Discretizing Unobserved Heterogeneity*

Stéphane Bonhomme† Thibaut Lamadon‡ Elena Manresa§

May 2021

Abstract

We study discrete panel data methods where unobserved heterogeneity is revealed in a first step, in environments where population heterogeneity is not discrete. We focus on *two-step grouped fixed-effects* (GFE) estimators, where individuals are first classified into groups using *kmeans* clustering, and the model is then estimated allowing for group-specific heterogeneity. Our framework relies on two key properties: heterogeneity is a function — possibly nonlinear and time-varying — of a low-dimensional continuous latent type, and informative moments are available for classification. We illustrate the method in a model of wages and labor market participation, and in a probit model with time-varying heterogeneity. We derive asymptotic expansions of two-step GFE estimators as the number of groups grows with the two dimensions of the panel. We propose a data-driven rule for the number of groups, and discuss bias reduction and inference.

JEL codes: C23, C38.

Keywords: Unobserved heterogeneity, panel data, kmeans clustering, dimension reduction.

*We thank four anonymous referees, Manuel Arellano, Neele Balke, Jesus Carro, Gary Chamberlain, Tim Christensen, Liran Einav, Alfred Galichon, Chris Hansen, Joe Hotz, Grégory Jolivet, Arthur Lewbel, Anna Mikusheva, Roger Moon, Whitney Newey, Juan Pantano, Philippe Rigollet, Anna Simoni, Martin Weidner, and seminar audiences at various places for comments. The authors acknowledge support from the NSF grant number SES-1658920. Replication codes are available at: https://github.com/tlamadon/BonhommeLamadonManresa2021

†University of Chicago, sbonhomme@uchicago.edu

‡University of Chicago, lamadon@uchicago.edu

§New York University, elena.manresa@nyu.edu

1 Introduction

In both reduced-form and structural work in economics, it is common to model unobserved heterogeneity as a small number of discrete types. Various estimation strategies are available, including discrete-type random-effects (as in Keane and Wolpin, 1997, and many other applications) and grouped fixed-effects (as recently studied by Hahn and Moon, 2010, and Bonhomme and Manresa, 2015). These methods require the researcher to jointly estimate individual heterogeneity and model parameters.[1] In addition, little is known about their properties when individual heterogeneity is not discrete in the population. In this paper, we study two-step discrete estimators for panel data, and provide conditions for their validity when heterogeneity is continuous.

We focus on *two-step grouped fixed-effects* (GFE) estimators. In a first step, we classify individuals based on a set of individual-specific moments, using the *kmeans* clustering algorithm. The aim of the kmeans classification is to group together individuals whose latent types are most similar.[2] In a second step, we estimate the model by allowing for group-specific heterogeneity. This second step is similar to fixed-effects (FE) estimation, albeit it involves a smaller number of parameters that are group-specific instead of individual-specific. We analyze the properties of these two-step estimators in panel data models where heterogeneity is continuous. Hence, in contrast with existing theoretical justifications for discrete-type methods, here we use discrete heterogeneity as a dimension reduction device rather than as a substantive assumption about population unobservables.

Our approach is targeted to environments with two key properties. *First*, unobserved heterogeneity is a function of a low-dimensional latent variable. We do not restrict this latent *type* to be discrete. In many economic models, agents' heterogeneity in preferences or technology is driven by a low-dimensional type, which enters the model nonlinearly and may affect multiple outcomes. As an example, we study a model of participation in the labor market where the worker's

[1] Also related, nonparametric maximum likelihood methods (e.g., Heckman and Singer, 1984) rely on joint estimation of the distribution of heterogeneity and the parameters.

[2] Buchinsky *et al.* (2005) also propose to group individuals in a first step using kmeans.

utility is a function of her productivity type, which in turn determines her wage. GFE provides a tool to exploit such nonlinear factor structures.

Second, the first-step moments satisfy an injectivity condition, which requires any two individuals with the same population moments to have the same type. The choice of moments is important to ensure good performance. In examples, we show how suitable moments arise naturally. In models with exogenous covariates, we propose and analyze the use of conditional moments to recover latent types.

Our setup also covers models where heterogeneity varies over time. Unlike additive FE methods and interactive FE methods based on linear factor structures (Bai, 2009), GFE does not require heterogeneity to take an additive or interactive form. As an illustration, we compare GFE and FE estimators in a probit model where heterogeneity is a nonlinear function of a time-invariant factor loading and a time-specific factor.

Our main results are large-N, T asymptotic expansions of two-step GFE estimators under time-invariant and time-varying continuous heterogeneity. In both settings, GFE is consistent as the number of groups grows with the sample size, under conditions that we provide. We find that, when the population heterogeneity is not discrete, estimating group membership induces an incidental parameter bias, similarly to FE methods. Moreover, since discreteness is an approximation in our setting, GFE is affected by approximation error.[3] We propose a simple data-driven rule for the number of groups that controls the approximation error, and discuss how to reduce incidental parameter bias for inference.

The outline of the paper is as follows. We introduce the setup and two-step GFE estimators in Section 2, study their asymptotic properties in Section 3, and outline several extensions in Section 4. The main proofs may be found in the appendix, and the supplemental material contains additional results. Codes to implement the method are available online.

[3]In a network context, Gao *et al.* (2015) provide results for stochastic blockmodels under continuous heterogeneity.

2 Two-step grouped fixed-effects (GFE)

We consider a panel data setup, where we denote outcome variables and exogenous covariates as $Y_i=(Y_{i1}', ..., Y_{iT}')'$ and $X_i=(X_{i1}', ..., X_{iT}')'$, respectively, for $i=1, ..., N$. In our theory we cover two models. In the first one, unobserved heterogeneity is time-invariant. In this case, the conditional log-density of Y_i given X_i is given by:[4]

$$\ln f_i(\alpha_{i0}, \theta_0) = \sum_{t=1}^{T} \ln f(Y_{it} \,|\, Y_{i,t-1}, X_{it}, \alpha_{i0}, \theta_0), \tag{1}$$

and the log-density of exogenous covariates X_i takes the form:

$$\ln g_i(\mu_{i0}) = \sum_{t=1}^{T} \ln g(X_{it} \,|\, X_{i,t-1}, \mu_{i0}),$$

where θ_0 is a vector of common parameters, and α_{i0} and μ_{i0} are individual-specific parameters. We leave the form of g unrestricted, and in estimation we will use a conditional likelihood approach based on f_i alone. In other words, in applications the researcher only needs to specify the parametric form of $f_i(\alpha_{i0}, \theta_0)$ in (1). However, the heterogeneity μ_{i0} in covariates plays an important role in our theory.

In the second model, unobserved heterogeneity varies over time. Such variation in unobservables over calendar time (e.g., business cycle), age (e.g., life cycle), counties, or markets, is of interest in many applications. In the time-varying case, log-densities take the form:

$$\ln f_i(\alpha_{i0}, \theta_0) = \sum_{t=1}^{T} \ln f(Y_{it} \,|\, Y_{i,t-1}, X_{it}, \alpha_{it0}, \theta_0),$$
$$\ln g_i(\mu_{i0}) = \sum_{t=1}^{T} \ln g(X_{it} \,|\, X_{i,t-1}, \mu_{it0}),$$

where $\alpha_{i0} = (\alpha_{i10}', ..., \alpha_{iT0}')'$ and $\mu_{i0} = (\mu_{i10}', ..., \mu_{iT0}')'$. In both models we are interested in estimating θ_0, as well as average effects depending on $\alpha_{10}, ..., \alpha_{N0}$.

[4] In models with first-order dependence, we assume that Y_{i0} is observed and we condition on it. Higher-order dependence can be accommodated similarly. In dynamic settings, Y_{it} may contain sequentially exogenous covariates in addition to outcome variables.

2.1 Main assumptions

GFE relies on two key assumptions that we now present. We defer the presentation of regularity conditions until Section 3. *First*, we assume that unobserved heterogeneity is a function of a low-dimensional vector ξ_{i0}.

Assumption 1. *(heterogeneity)*
(a) Time-invariant heterogeneity: There exist ξ_{i0} of fixed dimension d, and two Lipschitz-continuous functions α and μ, such that $\alpha_{i0} = \alpha(\xi_{i0})$ and $\mu_{i0} = \mu(\xi_{i0})$.
(b) Time-varying heterogeneity: There exist ξ_{i0} of fixed dimension d, λ_{t0} of dimension d_λ, and two functions α and μ that are Lipschitz-continuous in their first argument, such that $\alpha_{it0} = \alpha(\xi_{i0}, \lambda_{t0})$ and $\mu_{it0} = \mu(\xi_{i0}, \lambda_{t0})$.

We will refer to ξ_{i0} as an individual *type*, and to d as the *dimension* of heterogeneity. The researcher does not need to know d, α, or μ in applications. In models with time-varying unobserved heterogeneity, Assumption 1 requires unobservables to follow a factor structure. The link between α_{it0}, ξ_{i0} and λ_{t0} may be nonlinear, the linear structure $\alpha_{it0} = \xi_{i0}'\lambda_{t0}$ (Bai, 2009) being covered as a special case. Moreover, the dimension of λ_{t0} is unrestricted. Our theory will show that the performance of two-step GFE crucially relies on ξ_{i0} being low-dimensional, a leading case being $d = 1$. We provide examples in the next subsection.

