\left(\theta\right),\theta\right)\right|=o_p(1).$$

We have, using similar arguments as above:

$$\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}(\widehat{k}_i,\theta),\theta\right)-\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}_i\left(\theta\right),\theta\right)\right|\leq\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(a(\widehat{k}_i,\theta),\theta\right)-\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}_i\left(\theta\right),\theta\right)\right|$$
$$\leq\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(a(\widehat{k}_i,\theta),\theta\right)-\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\overline{\alpha}_i\left(\theta\right),\theta\right)\right|+o_p(1)=o_p(1),$$

where the last inequality comes from a first-order expansion around $\overline{\alpha}_i(\theta)$, Assumption 2 (iii), and the fact that: $\sup_{\theta\in\Theta}\frac{1}{N}\sum_{i=1}^{N}\|a(\widehat{k}_i,\theta)-\overline{\alpha}_i(\theta)\|^2=O_p(\delta)=o_p(1)$.

This implies (A5) and consistency of $\widehat{\theta}$ as N,T,K tend to infinity.

Proof of (A3). From (A8) evaluated at θ_0 we have, for some a_i between $\widehat{\alpha}(\widehat{k}_i,\theta_0)$ and $\widehat{\alpha}_i\left(\theta_0\right)$, and omitting from now on the reference to θ_0 for conciseness:

$$O_p(\delta)=\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}_i\right)-\frac{1}{N}\sum_{i=1}^{N}\ell_i(\widehat{\alpha}(\widehat{k}_i))=\frac{1}{2N}\sum_{i=1}^{N}\left(\widehat{\alpha}(\widehat{k}_i)-\widehat{\alpha}_i\right)'\left(-v_i^{\alpha}\left(a_i\right)\right)\left(\widehat{\alpha}(\widehat{k}_i)-\widehat{\alpha}_i\right)\geq 0.\quad\text{(A9)}$$

By parts (*ii*) and (*iii*) in Assumption 2 there exists a constant $\eta>0$ and a positive definite matrix $\underline{\Sigma}$ such that:

$$\inf_{\alpha_{i0}}\inf_{\|\alpha_i-\alpha_{i0}\|\leq\eta}\mathbb{E}\left(-v_i^{\alpha}(\alpha_i)\right)\geq\underline{\Sigma}.$$

For this η we will first show that:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{\|\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\|>\eta\right\}=O_p(\delta).\quad\text{(A10)}$$

Showing (A10) will allow us to control the difference $\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}$ in an average sense. This is important since, unlike for fixed-effects, we conjecture that $\max_{i=1,\ldots,N}\|\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\|$ may *not* be $o_p(1)$ in general.

To see that (A10) holds, let $\iota_i=\mathbf{1}\left\{\|\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\|\leq\eta\right\}$, and note that by (A9) we have, since $\ell_i\left(\widehat{\alpha}_i\right)\geq\ell_i(\widehat{\alpha}(\widehat{k}_i))$ for all i:

$$0\leq\frac{1}{N}\sum_{i=1}^{N}(1-\iota_i)\left(\ell_i\left(\widehat{\alpha}_i\right)-\ell_i(\widehat{\alpha}(\widehat{k}_i))\right)=O_p(\delta).$$

Now, by parts (*ii*) and (*iii*) in Assumption 2, and using that $\max_{i=1,\ldots,N}\|\widehat{\alpha}_i-\alpha_{i0}\|=o_p(1)$:

$$\min_{i=1,\ldots,N}\inf_{\|\alpha_i-\alpha_{i0}\|>\eta}\ell_i\left(\widehat{\alpha}_i\right)-\ell_i(\alpha_i)\geq\inf_{\alpha_{i0}}\inf_{\|\alpha_i-\alpha_{i0}\|>\eta}\mathbb{E}\left[\ell_i\left(\alpha_{i0}\right)\right]-\mathbb{E}\left[\ell_i(\alpha_i)\right]+o_p(1)\geq\zeta+o_p(1),$$

where $\zeta > 0$ is a constant, and the $o_p(1)$ term is uniform in i and α_i. Hence $\frac{1}{N}\sum_{i=1}^N (1-\iota_i)(\zeta + o_p(1)) = O_p(\delta)$, from which (A10) follows.

Next, by part (*iii*) in Assumption 2 $\|v_i^\alpha(\alpha_i) - \mathbb{E}\left(v_i^\alpha(\alpha_i)\right)\|$ is $o_p\left(1\right)$ uniformly in i and α_i. We thus have:

$$\min_{i=1,\ldots,N} \inf_{\|\alpha_i - \alpha_{i0}\| \leq \eta} \left(-v_i^\alpha(\alpha_i)\right) \geq \underline{\Sigma} + o_p(1). \tag{A11}$$

Using (A9) this implies that: $\frac{1}{N}\sum_{i=1}^N \iota_i \left\|\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right\|^2 = O_p(\delta)$. Hence, using in addition (A10) and the fact that $\mathcal{A}$ is compact, we have:

$$\frac{1}{N}\sum_{i=1}^N \left\|\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right\|^2 = \frac{1}{N}\sum_{i=1}^N \iota_i \left\|\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right\|^2 + \frac{1}{N}\sum_{i=1}^N (1-\iota_i) \left\|\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right\|^2 = O_p(\delta). \tag{A12}$$

We are now going to show (A3). It follows from (A12) and a second-order expansion that:

$$\frac{1}{N}\sum_{i=1}^N \frac{\partial \ell_i(\widehat{\alpha}(\widehat{k}_i))}{\partial \theta} = \frac{1}{N}\sum_{i=1}^N \frac{\partial \ell_i(\alpha_{i0})}{\partial \theta} + \frac{1}{N}\sum_{i=1}^N v_i^\theta \left(\widehat{\alpha}(\widehat{k}_i) - \alpha_{i0}\right) + O_p\left(\delta\right).$$

By Cauchy Schwarz, using (A12) and the fact that $\frac{1}{N}\sum_{i=1}^N \|v_i^\theta - \mathbb{E}\left(v_i^\theta\right)\|^2 = O_p(T^{-1})$ by Assumption 2 (*iii*), we have:

$$\frac{1}{N}\sum_{i=1}^N v_i^\theta \left(\widehat{\alpha}(\widehat{k}_i) - \alpha_{i0}\right) = \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left(v_i^\theta\right)\left(\widehat{\alpha}(\widehat{k}_i) - \alpha_{i0}\right) + O_p\left(\delta\right).$$

We are going to show that:

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left(v_i^\theta\right)\left(\widehat{\alpha}(\widehat{k}_i) - \alpha_{i0} - \left[\mathbb{E}\left(-v_i^\alpha\right)\right]^{-1} v_i\right) = O_p\left(\delta\right). \tag{A13}$$

Expanding $v_i(\widehat{\alpha}_i) = 0$ around α_{i0} we have, by Assumption 2:

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left(v_i^\theta\right)\left(\widehat{\alpha}_i - \alpha_{i0} - \left[\mathbb{E}\left(-v_i^\alpha\right)\right]^{-1} v_i\right) = O_p\left(\frac{1}{T}\right).$$

It will thus be enough to show that:

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left(v_i^\theta\right)\left(\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right) = O_p\left(\delta\right). \tag{A14}$$

Next, expanding to second order each $v_i(\widehat{\alpha}(k))$ around $\widehat{\alpha}_i$ in the score equation: $\sum_{i=1}^N \mathbf{1}\{\widehat{k}_i = k\} v_i(\widehat{\alpha}(k)) = 0$, we have:

$$\widehat{\alpha}(k) = \left(\sum_{i=1}^N \mathbf{1}\{\widehat{k}_i = k\}(-\widetilde{v}_i^\alpha)\right)^{-1} \left(\sum_{i=1}^N \mathbf{1}\{\widehat{k}_i = k\}\left[(-\widetilde{v}_i^\alpha)\widehat{\alpha}_i + \frac{1}{2} v_i^{\alpha\alpha}\left(a_i(k)\right)\left(\widehat{\alpha}(k) - \widehat{\alpha}_i\right) \otimes \left(\widehat{\alpha}(k) - \widehat{\alpha}_i\right)\right]\right),$$

where $\widetilde{v}_i^\alpha = v_i^\alpha(\widehat{\alpha}_i)$, and $a_i(k)$ lies between $\widehat{\alpha}_i$ and $\widehat{\alpha}(k)$. Let us also define:

$$\widetilde{\alpha}(k) = \left(\sum_{i=1}^N \mathbf{1}\{\widehat{k}_i = k\}(-\widetilde{v}_i^\alpha)\right)^{-1} \left(\sum_{i=1}^N \mathbf{1}\{\widehat{k}_i = k\}(-\widetilde{v}_i^\alpha)\widehat{\alpha}_i\right).$$

We start by noting that, since:

$$\widetilde{\alpha}(k) - \widehat{\alpha}(k) = -\frac{1}{2}\left(\sum_{i=1}^N \mathbf{1}\{\widehat{k}_i = k\}(-\widetilde{v}_i^\alpha)\right)^{-1} \sum_{i=1}^N \mathbf{1}\{\widehat{k}_i = k\} v_i^{\alpha\alpha}\left(a_i(k)\right)\left(\widehat{\alpha}(k) - \widehat{\alpha}_i\right) \otimes \left(\widehat{\alpha}(k) - \widehat{\alpha}_i\right), \tag{A15}$$

and since $\min_{i=1,\dots,N}\left(-\widetilde{v}_i^\alpha\right) \geq \underline{\Sigma} + o_p(1)$, it follows from (A12) that:

$$\begin{aligned}\left\| \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left(v_i^\theta\right)\left(\widetilde{\alpha}\left(\widehat{k}_i\right) - \widehat{\alpha}\left(\widehat{k}_i\right)\right)\right\| &\leq \frac{1}{2}\max_{i=1,\dots,N}\left\|\mathbb{E}\left(v_i^\theta\right)\right\| \max_{i=1,\dots,N}\left\|(-\widetilde{v}_i^\alpha)^{-1}\right\| \max_{i=1,\dots,N}\left\|\left(v_i^{\alpha\alpha}\right)\left(a_i(\widehat{k}_i)\right)\right\| \\ &\quad \times \frac{1}{N}\sum_{i=1}^N \left[\left(\sum_{j=1}^N \mathbf{1}\{\widehat{k}_j = \widehat{k}_i\}\right)^{-1} \sum_{j=1}^N \mathbf{1}\{\widehat{k}_j = \widehat{k}_i\}\left\|\widehat{\alpha}(\widehat{k}_j) - \widehat{\alpha}_j\right\|^2\right] \\ &= O_p\left(\frac{1}{N}\sum_{i=1}^N \left\|\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right\|^2\right) = O_p(\delta),\end{aligned}$$

where we have used (A12) in the last step. To show (A14) it will thus suffice to show that:

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left(v_i^\theta\right)\left(\widetilde{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right) = O_p\left(\delta\right). \tag{A16}$$

Before continuing, note also that, by a similar argument and using that $\mathcal{A}$ is compact:

$$\frac{1}{N}\sum_{i=1}^N \left\|\widetilde{\alpha}\left(\widehat{k}_i\right) - \widehat{\alpha}\left(\widehat{k}_i\right)\right\|^2 = O_p(\delta),$$

hence, by (A12):

$$\frac{1}{N}\sum_{i=1}^N \left\|\widetilde{\alpha}\left(\widehat{k}_i\right) - \widehat{\alpha}_i\right\|^2 = O_p(\delta). \tag{A17}$$

Let now $z_i' = \mathbb{E}\left(v_i^\theta\right)\left[\mathbb{E}\left(-v_i^\alpha\right)\right]^{-1}$, and let $\widetilde{z}(k)$ be the weighted mean:

$$\widetilde{z}(k) = \left(\sum_{i=1}^N \mathbf{1}\{\widehat{k}_i = k\}(-\widetilde{v}_i^\alpha)\right)^{-1} \left(\sum_{i=1}^N \mathbf{1}\{\widehat{k}_i = k\}(-\widetilde{v}_i^\alpha) z_i\right).$$

We have, by Assumption 2:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(v_i^{\theta}\right)\left(\widetilde{\alpha}(\widehat{k}_i)-\widehat{\alpha}_i\right)=\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(v_i^{\theta}\right)\left[\mathbb{E}\left(-v_i^{\alpha}\right)\right]^{-1}\left(-v_i^{\alpha}\right)\left(\widetilde{\alpha}(\widehat{k}_i)-\widehat{\alpha}_i\right)+O_p\left(\delta\right)$$

$$=\frac{1}{N}\sum_{i=1}^{N}\underbrace{\mathbb{E}\left(v_i^{\theta}\right)\left[\mathbb{E}\left(-v_i^{\alpha}\right)\right]^{-1}}_{=z_i'}\left(-\widetilde{v}_i^{\alpha}\right)\left(\widetilde{\alpha}(\widehat{k}_i)-\widehat{\alpha}_i\right)+O_p\left(\delta\right)$$

$$=\frac{1}{N}\sum_{i=1}^{N}\left(z_i-\widetilde{z}(\widehat{k}_i)\right)'\left(-\widetilde{v}_i^{\alpha}\right)\left(\widetilde{\alpha}(\widehat{k}_i)-\widehat{\alpha}_i\right)+O_p\left(\delta\right), \tag{A18}$$

where the first equality comes from parts (*ii*) and (*iii*), the second equality comes from combining (A17) with: $\frac{1}{N}\sum_{i=1}^{N}\|v_i^{\alpha}-\widetilde{v}_i^{\alpha}\|^2=O_p\left(\frac{1}{N}\sum_{i=1}^{N}\|\alpha_{i0}-\widehat{\alpha}_i\|^2\right)=O_p(1/T)$, and the last equality comes from $\widetilde{\alpha}(k)$ and $\widetilde{z}(k)$ being weighted means of $\widehat{\alpha}_i$ and z_i with weights $(-\widetilde{v}_i^{\alpha})$.