Second, we rely on individual-specific moment vectors h_i that are informative about the types ξ_{i0}. The moments h_i can be functions of Y_i, X_i, or additional data on the individual, and their dimension is kept fixed as the sample size grows. Formally, we now state our second main assumption, where $\|\cdot\|$ denotes an Euclidean norm.

Assumption 2. *(injective moments)*
There exist vectors h_i of fixed dimension, and a Lipschitz-continuous function φ, such that $\operatorname{plim}_{T\to\infty} h_i = \varphi(\xi_{i0})$, and $\frac{1}{N}\sum_{i=1}^{N} \|h_i - \varphi(\xi_{i0})\|^2 = O_p(1/T)$ as N, T tend to infinity. Moreover, there exists a Lipschitz-continuous function ψ such that $\xi_{i0} = \psi(\varphi(\xi_{i0}))$.

Assumption 2 requires the individual moment vector h_i to be informative about

ξ_{i0}, in the sense that, for large T, ξ_{i0} can be uniquely recovered from h_i. Neither φ nor ψ (which may depend on θ_0) need to be known to the econometrician. Intuitively, injectivity guarantees that one can separate the types of two individuals ξ_{i0} and $\xi_{i'0}$ by comparing their moments h_i and $h_{i'}$. Formally, if h_i and $h_{i'}$ have the same large-T limit, then $\xi_{i0} = \xi_{i'0}$. For example, an average $h_i = \frac{1}{T}\sum_{t=1}^{T} h(Y_{it}, X_{it})$ will, under Assumption 1 and suitable regularity conditions, converge as T tends to infinity to a function $\varphi(\xi_{i0})$ of the type ξ_{i0}.

The convergence rate in Assumption 2 requires appropriate conditions on the serial dependence of Y_{it} and X_{it}. In models with time-varying heterogeneity, φ will also depend on the λ_{t0} process. In such models, Assumption 2 requires the moments to be informative about ξ_{i0}, and not λ_{t0}. Injectivity is a key requirement for consistency of two-step GFE estimators. More generally, the choice of moments h_i is important for finite-sample performance.

2.2 Examples

To illustrate the framework we now describe two examples, for which we will provide illustrative simulations in Subsection 2.4. First, consider a dynamic model of wages W_{it}^* and labor force participation Y_{it}:

$$\begin{cases} Y_{it} &= \mathbf{1}\left\{u(\alpha_{i0}) \geq c(Y_{i,t-1};\theta_0) + U_{it}\right\}, \\ W_{it}^* &= \alpha_{i0} + V_{it}, \\ W_{it} &= Y_{it} W_{it}^*, \end{cases} \tag{2}$$

where the wage W_{it}^* is only observed when i works, U_{it} are i.i.d. standard normal, independent of the past Y_{it}'s and α_{i0}, and V_{it} are i.i.d. independent of all U_{it}'s, Y_{i0}, and α_{i0}. Here the same scalar expected payoff $\alpha_{i0} = \xi_{i0}$, unobserved to the econometrician, drives the wage and the decision to work. Individuals have common preferences denoted by the utility function u, the cost function c is state-dependent, and both u and c are unknown to the econometrician.

In this setting, GFE provides a natural approach to exploit the functional link between α_{i0} and $u(\alpha_{i0})$, and to learn about the type α_{i0} using both wages and participation. For instance, when $h_i = (\overline{W}_i, \overline{Y}_i)'$, where $\overline{Z}_i = \frac{1}{T}\sum_{t=1}^{T} Z_{it}$ denotes

the individual mean of Z_{it}, injectivity is satisfied under mild conditions, provided $\overline{W}_i = \alpha_{i0}\overline{Y}_i + o_p(1)$ and $\text{plim}_{T\to\infty} \overline{Y}_i > 0$.

Fixed-effects (FE) is a possible approach to estimate θ_0 in (2). However, a conventional FE estimator would treat α_{i0} and $u_{i0} = u(\alpha_{i0})$ as unrelated parameters, so the FE estimate of θ_0 would be solely based on the binary participation decisions. Another strategy would be to rely on discrete-type random-effects methods, which are typically based on joint estimation. In contrast, we implement GFE in two steps with no need for iterative estimation, and we justify the estimator in environments where heterogeneity is not restricted to be discrete.

As a second example, consider the following probit model with time-varying heterogeneity:

$$\begin{cases} Y_{it} &= \mathbf{1}\left\{X_{it}'\theta_0 + \alpha_{it0} + U_{it} \geq 0\right\}, \\ X_{it} &= \mu_{it0} + V_{it}, \end{cases} \tag{3}$$

where U_{it} are i.i.d. standard normal, independent of all V_{it}'s, α_{it0}'s, and μ_{it0}'s, and V_{it} are i.i.d. independent of all α_{it0}'s and μ_{it0}'s. Under Assumption 1, α_{it0} and μ_{it0} depend on a low-dimensional vector ξ_{i0} of factor loadings, so $\alpha_{it0} = \alpha(\xi_{i0}, \lambda_{t0})$ and $\mu_{it0} = \mu(\xi_{i0}, \lambda_{t0})$. Here d is the dimension of the type ξ_{i0} governing both α_{it0} and μ_{it0}.

To motivate why, in static models with covariates such as (3), α_{it0} and μ_{it0} may depend on a common low-dimensional type ξ_{i0}, suppose that, in every period, agent i chooses X_{it} based on expected utility or profit maximization. She observes ξ_{i0} and λ_{t0} — which enter outcomes through α_{it0} — and takes her decision before the i.i.d. shock U_{it} is realized. In such a case, X_{it} will be a function of ξ_{i0} and λ_{t0}, as well as idiosyncratic factors V_{it} in the agent's information set. Here we assume that the agent's information set, and primitives such as preferences or costs, do not include other i-specific elements beyond ξ_{i0}.[5]

When $\alpha(\cdot,\cdot)$ is additive or multiplicative in its arguments, model (3) can be estimated using two-way FE (Fernández-Val and Weidner, 2016) or interactive

[5]This example is reminiscent of Mundlak's (1961) classic analysis of farm production functions, where soil quality ξ_{i0} is observed to the farmer but latent to the analyst.

FE (Bai, 2009, Chen *et al.*, 2020), respectively. However, when $\alpha(\cdot,\cdot)$ is unknown, these fixed-effects estimators are inconsistent in general. In contrast, GFE will remain consistent when unobservables are unknown nonlinear functions of factor loadings ξ_{i0} and factors λ_{t0}, and injectivity holds. Taking $h_i = (\overline{Y}_i, \overline{X}_i')'$ as moments in model (3), injectivity is satisfied when types have monotone effects on the heterogeneity components.[6] More generally, in Assumption 2 we require that the latent type ξ_{i0} can be asymptotically recovered from a moment vector whose dimension is not growing with the sample size.

2.3 Estimator

Two-step GFE consists of a *classification* step and an *estimation* step.

First step: classification. We rely on the individual-specific moments h_i to learn about the individual types ξ_{i0}. Specifically, we partition individuals into K groups, corresponding to group indicators $\widehat{k}_i \in \{1, ..., K\}$, by computing:

$$\left(\widehat{h}(1), ..., \widehat{h}(K), \widehat{k}_1, ..., \widehat{k}_N\right) = \underset{\left(\widetilde{h}(1),...,\widetilde{h}(K),k_1,...,k_N\right)}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| h_i - \widetilde{h}(k_i) \right\|^2, \qquad (4)$$

where $\{k_i\}$ are partitions of $\{1, ..., N\}$ into K groups, and $\widetilde{h}(k)$ is a vector. Note that $\widehat{h}(k)$ is simply the mean of h_i in group $\widehat{k}_i = k$.

In the *kmeans* optimization problem (4), the minimum is taken with respect to all possible partitions $\{k_i\}$. Fast and stable optimization methods such as Lloyd's algorithm are available, although computing a global minimum may be challenging; see Bonhomme and Manresa (2015) for references. Following the literature, we will focus on the asymptotic properties of the global minimum and abstract from optimization error. Lastly, note that the quadratic loss function in (4) can accommodate weights on different components of h_i, although here for simplicity we present the unweighted case.

[6] To see this, consider the case where α_{it0} is the only component of heterogeneity (i.e., $\mu_{it0} = 0$ in (3)), and take $h_i = \overline{Y}_i$. Letting G denote the cdf of $-(V_{it}'\theta_0 + U_{it})$, injectivity will hold when $\alpha(\cdot,\cdot)$ is strictly increasing in its first argument and G is strictly increasing, since then $\varphi(\xi) = \operatorname{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} G(\alpha(\xi, \lambda_{t0}))$ is strictly increasing.