Let now $\overline{z}(k)=(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\})^{-1}(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}z_i)$ be the *unweighted* mean of z_i in group $\widehat{k}_i=k$. Since $\min_{i=1,\ldots,N}\left(-\widetilde{v}_i^{\alpha}\right)\geq\underline{\Sigma}+o_p(1)$, we have:

$$\frac{1}{N}\sum_{i=1}^{N}\left(z_i-\widetilde{z}(\widehat{k}_i)\right)'\left(-\widetilde{v}_i^{\alpha}\right)\left(z_i-\widetilde{z}(\widehat{k}_i)\right)=O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left(z_i-\overline{z}(\widehat{k}_i)\right)'\left(-\widetilde{v}_i^{\alpha}\right)\left(z_i-\overline{z}(\widehat{k}_i)\right)\right),$$

where we have used that $\widetilde{z}(k)$ is the weighted mean of z_i. Using Assumption 2 (*iii*) then gives:

$$\frac{1}{N}\sum_{i=1}^{N}\left(z_i-\widetilde{z}(\widehat{k}_i)\right)'\left(-\widetilde{v}_i^{\alpha}\right)\left(z_i-\widetilde{z}(\widehat{k}_i)\right)=O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left\|z_i-\overline{z}(\widehat{k}_i)\right\|^2\right).$$

Moreover, by parts (ii) and (iii) in Assumption 2, $z_i=g(\alpha_{i0})$ is a Lipschitz function of α_{i0}. We thus have:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|z_i-\overline{z}(\widehat{k}_i)\right\|^2\leq\frac{1}{N}\sum_{i=1}^{N}\left\|g(\alpha_{i0})-g\left(a(\widehat{k}_i,\theta_0)\right)\right\|^2=O_p(B_{\alpha}(K))=O_p(\delta),$$

where we have used (A7) at $\theta=\theta_0$.[47]

Hence, using again that: $\min_{i=1,\ldots,N}\left(-\widetilde{v}_i^{\alpha}\right)\geq\underline{\Sigma}+o_p(1)$, we obtain:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|z_i-\widetilde{z}(\widehat{k}_i)\right\|^2=O_p(\delta). \tag{A19}$$

Applying Cauchy Schwarz to the right-hand side of (A18), and using (A17) and Assumption 2 (*iii*), then shows (A16), hence (A13), hence (A3).

[47] In conditional models: $\frac{1}{N}\sum_{i=1}^{N}\|g(\alpha_{i0},\mu_{i0})-g(a(\widehat{k}_i,\theta_0))\|^2=O_p(B_{(\alpha,\mu)}(K))=O_p(\delta)$.

Proof of (A4). Let $\widetilde{\iota}(k)=\mathbf{1}\left\{\sum_{j=1}^N \mathbf{1}\{\widehat{k}_j=k\}\left(-v_j^\alpha\left(\widehat{\alpha}(\widehat{k}_j)\right)\right)\geq \frac{1}{2}\underline{\Sigma}\right\}$. We are first going to show that:

$$\frac{1}{N}\sum_{i=1}^N\left(1-\widetilde{\iota}\left(\widehat{k}_i\right)\right)=O_p(\delta). \tag{A20}$$

Similarly as (A10), showing (A20) is needed since we have not established that $\max_{i=1,\ldots,N}\|\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\|$ is $o_p(1)$ (in fact, we conjecture that uniform consistency may not hold in general).

Let $\eta>0$ as in (A10), and define ι_i accordingly. From (A10) it suffices to show that:

$$\frac{1}{N}\sum_{i=1}^N\iota_i\left(1-\widetilde{\iota}\left(\widehat{k}_i\right)\right)=O_p(\delta).$$

With probability approaching one we have: $\min_{i,\iota_i=1}\left(-v_i^\alpha\left(\widehat{\alpha}(\widehat{k}_i)\right)\right)\geq\frac{2}{3}\underline{\Sigma}$. When this condition is satisfied we have:

$$\begin{aligned}
\iota_i\left(1-\widetilde{\iota}\left(\widehat{k}_i\right)\right)&=\iota_i\left(1-\mathbf{1}\left\{\sum_{j=1}^N\mathbf{1}\{\widehat{k}_j=\widehat{k}_i\}\left(-v_j^\alpha\left(\widehat{\alpha}(\widehat{k}_j)\right)\right)-\frac{1}{2}\underline{\Sigma}\geq 0\right\}\right)\\
&\leq\iota_i\left(1-\mathbf{1}\left\{\left(-v_i^\alpha\left(\widehat{\alpha}(\widehat{k}_i)\right)\right)-\frac{1}{2}\underline{\Sigma}+\sum_{j\neq i}\mathbf{1}\{\widehat{k}_j=\widehat{k}_i\}\left(-v_j^\alpha\left(\widehat{\alpha}(\widehat{k}_j)\right)\right)\geq 0\right\}\right)\\
&\leq\iota_i\left(1-\mathbf{1}\left\{\sum_{j=1}^N\iota_j\mathbf{1}\{\widehat{k}_j=\widehat{k}_i\}\frac{1}{6}\underline{\Sigma}\geq-\sum_{j=1}^N(1-\iota_j)\mathbf{1}\{\widehat{k}_j=\widehat{k}_i\}\left(-v_j^\alpha\left(\widehat{\alpha}(\widehat{k}_j)\right)\right)\right\}\right)\\
&\leq\mathbf{1}\left\{\sum_{j=1}^N\iota_j\mathbf{1}\{\widehat{k}_j=\widehat{k}_i\}\leq\frac{6}{\underline{\sigma}}\left(\max_{j=1,\ldots,N}\left\|-v_j^\alpha\left(\widehat{\alpha}(\widehat{k}_j)\right)\right\|\right)\sum_{j=1}^N(1-\iota_j)\mathbf{1}\{\widehat{k}_j=\widehat{k}_i\}\right\}\\
&\leq\mathbf{1}\left\{\sum_{j=1}^N\mathbf{1}\{\widehat{k}_j=\widehat{k}_i\}\leq\left(1+\frac{6}{\underline{\sigma}}\max_{j=1,\ldots,N}\left\|-v_j^\alpha\left(\widehat{\alpha}(\widehat{k}_j)\right)\right\|\right)\sum_{j=1}^N(1-\iota_j)\mathbf{1}\{\widehat{k}_j=\widehat{k}_i\}\right\},
\end{aligned}$$

where $\underline{\sigma}$ denotes the minimum eigenvalue of $\underline{\Sigma}$. Hence we have, with probably approaching one:

$$\begin{aligned}
0&\leq\frac{1}{N}\sum_{i=1}^N\iota_i\left(1-\widetilde{\iota}\left(\widehat{k}_i\right)\right)\\
&\leq\frac{1}{N}\sum_{i=1}^N\mathbf{1}\left\{\sum_{j=1}^N\mathbf{1}\{\widehat{k}_j=\widehat{k}_i\}\leq\left(1+\frac{6}{\underline{\sigma}}\max_{j=1,\ldots,N}\left\|-v_j^\alpha\left(\widehat{\alpha}(\widehat{k}_j)\right)\right\|\right)\sum_{j=1}^N(1-\iota_j)\mathbf{1}\{\widehat{k}_j=\widehat{k}_i\}\right\}\\
&=\frac{1}{N}\sum_{k=1}^K\sum_{i=1}^N\mathbf{1}\{\widehat{k}_i=k\}\mathbf{1}\left\{\sum_{j=1}^N\mathbf{1}\{\widehat{k}_j=k\}\leq\left(1+\frac{6}{\underline{\sigma}}\max_{j=1,\ldots,N}\left\|-v_j^\alpha\left(\widehat{\alpha}(\widehat{k}_j)\right)\right\|\right)\sum_{j=1}^N(1-\iota_j)\mathbf{1}\{\widehat{k}_j=k\}\right\}\\
&\leq\frac{1}{N}\sum_{k=1}^K\left(1+\frac{6}{\underline{\sigma}}\max_{j=1,\ldots,N}\left\|-v_j^\alpha\left(\widehat{\alpha}(\widehat{k}_j)\right)\right\|\right)\sum_{j=1}^N(1-\iota_j)\mathbf{1}\{\widehat{k}_j=k\}=O_p\left(\frac{1}{N}\sum_{j=1}^N(1-\iota_j)\right)=O_p(\delta).
\end{aligned}$$

This shows (A20).

We are now going to show (A4). By part (*iv*) in Assumption 2, Cauchy Schwarz, and (A20), we have:

$$\left\| \frac{1}{N}\sum_{i=1}^{N}\left(1-\widetilde{\iota}(\widehat{k}_i)\right)\frac{\partial^2}{\partial\theta\partial\theta'}\bigg|_{\theta_0}\ell_i\left(\widehat{\alpha}(\widehat{k}_i,\theta),\theta\right)\right\|^2$$
$$\leq \frac{1}{N}\sum_{i=1}^{N}(1-\widetilde{\iota}(\widehat{k}_i))\times\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\partial^2}{\partial\theta\partial\theta'}\bigg|_{\theta_0}\ell_i\left(\widehat{\alpha}(\widehat{k}_i,\theta),\theta\right)\right\|^2=o_p(1).$$

Let k such that $\widetilde{\iota}(k)=1$. Differentiating with respect to θ: $\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}v_i(\widehat{\alpha}(k,\theta),\theta)=0$ we obtain, at $\theta=\theta_0$:

$$\frac{\partial\widehat{\alpha}(k)}{\partial\theta'}=\left(\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=k\}\left(-v_j^{\alpha}\left(\widehat{\alpha}(\widehat{k}_j)\right)\right)\right)^{-1}\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=k\}\left(v_j^{\theta}\left(\widehat{\alpha}(\widehat{k}_j)\right)\right)', \tag{A21}$$

where we note that since, $\widetilde{\iota}(k)=1$, $\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=k\}\left(-v_j^{\alpha}\left(\widehat{\alpha}(\widehat{k}_j)\right)\right)$ is bounded from below by $\underline{\Sigma}/2$.

Let now:

$$\Delta^2S(\theta_0)\equiv\frac{1}{N}\sum_{i=1}^{N}\widetilde{\iota}(\widehat{k}_i)\frac{\partial^2}{\partial\theta\partial\theta'}\bigg|_{\theta_0}\ell_i\left(\widehat{\alpha}(\widehat{k}_i,\theta),\theta\right)-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2}{\partial\theta\partial\theta'}\bigg|_{\theta_0}\ell_i\left(\overline{\alpha}_i(\theta),\theta\right).$$

We have, at θ_0 (omitting again the reference to θ_0 from the notation):

$$\begin{aligned}\Delta^2S(\theta_0) &= \frac{1}{N}\sum_{i=1}^{N}\Bigg\{\widetilde{\iota}(\widehat{k}_i)\frac{\partial^2\ell_i\left(\widehat{\alpha}(\widehat{k}_i)\right)}{\partial\theta\partial\theta'}+\widetilde{\iota}(\widehat{k}_i)v_i^{\theta}\left(\widehat{\alpha}(\widehat{k}_i)\right)\frac{\partial\widehat{\alpha}(\widehat{k}_i)}{\partial\theta'}-\frac{\partial^2\ell_i\left(\alpha_{i0},\theta_0\right)}{\partial\theta\partial\theta'}-v_i^{\theta}\frac{\partial\overline{\alpha}_i}{\partial\theta'}\\ &\qquad -\left(\frac{\partial\overline{\alpha}_i}{\partial\theta'}\right)'(v_i^{\theta})'-\left(\frac{\partial\overline{\alpha}_i}{\partial\theta'}\right)'v_i^{\alpha}\frac{\partial\overline{\alpha}_i}{\partial\theta'}-\frac{\partial^2}{\partial\theta\partial\theta'}\bigg|_{\theta_0}\left(\overline{\alpha}_i(\theta)'v_i\right)\Bigg\}\\ &= \frac{1}{N}\sum_{i=1}^{N}\widetilde{\iota}(\widehat{k}_i)\frac{\partial^2\ell_i\left(\widehat{\alpha}(\widehat{k}_i)\right)}{\partial\theta\partial\theta'}+\widetilde{\iota}(\widehat{k}_i)v_i^{\theta}\left(\widehat{\alpha}(\widehat{k}_i)\right)\frac{\partial\widehat{\alpha}(\widehat{k}_i)}{\partial\theta'}-\frac{\partial^2\ell_i\left(\alpha_{i0},\theta_0\right)}{\partial\theta\partial\theta'}-v_i^{\theta}\frac{\partial\overline{\alpha}_i}{\partial\theta'}+o_p(1),\end{aligned}$$

where we have used that $\mathbb{E}\left(v_i\right)=0$ and: $\frac{\partial\overline{\alpha}_i}{\partial\theta'}=\left[\mathbb{E}\left(-v_i^{\alpha}\right)\right]^{-1}\mathbb{E}\left(v_i^{\theta}\right)'$. Hence, using (A12) and (A20):

$$\Delta^2S(\theta_0) = \frac{1}{N}\sum_{i=1}^{N}\widetilde{\iota}(\widehat{k}_i)v_i^{\theta}\left(\frac{\partial\widehat{\alpha}(\widehat{k}_i,\theta_0)}{\partial\theta'}-\frac{\partial\overline{\alpha}_i(\theta_0)}{\partial\theta'}\right)+o_p\left(1\right),$$

so:

$$\Delta^2S(\theta_0) = \frac{1}{N}\sum_{i=1}^{N}\widetilde{\iota}(\widehat{k}_i)\mathbb{E}\left(v_i^{\theta}\right)\left(\frac{\partial\widehat{\alpha}(\widehat{k}_i,\theta_0)}{\partial\theta'}-\frac{\partial\overline{\alpha}_i(\theta_0)}{\partial\theta'}\right)+o_p\left(1\right).$$

Next, defining $z_i' = \mathbb{E}\left(v_i^{\theta}\right)\left[\mathbb{E}\left(-v_i^{\alpha}\right)\right]^{-1}$ and $\widetilde{z}(k)$ as above we have:

$$\Delta^2 S(\theta_0) = \frac{1}{N}\sum_{i=1}^{N}\widetilde{\iota}(\widehat{k}_i)z_i'\mathbb{E}\left(-v_i^{\alpha}\right)\left(\frac{\partial\widehat{\alpha}(\widehat{k}_i)}{\partial\theta'} - \frac{\partial\overline{\alpha}_i}{\partial\theta'}\right) + o_p\left(1\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\widetilde{\iota}(\widehat{k}_i)z_i'\left(-v_i^{\alpha}\right)\left(\frac{\partial\widehat{\alpha}(\widehat{k}_i)}{\partial\theta'} - \frac{\partial\overline{\alpha}_i}{\partial\theta'}\right) + o_p\left(1\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\widetilde{\iota}(\widehat{k}_i)z_i'\left(-v_i^{\alpha}\left(\widehat{\alpha}(\widehat{k}_i)\right)\right)\left(\frac{\partial\widehat{\alpha}(\widehat{k}_i)}{\partial\theta'} - \frac{\partial\overline{\alpha}_i}{\partial\theta'}\right) + o_p\left(1\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\widetilde{\iota}(\widehat{k}_i)\widetilde{z}(\widehat{k}_i)'\left(-v_i^{\alpha}\left(\widehat{\alpha}(\widehat{k}_i)\right)\right)\left(\frac{\partial\widehat{\alpha}(\widehat{k}_i)}{\partial\theta'} - \frac{\partial\overline{\alpha}_i}{\partial\theta'}\right) + o_p\left(1\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\widetilde{\iota}(\widehat{k}_i)\widetilde{z}(\widehat{k}_i)'\left(\left(v_i^{\theta}\left(\widehat{\alpha}(\widehat{k}_i)\right)\right)' - \left(-v_i^{\alpha}\left(\widehat{\alpha}(\widehat{k}_i)\right)\right)\frac{\partial\overline{\alpha}_i}{\partial\theta'}\right) + o_p\left(1\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\widetilde{\iota}(\widehat{k}_i)\widetilde{z}(\widehat{k}_i)'\left(\underbrace{\left(\mathbb{E}\left(v_i^{\theta}\right)\right)' - \left(\mathbb{E}\left(-v_i^{\alpha}\right)\right)\frac{\partial\overline{\alpha}_i}{\partial\theta'}}_{=0}\right) + o_p\left(1\right) = o_p\left(1\right),$$

where we have used (A12) in the third equality, (A19) in the fourth one, (A21) and the expression of $\partial\widehat{\alpha}(k)/\partial\theta'$ in the fifth one, and we have expanded around α_{i0} and used (A12) in the last equality.

Proof of the second part of Theorem 1. Finally, to show (7) let us define, analogously to the beginning of the proof of (A3):

$$\widehat{\iota}_i = \mathbf{1}\left\{\|\widehat{\alpha}(\widehat{k}_i,\widehat{\theta}) - \alpha_{i0}\| \leq \eta\right\},$$

where η is such that: $\inf_{\alpha_{i0}}\inf_{\|(\alpha_i,\theta)-(\alpha_{i0},\theta_0)\|\leq 2\eta}\mathbb{E}\left(-v_i^{\alpha}(\alpha_i,\theta)\right) \geq \underline{\Sigma}$. Using that $\widehat{\theta}$ is consistent it is easy to verify that:

$$\left|\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}(\widehat{k}_i,\widehat{\theta}),\widehat{\theta}\right) - \frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}_i\left(\widehat{\theta}\right),\widehat{\theta}\right)\right| \leq \left|\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(a(\widehat{k}_i,\widehat{\theta}),\widehat{\theta}\right) - \frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}_i\left(\widehat{\theta}\right),\widehat{\theta}\right)\right| = O_p(\delta). \tag{A22}$$

Using similar arguments as at the beginning of the proof of (A3), but now at $\widehat{\theta}$, it can be shown that:

$$\frac{1}{N}\sum_{i=1}^{N}(1-\widehat{\iota}_i) = O_p(\delta), \quad \frac{1}{N}\sum_{i=1}^{N}\widehat{\iota}_i\left\|\widehat{\alpha}(\widehat{k}_i,\widehat{\theta}) - \widehat{\alpha}_i\left(\widehat{\theta}\right)\right\|^2 = O_p(\delta),$$

hence that:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\alpha}(\widehat{k}_i,\widehat{\theta}) - \widehat{\alpha}_i\left(\widehat{\theta}\right)\right\|^2 = O_p(\delta).$$

(7) then comes from the fact that $\frac{1}{N}\sum_{i=1}^{N}\|\widehat{\alpha}_i(\widehat{\theta}) - \alpha_{i0}\|^2 = O_p(T^{-1})$.

This ends the proof of Theorem 1.

A.4 Proof of Corollary 2

We follow a likelihood approach as in Arellano and Hahn (2007, 2016). Consider the difference between the grouped fixed-effects and fixed-effects concentrated likelihoods:

$$\Delta L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell_i(\widehat{\alpha}(\widehat{k}_i, \theta), \theta) - \frac{1}{N}\sum_{i=1}^{N} \ell_i(\widehat{\alpha}_i(\theta), \theta).$$

We are going to derive an expansion for the derivative of $\Delta L(\theta)$ at θ_0. From there, we will characterize the first-order bias of the grouped fixed-effects estimator $\widehat{\theta}$.

For any z_i let us denote as $\overline{\mathbb{E}}\left(z_i \,|\, h_i\right)$ the conditional expectation of z_i given h_i *across* individuals; that is, the function of h_i which minimizes:

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\alpha_{i0}}\left[\left\| z_i - \overline{\mathbb{E}}\left(z_i \,|\, h_i\right)\right\|^2\right].$$

Let: $\nu_i(\theta) = \widehat{\alpha}_i(\theta) - \overline{\mathbb{E}}\left(\widehat{\alpha}_i(\theta) \,|\, h_i\right)$. We are going to show that:

$$\frac{\partial}{\partial\theta}\Big|_{\theta_0} \Delta L(\theta) \quad = \quad -\frac{\partial}{\partial\theta}\Big|_{\theta_0} \frac{1}{2N}\sum_{i=1}^{N} \nu_i(\theta)' \mathbb{E}\left[-v_i^{\alpha}\left(\overline{\alpha}_i(\theta), \theta\right)\right] \nu_i(\theta) + o_p\left(\frac{1}{T}\right). \tag{A23}$$

To show (A23) we are first going to establish several preliminary results. Together with fourth-order differentiability, those will allow us to derive the required expansions. In the following we will evaluate all functions at θ_0, and omit θ_0 for the notation.[48] First, note that from the proof of Theorem 1 and using the fact that $\frac{1}{N}\sum_{i=1}^{N} \|h_i - \widehat{h}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right)$ we have:

$$\frac{1}{N}\sum_{i=1}^{N} \|\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\|^2 = O_p\left(\frac{1}{T}\right). \tag{A24}$$

Next, let $\widehat{\alpha}_i = \gamma(h_i) + \nu_i$, where $\gamma(h_i) = \overline{\mathbb{E}}\left(\widehat{\alpha}_i \,|\, h_i\right)$. We have:

$$\widehat{\alpha}(k) = \left(\sum_{i=1}^{N} \mathbf{1}\{\widehat{k}_i = k\}(-v_i^{\alpha}(a_i(k)))\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{1}\{\widehat{k}_i = k\}(-v_i^{\alpha}(a_i(k)))\widehat{\alpha}_i\right), \tag{A25}$$

for some $a_i(k)$ between $\widehat{\alpha}_i$ and $\widehat{\alpha}(k)$. Note that, by condition (ii) in Corollary 2 and Assumption 2 (iii), $(-v_i^{\alpha}(\alpha_i))$ is uniformly bounded away from zero with probability approaching one. Let $\widehat{\gamma}(k)$ and $\widehat{\nu}(k)$ denote the weighted means of $\gamma(h_i)$ and ν_i in group $\widehat{k}_i = k$, respectively, where the weight is $(-v_i^{\alpha}(a_i(k)))$. Note that $\widehat{\alpha}(k) = \widehat{\gamma}(k) + \widehat{\nu}(k)$. Since $\frac{1}{N}\sum_{i=1}^{N} \|h_i - \widehat{h}(\widehat{k}_i)\|^2 = o_p\left(1/T\right)$ and γ is uniformly Lipschitz, we have:

$$\frac{1}{N}\sum_{i=1}^{N} \|\gamma(h_i) - \widehat{\gamma}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right). \tag{A26}$$

[48] In particular, $\widehat{\alpha}_i$ will be a shorthand for $\widehat{\alpha}_i(\theta_0)$.

Moreover, since by condition (iii) in Corollary 2 the $\sqrt{T}\nu_i$, which are mean independent of the $\widehat{k}_j$'s and have zero mean, have bounded conditional variance, and denoting as $\overline{\nu}(k)$ the unweighted mean of ν_i in group $\widehat{k}_i = k$, we have: $\frac{1}{N}\sum_{i=1}^N \|\overline{\nu}(\widehat{k}_i)\|^2 = O_p\left(\frac{K}{NT}\right) = o_p\left(\frac{1}{T}\right)$, where we have used that K/N tends to zero. Hence:

$$\frac{1}{N}\sum_{i=1}^N \|\widehat{\nu}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right). \tag{A27}$$

Let $\widehat{g}_i = v_i^\theta(\widehat{\alpha}_i)(-v_i^\alpha(\widehat{\alpha}_i))^{-1} = \lambda(h_i) + \xi_i$, where $\lambda(h_i) = \overline{\mathbb{E}}\left(\widehat{g}_i \,|\, h_i\right)$. Similarly we have, using analogous notations for weighted group means:

$$\frac{1}{N}\sum_{i=1}^N \|\lambda(h_i) - \widehat{\lambda}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right), \quad \frac{1}{N}\sum_{i=1}^N \|\widehat{\xi}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right). \tag{A28}$$

Further, denote as $\widetilde{\gamma}(k)$, $\widetilde{\nu}(k)$, $\widetilde{\lambda}(k)$, and $\widetilde{\xi}(k)$ the weighted means of $\gamma(h_i)$, ν_i, $\lambda(h_i)$, and ξ_i in group $\widehat{k}_i = k$, respectively, where the weight is $(-v_i^\alpha(\widehat{\alpha}_i))$. By similar arguments we have:

$$\frac{1}{N}\sum_{i=1}^N \|\gamma(h_i) - \widetilde{\gamma}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right), \quad \frac{1}{N}\sum_{i=1}^N \|\widetilde{\nu}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right), \tag{A29}$$

$$\frac{1}{N}\sum_{i=1}^N \|\lambda(h_i) - \widetilde{\lambda}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right), \quad \frac{1}{N}\sum_{i=1}^N \|\widetilde{\xi}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right). \tag{A30}$$

Next, using (A25), (A26), and (A27), in addition to $\mathcal{A}$ being compact and γ being bounded, we have that: $\frac{1}{N}\sum_{i=1}^N \|\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\|^3 = -\frac{1}{N}\sum_{i=1}^N \|\nu_i\|^3 + o_p(1/T)$. Hence, by condition (iii) in Corollary 2:

$$\frac{1}{N}\sum_{i=1}^N \|\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\|^3 = o_p\left(\frac{1}{T}\right). \tag{A31}$$

To see that (A23) holds, first note that, denoting $a^{\otimes 2} = a \otimes a$:

$$\begin{aligned}
\frac{\partial}{\partial\theta}\Big|_{\theta_0} \Delta L(\theta) &= \frac{1}{N}\sum_{i=1}^N \frac{\partial \ell_i(\widehat{\alpha}(\widehat{k}_i))}{\partial\theta} - \frac{1}{N}\sum_{i=1}^N \frac{\partial \ell_i(\widehat{\alpha}_i)}{\partial\theta} \\
&= \frac{1}{N}\sum_{i=1}^N v_i^\theta(\widehat{\alpha}_i)\left(\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right) + \frac{1}{2N}\sum_{i=1}^N v_i^{\theta\alpha}(a_i)\left(\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right)^{\otimes 2} \\
&= \underbrace{\frac{1}{N}\sum_{i=1}^N v_i^\theta(\widehat{\alpha}_i)\left(\widetilde{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right)}_{\equiv A_1} + \underbrace{\frac{1}{2N}\sum_{i=1}^N v_i^{\theta\alpha}(a_i)\left(\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right)^{\otimes 2}}_{\equiv A_2} \\
&\quad \underbrace{+\frac{1}{2N}\sum_{i=1}^N v_i^\theta(\widehat{\alpha}_i)\left(\sum_{j=1}^N \mathbf{1}\{\widehat{k}_j = \widehat{k}_i\}(-v_j^\alpha(\widehat{\alpha}_j))\right)^{-1} \sum_{j=1}^N \mathbf{1}\{\widehat{k}_j = \widehat{k}_i\} v_j^{\alpha\alpha}\left(a_j(\widehat{k}_j)\right)\left(\widehat{\alpha}(\widehat{k}_j) - \widehat{\alpha}_j\right)^{\otimes 2}}_{\equiv A_3},
\end{aligned}$$

where we have used the notation of the proof of Theorem 1, a_i lies between $\widehat{\alpha}_i$ and $\widehat{\alpha}(\widehat{k}_i)$ and so does $a_i(\widehat{k}_i)$, $v_i^{\theta\alpha}(a_i)$ is a matrix of third derivatives with q^2 columns, and the last equality comes from (A15), where note that $(-v_i^{\alpha}(\widehat{\alpha}_i))$ is uniformly bounded away from zero with probability approaching one.