Second step: estimation. We maximize the log-likelihood function with respect to common parameters θ and group-specific effects α, where the groups are given by the $\widehat{k}_i$ estimated in the first step. We define the two-step GFE estimator as:

$$\left(\widehat{\theta}, \widehat{\alpha}(1), ..., \widehat{\alpha}(K)\right) = \underset{(\theta,\alpha(1),...,\alpha(K))}{\operatorname{argmax}} \sum_{i=1}^{N} \ln f_i \left(\alpha\left(\widehat{k}_i\right), \theta\right). \tag{5}$$

Note that, in contrast to fixed-effects (FE) maximum likelihood, this second step involves a maximization with respect to K group-specific parameters instead of N individual-specific ones. In models with time-varying heterogeneity, $\alpha(k)$ will simply be a vector $(\alpha_1(k)', ..., \alpha_T(k)')'$.

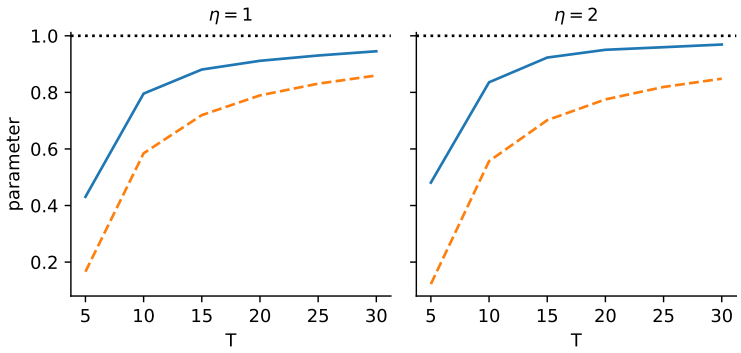
Choice of K. Two-step GFE estimation requires setting a number of groups K. We propose a simple data-driven selection rule based on the first step. The convergence rate of the kmeans estimator (and the rate of the GFE estimator) will be governed by two quantities: the kmeans objective function $\widehat{Q}(K) = \frac{1}{N}\sum_{i=1}^{N} \|h_i - \widehat{h}(\widehat{k}_i)\|^2$, which decreases as K gets larger and the group approximation becomes more accurate, and the variability $V_h = \mathbb{E}[\|h_i - \varphi(\xi_{i0})\|^2]$ of the moment h_i, which does not depend on K. Given our goal to approximate the heterogeneity while limiting the number of groups, we take the smallest K that guarantees that $\widehat{Q}(K)$ is of the same or lower order as V_h. That is, letting $\widehat{V}_h = V_h + o_p(1/T)$, we suggest setting:

$$\widehat{K} = \min_{K \geq 1} \left\{ K \,:\, \widehat{Q}(K) \leq \gamma \widehat{V}_h \right\}, \tag{6}$$

where $\gamma \in (0, 1]$ is a user-specified parameter.[7] In the simulations in the next subsection we will set $\gamma = 1$, although smaller γ values corresponding to larger K's will also be supported by our theory.

[7] When $h_i = \frac{1}{T}\sum_{t=1}^{T} h(Y_{it}, X_{it})$ and observations are independent over time, one may take $\widehat{V}_h = \frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T} \|h(Y_{it}, X_{it}) - h_i\|^2$. With dependent data, one can use trimming or the bootstrap to estimate V_h (Hahn and Kuersteiner, 2011, Arellano and Hahn, 2007).

Figure 1: Model (2) of wages and participation



Notes: Means of $c(0;\widehat{\theta}) - c(1;\widehat{\theta})$ *over 1000 simulations. GFE is indicated in solid, FE is in dashed, and the truth* $c(0;\theta_0) - c(1;\theta_0) = 1$ *is in dotted.* $N = 1000$*, and* T *is indicated on the x-axis.* η *is the risk aversion parameter in* $u(\cdot)$*. See the supplemental material for details.*

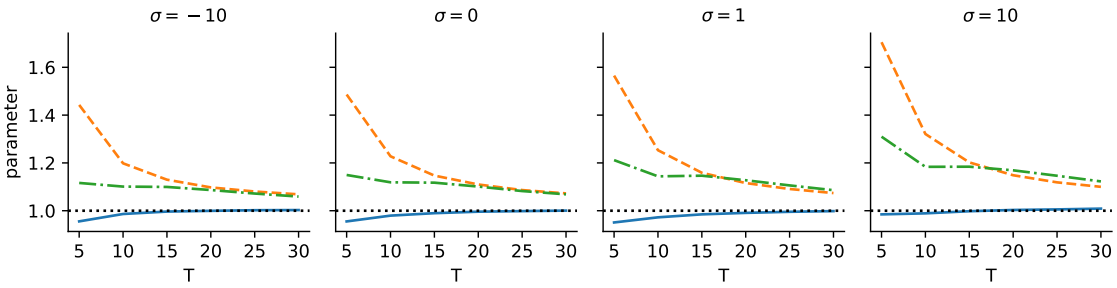
2.4 Illustrative simulations

To illustrate the performance of GFE in models where heterogeneity follows a nonlinear factor structure, we present the results of a small-scale simulation study based on our two examples (2) and (3). In both cases, we assume that the type ξ_{i0} governing heterogeneity is scalar. We compare the bias of GFE to that of FE and interactive FE estimators. In the supplemental material, we provide details on the simulations and report additional results.[8]

In Figure 1, we compare GFE and FE in model (2), using a CRRA functional form: $u(\alpha) = \frac{e^{\alpha(1-\eta)}-1}{1-\eta}$, with a risk aversion parameter $\eta \in \{1, 2\}$. We focus on the difference in costs $c(0;\theta_0) - c(1;\theta_0)$, which measures the degree of state dependence in participation decisions. We take $h_i = (\overline{W}_i, \overline{Y}_i)'$ as moments for GFE, and report average parameter estimates over 1000 simulations. We set N=1000 and show results for T between 5 and 30. The data generating process implies a participation rate of approximately 85% (respectively, 20%) for those who participated (resp., did not participate) last period, and a mean gap between log-wages of participants and log-potential wages equal to one third of the standard deviation of log-potential wages. We find that FE is more biased than GFE for both values of risk aversion. This is consistent with wages and participation

[8] In Bonhomme *et al.* (2017) we report simulations calibrated to two empirical settings.

Figure 2: Probit model (3) with time-varying heterogeneity



Notes: Means of $\widehat{\theta}$ over 1000 simulations. GFE is indicated in solid, FE is in dashed, interactive FE is in dash-dotted, and the truth $\theta_0 = 1$ is in dotted. $N = 1000$, and T is indicated on the x-axis. σ is the substitution parameter in $\alpha(\cdot,\cdot)$. See the supplemental material for details.

providing informative moments about the latent type in this setting.

In Figure 2 we compare GFE, FE, and interactive FE in model (3) with X_{it} scalar, using a CES specification: $\alpha_{it0} = (a\xi_{i0}^{\sigma} + (1-a)\lambda_{t0}^{\sigma})^{\frac{1}{\sigma}}$, for $\sigma \in \{-10, 0, 1, 10\}$ and a=0.5, and $\mu_{it0} = \alpha_{it0}$. The factors λ_{t0} and the individual loadings ξ_{i0} enter heterogeneity in a nonlinear way. We show estimates of θ_0 for various estimators: GFE, FE with additive individual and time effects, and interactive FE with a single multiplicative factor. We use $(\overline{Y}_i, \overline{X}_i)'$ as moments for GFE. Note that both $\overline{Y}_i$ and $\overline{X}_i$ are informative about ξ_{i0} in this data generating process. We report parameter averages over 1000 simulations, for N=1000. We find that, while GFE, FE, and interactive FE are all biased, the bias of GFE is smaller across all σ values.[9]

3 Asymptotic properties

In this section we provide asymptotic expansions for two-step GFE estimators. Our first result is a rate of convergence for kmeans. Let us define the *approximation*

[9]Large-N,T theory implies that additive and interactive FE are consistent when $\sigma = 1$ and $\sigma = 0$, respectively. Figure 2 shows that, despite being large-N,T consistent in these specifications, in our simulations, additive and interactive FE have larger biases than GFE for the N and T values we consider.

error one would make if one were to discretize the latent types ξ_{i0} directly, as:

$$B_\xi(K) = \min_{\left(\widetilde{\xi}(1),...,\widetilde{\xi}(K),k_1,...,k_N\right)} \frac{1}{N}\sum_{i=1}^{N}\left\|\xi_{i0} - \widetilde{\xi}(k_i)\right\|^2, \tag{7}$$

where, similarly to (4), the minimum is taken with respect to all partitions $\{k_i\}$ and vectors $\widetilde{\xi}(k)$. In the following result we let $T = T_N$ and $K = K_N$ tend to infinity jointly with N.

Lemma 1. *Let Assumption 2 hold. Let $\widehat{h}(1), ..., \widehat{h}(K)$ and $\widehat{k}_1, ..., \widehat{k}_N$ given by (4). Then, as N, T, K tend to infinity we have:*

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

The bound in Lemma 1 has two terms: an $O_p(1/T)$ term that depends on the number of periods used to construct the moments h_i, and an $O_p\left(B_\xi(K)\right)$ term that reflects the presence of an approximation error. The rate at which $B_\xi(K)$ tends to zero depends on the dimension of ξ_{i0}. Graf and Luschgy (2002, Theorem 5.3) provide explicit characterizations in the case where ξ_{i0} has compact support.[10] For example, the following lemma implies that $B_\xi(K) = O_p(K^{-2})$ when ξ_{i0} is one-dimensional, and $B_\xi(K) = O_p(K^{-1})$ when ξ_{i0} is two-dimensional.