Let us consider the three terms in turn. First, we have:

$$\begin{aligned}
A_1 &= \frac{1}{N}\sum_{i=1}^{N}\widehat{g}_i\left(-v_i^{\alpha}(\widehat{\alpha}_i)\right)\left(\widetilde{\alpha}(\widehat{k}_i)-\widehat{\alpha}_i\right)\\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{g}_i-\widetilde{g}(\widehat{k}_i)\right)\left(-v_i^{\alpha}(\widehat{\alpha}_i)\right)\left(\widetilde{\alpha}(\widehat{k}_i)-\widehat{\alpha}_i\right)\\
&= -\frac{1}{N}\sum_{i=1}^{N}\left(\lambda(h_i)-\widetilde{\lambda}(\widehat{k}_i)+\xi_i-\widetilde{\xi}(\widehat{k}_i)\right)\left(-v_i^{\alpha}(\widehat{\alpha}_i)\right)\left(\gamma(h_i)-\widetilde{\gamma}(\widehat{k}_i)+\nu_i-\widetilde{\nu}(\widehat{k}_i)\right)\\
&= -\frac{1}{N}\sum_{i=1}^{N}\xi_i(-v_i^{\alpha}(\widehat{\alpha}_i))\nu_i+o_p\left(\frac{1}{T}\right)\\
&= -\frac{1}{N}\sum_{i=1}^{N}\xi_i\mathbb{E}(-v_i^{\alpha}(\alpha_{i0}))\nu_i+o_p\left(\frac{1}{T}\right),
\end{aligned}$$

where we have used (A25), (A29), and (A30).

Next, we have, using in addition (A31):

$$\begin{aligned}
A_2 &= \frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left(v_i^{\theta\alpha}(\alpha_{i0})\right)\left(\widehat{\alpha}(\widehat{k}_i)-\widehat{\alpha}_i\right)^{\otimes 2}+o_p\left(\frac{1}{T}\right)\\
&= \frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left(v_i^{\theta\alpha}(\alpha_{i0})\right)\left(\widehat{\gamma}(\widehat{k}_i)-\gamma(h_i)+\widehat{\nu}(\widehat{k}_i)-\nu_i\right)^{\otimes 2}+o_p\left(\frac{1}{T}\right)\\
&= \frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left(v_i^{\theta\alpha}(\alpha_{i0})\right)\nu_i^{\otimes 2}+o_p\left(\frac{1}{T}\right).
\end{aligned}$$

Lastly, defining $\widetilde{g}(k)$ the weighted mean of $\widehat{g}_i$ in group $\widehat{k}_i = k$ with weight $(-v_i^{\alpha}(\widehat{\alpha}_i))$, we have:

$$
\begin{aligned}
A_3 = \frac{1}{2N}\sum_{i=1}^{N}\widehat{g}_i(-v_i^{\alpha}(\widehat{\alpha}_i))\left(\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=\widehat{k}_i\}(-v_j^{\alpha}(\widehat{\alpha}_j))\right)^{-1} \\
\times\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=\widehat{k}_i\}v_j^{\alpha\alpha}\left(a_j(\widehat{k}_j)\right)\left(\widehat{\alpha}(\widehat{k}_j)-\widehat{\alpha}_j\right)^{\otimes 2} \\
= \frac{1}{2N}\sum_{i=1}^{N}\widetilde{g}(\widehat{k}_i)(-v_i^{\alpha}(\widehat{\alpha}_i))\left(\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=\widehat{k}_i\}(-v_j^{\alpha}(\widehat{\alpha}_j))\right)^{-1} \\
\times\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=\widehat{k}_i\}v_j^{\alpha\alpha}\left(a_j(\widehat{k}_j)\right)\left(\widehat{\alpha}(\widehat{k}_j)-\widehat{\alpha}_j\right)^{\otimes 2}+o_p\left(\frac{1}{T}\right) \\
= \frac{1}{2N}\sum_{i=1}^{N}\widetilde{g}(\widehat{k}_i)v_i^{\alpha\alpha}\left(a_i(\widehat{k}_i)\right)\left(\widehat{\alpha}(\widehat{k}_i)-\widehat{\alpha}_i\right)^{\otimes 2}+o_p\left(\frac{1}{T}\right) \\
= \frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left(v_i^{\theta}(\alpha_{i0})\right)\left[\mathbb{E}(-v_i^{\alpha}(\alpha_{i0}))\right]^{-1}\mathbb{E}\left[v_i^{\alpha\alpha}\left(\alpha_{i0}\right)\right]\nu_i^{\otimes 2}+o_p\left(\frac{1}{T}\right).
\end{aligned}
$$

Combining results, we get:

$$
\begin{aligned}
\left.\frac{\partial}{\partial\theta}\right|_{\theta_0}\Delta L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\xi_i\mathbb{E}(-v_i^{\alpha}(\alpha_{i0}))\nu_i \\
+\frac{1}{2N}\sum_{i=1}^{N}\left[\mathbb{E}\left(v_i^{\theta\alpha}(\alpha_{i0})\right)+\mathbb{E}\left(v_i^{\theta}(\alpha_{i0})\right)\left[\mathbb{E}(-v_i^{\alpha}(\alpha_{i0}))\right]^{-1}\mathbb{E}\left[v_i^{\alpha\alpha}\left(\alpha_{i0}\right)\right]\right]\nu_i^{\otimes 2}+o_p\left(\frac{1}{T}\right).
\end{aligned}
$$

This shows (A23), since $\frac{\partial\widehat{\alpha}_i(\theta_0)}{\partial\theta'}=\widehat{g}_i'$, and:

$$
\left.\frac{\partial}{\partial\theta'}\right|_{\theta_0}\operatorname{vec}\mathbb{E}\left[-v_i^{\alpha}\left(\overline{\alpha}_i(\theta),\theta\right)\right] = -\left(\mathbb{E}\left(v_i^{\theta\alpha}(\alpha_{i0})\right)+\mathbb{E}\left(v_i^{\theta}(\alpha_{i0})\right)\left[\mathbb{E}(-v_i^{\alpha}(\alpha_{i0}))\right]^{-1}\mathbb{E}\left[v_i^{\alpha\alpha}\left(\alpha_{i0}\right)\right]\right)'.
$$

As an example, consider the case where φ in Assumption 1 is one-to-one. Note that:

$$
\widehat{\alpha}_i(\theta) \quad = \quad \underbrace{\overline{\alpha}_i(\theta)}_{=\overline{\alpha}(\theta,\alpha_{i0})}+\mathbb{E}\left[-v_i^{\alpha}\left(\overline{\alpha}_i(\theta),\theta\right)\right]^{-1}v_i\left(\overline{\alpha}_i(\theta),\theta\right)+o_p\left(\frac{1}{\sqrt{T}}\right).
$$

In this case it can be shown that: $\overline{\mathbb{E}}\left(\widehat{\alpha}_i(\theta)\,|\,h_i\right)=\overline{\alpha}\left(\theta,\varphi^{-1}(h_i)\right)+o_p\left(\frac{1}{\sqrt{T}}\right)$. Hence, under suitable differentiability conditions we have the following explicit expression for $\nu_i(\theta)$ up to smaller order terms:

$$
\begin{aligned}
\nu_i(\theta) \quad &= \quad \overline{\alpha}(\theta,\alpha_{i0})-\overline{\alpha}\left(\theta,\varphi^{-1}(h_i)\right)+\mathbb{E}\left[-v_i^{\alpha}\left(\overline{\alpha}_i(\theta),\theta\right)\right]^{-1}v_i\left(\overline{\alpha}_i(\theta),\theta\right)+o_p\left(\frac{1}{\sqrt{T}}\right) \\
&= \quad \mathbb{E}\left[-v_i^{\alpha}\left(\overline{\alpha}_i(\theta),\theta\right)\right]^{-1}v_i\left(\overline{\alpha}_i(\theta),\theta\right)-\frac{\partial\overline{\alpha}\left(\theta,\alpha_{i0}\right)}{\partial\alpha_i'}\left(\frac{\partial\varphi(\alpha_{i0})}{\partial\alpha_i'}\right)^{-1}\varepsilon_i+o_p\left(\frac{1}{\sqrt{T}}\right),
\end{aligned}
$$

where recall that $\varepsilon_i=h_i-\varphi(\alpha_{i0})$, and the presence of $\left(\frac{\partial\varphi(\alpha_{i0})}{\partial\alpha_i'}\right)^{-1}$ shows the need for φ to be injective.

Equation (A23) readily delivers an expression for the first-order bias term of the grouped fixed-effects estimator. Focusing first on the case where φ is one-to-one, (A23) implies that:

$$\begin{aligned}
&\left.\frac{\partial}{\partial\theta}\right|_{\theta_0} \frac{1}{N}\sum_{i=1}^{N} \ell_i(\widehat{\alpha}(\widehat{k}_i,\theta),\theta) - \frac{1}{N}\sum_{i=1}^{N} \ell_i(\overline{\alpha}_i(\theta),\theta) \\
&= -\left.\frac{\partial}{\partial\theta}\right|_{\theta_0} \frac{1}{2N}\sum_{i=1}^{N} \varepsilon_i' \left(\frac{\partial\overline{\alpha}\left(\theta,\alpha_{i0}\right)}{\partial\alpha_i'}\left(\frac{\partial\varphi(\alpha_{i0})}{\partial\alpha_i'}\right)^{-1}\right)' \mathbb{E}\left[-v_i^{\alpha}\left(\overline{\alpha}_i(\theta),\theta\right)\right] \frac{\partial\overline{\alpha}\left(\theta,\alpha_{i0}\right)}{\partial\alpha_i'}\left(\frac{\partial\varphi(\alpha_{i0})}{\partial\alpha_i'}\right)^{-1}\varepsilon_i \\
&\qquad + \left.\frac{\partial}{\partial\theta}\right|_{\theta_0} \frac{1}{N}\sum_{i=1}^{N} v_i\left(\overline{\alpha}_i(\theta),\theta\right)' \frac{\partial\overline{\alpha}\left(\theta,\alpha_{i0}\right)}{\partial\alpha_i'}\left(\frac{\partial\varphi(\alpha_{i0})}{\partial\alpha_i'}\right)^{-1}\varepsilon_i + o_p\left(\frac{1}{T}\right),
\end{aligned}$$

where we have used that (e.g., Arellano and Hahn, 2007):

$$\begin{aligned}
&\left.\frac{\partial}{\partial\theta}\right|_{\theta_0} \frac{1}{N}\sum_{i=1}^{N} \ell_i(\overline{\alpha}_i(\theta),\theta) - \frac{1}{N}\sum_{i=1}^{N} \ell_i(\widehat{\alpha}_i(\theta),\theta) \\
&\quad = -\left.\frac{\partial}{\partial\theta}\right|_{\theta_0} \frac{1}{2N}\sum_{i=1}^{N} v_i\left(\overline{\alpha}_i(\theta),\theta\right)' \mathbb{E}\left[-v_i^{\alpha}\left(\overline{\alpha}_i(\theta),\theta\right)\right]^{-1} v_i\left(\overline{\alpha}_i(\theta),\theta\right) + o_p\left(\frac{1}{T}\right).
\end{aligned}$$

It thus follows that Corollary 2 holds, with:

$$B = H^{-1} \lim_{N,T\rightarrow\infty} \frac{1}{N}\sum_{i=1}^{N} \left.\frac{\partial}{\partial\theta}\right|_{\theta_0} T\, b_i(\theta),$$

and:

$$\begin{aligned}
b_i(\theta) = &-\frac{1}{2}\varepsilon_i' \left(\frac{\partial\overline{\alpha}\left(\theta,\alpha_{i0}\right)}{\partial\alpha_i'}\left(\frac{\partial\varphi(\alpha_{i0})}{\partial\alpha_i'}\right)^{-1}\right)' \mathbb{E}\left[-v_i^{\alpha}\left(\overline{\alpha}_i(\theta),\theta\right)\right] \frac{\partial\overline{\alpha}\left(\theta,\alpha_{i0}\right)}{\partial\alpha_i'}\left(\frac{\partial\varphi(\alpha_{i0})}{\partial\alpha_i'}\right)^{-1}\varepsilon_i \\
&+ v_i\left(\overline{\alpha}_i(\theta),\theta\right)' \frac{\partial\overline{\alpha}\left(\theta,\alpha_{i0}\right)}{\partial\alpha_i'}\left(\frac{\partial\varphi(\alpha_{i0})}{\partial\alpha_i'}\right)^{-1}\varepsilon_i.
\end{aligned}$$

More generally, when φ is not surjective:

$$\begin{aligned}
b_i(\theta) = &-\frac{1}{2}\left(\widehat{\alpha}_i(\theta) - \overline{\mathbb{E}}\left(\widehat{\alpha}_i(\theta)\,|\,h_i\right)\right)' \mathbb{E}\left[-v_i^{\alpha}\left(\overline{\alpha}_i(\theta),\theta\right)\right]\left(\widehat{\alpha}_i(\theta) - \overline{\mathbb{E}}\left(\widehat{\alpha}_i(\theta)\,|\,h_i\right)\right) \\
&+ \frac{1}{2}v_i\left(\overline{\alpha}_i(\theta),\theta\right)' \mathbb{E}\left[-v_i^{\alpha}\left(\overline{\alpha}_i(\theta),\theta\right)\right]^{-1} v_i\left(\overline{\alpha}_i(\theta),\theta\right).
\end{aligned}$$

Bias in the regression example. In Example 2, $\widehat{\alpha}_i(\theta) = (1-\rho)\overline{Y}_i - \overline{X}_i'\beta + o_p\left(T^{-\frac{1}{2}}\right)$. Hence, when classifying individuals based on $h_i = \left(\overline{Y}_i, \overline{X}_i'\right)'$, $\widehat{\alpha}_i(\theta)$ belongs to the span of h_i, up to small order terms. Hence B/T is identical to the first-order bias B^{FE}/T of fixed effects, and fixed-effects and two-step grouped fixed-effects are first-order equivalent.

This equivalence does not hold generally. As an example, suppose the unobservables $(\alpha_{i0}, \mu_{i0}')'$ follow a one-factor structure with $\mu_{i0} = \lambda\alpha_{i0}$ for a vector λ, and base the classification on $h_i = \overline{Y}_i$

only. Injectivity is satisfied in this example, due to the low underlying dimensionality of $(\alpha_{i0},\mu_{i0}')'$. In this case it can be verified that:

$$\overline{\mathbb{E}}\left(\widehat{\alpha}_i(\theta)\,|\,h_i\right)=\left(\frac{\frac{1-\rho}{1-\rho_0}+\left(\frac{1-\rho}{1-\rho_0}\beta_0-\beta\right)'\lambda}{\frac{1}{1-\rho_0}+\frac{\beta_0'\lambda}{1-\rho_0}}\right)\overline{Y}_i+o_p\left(\frac{1}{\sqrt{T}}\right),$$

and, letting $V_{it}=X_{it}-\lambda\alpha_{i0}$:

$$\nu_i(\theta)=\beta'\frac{\lambda\overline{U}_i-\overline{V}_i}{1+\beta_0'\lambda}+o_p\left(\frac{1}{\sqrt{T}}\right).$$

As a result, the first-order bias term on ρ_0 is the same for grouped fixed-effects and fixed-effects, while for β_0 we have, letting $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(V_{it}V_{it}')=\Sigma>0$:

$$B=B^{FE}-\Sigma^{-1}\lim_{T\to\infty}\mathbb{E}\left[T\left(\lambda\overline{U}_i-\overline{V}_i\right)\left(\lambda\overline{U}_i-\overline{V}_i\right)'\right]\frac{\beta_0}{\left(1+\beta_0'\lambda\right)^2},$$

so $B\neq B^{FE}$ in general.

A.5 Proof of Corollary 3

We have, by the two parts of Theorem 1 and Assumption 4:

$$\begin{aligned}\widehat{M}-M_0 &= \frac{1}{N}\sum_{i=1}^{N}m_i\left(\widehat{\alpha}(\widehat{k}_i,\widehat{\theta}),\widehat{\theta}\right)-\frac{1}{N}\sum_{i=1}^{N}m_i\left(\alpha_{i0},\theta_0\right)\\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial m_i\left(\alpha_{i0},\theta_0\right)}{\partial\alpha_i'}\left(\widehat{\alpha}(\widehat{k}_i,\widehat{\theta})-\alpha_{i0}\right)+\frac{1}{N}\sum_{i=1}^{N}\frac{\partial m_i\left(\alpha_{i0},\theta_0\right)}{\partial\theta'}\left(\widehat{\theta}-\theta_0\right)+O_p\left(\delta\right).\end{aligned}$$

Using similar arguments to those used to establish (A14) in the proof of Theorem 1, we can show that under Assumption 4:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial m_i\left(\alpha_{i0},\theta_0\right)}{\partial\alpha_i'}\left(\widehat{\alpha}(\widehat{k}_i,\widehat{\theta})-\widehat{\alpha}_i(\widehat{\theta})\right)=O_p(\delta). \tag{A32}$$

The result then comes from substituting $\widehat{\theta}-\theta_0$ by its influence function, and differentiating the identity: $v_i(\widehat{\alpha}_i(\theta),\theta)=0$ with respect to θ.

A.6 Proof of Corollary 4

Let $K\geq\widehat{K}$. Let $(\widehat{h},\{\widehat{k}_i\})$ be given by (1). We have, by (12):

$$\frac{1}{N}\sum_{i=1}^{N}\left\|h_i-\widehat{h}(\widehat{k}_i)\right\|^2\leq\xi\,\frac{1}{N}\sum_{i=1}^{N}\|h_i-\varphi(\alpha_{i0})\|^2+o_p\left(\frac{1}{T}\right).$$

Hence, by the triangular inequality we get:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{h}(\widehat{k}_i)-\varphi(\alpha_{i0})\right\|^2=O_p\left(\frac{1}{N}\sum_{i=1}^{N}\|h_i-\varphi(\alpha_{i0})\|^2\right)+o_p\left(\frac{1}{T}\right)=O_p\left(\frac{1}{T}\right).$$

Following the steps of the proof of Theorem 1 then gives the desired result.

A.7 Proof of Theorem 2

The proof shares some similarities with the proof of Theorem 1, with some important differences. The outline of the proof is identical. Throughout the proof we let $\delta=\frac{1}{T}+B_{\alpha}(K)+\frac{K}{NS}$ (or more generally $\delta=\frac{1}{T}+B_{(\alpha,\mu)}(K)+\frac{K}{NS}$ in conditional models).

Consistency of $\widehat{\theta}$. Let $a(k,\theta)=\overline{\alpha}\left(\theta,\psi\left(\widehat{h}\left(k\right)\right)\right)$. We have:

$$\sum_{i=1}^{N}\ell_i(\widehat{\alpha}(\widehat{k}_i,\theta),\theta)\geq\sum_{i=1}^{N}\ell_i\left(a\left(\widehat{k}_i,\theta\right),\theta\right).$$

Expanding, we have:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\ell_i(\widehat{\alpha}(\widehat{k}_i,\theta),\theta) &= \frac{1}{N}\sum_{i=1}^{N}\ell_i(\overline{\alpha}_i(\theta),\theta)+\frac{1}{N}\sum_{i=1}^{N}v_i(\overline{\alpha}_i(\theta),\theta)'\left(\widehat{\alpha}(\widehat{k}_i,\theta)-\overline{\alpha}_i(\theta)\right)\\ &\quad+\frac{1}{2N}\sum_{i=1}^{N}\left(\widehat{\alpha}(\widehat{k}_i,\theta)-\overline{\alpha}_i(\theta)\right)'v_i^{\alpha}(a_i(\theta),\theta)\left(\widehat{\alpha}(\widehat{k}_i,\theta)-\overline{\alpha}_i(\theta)\right),\end{aligned}$$

and:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(a\left(\widehat{k}_i,\theta\right),\theta\right) &= \frac{1}{N}\sum_{i=1}^{N}\ell_i(\overline{\alpha}_i(\theta),\theta)+\frac{1}{N}\sum_{i=1}^{N}v_i(\overline{\alpha}_i(\theta),\theta)'\left(a\left(\widehat{k}_i,\theta\right)-\overline{\alpha}_i(\theta)\right)\\ &\quad+\frac{1}{2N}\sum_{i=1}^{N}\left(a\left(\widehat{k}_i,\theta\right)-\overline{\alpha}_i(\theta)\right)'v_i^{\alpha}(b_i(\theta),\theta)\left(a\left(\widehat{k}_i,\theta\right)-\overline{\alpha}_i(\theta)\right)\\ &= \frac{1}{N}\sum_{i=1}^{N}\ell_i(\overline{\alpha}_i(\theta),\theta)+\frac{1}{N}\sum_{i=1}^{N}v_i(\overline{\alpha}_i(\theta),\theta)'\left(a\left(\widehat{k}_i,\theta\right)-\overline{\alpha}_i(\theta)\right)+O_p\left(\frac{1}{T}\right)+O_p(B_{\alpha}(K)),\end{aligned}$$

where we have used Lemma 1, that $\overline{\alpha}$ and ψ are Lipschitz, and that $(-v_i^{\alpha}(\alpha_i,\theta))$ is uniformly bounded. The O_p terms are uniform in θ.

Hence, using that $(-v_i^{\alpha}(a_i(\theta),\theta))$ is uniformly bounded away from zero:

$$\sup_{\theta\in\Theta}\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\alpha}(\widehat{k}_i,\theta)-\overline{\alpha}_i(\theta)\right\|^2 = O_p\left(\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{i=1}^{N}v_i(\overline{\alpha}_i(\theta),\theta)'\left(\widehat{\alpha}(\widehat{k}_i,\theta)-a\left(\widehat{k}_i,\theta\right)\right)\right|\right)+O_p(\delta).$$

Let $\overline{v}(\theta,k)$ denote the mean of $v_i(\overline{\alpha}_i(\theta),\theta)$ in group $\widehat{k}_i=k$. We are going to bound the following quantity:

$$\frac{1}{N}\sum_{i=1}^{N}v_i(\overline{\alpha}_i(\theta),\theta)'\left(\widehat{\alpha}(\widehat{k}_i,\theta)-a\left(\widehat{k}_i,\theta\right)\right)=\frac{1}{N}\sum_{i=1}^{N}\overline{v}(\widehat{k}_i,\theta)'\left(\widehat{\alpha}(\widehat{k}_i,\theta)-a\left(\widehat{k}_i,\theta\right)\right).$$

We have, for all $\theta\in\Theta$ (that is, pointwise):

$$\begin{aligned}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\|\overline{v}(\widehat{k}_i,\theta)\|^2\right] &= \frac{1}{N}\sum_{k=1}^{K}\mathbb{E}\left[\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}\right)\|\overline{v}(k,\theta)\|^2\right]\\ &= \frac{1}{N}\sum_{k=1}^{K}\mathbb{E}\left[\frac{\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}\mathbb{E}\left(v_i(\overline{\alpha}_i(\theta),\theta)'v_i(\overline{\alpha}_i(\theta),\theta)\,|\,\{\alpha_{i0}\}\right)}{\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}}\right]=O\left(\frac{K}{NS}\right),\end{aligned}$$

where we have used that, by Assumptions 5 and 6 (iii), the $v_i(\overline{\alpha}_i(\theta), \theta)$ are independent of each other and independent of the $\widehat{k}_j$'s conditional on the α_{j0}'s, with conditional variances that are $O(1/S)$. Hence:

$$\frac{1}{N}\sum_{i=1}^{N} \|\overline{v}(\widehat{k}_i, \theta)\|^2 = O_p\left(\frac{K}{NS}\right). \tag{A33}$$

Hence, by the Cauchy Schwarz and triangular inequalities:

$$A \equiv \frac{1}{N}\sum_{i=1}^{N} \left\|\widehat{\alpha}(\widehat{k}_i, \theta) - \overline{\alpha}_i(\theta)\right\|^2 \quad \leq \quad O_p\left(\sqrt{\frac{K}{NS}}\right)\left(\sqrt{A} + \sqrt{O_p\left(\delta\right)}\right) + O_p(\delta),$$

so, solving for $\sqrt{A}$ we get:

$$\frac{1}{N}\sum_{i=1}^{N} \left\|\widehat{\alpha}(\widehat{k}_i, \theta) - \overline{\alpha}_i(\theta)\right\|^2 \quad = \quad O_p\left(\delta\right). \tag{A34}$$

We are now going to show that:

$$\sup_{\theta \in \Theta} \frac{1}{N}\sum_{i=1}^{N} \|\overline{v}(\widehat{k}_i, \theta)\|^2 = o_p\left(1\right). \tag{A35}$$

Using a similar bounding argument as above will then imply that:

$$\sup_{\theta \in \Theta} \frac{1}{N}\sum_{i=1}^{N} \left\|\widehat{\alpha}(\widehat{k}_i, \theta) - \overline{\alpha}_i(\theta)\right\|^2 = o_p\left(1\right).$$

To see that (A35) holds, let $Z(\theta) = \frac{1}{N}\sum_{i=1}^{N} \|\overline{v}(\widehat{k}_i, \theta)\|^2$. We have shown that $Z(\theta) = O_p(K/NS)$ for all θ. Moreover, $\frac{\partial Z(\theta)}{\partial \theta} = \frac{2}{N}\sum_{i=1}^{N} \overline{v}^{\theta}(\widehat{k}_i, \theta)\overline{v}(\widehat{k}_i, \theta) = O_p(\sqrt{\sup_{\theta \in \Theta} Z(\theta)})$ uniformly in θ by Cauchy Schwarz and Assumption 6 (ii). Since the parameter space is compact it follows that $\sup_{\theta} Z(\theta) = o_p(1)$.[49]

The above shows that, as N, T, K tend to infinity such that $\frac{K}{NS}$ tends to zero, $\frac{1}{N}\sum_{i=1}^{N} \ell_i(\widehat{\alpha}(\widehat{k}_i, \theta), \theta)$ is uniformly consistent to: $\overline{\ell}(\theta) = \lim_{N \to \infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\ell_i(\overline{\alpha}_i(\theta), \theta)\right]$, which is uniquely maximized at θ_0 by Assumption 6 (i). Consistency of $\widehat{\theta}$ follows since the parameter space for θ is compact.

Rate of the score. We are now going to show that:

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\partial \ell_i(\widehat{\alpha}(\widehat{k}_i, \theta_0), \theta_0)}{\partial \theta} \quad = \quad \frac{1}{N}\sum_{i=1}^{N} s_i + O_p\left(\delta\right). \tag{A36}$$

[49] Let $\eta > 0, \epsilon > 0$. There is a constant $M > 0$ such that $\Pr\left(\sup_{\theta \in \Theta} \left\|\frac{\partial Z(\theta)}{\partial \theta}\right\| > M\sqrt{\sup_{\theta \in \Theta} Z(\theta)}\right) < \frac{\epsilon}{2}$. Take a finite cover of $\Theta = B_1 \cup ... \cup B_R$, where B_r are balls with centers θ_r and $\operatorname{diam} B_r \leq \frac{1}{2M}\sqrt{\eta}$. Since: $\sup_{\theta \in \Theta} Z(\theta) \leq \max_r Z(\theta_r) + \sup_{\theta} \left\|\frac{\partial Z(\theta)}{\partial \theta}\right\| \frac{1}{2M}\sqrt{\eta}$, and since: $a > \eta \Rightarrow a - \sqrt{a}\frac{1}{2}\sqrt{\eta} > \frac{\eta}{2}$, we have: $\Pr\left(\sup_{\theta \in \Theta} Z(\theta) > \eta\right) \leq \frac{\epsilon}{2} + \Pr\left(\max_r Z(\theta_r) > \frac{\eta}{2}\right)$, which, by (A33), is smaller than ϵ for N, T, K large enough.