Lemma 2. *(Graf and Luschgy, 2002) Let ξ_{i0} be random vectors with compact support in $\mathbb{R}^d$. Then, as N, K tend to infinity we have $B_\xi(K) = O_p(K^{-\frac{2}{d}})$.*

We now use these results to study the properties of GFE in models with time-invariant and time-varying heterogeneity, in turn. We use the shorthand notation $\mathbb{E}_Z(W)$ and $\mathbb{E}_{Z=z}(W)$ for the conditional expectations of W given Z and $Z = z$, respectively. In the time-varying case, we denote as λ_0 the process of λ_{t0}'s, and as $\mathbb{E}_{\lambda_0=\lambda}(W)$ the conditional expectation of W given $\lambda_0 = \lambda$. We use a similar notation for variances. Finally, when M is a matrix $\|M\|$ denotes its spectral norm, and we use $M > 0$ to denote that M is positive definite.

[10] See Graf and Luschgy (2002, p. 875) for a discussion of the compact support assumption.

3.1 Time-invariant heterogeneity

To state our first main theorem, where heterogeneity is time-invariant, we make the following assumptions, where $\ell_{it}(\alpha_i,\theta) = \ln f(Y_{it} \,|\, Y_{i,t-1}, X_{it}, \alpha_i, \theta)$, $\ell_i(\alpha_i,\theta) = \frac{1}{T}\sum_{t=1}^{T} \ell_{it}(\alpha_i,\theta)$, and $\overline{\alpha}(\theta,\xi) = \text{argmax}_\alpha \, \mathbb{E}_{\xi_{i0}=\xi}(\ell_i(\alpha,\theta))$ for all θ, ξ.

Assumption 3. *(regularity, time-invariant heterogeneity)*

(i) $(Y_i', X_i', \xi_{i0}', h_i')'$ *are i.i.d.;* $(Y_{it}', X_{it}')'$ *are stationary for all* i*;* $\ell_{it}(\alpha,\theta)$ *is three times differentiable in* (α,θ) *for all* i,t*;*[11] *and the parameter space* Θ *for* θ_0 *is compact, the supports of* ξ_{i0} *and* α_{i0} *are compact, and* θ_0 *belongs to the interior of* Θ*.*

(ii) N,T *tend jointly to infinity;* $\sup_{\xi,\alpha,\theta} |\mathbb{E}_{\xi_{i0}=\xi}(\ell_{it}(\alpha,\theta))| = O(1)$*, and similarly for the first three derivatives of* ℓ_{it}*;* $\inf_{\xi,\alpha,\theta} \mathbb{E}_{\xi_{i0}=\xi}(-\frac{\partial^2 \ell_{it}(\alpha,\theta)}{\partial\alpha\partial\alpha'}) > 0$*; and* $\max_i \sup_{\alpha,\theta} \left|\ell_i(\alpha,\theta) - \mathbb{E}_{\xi_{i0}}\left(\ell_i(\alpha,\theta)\right)\right| = o_p(1)$*, and similarly for the first three derivatives of* ℓ_i*.*

(iii) $\inf_{\xi,\theta} \mathbb{E}_{\xi_{i0}=\xi}(-\frac{\partial^2 \ell_{it}(\overline{\alpha}(\theta,\xi),\theta)}{\partial\alpha\partial\alpha'}) > 0$*;* $\mathbb{E}[\frac{1}{T}\sum_{t=1}^{T} \ell_{it}(\overline{\alpha}(\theta,\xi_{i0}),\theta)]$ *has a unique maximum at* θ_0 *on* Θ*, and its matrix of second derivatives is* $-H < 0$*; and* $\sup_\theta \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \|\frac{\partial^2 \ell_{it}(\overline{\alpha}(\theta,\xi_{i0}),\theta)}{\partial\theta\partial\alpha'}\|^2 = O_p(1)$*.*

(iv) $\sup_{\widetilde{\xi},\alpha} \|\frac{\partial}{\partial\xi'}\big|_{\xi=\widetilde{\xi}} \mathbb{E}_{\xi_{i0}=\xi}(\text{vec}\, \frac{\partial^2 \ell_{it}(\alpha,\theta_0)}{\partial\theta\partial\alpha'})\|$*,* $\sup_{\widetilde{\xi},\alpha} \|\frac{\partial}{\partial\xi'}\big|_{\xi=\widetilde{\xi}} \mathbb{E}_{\xi_{i0}=\xi}(\text{vec}\, \frac{\partial^2 \ell_{it}(\alpha,\theta_0)}{\partial\alpha\partial\alpha'})\|$*, and* $\sup_{\widetilde{\xi},\theta} \|\frac{\partial}{\partial\xi'}\big|_{\xi=\widetilde{\xi}} \mathbb{E}_{\xi_{i0}=\xi}(\frac{\partial \ell_{it}(\overline{\alpha}(\theta,\widetilde{\xi}),\theta)}{\partial\alpha})\|$ *are* $O(1)$*.*

In part (i) in Assumption 3 we treat heterogeneity as random in order to use Lemma 2, which requires ξ_{i0} to be i.i.d. draws from a distribution. However, note we do not restrict how α_{i0} and μ_{i0} depend on each other. Moreover, while our results require asymptotic stationarity of the time-series processes, the theorem could be extended to allow for nonstationary initial conditions.

In part (ii) we require strict concavity of the log-likelihood as a function of α. Concavity holds in a number of nonlinear panel data models such as probit and logit models, tobit, Poisson, or multinomial logit; see Fernández-Val and Weidner (2016) and Chen *et al.* (2020). One can show that Theorem 1 continues to hold without concavity, under an identification condition and an assumption bounding

[11] That is, $\ln f(y_{it} \,|\, y_{i,t-1}, x_{it}, \alpha, \theta)$ is three times differentiable in (α,θ), for almost all $(y_{it}, y_{it-1}, x_{it})$.

the derivatives of the empirical GFE objective function. Importantly, note that H^{-1} is the asymptotic variance of the FE estimator. As a result, H being positive definite rules out models that are not identified under FE, such as a linear model with a time-invariant covariate and a heterogeneous intercept.

In part (iii) we introduce the *target* log-likelihood $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\ell_{it}(\overline{\alpha}(\theta,\xi_{i0}),\theta)$ (Arellano and Hahn, 2007), which we will show approximates the GFE log-likelihood in large samples under our assumptions (note that $\overline{\alpha}(\theta_0,\xi_{i0})=\alpha_{i0}$). In part (iv) we require some moments to be bounded asymptotically.

We now state our first main result, where we denote, evaluating all quantities at true values (θ_0,α_{i0}) and omitting the dependence from the notation:

$$s_i = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\partial\ell_{it}}{\partial\theta}+\mathbb{E}_{\xi_{i0}}\left(\frac{\partial^2\ell_{it}}{\partial\theta\partial\alpha'}\right)\left[\mathbb{E}_{\xi_{i0}}\left(-\frac{\partial^2\ell_{it}}{\partial\alpha\partial\alpha'}\right)\right]^{-1}\frac{\partial\ell_{it}}{\partial\alpha}\right), \tag{8}$$

$$\begin{aligned} H = \operatorname*{plim}_{N,T\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Bigg(&\mathbb{E}_{\xi_{i0}}\left(-\frac{\partial^2\ell_{it}}{\partial\theta\partial\theta'}\right)\\ &-\mathbb{E}_{\xi_{i0}}\left(\frac{\partial^2\ell_{it}}{\partial\theta\partial\alpha'}\right)\left[\mathbb{E}_{\xi_{i0}}\left(-\frac{\partial^2\ell_{it}}{\partial\alpha\partial\alpha'}\right)\right]^{-1}\mathbb{E}_{\xi_{i0}}\left(\frac{\partial^2\ell_{it}}{\partial\alpha\partial\theta'}\right)\Bigg). \end{aligned} \tag{9}$$

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

$$\widehat{\theta} \;=\; \theta_0 + H^{-1}\frac{1}{N}\sum_{i=1}^{N}s_i + O_p\left(\frac{1}{T}\right) + O_p\left(K^{-\frac{2}{d}}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right). \tag{10}$$

The first three terms in (10) also appear in large-N,T expansions of FE estimators (e.g., Hahn and Newey, 2004).[12] Similarly to FE, GFE is subject to incidental parameter $O_p(1/T)$ bias. This contrasts with the properties of GFE estimators under discrete heterogeneity (e.g., Hahn and Moon, 2010, Bonhomme and Manresa, 2015). Indeed, when heterogeneity is *not* restricted to have a small number of points of support, classification noise affects the properties of second-step estimators in general. This motivates using bias reduction techniques for inference

[12] In the supplemental material we provide a similar expansion for GFE estimators of average effects $M_0 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}m\left(X_{it},\alpha_{i0},\theta_0\right)$, which are functions of both common parameters and individual heterogeneity.

analogous to those used in FE, as we will discuss in the next section.

The $O_p(K^{-\frac{2}{d}})$ term in (10) reflects the approximation error, which depends on the number of groups. Setting $K = \widehat{K}$ according to (6) guarantees that the approximation error is $O_p(1/T)$. Formally, we have the following result.