We have, omitting references to θ_0 and α_{i0} for conciseness:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \ell_i(\widehat{\alpha}(\widehat{k}_i))}{\partial \theta} \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \ell_i(\alpha_{i0})}{\partial \theta}+\frac{1}{N}\sum_{i=1}^{N} v_i^{\theta}\left(\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\right)+O_p\left(\delta\right),$$

where we have used (A34) evaluated at $\theta=\theta_0$, and part (ii) in Assumption 6.

Expanding $\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}v_i(\widehat{\alpha}(k))=0$, we have: $\widehat{\alpha}(k)=\widetilde{\alpha}(k)+\widetilde{v}(k)+\widetilde{w}(k)$, where:

$$\widetilde{\alpha}(k)=\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}(-v_i^{\alpha})\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}(-v_i^{\alpha})\alpha_{i0}\right),$$

$$\widetilde{v}(k)=\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}(-v_i^{\alpha})\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}v_i\right),$$

and:

$$\widetilde{w}(k)=\frac{1}{2}\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}(-v_i^{\alpha})\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}v_i^{\alpha\alpha}(a_i)\left(\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\right)\otimes\left(\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\right)\right), \quad \text{(A37)}$$

where a_i lies between α_{i0} and $\widehat{\alpha}(\widehat{k}_i)$.

For all functions of α_{i0}, say $z_i=g(\alpha_{i0})$, we will denote:

$$\widetilde{z}(k)=\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}(-v_i^{\alpha})\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}(-v_i^{\alpha})z_i\right),$$

and:

$$z^{*}(k)=\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}\mathbb{E}\left(-v_i^{\alpha}\right)\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}\mathbb{E}\left(-v_i^{\alpha}\right)z_i\right).$$

To establish (A36) we are going to show that:

$$\frac{1}{N}\sum_{i=1}^{N} v_i^{\theta}\left(\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\right)+\mathbb{E}\left(v_i^{\theta}\right)\left[\mathbb{E}\left(v_i^{\alpha}\right)\right]^{-1}v_i=O_p\left(\delta\right). \quad \text{(A38)}$$

For this we will bound, in turn:

$$\begin{aligned} A &\equiv \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(v_i^{\theta}\right)\left[\mathbb{E}\left(v_i^{\alpha}\right)\right]^{-1}v_i^{\alpha}\left(\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}+(v_i^{\alpha})^{-1}v_i\right), \\ B &\equiv \frac{1}{N}\sum_{i=1}^{N}\left(v_i^{\theta}\left(v_i^{\alpha}\right)^{-1}-\mathbb{E}\left(v_i^{\theta}\right)\left[\mathbb{E}\left(v_i^{\alpha}\right)\right]^{-1}\right)v_i^{\alpha}\left(\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\right). \end{aligned}$$

Let us start with A. We have:

$$A \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(v_i^{\theta}\right)\left[\mathbb{E}\left(v_i^{\alpha}\right)\right]^{-1}v_i^{\alpha}\left(\widetilde{w}(\widehat{k}_i)+\widetilde{\alpha}(\widehat{k}_i)-\alpha_{i0}+\widetilde{v}(\widehat{k}_i)+(v_i^{\alpha})^{-1}v_i\right).$$

Note first that:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(v_i^{\theta}\right)\left[\mathbb{E}\left(v_i^{\alpha}\right)\right]^{-1}v_i^{\alpha}\widetilde{w}(\widehat{k}_i) \quad = \quad O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\right\|^2\right)=O_p(\delta),$$

where we have used (A37), (A34) at $\theta=\theta_0$, and parts (i) and (ii) in Assumption 6.

Let $z_i'=\mathbb{E}\left(v_i^{\theta}\right)\left[\mathbb{E}\left(v_i^{\alpha}\right)\right]^{-1}$. We have (with probability approaching one):

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(v_i^{\theta}\right)\left[\mathbb{E}\left(v_i^{\alpha}\right)\right]^{-1}v_i^{\alpha}\left(\widetilde{\alpha}(\widehat{k}_i)-\alpha_{i0}\right) \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\left(z_i'-\widetilde{z}\left(\widehat{k}_i\right)'\right)v_i^{\alpha}\left(\widetilde{\alpha}(\widehat{k}_i)-\alpha_{i0}\right).$$

Now, from the assumptions on derivatives, strict concavity of ℓ_i, and (A34), we have, since $\widetilde{\alpha}=\operatorname{argmin}_{(\alpha(1),\ldots,\alpha(K))}\sum_{i=1}^{N}\left(\alpha(\widehat{k}_i)-\alpha_{i0}\right)'(-v_i^{\alpha})\left(\alpha(\widehat{k}_i)-\alpha_{i0}\right)$:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\widetilde{\alpha}(\widehat{k}_i)-\alpha_{i0}\right\|^2=O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left(\widetilde{\alpha}(\widehat{k}_i)-\alpha_{i0}\right)'(-v_i^{\alpha})\left(\widetilde{\alpha}(\widehat{k}_i)-\alpha_{i0}\right)\right)$$
$$=O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\right)'(-v_i^{\alpha})\left(\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\right)\right)=O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\right\|^2\right)=O_p(\delta).$$

Likewise, for any $z_i=g(\alpha_{i0})$ with g Lipschitz:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\left\|\widetilde{z}(\widehat{k}_i)-z_i\right\|^2&=O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left(\widetilde{z}(\widehat{k}_i)-z_i\right)'(-v_i^{\alpha})\left(\widetilde{z}(\widehat{k}_i)-z_i\right)\right)\\&=O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left(g\left(a(\widehat{k}_i,\theta_0)\right)-g(\alpha_{i0})\right)'(-v_i^{\alpha})\left(g\left(a(\widehat{k}_i,\theta_0)\right)-g(\alpha_{i0})\right)\right)\\&=O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left\|a(\widehat{k}_i,\theta_0)-\alpha_{i0}\right\|^2\right)=O_p(\delta).\end{aligned}\tag{A39}$$

Combining, using Cauchy Schwarz we get:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(v_i^{\theta}\right)\left[\mathbb{E}\left(v_i^{\alpha}\right)\right]^{-1}v_i^{\alpha}\left(\widetilde{\alpha}(\widehat{k}_i)-\alpha_{i0}\right) \quad = \quad O_p(\delta).$$

The last term in A is:

$$A_3=\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(v_i^{\theta}\right)\left[\mathbb{E}\left(v_i^{\alpha}\right)\right]^{-1}(-v_i^{\alpha})\left((-v_i^{\alpha})^{-1}v_i-\widetilde{v}(\widehat{k}_i)\right).$$

Note that:

$$\widetilde{v}(k)=\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}(-v_i^{\alpha})\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}(-v_i^{\alpha})(-v_i^{\alpha})^{-1}v_i\right).$$

Letting as before $z_i' = \mathbb{E}\left(v_i^\theta\right)\left[\mathbb{E}\left(v_i^\alpha\right)\right]^{-1}$, we thus have:

$$\begin{aligned} A_3 &= \frac{1}{N}\sum_{i=1}^N \left(z_i' - \widetilde{z}\left(\widehat{k}_i\right)'\right)(-v_i^\alpha)(-v_i^\alpha)^{-1}v_i = \frac{1}{N}\sum_{i=1}^N \left(z_i' - \widetilde{z}\left(\widehat{k}_i\right)'\right)v_i \\ &= \frac{1}{N}\sum_{i=1}^N \left(z_i' - z^*\left(\widehat{k}_i\right)'\right)v_i + \frac{1}{N}\sum_{i=1}^N \left(z^*\left(\widehat{k}_i\right)' - \widetilde{z}\left(\widehat{k}_i\right)'\right)v_i. \end{aligned} \tag{A40}$$

The first term in (A40) is $O_p(\delta)$ due to the fact that, conditionally on all α_{j0}'s, the v_i are independent of each other with zero mean, and independent of all $\widehat{k}_j$'s, so:

$$\begin{aligned} \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \left(z_i' - z^*\left(\widehat{k}_i\right)'\right)v_i\right\|^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \left(z_i' - z^*\left(\widehat{k}_i\right)'\right)v_i\right\|^2 \middle| \{\alpha_{i0}\}\right]\right] \\ &= \mathbb{E}\left[\frac{1}{N^2}\sum_{i=1}^N \left(z_i' - z^*\left(\widehat{k}_i\right)'\right)\mathbb{E}\left[v_iv_i' \middle| \{\alpha_{i0}\}\right]\left(z_i - z^*\left(\widehat{k}_i\right)\right)\right], \end{aligned}$$

which is $O(\delta/NS) = O(\delta^2)$ since by part (iii) in Assumption 6 $\mathbb{E}[v_iv_i' \,|\, \{\alpha_{i0}\}]$ is uniformly $O(1/S)$, and $\frac{1}{N}\sum_{i=1}^N \|z_i - z^*(\widehat{k}_i)\|^2$ is $O_p(\delta)$ by a similar argument as (A39), since $\mathbb{E}(-v_i^\alpha)$ is bounded away from zero.

As for the second term in (A40) we have:

$$\frac{1}{N}\sum_{i=1}^N \left(z^*\left(\widehat{k}_i\right)' - \widetilde{z}\left(\widehat{k}_i\right)'\right)v_i = \frac{1}{N}\sum_{i=1}^N \left(z^*\left(\widehat{k}_i\right)' - \widetilde{z}\left(\widehat{k}_i\right)'\right)\overline{v}\left(\widehat{k}_i\right),$$

where by (A33) evaluated at $\theta = \theta_0$ we have: $\frac{1}{N}\sum_{i=1}^N \|\overline{v}(\widehat{k}_i)\|^2 = O_p(K/NS) = O_p(\delta)$.

Moreover:

$$\frac{1}{N}\sum_{i=1}^N \left\|z^*\left(\widehat{k}_i\right) - \widetilde{z}\left(\widehat{k}_i\right)\right\|^2 = O_p\left(\frac{1}{N}\sum_{i=1}^N \left\|z_i - z^*\left(\widehat{k}_i\right)\right\|^2 + \frac{1}{N}\sum_{i=1}^N \left\|z_i - \widetilde{z}\left(\widehat{k}_i\right)\right\|^2\right),$$

where the second term on the right-hand side is $O_p(\delta)$ due to (A39), and the first term is also $O_p(\delta)$. This establishes that $A = O_p(\delta)$.

Let us now turn to B. Letting: $\eta_i' = v_i^\theta\left(v_i^\alpha\right)^{-1} - \mathbb{E}\left(v_i^\theta\right)\left[\mathbb{E}\left(v_i^\alpha\right)\right]^{-1}$, we have:

$$B = \frac{1}{N}\sum_{i=1}^N \eta_i' v_i^\alpha\left(\widetilde{w}(\widehat{k}_i) + \widetilde{v}(\widehat{k}_i) + \widetilde{\alpha}(\widehat{k}_i) - \alpha_{i0}\right).$$

Similarly as above we have, using part (ii) in Assumption 6: $\frac{1}{N}\sum_{i=1}^N \eta_i' v_i^\alpha \widetilde{w}(\widehat{k}_i) = O_p(\delta)$. Next, we have:

$$\frac{1}{N}\sum_{i=1}^N \eta_i' v_i^\alpha \widetilde{v}(\widehat{k}_i) = \frac{1}{N}\sum_{i=1}^N \widetilde{\eta}(\widehat{k}_i)' v_i^\alpha \widetilde{v}(\widehat{k}_i).$$

To see that the right-hand side is $O_p(K/NS)$, first note that, by strict concavity of the likelihood:[50]

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\widetilde{v}(\widehat{k}_i)\right\|^2=\frac{1}{N}\sum_{i=1}^{N}\left\|\left(\frac{\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=\widehat{k}_i\}(-v_j^{\alpha})}{\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=\widehat{k}_i\}}\right)^{-1}\overline{v}(\widehat{k}_i)\right\|^2=O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\overline{v}(\widehat{k}_i)\right\|^2\right).$$

Moreover, letting $\tau_i=\eta_i'v_i^{\alpha}$ we have:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\widetilde{\eta}(\widehat{k}_i)\right\|^2=\frac{1}{N}\sum_{i=1}^{N}\left\|\left(\frac{\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=\widehat{k}_i\}(-v_j^{\alpha})}{\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=\widehat{k}_i\}}\right)^{-1}\overline{\tau}(\widehat{k}_i)\right\|^2=O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\overline{\tau}(\widehat{k}_i)\right\|^2\right),$$

where the τ_i's are independent of each other conditional on α_{j0}'s, and independent of $\widehat{k}_j$'s, with mean:

$$\mathbb{E}\left(\eta_i'v_i^{\alpha}\right)=\mathbb{E}\left(\left(v_i^{\theta}\left(v_i^{\alpha}\right)^{-1}-\mathbb{E}\left(v_i^{\theta}\right)\left[\mathbb{E}\left(v_i^{\alpha}\right)\right]^{-1}\right)v_i^{\alpha}\right)=0,$$

and bounded conditional variances. It thus follows that $\frac{1}{N}\sum_{i=1}^{N}\|\widetilde{\eta}(\widehat{k}_i)\|^2=O_p(\delta)$, by a similar argument as in (A33).