Corollary 1. *Let the conditions in Theorem 1 hold. Let $K = \widehat{K}$ given by (6), with $\gamma = O(1)$. Then, as N, T tend to infinity we have:*

$$\widehat{\theta} \;=\; \theta_0 + H^{-1}\frac{1}{N}\sum_{i=1}^{N} s_i + O_p\left(\frac{1}{T}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right). \tag{11}$$

Under Corollary 1, the biases of FE and GFE have the same order of magnitude. However, the required value of K depends on the dimension d of individual heterogeneity. Specifically, when ξ_{i0} follows a continuous distribution of dimension d, setting K proportional to or greater than $\min(T^{\frac{d}{2}}, N)$ will ensure that the approximation error is $O_p(1/T)$. For small d (e.g., when $d = 1$) this will typically require a small number of groups (of the order of $\sqrt{T}$).

GFE can have advantages compared to FE, for two reasons. First, the two-step method can allow researchers to select moments that are particularly informative about the unobserved heterogeneity. To provide intuition, consider a setting where the number of groups is sufficiently large for the approximation error to be of smaller order compared to $1/T$, yet K/N tends to zero. We have the following.

Corollary 2. *Let the conditions in Theorem 1 hold. Let $K = \widehat{K}$ given by (6), with $\gamma = o(1)$. Suppose that K/N tends to zero, and that Assumption A1 in the appendix holds. Then the $O_p(1/T)$ term in (11) takes the explicit form $C/T + o_p(1/T)$, where:*

$$\frac{C}{T} = H^{-1}\frac{\partial}{\partial\theta}\bigg|_{\theta_0}\mathbb{E}\left[\frac{1}{2}\left\|\widehat{\alpha}_i(\theta)-\mathbb{E}_{\xi_{i0}}(\widehat{\alpha}_i(\theta))\right\|^2_{\Omega_i(\theta)} - \frac{1}{2}\left\|\widehat{\alpha}_i(\theta)-\mathbb{E}_{h_i}(\widehat{\alpha}_i(\theta))\right\|^2_{\Omega_i(\theta)}\right], \tag{12}$$

with $\widehat{\alpha}_i(\theta) = \operatorname{argmax}_\alpha \ell_i(\theta,\alpha)$, $\Omega_i(\theta) = \mathbb{E}_{\xi_{i0}}(-\frac{\partial^2 \ell_{it}(\overline{\alpha}(\theta,\xi_{i0}),\theta)}{\partial\alpha\partial\alpha'})$, and $\|V\|^2_\Omega = V'\Omega V$.

Corollary 2 shows that the first-order asymptotic bias of GFE is the difference between two terms. The bias is zero when h_i is an injective function of ξ_{i0}; i.e.,

when $\varepsilon_i = h_i - \varphi(\xi_{i0}) = 0$. More generally, the bias can be expanded in ε_i, and it is small when moments provide accurate estimates of the latent types. Moreover, the first term on the right-hand side of (12) coincides with the bias of FE (e.g., Arellano and Hahn, 2007). The form of (12) implies that the biases of FE and GFE are equal when the moments are the FE estimates $h_i = \widehat{\alpha}_i(\theta_0)$, however other moment choices can lead to smaller biases. From this perspective, GFE provides flexibility to use well-suited proxies of the latent types. As an example, our simulations of the labor force participation model (2) show that, by jointly exploiting wages and participation to construct moments that are informative about the latent type, GFE can have smaller bias than FE (and smaller mean squared error as well, as shown in the supplemental material).

A second advantage of GFE comes from the use of grouping, and from the resulting regularization. Indeed, individual FE estimates can be highly variable whenever the number of parameters per individual is large. In such cases, reducing the number of parameters through grouping can improve performance. For instance, the ability to handle multiple components of heterogeneity is central to the performance of GFE in models with time-varying unobserved heterogeneity. This is the case we focus on next.

3.2 Time-varying heterogeneity

To state our second main theorem, where heterogeneity is time-varying, we make the following assumptions, where $\ell_{it}(\alpha_{it}, \theta) = \ln f(Y_{it} \,|\, Y_{i,t-1}, X_{it}, \alpha_{it}, \theta)$, $\ell_i(\alpha_i, \theta) = \frac{1}{T}\sum_{t=1}^{T} \ell_{it}(\alpha_{it}, \theta)$, and $\overline{\alpha}^t(\theta, \xi) = \text{argmax}_\alpha \, \mathbb{E}_{\xi_{i0}=\xi, \lambda_0=\lambda}(\ell_{it}(\alpha, \theta))$.[13]

Assumption 4. *(regularity, time-varying heterogeneity)*

(i) $(Y_i', X_i', \xi_{i0}', h_i')'$ *are i.i.d. across* i *conditional on* λ_0*;* $(Y_{it}', X_{it}', \lambda_{t0}')'$ *are stationary for all* i*;* $\ell_{it}(\alpha_{it}, \theta)$ *is three times differentiable, for all* i, t*; and* Θ *and the supports of* ξ_{i0} *and* α_{it0} *are compact, and* θ_0 *belongs to the interior of* Θ*.*

(ii) N, T *tend jointly to infinity;* $\max_t \sup_{\xi,\lambda,\alpha,\theta} |\mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}(\ell_{it}(\alpha, \theta))| = O(1)$*, and similarly for the first three derivatives of* ℓ_{it}*; the minimum (respectively,*

[13] Note that $\overline{\alpha}^t(\theta, \xi_{i0})$ depends on the process λ_0 in addition to the type ξ_{i0}, although we leave the dependence on λ_0 implicit in the notation. In a static model, $\overline{\alpha}^t(\theta, \xi_{i0})$ is a function of ξ_{i0} and λ_{t0}, while in a dynamic model it also depends on the history of the time effects $(\lambda_{t0}, \lambda_{t-1,0}, ...)$.

maximum) eigenvalue of $(-\frac{\partial^2 \ell_{it}(\alpha,\theta)}{\partial\alpha\partial\alpha'})$ *is bounded away from zero (resp., infinity) with probability one, uniformly in* i, t, α, θ*; the third derivatives of* $\ell_{it}(\alpha,\theta)$ *are* $O_p(1)$*, uniformly in* i, t, α, θ*; and* $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}[\ell_{it}(\alpha_{it0},\theta_0) - \mathbb{E}_{\xi_{i0},\lambda_0}(\ell_{it}(\alpha_{it0},\theta_0))]^2 = O_p(1)$*, and similarly for the first three derivatives.*

(iii) $\min_t \inf_{\xi,\lambda,\theta} \mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}(-\frac{\partial^2 \ell_{it}(\overline{\alpha}^t(\theta,\xi),\theta)}{\partial\alpha\partial\alpha'}) > 0$*;* $\mathbb{E}[\frac{1}{T}\sum_{t=1}^{T}\ell_{it}(\overline{\alpha}^t(\theta,\xi_{i0}),\theta)]$ *has a unique maximum at* θ_0 *on* Θ*, and its matrix of second derivatives is* $-H < 0$*; and* $\sup_\theta \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\frac{\partial^2 \ell_{it}(\overline{\alpha}^t(\theta,\xi_{i0}),\theta)}{\partial\theta\partial\alpha'}\|^2 = O_p(1)$*.*

(iv) $\|\frac{\partial}{\partial\xi'}\big|_{\xi=\widetilde{\xi}}\mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}(\mathrm{vec}\,\frac{\partial^2 \ell_{it}(\alpha,\theta_0)}{\partial\theta\partial\alpha'})\|$*,* $\|\frac{\partial}{\partial\xi'}\big|_{\xi=\widetilde{\xi}}\mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}(\mathrm{vec}\,\frac{\partial^2 \ell_{it}(\alpha,\theta_0)}{\partial\alpha\partial\alpha'})\|$*, and* $\|\frac{\partial}{\partial\xi'}\big|_{\xi=\widetilde{\xi}}\mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}(\frac{\partial \ell_{it}(\overline{\alpha}^t(\theta,\widetilde{\xi}),\theta)}{\partial\alpha})\|$ *are* $O(1)$*, uniformly in* t*,* $\widetilde{\xi}$*,* λ*,* α*, and* θ*.*

(v) $\mathbb{E}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}(\frac{\partial \ell_{it}(\overline{\alpha}^t(\theta,\xi),\theta)}{\partial\alpha})$ *and* $\mathbb{E}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}(\mathrm{vec}\,\frac{\partial}{\partial\theta'}\big|_{\theta_0}\frac{\partial \ell_{it}(\overline{\alpha}^t(\theta,\xi),\theta)}{\partial\alpha})$ *are twice differentiable with respect to* h*, with first and second derivatives that are uniformly bounded in* t*,* ξ*,* λ*,* h *in the support of* h_i *given* $\lambda_0 = \lambda$*, and* $\theta \in \Theta$*; and* $\|\mathrm{Var}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}(\frac{\partial \ell_{it}(\overline{\alpha}^t(\theta,\xi),\theta)}{\partial\alpha})\|$ *and* $\|\mathrm{Var}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}(\mathrm{vec}\,\frac{\partial}{\partial\theta'}\big|_{\theta_0}\frac{\partial \ell_{it}(\overline{\alpha}^t(\theta,\xi),\theta)}{\partial\alpha})\|$ *are* $O(1)$*, uniformly in* t*,* ξ*,* λ*,* h*, and* θ*.*

In part (ii) in Assumption 4, we impose a stronger concavity condition than in Assumption 3.[14] The other parts are similar to Assumption 3, except part (v) where we require regularity of certain conditional expectations and variances.

We next state our second main result, where, differently from Theorem 1, s_i in (8) and H in (9) are now evaluated at (θ_0, α_{it0}), and expectations are conditional on (ξ_{i0}, λ_0).

Theorem 2. *Let Assumptions 1, 2 and 4 hold. Then, as* N, T, K *tend to infinity such that* K/N *tends to zero, we have:*

$$\widehat{\theta} = \theta_0 + H^{-1}\frac{1}{N}\sum_{i=1}^{N} s_i + O_p\left(\frac{1}{T}\right) + O_p\left(\frac{K}{N}\right) + O_p\left(K^{-\frac{2}{d}}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right). \quad (13)$$

Theorem 2 shows that GFE is consistent as N, T, K tend to infinity and K/N tends to zero. This requires no parametric assumption about how ξ_{i0} and λ_{t0} affect individual and time heterogeneity, unlike additive or interactive FE methods.

[14] In particular, we use part (ii) in Assumption 4 to establish consistency. Note that this condition can be restrictive in models with time-varying random coefficients.

To give intuition, consider the probit model (3) with time-varying unobservables. Under Assumption 2, in the first step, GFE consistently estimates an injective function $\varphi_{i0} = \varphi(\xi_{i0})$ of the type. One can then rewrite the outcome equation in (3) as $Y_{it} = \mathbf{1}\{X_{it}'\theta_0 + \alpha(\psi(\varphi_{i0}), \lambda_{t0}) + U_{it} \geq 0\}$, where ψ is the function introduced in Assumption 2, and $\alpha_{it0} = \alpha(\psi(\varphi_{i0}), \lambda_{t0})$ is simply a time-varying function of φ_{i0}. In the second step, GFE estimates this function by including group-time indicators in the probit regression.

As in Theorem 1, the expansion in Theorem 2 features a combination of incidental parameter bias and approximation error. When using the rule (6) for K, in the spirit of Corollary 1, the approximation error is of the same or lower order compared to 1/T. However, the $O_p(K/N)$ term is a new contribution relative to the time-invariant case, which reflects the estimation of KT group-specific parameters using NT observations. As an example, when $d = 1$ and K is chosen of the order of $\sqrt{T}$, the O_p terms in (13) are $O_p\left(1/T + \sqrt{T}/N\right)$.[15] Although this rate of convergence can be fast when N is sufficiently large relative to T, it is too slow to apply conventional bias-reduction methods for inference. In the next section, under the additional assumption that time heterogeneity λ_{t0} is low-dimensional, we describe how to obtain a faster convergence rate by grouping both individuals and time periods.

4 Complements and extensions

4.1 Bias reduction and inference

In models with time-invariant heterogeneity, Corollary 1 can be used to characterize the asymptotic distribution of GFE estimators. However, as in FE, the presence of the $O_p(1/T)$ term in (11) shifts the distribution of $\widehat{\theta}$ away from θ_0 whenever T is not large relative to N. A variety of methods are available to bias-correct FE estimators and construct asymptotically valid confidence intervals; see Arellano and Hahn (2007) for a review. Consider the setup of Corollary 1, under the additional assumption that the $O_p(1/T)$ term in (11) is equal to $C/T + o_p(1/T)$

[15] When $N/T^{\frac{3}{2}} \to 0$, one could obtain a faster rate in (13) by choosing another rule for K.

for some constant C. In this case, one can show that half-panel jackknife (Dhaene and Jochmans, 2015) gives asymptotically valid inference based on GFE as N and T tend to infinity at the same rate.[16] The distribution of the bias-corrected GFE estimator is then asymptotically normal centered at the truth, and the asymptotic variance H^{-1} can be consistently estimated by replacing the expectations in (8) and (9) by group-specific means.

In settings where heterogeneity varies over time, it can be desirable to group not only individuals as in (4), but also time periods (or alternatively counties or markets, depending on the application). We now describe such a method, and discuss its potential for performing inference in models with time-varying heterogeneity. In the *two-way GFE* approach, we classify time periods based on cross-sectional moments $w_t = \frac{1}{N}\sum_{i=1}^{N} w(Y_{it}, X_{it})$, and compute:

$$\left(\widehat{w}(1), ..., \widehat{w}(L), \widehat{l}_1, ..., \widehat{l}_T\right) = \underset{(\widetilde{w}(1),...,\widetilde{w}(L),l_1,...,l_T)}{\operatorname{argmin}} \sum_{t=1}^{T} \left\| w_t - \widetilde{w}(l_t) \right\|^2, \tag{14}$$

where $\{l_t\}$ are partitions of $\{1, ..., T\}$ into L groups. Given the group indicators $\widehat{k}_i$ and $\widehat{l}_t$, we then maximize $\sum_{i=1}^{N}\sum_{t=1}^{T} \ln f(Y_{it} \,|\, X_{it}, \alpha(\widehat{k}_i, \widehat{l}_t), \theta)$, with respect to θ and the KL group-specific parameters $\alpha(k, l)$.

Two-way GFE estimators can be expanded similarly to Theorem 2, under two main additional assumptions: the model is *static* and observations are independent across i and t, and the dimensions d_λ of time heterogeneity λ_{t0} and d of individual heterogeneity ξ_{i0} are *both* small. Then, for s_i and H as in Theorem 2, we show in the supplemental material that:

$$\widehat{\theta} = \theta_0 + H^{-1}\frac{1}{N}\sum_{i=1}^{N} s_i + O_p\left(\frac{1}{T} + \frac{1}{N} + \frac{KL}{NT}\right) + O_p\left(K^{-\frac{2}{d}} + L^{-\frac{2}{d_\lambda}}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right).$$

Suppose $d = d_\lambda = 1$, and K is given by (6) with γ asymptotically constant, with an analogous choice for L. Then the O_p terms in this expansion can be

[16] In particular, half-panel jackknife is valid under the conditions of Corollary 2, which requires taking $\gamma = o(1)$ in our rule (6) for K in order for the approximation error to be of small order. Deriving primitive conditions for the validity of half-panel jackknife and other bias-reduction methods for other choices of K is left for future work.

shown to be $O_p(1/T + 1/N)$. We leave to future work the formal study of the asymptotic validity of bias reduction methods for inference, such as two-way split panel jackknife (Fernández-Val and Weidner, 2016).

4.2 GFE with conditional moments

Our theory shows that the dimension d of heterogeneity plays a key role in the properties of GFE. While models with scalar latent types ξ_{i0}, such as model (2) of wages and labor force participation, are not uncommon in economics, many applications involve conditioning covariates. Under Assumptions 1 and 2, the moments h_i should, asymptotically, be injective functions of all the heterogeneity coming from both Y_i and X_i. However, when X_i depends on multiple components of heterogeneity, this might lead to a large dimension d.

We now show that GFE can still perform well under a weaker form of injectivity. Consider the case where Assumption 1 is replaced by $\alpha_{i0} = \alpha(\xi_{i0})$ and $\mu_{i0} = \mu(\xi_{i0}, \nu_{i0})$, where ν_{i0} is another latent component that affects covariates. Moreover, instead of requiring injectivity for both ξ_{i0} and ν_{i0}, let us maintain Assumption 2, which only requires h_i to be injective for ξ_{i0}. In other words, h_i needs to be directly informative about the unobserved heterogeneity component ξ_{i0} that appears in the *conditional distribution* of Y_i given X_i. We show in the supplemental material that, under regularity conditions otherwise similar to those of Corollary 1, the convergence rate of GFE is unaffected by the dimension of ν_{i0}. Specifically, when $K = \widehat{K}$ is given by (6) with $\gamma = O(1)$ (which adapts to the dimension of ξ_{i0} and not the one of ν_{i0}), we have:

$$\widehat{\theta} = \theta_0 + O_p\left(\frac{1}{T}\right) + O_p\left(\frac{1}{\sqrt{NT}}\right). \tag{15}$$

To prove (15) we assume that the rate condition $T^{1+\frac{d}{2}} = O(N)$ holds, where d is the (small) dimension of ξ_{i0}.[17]

[17] In the supplemental material, we provide an asymptotic expansion for GFE in a linear homoskedastic model under a small approximation error, as in Corollary 2. The argument requires no restriction on the relative rates of N and T. Interestingly, in this case the asymptotic variances of GFE and FE differ, since the within-group variation in ν_{i0} tends to decrease the variance, yet the expansion features an additional score term compared to Theorem 1.