We lastly bound the third term in B:

$$B_3=\frac{1}{N}\sum_{i=1}^{N}\eta_i'v_i^{\alpha}\left(\widetilde{\alpha}(\widehat{k}_i)-\alpha_{i0}\right)=\frac{1}{N}\sum_{i=1}^{N}\eta_i'v_i^{\alpha}\left(\alpha^*(\widehat{k}_i)-\alpha_{i0}\right)+\frac{1}{N}\sum_{i=1}^{N}\eta_i'v_i^{\alpha}\left(\widetilde{\alpha}(\widehat{k}_i)-\alpha^*(\widehat{k}_i)\right).$$

The first term is $O_p(\delta)$ since: $\frac{1}{N}\sum_{i=1}^{N}\|\alpha^*(\widehat{k}_i)-\alpha_{i0}\|^2=O_p(\delta)$, and the $\tau_i=\eta_i'v_i^{\alpha}$ are independent of each other conditional on α_{j0}'s, and independent of $\widehat{k}_j$'s, with zero mean and bounded conditional variances (using a similar argument as for the first term in (A40)). The second term is:

$$\frac{1}{N}\sum_{i=1}^{N}\eta_i'v_i^{\alpha}\left(\widetilde{\alpha}(\widehat{k}_i)-\alpha^*(\widehat{k}_i)\right)\quad=\quad\frac{1}{N}\sum_{i=1}^{N}\widetilde{\eta}(\widehat{k}_i)'v_i^{\alpha}\left(\widetilde{\alpha}(\widehat{k}_i)-\alpha^*(\widehat{k}_i)\right).$$

We have already shown that: $\frac{1}{N}\sum_{i=1}^{N}\|\widetilde{\eta}(\widehat{k}_i)\|^2=O_p(\delta)$. Moreover, using similar arguments as for $\frac{1}{N}\sum_{i=1}^{N}\|z^*(\widehat{k}_i)-\widetilde{z}(\widehat{k}_i)\|^2$ above, we have: $\frac{1}{N}\sum_{i=1}^{N}\|\widetilde{\alpha}(\widehat{k}_i)-\alpha^*(\widehat{k}_i)\|^2=O_p(\delta)$.

This shows that $B=O_p(\delta)$ and establishes (A36).

Consistency of the Hessian. We are finally going to show that:

$$\Delta^2S(\theta_0)\equiv\left.\frac{\partial^2}{\partial\theta\partial\theta'}\right|_{\theta_0}\frac{1}{N}\sum_{i=1}^{N}\left(\ell_i\left(\widehat{\alpha}(\widehat{k}_i,\theta),\theta\right)-\ell_i\left(\overline{\alpha}_i(\theta),\theta\right)\right)=o_p(1).\qquad\text{(A41)}$$

The proof of Theorem 2 will then follow from standard arguments as in the proof of Theorem 1. Similarly as in the proof of Theorem 1, we have:

$$\Delta^2S(\theta_0)\quad=\quad\frac{1}{N}\sum_{i=1}^{N}v_i^{\theta}\left(\frac{\partial\widehat{\alpha}(\widehat{k}_i,\theta_0)}{\partial\theta'}-\frac{\partial\overline{\alpha}_i(\theta_0)}{\partial\theta'}\right)+o_p\left(1\right).$$

[50] Recall that: $\widetilde{v}(k)=\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}(-v_i^{\alpha})\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}(-v_i^{\alpha})(-v_i^{\alpha})^{-1}v_i\right)$. Hence a more precise (though also more cumbersome) alternative notation for $\widetilde{v}(k)$ could be: $\widetilde{(-v^{\alpha})^{-1}v}\ (k)$.

We will now show that:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\partial\widehat{\alpha}(\widehat{k}_i,\theta_0)}{\partial\theta'}-\frac{\partial\overline{\alpha}_i(\theta_0)}{\partial\theta'}\right\|^2=o_p\left(1\right). \tag{A42}$$

We have:

$$\frac{\partial\widehat{\alpha}(k,\theta_0)}{\partial\theta'}=\left(\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=k\}\left(-v_j^{\alpha}\left(\widehat{\alpha}(k)\right)\right)\right)^{-1}\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=k\}\left(v_j^{\theta}\left(\widehat{\alpha}(k)\right)\right)'. \tag{A43}$$

Let us define, at true values:

$$\frac{\partial\widetilde{\alpha}(k,\theta_0)}{\partial\theta'}=\left(\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=k\}(-v_j^{\alpha})\right)^{-1}\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=k\}(v_j^{\theta})',$$

and:

$$\frac{\partial\widetilde{\alpha}^*(k,\theta_0)}{\partial\theta'}=\left(\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=k\}(-v_j^{\alpha})\right)^{-1}\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=k\}(-v_j^{\alpha})\underbrace{\left[\mathbb{E}(-v_j^{\alpha})\right]^{-1}\mathbb{E}(v_j^{\theta})'}_{=\frac{\partial\overline{\alpha}_j(\theta_0)}{\partial\theta'}}.$$

We have:

$$\begin{aligned}&\frac{\partial\widehat{\alpha}(\widehat{k}_i,\theta_0)}{\partial\theta'}-\frac{\partial\widetilde{\alpha}(\widehat{k}_i,\theta_0)}{\partial\theta'}\\&=\left(\frac{\partial}{\partial\alpha}\bigg|_{a_i}\left(\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=k\}(-v_j^{\alpha}\left(\alpha,\theta_0\right))\right)^{-1}\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=k\}\left(v_j^{\theta}\left(\alpha,\theta_0\right)\right)'\right)\left(\widehat{\alpha}(\widehat{k}_i,\theta_0)-\alpha_{i0}\right),\end{aligned}$$

where a_i lies between α_{i0} and $\widehat{\alpha}(\widehat{k}_i,\theta_0)$. By parts (i) and (ii) in Assumption 6 we thus have, using (A34):

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\partial\widehat{\alpha}(\widehat{k}_i,\theta_0)}{\partial\theta'}-\frac{\partial\widetilde{\alpha}(\widehat{k}_i,\theta_0)}{\partial\theta'}\right\|^2=o_p\left(1\right).$$

Moreover:

$$\begin{aligned}\frac{\partial\widetilde{\alpha}(k,\theta_0)}{\partial\theta'}-\frac{\partial\widetilde{\alpha}^*(k,\theta_0)}{\partial\theta'}&=\left(\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=k\}(-v_j^{\alpha})\right)^{-1}\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=k\}\tau_j'\\&=\left(\frac{\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=k\}(-v_j^{\alpha})}{\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=k\}}\right)^{-1}\left(\frac{\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=k\}\tau_j'}{\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=k\}}\right),\end{aligned}$$

where the $\tau_i'=(v_i^{\theta})'-(-v_i^{\alpha})\left[\mathbb{E}(-v_i^{\alpha})\right]^{-1}\mathbb{E}(v_i^{\theta})'$ are independent of each other conditional on α_{j0}'s, and independent of $\widehat{k}_j$'s, with zero mean and bounded conditional variances. Hence, since $(-v_i^{\alpha})$ is bounded away from zero:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\partial\widetilde{\alpha}(\widehat{k}_i,\theta_0)}{\partial\theta'}-\frac{\partial\widetilde{\alpha}^*(\widehat{k}_i,\theta_0)}{\partial\theta'}\right\|^2=o_p\left(1\right).$$

Lastly, using again that $(-v_i^{\alpha})$ is bounded away from zero we have, as in (A39):

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\partial\widetilde{\alpha}^*(\widehat{k}_i,\theta_0)}{\partial\theta'}-\frac{\partial\overline{\alpha}_i(\theta_0)}{\partial\theta'}\right\|^2=o_p\left(1\right).$$

Combining results shows (A42).

Finally, using the expression of $\frac{\partial\widehat{\alpha}(\widehat{k}_i,\theta)}{\partial\theta'}$ (see (A43)), and using parts (i) and (ii) in Assumption 6, we have: $\frac{1}{N}\sum_{i=1}^{N}\|\frac{\partial^2\widehat{\alpha}(\widehat{k}_i,\theta)}{\partial\theta'\otimes\partial\theta'}\|^2=O_p(1)$, uniformly around θ_0. This implies that the third derivative of $\frac{1}{N}\sum_{i=1}^{N}\widehat{\ell}_i(\theta)$ is uniformly $O_p(1)$ in a neighborhood of θ_0.

This ends the proof of Theorem 2.

A.8 Proof of Theorem 3

Let us start with a lemma.[51]

Lemma A1. *Let Assumption 7 hold. Then, as N,T,K tend to infinity:*

$$\frac{1}{NT}\sum_{i=1}^{N}\left\|\widehat{h}(\widehat{k}_i)-\varphi(\alpha_{i0})\right\|^2 \;=\; O_p\left(\frac{\ln K}{T}\right)+O_p\left(\frac{K}{N}\right)+O_p\left(\frac{B_\alpha(K)}{T}\right). \tag{A44}$$

Proof. Let $(h^*,\{k_i^*\})$ be defined similarly as in Lemma 1. Let $\overline{\varepsilon}(\widehat{k}_i,\widetilde{k}_i)$ denote the linear projection of ε_i on the indicators $\mathbf{1}\{\widehat{k}_i=k\}$ and $\mathbf{1}\{\widetilde{k}_i=k\}$, all of which are interacted with component indicators. Since: $\sum_{i=1}^{N}\|h_i-\widehat{h}(\widehat{k}_i)\|^2\leq\sum_{i=1}^{N}\|h_i-h^*(k_i^*)\|^2$ we have:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\left\|\varphi(\alpha_{i0})-\widehat{h}(\widehat{k}_i)\right\|^2 &\leq B_{\varphi(\alpha)}(K)+\frac{2}{N}\sum_{i=1}^{N}\varepsilon_i'\left(\widehat{h}(\widehat{k}_i)-h^*(k_i^*)\right)\\ &= B_{\varphi(\alpha)}(K)+\frac{2}{N}\sum_{i=1}^{N}\overline{\varepsilon}(\widehat{k}_i,\widetilde{k}_i)'\left(\widehat{h}(\widehat{k}_i)-h^*(k_i^*)\right)\\ &\leq B_{\varphi(\alpha)}(K)+2\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\overline{\varepsilon}(\widehat{k}_i,\widetilde{k}_i)\right\|^2\right)^{\frac{1}{2}}\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{h}(\widehat{k}_i)-h^*(k_i^*)\right\|^2\right)^{\frac{1}{2}}.\end{aligned}$$

Letting $A=\frac{1}{N}\sum_{i=1}^{N}\|\varphi(\alpha_{i0})-\widehat{h}(\widehat{k}_i)\|^2$ we thus have:

$$A\leq B_{\varphi(\alpha)}(K)+2\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\overline{\varepsilon}(\widehat{k}_i,\widetilde{k}_i)\right\|^2\right)^{\frac{1}{2}}\left(\sqrt{A}+\sqrt{B_{\varphi(\alpha)}(K)}\right).$$

Solving for $\sqrt{A}$ in this equation gives, using that $B_{\varphi(\alpha)}(K)=O_p(B_\alpha(K))$ since φ is Lipschitz:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\varphi(\alpha_{i0})-\widehat{h}(\widehat{k}_i)\right\|^2 \;=\; O_p(B_\alpha(K))+O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\overline{\varepsilon}(\widehat{k}_i,\widetilde{k}_i)\right\|^2\right).$$

[51] In conditional models Lemma A1 holds with $O_p\left(\frac{B_{(\alpha,\mu)}(K)}{T}\right)$ instead of $O_p\left(\frac{B_\alpha(K)}{T}\right)$.

We are now going to show that:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\overline{\varepsilon}(\widehat{k}_i,\widetilde{k}_i)\right\|^2 = O_p\left(\ln K\right) + O_p\left(\frac{KT}{N}\right). \tag{A45}$$

For this purpose we apply a version the Hanson-Wright tail inequality for quadratic forms, due to Hsu, Kakade and Zhang (2012, Theorem 2.1), which allows for dependent data.

Lemma A2. *(Hsu et al., 2012) Let Z be a m-dimensional random vector such that, for some $\lambda > 0$, $\mathbb{E}\left[\exp(\tau' Z)\right] \leq \exp(\lambda \left\|\tau\right\|^2)$ for all $\tau \in \mathbb{R}^m$. Let A be a positive semi-definite matrix. Then, for all $s > 0$:*

$$\Pr\left[Z'AZ > 2\lambda \operatorname{tr} A + 4\lambda\sqrt{s \operatorname{tr} A^2} + s4\lambda \left\|A\right\|\right] \leq \exp(-s).$$

Let, for given partitions $\{k_{i1}\},\{k_{i2}\}$: $\frac{1}{N}\sum_{i=1}^{N}\|\overline{\varepsilon}(k_{1i},k_{2i})\|^2 = \frac{\varepsilon' A\varepsilon}{N}$, where $\varepsilon = (\varepsilon_1',...,\varepsilon_N')'$, A is a $rN \times rN$ projection matrix with $\operatorname{tr} A = 2rK$, $A^2 = A$, and $\|A\| = 1$. By Lemma A2 and Assumption 7 we have:

$$\Pr\left[\varepsilon' A\varepsilon > 4\lambda rK + 4\lambda\sqrt{2rKs} + 4\lambda s\right] \leq \exp(-s),$$

so, using that $2\sqrt{ab} \leq a + b$:

$$\Pr\left[\varepsilon' A\varepsilon > 8\lambda rK + 6\lambda s\right] \leq \exp(-s),$$

hence, for all $b > 0$:

$$\Pr\left[\frac{\varepsilon' A\varepsilon}{N} > b\right] \leq \exp\left[-\left(\frac{bN}{6\lambda} - \frac{4rK}{3}\right)\right].$$

Lastly, by the union bound, given that the set of partitions $\{k_{i1}\} \cap \{k_{i2}\}$ has K^{2N} elements:

$$\begin{aligned}
\Pr\left[\frac{1}{N}\sum_{i=1}^{N}\left\|\overline{\varepsilon}(\widehat{k}_i,\widetilde{k}_i)\right\|^2 > b\right] &\leq K^{2N}\max_{(\{k_{i1}\},\{k_{i2}\})}\Pr\left[\frac{1}{N}\sum_{i=1}^{N}\|\overline{\varepsilon}(k_{1i},k_{2i})\|^2 > b\right] \\
&\leq \exp\left[2N\ln K + \frac{4rK}{3} - \frac{bN}{6\lambda}\right].
\end{aligned}$$

Using that r/T tends to a positive constant then implies (A45) and ends the proof of Lemma A1.

∎

The rest of the proof is similar to the proof of Theorem 2. Let us denote $v_{it} = \frac{\partial \ell_{it}}{\partial \alpha_i(t)}$, $v_{it}^{\alpha} = \frac{\partial^2 \ell_{it}}{\partial \alpha_i(t)\partial \alpha_i(t)'}$, $v_i = \frac{1}{T}\left(v_{i1}',...,v_{iT}'\right)'$, and $v_i^{\alpha} = \frac{1}{T}\operatorname{diag}\left(v_{i1}^{\alpha},...,v_{iT}^{\alpha}\right)$. Let also $\delta = \frac{\ln K}{T} + \frac{K}{N} + \frac{B_{\alpha}(K)}{T}$ (or $\delta = \frac{\ln K}{T} + \frac{K}{N} + \frac{B_{(\alpha,\mu)}(K)}{T}$ in conditional models).

Consistency of $\widehat{\theta}$. Letting $a(k,\theta)=\overline{\alpha}\left(\theta,\psi\left(\widehat{h}\left(k\right)\right)\right)$, we start by noting that, by Lemma A1 and since $\overline{\alpha}$ and ψ are Lipschitz by Assumptions 7 and 8 (i)-(ii):

$$\sup_{\theta\in\Theta}\frac{1}{NT}\sum_{i=1}^{N}\left\|a(\widehat{k}_i,\theta)-\overline{\alpha}_i(\theta)\right\|^2=O_p(\delta).$$

Proceeding as in the beginning of the proof of Theorem 2 we then have, using that $(-v_{it}^{\alpha})$ is bounded away from zero and infinity:

$$\sup_{\theta\in\Theta}\frac{1}{NT}\sum_{i=1}^{N}\left\|\widehat{\alpha}(\widehat{k}_i,\theta)-\overline{\alpha}_i(\theta)\right\|^2=O_p\left(\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{i=1}^{N}\overline{v}(\widehat{k}_i,\theta)'\left(\widehat{\alpha}(\widehat{k}_i,\theta)-a\left(\widehat{k}_i,\theta\right)\right)\right|\right)+O_p(\delta),$$

where $\overline{v}(k,\theta)$ denotes the mean of $v_i(\overline{\alpha}_i(\theta),\theta)$ in group $\widehat{k}_i=k$.

We are first going to show that, for all $\theta\in\Theta$ (pointwise):

$$A\equiv\frac{1}{NT}\sum_{i=1}^{N}\left\|\widehat{\alpha}(\widehat{k}_i,\theta)-\overline{\alpha}_i(\theta)\right\|^2=O_p\left(\delta\right). \tag{A46}$$

To see this, note that by Cauchy Schwarz and triangular inequalities:

$$A\leq O_p\left(\left(\frac{1}{N}\sum_{i=1}^{N}T\left\|\overline{v}(\widehat{k}_i,\theta)\right\|^2\right)^{\frac{1}{2}}\left(\sqrt{A}+\sqrt{O_p(\delta)}\right)\right)+O_p(\delta),$$

from which we get:

$$A=O_p\left(\frac{1}{N}\sum_{i=1}^{N}T\left\|\overline{v}(\widehat{k}_i,\theta)\right\|^2\right)+O_p(\delta).$$

Now, since $Tv_i(\overline{\alpha}_i(\theta),\theta)$ satisfies Definition 1 we have, as in the proof of Lemma A1 (see (A45)):

$$\frac{1}{N}\sum_{i=1}^{N}T\left\|\overline{v}(\widehat{k}_i,\theta)\right\|^2=O_p(\delta).$$

This shows (A46).

Proceeding in a similar way as in the proof of Theorem 2 then gives:

$$\sup_{\theta\in\Theta}\frac{1}{NT}\sum_{i=1}^{N}\left\|\widehat{\alpha}(\widehat{k}_i,\theta)-\overline{\alpha}_i(\theta)\right\|^2=o_p(1).$$

Uniform convergence of the objective function then comes from:

$$\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}(\widehat{k}_i,\theta),\theta\right)-\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\overline{\alpha}_i(\theta),\theta\right)\right|\leq\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{i=1}^{N}v_i\left(\overline{\alpha}_i(\theta),\theta\right)'\left(\widehat{\alpha}(\widehat{k}_i,\theta)-\overline{\alpha}_i(\theta)\right)\right|+o_p(1),$$

and using that $T\,v_i(\overline{\alpha}_i(\theta),\theta)$ satisfies Definition 1 for a common λ, so (e.g., Lemma 5.5 in Vershynin, 2010): $\sup_\theta\frac{1}{N}\sum_{i=1}^{N}T\|v_i\left(\overline{\alpha}_i(\theta),\theta\right)\|^2=O_p(1/T)$. Consistency of $\widehat{\theta}$ then follows similarly as in the proof of Theorem 2.

Rate of the score. The rest of the proof follows closely that of Theorem 2. To show that the grouped fixed-effects score is $O_p(\delta)$ it suffices to show that:

$$\frac{1}{N}\sum_{i=1}^N v_i^\theta\left(\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\right)+\mathbb{E}\left(v_i^\theta\right)\left[\mathbb{E}\left(v_i^\alpha\right)\right]^{-1}v_i=O_p\left(\delta\right). \tag{A47}$$

Following the steps of the proof of Theorem 2, and letting:

$$\widetilde{\alpha}(k,t)=\left(\sum_{i=1}^N \mathbf{1}\{\widehat{k}_i=k\}(-v_{it}^\alpha)\right)^{-1}\left(\sum_{i=1}^N \mathbf{1}\{\widehat{k}_i=k\}(-v_{it}^\alpha)\alpha_{i0}(t)\right),$$

we have, using in particular Assumption 8 (i):

$$\frac{1}{NT}\sum_{i=1}^N\left\|\widetilde{\alpha}(\widehat{k}_i)-\alpha_{i0}\right\|^2=\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T\left\|\widetilde{\alpha}(\widehat{k}_i,t)-\alpha_{i0}(t)\right\|^2=O_p(\delta).$$

So, using that $T(-v_i^\alpha)$ is bounded away from zero we have, similarly as in the proof of Theorem 2:

$$\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left(v_i^\theta\right)\left[\mathbb{E}\left(v_i^\alpha\right)\right]^{-1}v_i^\alpha\left(\widetilde{\alpha}(\widehat{k}_i)-\alpha_{i0}\right) \quad = \quad O_p(\delta).$$

Let $z_i'=\mathbb{E}\left(v_i^\theta\right)\left[\mathbb{E}\left(-v_i^\alpha\right)\right]^{-1}$. In order to bound the analog to the term A_3 in the proof of Theorem 2 we are first going to apply Lemma A2 to the following quadratic form:

$$\left\|\frac{1}{N}\sum_{i=1}^N\left(z_i'-z^*\left(\widehat{k}_i\right)'\right)v_i\right\|^2.$$

Let $\epsilon>0$. As in the proof of Theorem 2 we have, since z_i is a Lipschitz function of α_{i0}:

$$\text{Tr}\left(\frac{1}{NT}\sum_{i=1}^N\left(z_i-z^*\left(\widehat{k}_i\right)\right)\left(z_i'-z^*\left(\widehat{k}_i\right)'\right)\right)=O_p\left(\frac{1}{NT}\sum_{i=1}^N\|z_i-z^*(\widehat{k}_i)\|^2\right)=O_p(\delta).$$

Hence there exists a $c>0$ such that for N,T,K large enough:

$$\Pr\left[\text{Tr}\left(\frac{1}{NT}\sum_{i=1}^N\left(z_i-z^*\left(\widehat{k}_i\right)\right)\left(z_i'-z^*\left(\widehat{k}_i\right)'\right)\right)>c\delta\right]<\frac{\epsilon}{2}.$$

Let E denote this event, and E^c denote the complementary event (which happens with probability $\geq 1-\frac{\epsilon}{2}$).

Let $b>0$. Since $T\,v_i$ satisfies Definition 1, and using similar derivations as in the proof of Lemma

A1, we have:[52]

$$\begin{aligned}\Pr\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\left(z_i' - z^*\left(\widehat{k}_i\right)'\right)v_i\right\| > b\delta\right] &\leq \Pr(E) + \Pr\left(\left\|\frac{1}{NT}\sum_{i=1}^{N}\left(z_i' - z^*\left(\widehat{k}_i\right)'\right)Tv_i\right\|^2 > b^2\delta^2\,,\,E^c\right)\\
&\leq \Pr(E) + \Pr\left(\frac{\left\|\frac{1}{NT}\sum_{i=1}^{N}\left(z_i' - z^*\left(\widehat{k}_i\right)'\right)Tv_i\right\|^2}{\operatorname{Tr}\left(\frac{1}{NT}\sum_{i=1}^{N}\left(z_i - z^*\left(\widehat{k}_i\right)\right)\left(z_i' - z^*\left(\widehat{k}_i\right)'\right)\right)} > \frac{b^2\delta^2}{c\delta}\,,\,E^c\right)\\
&\leq \Pr(E) + K^N\max_{\{k_i\}}\Pr\left(\frac{\left\|\frac{1}{NT}\sum_{i=1}^{N}\left(z_i' - z^*\left(k_i\right)'\right)Tv_i\right\|^2}{\operatorname{Tr}\left(\frac{1}{NT}\sum_{i=1}^{N}\left(z_i - z^*\left(k_i\right)\right)\left(z_i' - z^*\left(k_i\right)'\right)\right)} > \frac{b^2\delta}{c}\right)\\
&\leq \frac{\epsilon}{2} + K^N \times C\exp\left(-b^2\frac{\delta NT}{6c\lambda}\right),\end{aligned}$$

where we have used the union bound in the next-to-last inequality and Lemma A2 in the last inequality, and $C>0$ is a constant. Taking $b > \sqrt{6c\lambda}$ we thus obtain that $\Pr\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\left(z_i' - z^*\left(\widehat{k}_i\right)'\right)v_i\right\| > b\delta\right] < \epsilon$ for N, T, K large enough. Hence we obtain that:

$$\frac{1}{N}\sum_{i=1}^{N}\left(z_i' - z^*\left(\widehat{k}_i\right)'\right)v_i = O_p\left(\delta\right).$$

Turning to the second part in the analog to A_3, we apply Lemma A2 to the quadratic form $\frac{1}{N}\sum_{i=1}^{N}T\|\overline{v}(\widehat{k}_i)\|^2$, we obtain that:

$$\frac{1}{N}\sum_{i=1}^{N}T\left\|\overline{v}(\widehat{k}_i)\right\|^2 = O_p\left(\frac{\ln K}{T}\right) + O_p\left(\frac{K}{N}\right) = O_p(\delta).$$

Proceeding as in the proof of Theorem 2 then implies that $A_3 = O_p(\delta)$, hence that $A = O_p(\delta)$.

From similar derivations, in particular relying for the analog to B_3 on the fact that $T\tau_i$ satisfies Definition 1 where:

$$\tau_i = v_i^\theta - \mathbb{E}\left(v_i^\theta\right)\left[\mathbb{E}\left(v_i^\alpha\right)\right]^{-1}\left(v_i^\alpha\right),$$

it then follows that $B = O_p(\delta)$.

Consistency of the Hessian. This is essentially the same proof as for Theorem 2, except for the argument that leads to bounding:

$$\frac{\partial\widetilde{\alpha}(k,\theta_0)}{\partial\theta'} - \frac{\partial\widetilde{\alpha}^*(k,\theta_0)}{\partial\theta'} = \left(\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j = k\}(-v_j^\alpha)\right)^{-1}\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j = k\}\tau_j',$$

where here we apply Lemma A2 to $T\tau_i$, which satisfies Definition 1 by Assumption 8 (iii), and we use again that $T(-v_i^\alpha)$ is bounded away from zero.

[52] We are implicitly assuming that $\operatorname{Tr}\left(\frac{1}{NT}\sum_{i=1}^{N}\left(z_i - z^*\left(\widehat{k}_i\right)\right)\left(z_i' - z^*\left(\widehat{k}_i\right)'\right)\right) \neq 0$. The event that the trace is zero can easily be taken care of.

Convergence rate of time-varying individual effects. Finally, the rate of convergence of $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\widehat{\alpha}(\widehat{k}_i,t)-\alpha_{i0}(t)\|^2$ then comes from expanding:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\widehat{\alpha}(\widehat{k}_i,\widehat{\theta},t)-\alpha_{i0}(t)\|^2=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\widehat{\alpha}(\widehat{k}_i,\theta_0,t)-\alpha_{i0}(t)+\frac{\partial\widehat{\alpha}(\widehat{k}_i,\widetilde{\theta},t)}{\partial\theta'}\left(\widehat{\theta}-\theta_0\right)\right\|^2$$

$$=O_p\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\widehat{\alpha}(\widehat{k}_i,\theta_0,t)-\alpha_{i0}(t)\right\|^2\right)+O_p\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\frac{\partial\widehat{\alpha}(\widehat{k}_i,\widetilde{\theta},t)}{\partial\theta'}\right\|^2\left\|\widehat{\theta}-\theta_0\right\|^2\right)=O_p(\delta),$$

where we have used (A46), (15), and the fact that by the expression of $\frac{\partial\widehat{\alpha}(k,\theta,t)}{\partial\theta'}$ (which is analogous to (A21)), and by Assumption 8 (ii), we have:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\frac{\partial\widehat{\alpha}(\widehat{k}_i,\widetilde{\theta},t)}{\partial\theta'}\right\|^2=O_p(1).$$

This ends the proof of Theorem 3.