In models with time-varying conditioning covariates, a simple way to target moments to ξ_{i0} is to construct h_i using the conditional distribution of Y_i given X_i. To see this, consider a static model $f(Y_{it} \,|\, X_{it}, \alpha_{i0}, \theta_0)$ where X_{it} has finite support. In this case, we have under appropriate conditions:

$$\underbrace{\frac{\sum_{t=1}^{T} \mathbf{1}\{X_{it} = x\} h(Y_{it}, X_{it})}{\sum_{t=1}^{T} \mathbf{1}\{X_{it} = x\}}}_{=h_i(x)} = \underbrace{\mathbb{E}_{X_{it}=x,\xi_{i0}}[h(Y_{it}, X_{it})]}_{=\varphi(x,\xi_{i0})} + o_p\left(1\right),$$

where $h_i(x)$ is only defined when $\sum_{t=1}^{T} \mathbf{1}\{X_{it} = x\} \neq 0$, and, importantly, $\varphi(x, \xi_{i0})$ does not depend on ν_{i0}. In the supplemental material we discuss implementation, and we report simulation results in a probit model with binary covariates. We find that using conditional moments can enhance the performance of GFE in such settings. We leave the analysis of conditional moments in the presence of continuous covariates to future work.

5 Conclusion

In this paper, we analyze some properties of two-step grouped fixed-effects (GFE) methods in settings where population heterogeneity is not discrete. Our framework relies on two main assumptions: low-dimensional individual heterogeneity, and the availability of moments to approximate the latent types. In many economic models, individual types are low-dimensional. By taking advantage of this feature, GFE can allow for flexible forms of heterogeneity across individuals and over time.

GFE methods are of interest in various applied settings. In a previous version of this paper (Bonhomme *et al.*, 2017), we used two-step GFE to estimate a dynamic structural model of location choice in the spirit of Kennan and Walker (2011), and we analyzed the performance of the discrete estimator of Bonhomme *et al.* (2019) for matched employer-employee data in the presence of continuous firm heterogeneity. Other potential applications include nonlinear factor models, nonparametric and semi-parametric panel data models such as quantile regression with individual effects, and network models.

References

[1] 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.

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

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

[4] Bonhomme, S., T. Lamadon, and E. Manresa (2017): "Discretizing Unobserved Heterogeneity," Becker Friedman Working Paper No 2019-16.

[5] Bonhomme, S., T. Lamadon, and E. Manresa (2019): "A Distributional Framework for Matched Employer-Employee Data," *Econometrica*, 87(3), 699–739.

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

[7] Chen, M., I. Fernández-Val, and M. Weidner (2020): "Nonlinear Panel Models with Interactive Effects," to appear in the *Journal of Econometrics*.

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

[9] 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.

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

[11] Graf, S., and H. Luschgy (2002): "Rates of Convergence for the Empirical Quantization Error", *Annals of Probability*, 30(2), 874–897.

[12] Hahn, J., and G. Kuersteiner (2011): "Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects," *Econometric Theory*, 27(6), 1152–1191.

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

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

[15] 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.

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

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

[18] Mundlak, Y. (1961): "Empirical Production Function Free of Management Bias," *Journal of Farm Economics*, 43(1), 44–56.

APPENDIX

Proof of Lemma 1. Define $B_{\varphi(\xi)}(K) = \min_{(\widetilde{h},\{k_i\})} \frac{1}{N}\sum_{i=1}^N \|\varphi(\xi_{i0}) - \widetilde{h}(k_i)\|^2$, similarly to (7), and denote: $(\underline{h}, \{\underline{k}_i\}) = \operatorname{argmin}_{(\widetilde{h},\{k_i\})} \sum_{i=1}^N \|\varphi(\xi_{i0}) - \widetilde{h}(k_i)\|^2$. By definition of $(\widehat{h}, \{\widehat{k}_i\})$, we have: $\sum_{i=1}^N \|h_i - \widehat{h}(\widehat{k}_i)\|^2 \leq \sum_{i=1}^N \|h_i - \underline{h}(\underline{k}_i)\|^2$ (almost surely). Letting $\varepsilon_i = h_i - \varphi(\xi_{i0})$, we thus have, using the triangle inequality twice:

$$\frac{1}{N}\sum_{i=1}^N \left\|\varphi(\xi_{i0}) - \widehat{h}(\widehat{k}_i)\right\|^2 \leq \frac{2}{N}\sum_{i=1}^N \left\|h_i - \widehat{h}(\widehat{k}_i)\right\|^2 + \frac{2}{N}\sum_{i=1}^N \|h_i - \varphi(\xi_{i0})\|^2$$
$$\leq \frac{2}{N}\sum_{i=1}^N \|h_i - \underline{h}(\underline{k}_i)\|^2 + \frac{2}{N}\sum_{i=1}^N \|\varepsilon_i\|^2 \leq 4\underbrace{\left(\frac{1}{N}\sum_{i=1}^N \|\varphi(\xi_{i0}) - \underline{h}(\underline{k}_i)\|^2\right)}_{=B_{\varphi(\xi)}(K)} + \frac{6}{N}\sum_{i=1}^N \|\varepsilon_i\|^2.$$

By Assumption 2, $\frac{1}{N}\sum_{i=1}^N \|\varepsilon_i\|^2 = O_p(1/T)$. In addition, since φ is Lipschitz-continuous, there exists a constant τ such that $\|\varphi(\xi') - \varphi(\xi)\| \leq \tau\|\xi' - \xi\|$ for all (ξ, ξ'). This implies that $B_{\varphi(\xi)}(K) \leq \tau^2 B_\xi(K)$, and Lemma 1 follows.

Proofs of Theorems 1 and 2. It is convenient to use a common notation for Theorems 1 and 2. Let p denote the number of individual-specific vectors α_i^j, $j \in \{1, ..., p\}$. In the time-invariant case: $p = 1$, $j = 1$, and $\alpha_i^j = \alpha_i$. In the time-varying case: $p = T$, $j \in \{1, ..., T\}$, and $\alpha_i^j = \alpha_{it}$. Denote $\ell_{ij} = \ell_i$ in the time-invariant case, and $\ell_{ij} = \ell_{it}$ in the time-varying case. Let $v_{ij} = \frac{\partial \ell_{ij}}{\partial \alpha}$, $v_{ij}^\alpha = \frac{\partial^2 \ell_{ij}}{\partial \alpha \partial \alpha'}$, $v_{ij}^\theta = \frac{\partial^2 \ell_{ij}}{\partial \theta \partial \alpha'}$, and $v_{ij}^{\alpha\alpha} = \frac{\partial^3 \ell_{ij}}{\partial \alpha \partial \alpha' \otimes \partial \alpha'}$ (which is a $\dim \alpha_{i0}^j \times (\dim \alpha_{i0}^j)^2$ matrix). Let,

for all $\theta \in \Theta$, $j \in \{1, ..., p\}$, and $k \in \{1, ..., K\}$, $\widehat{\alpha}^j(k, \theta)$=argmax$_\alpha \sum_{i=1}^N \mathbf{1}\{\widehat{k}_i = k\}\ell_{ij}(\alpha, \theta)$. Likewise, denote $\overline{\alpha}^j(\theta, \xi)$=argmax$_\alpha \mathbb{E}_{\xi_{i0}=\xi, \lambda_0=\lambda}(\ell_{ij}(\alpha, \theta))$. We will index expectations by ξ_{i0} and λ_0, although the conditioning on λ_0 is not needed in the time-invariant case of Theorem 1. Finally, let $\delta = \frac{1}{T} + K^{-\frac{2}{d}}$ in the time-invariant case, and let $\delta = \frac{1}{T} + \frac{K}{N} + K^{-\frac{2}{d}}$ in the time-varying case.

To show consistency of $\widehat{\theta}$, we first establish the next technical lemma (see the supplemental material for the proof):

Lemma A1. *Under the conditions of either Theorem 1 or Theorem 2 we have:*

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\widehat{\alpha}^j(\widehat{k}_i, \theta) - \overline{\alpha}^j(\theta, \xi_{i0})\right\|^2 = O_p(\delta), \quad \forall \theta \in \Theta, \tag{A1}$$

$$\sup_{\theta\in\Theta}\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\widehat{\alpha}^j(\widehat{k}_i, \theta) - \overline{\alpha}^j(\theta, \xi_{i0})\right\|^2 = o_p(1). \tag{A2}$$

From (A2) we then verify using a Taylor expansion that:

$$\sup_{\theta\in\Theta}\left|\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\ell_{ij}\left(\widehat{\alpha}^j(\widehat{k}_i, \theta), \theta\right) - \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\ell_{ij}\left(\overline{\alpha}^j(\theta, \xi_{i0}), \theta\right)\right| = o_p(1).$$

Consistency of $\widehat{\theta}$ then follows by standard arguments.

Next, the two key steps in the proof consist in showing the following two expansions:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\frac{\partial\ell_{ij}(\widehat{\alpha}^j(\widehat{k}_i, \theta_0), \theta_0)}{\partial\theta} = \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\frac{\partial}{\partial\theta}\bigg|_{\theta_0}\ell_{ij}\left(\overline{\alpha}^j(\theta, \xi_{i0}), \theta\right) + O_p(\delta), \tag{A3}$$

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\frac{\partial^2}{\partial\theta\partial\theta'}\bigg|_{\theta_0}\left(\ell_{ij}\left(\widehat{\alpha}^j(\widehat{k}_i, \theta), \theta\right) - \ell_{ij}\left(\overline{\alpha}^j(\theta, \xi_{i0}), \theta\right)\right) = o_p(1). \tag{A4}$$

To show (A3), we show the following technical lemma, where we omit references to the evaluation points θ_0 and α^j_{i0} for conciseness:

Lemma A2. *Under the conditions of either Theorem 1 or Theorem 2 we have:*

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\mathbb{E}_{\xi_{i0},\lambda_0}\left(v^\theta_{ij}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_0}\left(v^\alpha_{ij}\right)\right]^{-1}v^\alpha_{ij}\left(\widehat{\alpha}^j(\widehat{k}_i, \theta_0) - \alpha^j_{i0} + (v^\alpha_{ij})^{-1}v_{ij}\right) = O_p(\delta),$$

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left(v_{ij}^{\theta}\left(v_{ij}^{\alpha}\right)^{-1}-\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\theta}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\alpha}\right)\right]^{-1}\right)v_{ij}^{\alpha}\left(\widehat{\alpha}^{j}(\widehat{k}_i,\theta_0)-\alpha_{i0}^{j}\right)=O_p(\delta).$$

Now, expanding $v_{ij}^{\theta}(\widehat{\alpha}_j(\widehat{k}_i,\theta_0),\theta_0)$ around $\overline{\alpha}^{j}(\theta_0,\xi_{i0})=\alpha_{i0}^{j}$, and using the identity $\frac{\partial\overline{\alpha}^{j}(\theta_0,\xi_{i0})}{\partial\theta'}=\left[\mathbb{E}_{\xi_{i0},\lambda_0}\left(-v_{ij}^{\alpha}\right)\right]^{-1}\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\theta}\right)'$, we obtain:

$$\begin{aligned}&\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\frac{\partial\ell_{ij}(\widehat{\alpha}^{j}(\widehat{k}_i,\theta_0),\theta_0)}{\partial\theta}-\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\frac{\partial}{\partial\theta}\Big|_{\theta_0}\ell_{ij}\left(\overline{\alpha}^{j}(\theta,\xi_{i0}),\theta\right)\\&=\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\{v_{ij}^{\theta}\left(\widehat{\alpha}^{j}(\widehat{k}_i,\theta_0)-\alpha_{i0}^{j}\right)+\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\theta}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\alpha}\right)\right]^{-1}v_{ij}\right\}+O_p(\delta),\end{aligned}$$

and summing the two parts in Lemma A2 shows that the last expression is $O_p(\delta)$. It follows that (A3) is satisfied.

To show (A4), we show the next technical lemma:

Lemma A3. *Under the conditions of either Theorem 1 or Theorem 2 we have:*

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\frac{\partial\widehat{\alpha}^{j}(\widehat{k}_i,\theta_0)}{\partial\theta'}-\frac{\partial\overline{\alpha}^{j}(\theta_0,\xi_{i0})}{\partial\theta'}\right\|^{2}=o_p(1).\tag{A5}$$

Using (A1) and the identity $\frac{\partial\overline{\alpha}^{j}(\theta_0,\xi_{i0})}{\partial\theta'}=\left[\mathbb{E}_{\xi_{i0},\lambda_0}\left(-v_{ij}^{\alpha}\right)\right]^{-1}\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\theta}\right)'$, we thus have, under the conditions of either Theorem 1 or 2:

$$\begin{aligned}&\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\frac{\partial^2}{\partial\theta\partial\theta'}\Big|_{\theta_0}\ell_{ij}\left(\widehat{\alpha}^{j}(\widehat{k}_i,\theta),\theta\right)-\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\frac{\partial^2}{\partial\theta\partial\theta'}\Big|_{\theta_0}\ell_{ij}\left(\overline{\alpha}^{j}(\theta,\xi_{i0}),\theta\right)\\&=\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}v_{ij}^{\theta}\left(\frac{\partial\widehat{\alpha}^{j}(\widehat{k}_i,\theta_0)}{\partial\theta'}-\frac{\partial\overline{\alpha}^{j}(\theta_0,\xi_{i0})}{\partial\theta'}\right)+o_p(1)=o_p(1),\end{aligned}$$

where we have used Lemma A3 in the last equality.

Finally, to show Theorems 1 and 2 we expand the GFE score as:

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

where $\widetilde{\theta}$ lies between θ_0 and $\widehat{\theta}$, and further expand $\frac{\partial}{\partial\theta'}\big|_{\widetilde{\theta}}\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\frac{\partial\ell_{ij}(\widehat{\alpha}^{j}(\widehat{k}_i,\theta),\theta)}{\partial\theta}$

around θ_0 using that $\widetilde{\theta}$ is consistent. Lastly, we use (A3) and (A4), and note that, if $\overline{\ell}_i(\theta) = \frac{1}{p}\sum_{j=1}^{p} \ell_{ij}\left(\overline{\alpha}^j\left(\theta, \xi_{i0}\right), \theta\right)$ denotes the individual target log-likelihood, then $s_i = \frac{\partial \overline{\ell}_i(\theta_0)}{\partial \theta}$ and $H = \text{plim}_{N,T\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\xi_{i0},\lambda_0}(-\frac{\partial^2 \overline{\ell}_i(\theta_0)}{\partial\theta\partial\theta'})$.

Proof of Corollary 1. By the triangle inequality: $\frac{1}{N}\sum_{i=1}^{N} \|\widehat{h}(\widehat{k}_i) - \varphi(\xi_{i0})\|^2 \leq 2\widehat{Q}(\widehat{K}) + O_p(\frac{1}{T}) = O_p(\frac{1}{T})$. The proof of Theorem 1 is then unchanged, simply redefining δ=1/T (since heterogeneity is time-invariant here). This shows (11).

Proof of Corollary 2. To prove Corollary 2, we follow a likelihood approach (see Arellano and Hahn, 2007). Consider the difference between the GFE and FE profile log-likelihoods: $\Delta L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell_i(\widehat{\alpha}(\widehat{k}_i, \theta), \theta) - \frac{1}{N}\sum_{i=1}^{N} \ell_i(\widehat{\alpha}_i(\theta), \theta)$.

Assumption A1. *(regularity) Let* $\widehat{w}_i = -\frac{\partial^2 \ell_i(\widehat{\alpha}_i(\theta_0),\theta_0)}{\partial\alpha\partial\alpha'}$, *and* $\widehat{g}_i = \frac{\partial^2 \ell_i(\widehat{\alpha}_i(\theta_0),\theta_0)}{\partial\theta\partial\alpha'}\widehat{w}_i^{-1}$.

(i) $\ell_{it}(\alpha_i, \theta)$ *is four times differentiable, and its fourth derivatives satisfy similar properties to the first three.*

(ii) $\gamma(h)$=$\{\mathbb{E}_{h_i=h}(\widehat{w}_i)\}^{-1}\mathbb{E}_{h_i=h}(\widehat{w}_i\widehat{\alpha}_i(\theta_0))$ *and* $\lambda(h)$=$\mathbb{E}_{h_i=h}(\widehat{g}_i\widehat{w}_i)\{\mathbb{E}_{h_i=h}(\widehat{w}_i)\}^{-1}$ *are Lipschitz-continuous in* h*; and* $\text{Var}_{h_i=h}(\widehat{w}_i(\widehat{\alpha}_i(\theta_0) - \gamma(h_i))) = O(\frac{1}{T})$ *and* $\text{Var}_{h_i=h}((\widehat{g}_i - \lambda(h_i))\widehat{w}_i) = O(\frac{1}{T})$, *uniformly in* h.

Lemma A4. *Let the conditions of Corollary 2 hold, and let* $\nu_i(\theta)$=$\widehat{\alpha}_i(\theta)$−$\mathbb{E}_{h_i}(\widehat{\alpha}_i(\theta))$. *We have:*

$$\frac{\partial}{\partial\theta}\Big|_{\theta_0} \Delta L(\theta) = -\frac{\partial}{\partial\theta}\Big|_{\theta_0} \frac{1}{2N}\sum_{i=1}^{N} \nu_i(\theta)'\mathbb{E}_{\xi_{i0}}\left[-v_i^{\alpha}\left(\overline{\alpha}(\theta, \xi_{i0}), \theta\right)\right]\nu_i(\theta) + o_p\left(\frac{1}{T}\right). \quad \text{(A6)}$$

Corollary 2 follows, since the bias of the FE score is: $\frac{\partial}{\partial\theta}\big|_{\theta_0}\left[\frac{1}{N}\sum_{i=1}^{N} \ell_i(\widehat{\alpha}_i(\theta), \theta) - \frac{1}{N}\sum_{i=1}^{N} \ell_i(\overline{\alpha}(\theta, \xi_{i0}), \theta)\right] = \frac{\partial}{\partial\theta}\big|_{\theta_0}\frac{1}{2N}\sum_{i=1}^{N} \widehat{\nu}_i(\theta)'\mathbb{E}_{\xi_{i0}}[-v_i^{\alpha}(\overline{\alpha}(\theta, \xi_{i0}), \theta)]\widehat{\nu}_i(\theta) + o_p(\frac{1}{T})$, where $\widehat{\nu}_i(\theta) = \widehat{\alpha}_i(\theta) - \mathbb{E}_{\xi_{i0}}(\widehat{\alpha}_i(\theta))$; see, e.g., Arellano and Hahn (2007).