Discretizing Unobserved Heterogeneity[*]

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

Revised draft: January 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.

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.[2] 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.[3] 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

---

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

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

example, we study a model of participation in the labor market where the worker's 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. 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.

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]

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

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

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

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:

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

$$\ln g_i(\mu_{i0}) = \sum_{t=1}^{T} \ln g(X_{it} \mid X_{i,t-1}, \mu_{it0}),$$

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 $\alpha_{10}, ..., \alpha_{N0}$.

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}.

---

[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.

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}. We state this formally as 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 $\mathrm{plim}_{T \to \infty} h_i = \varphi(\xi_{i0})$, 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 $h_{i'}$. 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}. We require φ to be injective.

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 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

---

[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.

[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

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.

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:

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

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

---

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.

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. 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.

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.

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 vary T

---

[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).
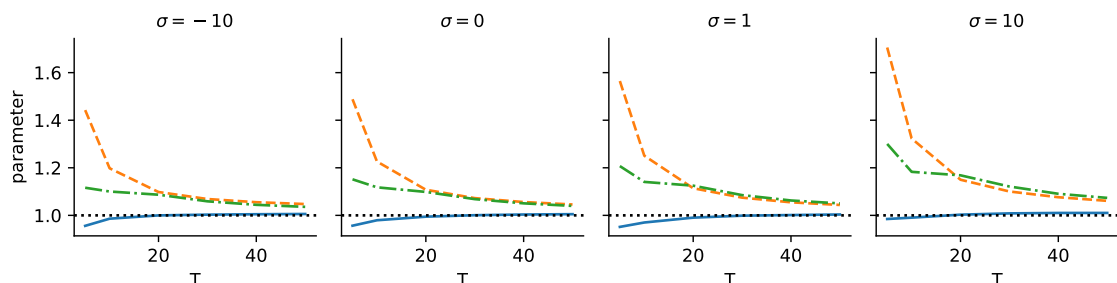
Figure 1: Model (2) of wages and participation
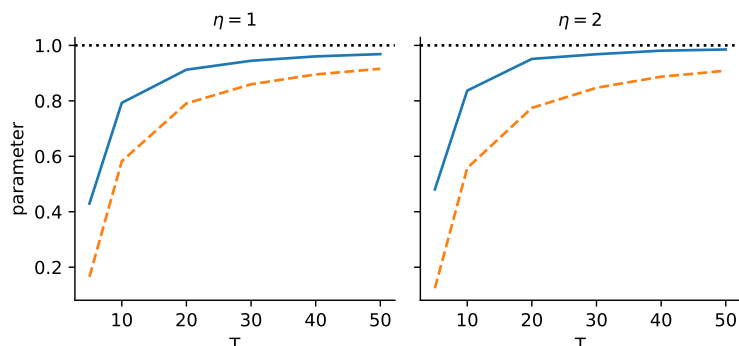
Notes: Means of $c(0; \widehat{\theta}) - c(1; \widehat{\theta})$ over 1000 simulations. GFE is indicated in solid, FE is in dashed, and the truth $c(0; \theta_0) - c(1; \theta_0) = 1$ is in dotted. $N = 1000$, and T is indicated on the x-axis. η is the risk aversion parameter in $u(\cdot)$. See the supplemental material for details.

between 5 and 50. We find that FE is more biased than GFE for both values of risk aversion. This is consistent with wages and participation 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

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.

$\overline{Y}_i$ 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.[8]

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 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 asymptotic analysis 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 $O_p(B_\xi(K))$ 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.[9] 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.

---

[8] 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.

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

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, $\|M\|$ denotes the spectral norm of a matrix M.

3.1 Time-invariant heterogeneity

To state our first main theorem, where heterogeneity is time-invariant, we make the following assumptions, where $\ell_{it}(\alpha_i, \theta) = \ln f(Y_{it} \mid Y_{i,t-1}, X_{it}, \alpha_i, \theta)$, $\ell_i(\alpha_i, \theta) = \frac{1}{T} \sum_{t=1}^T \ell_{it}(\alpha_i, \theta)$, and $\overline{\alpha}(\theta, \xi) = \operatorname{argmax}_\alpha \mathbb{E}_{\xi_{i0}=\xi}(\ell_i(\alpha, \theta))$ 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;[10] and the parameter space Θ for θ_0 is compact, the space for α_{i0} is compact, and θ_0 belongs to the interior of Θ.*

(ii) N, T, K *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 $\max_i \sup_{\alpha,\theta} |\ell_i(\alpha, \theta) - \mathbb{E}_{\xi_{i0}}(\ell_i(\alpha, \theta))| = o_p(1)$, 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) $\sup_{\widetilde{\xi},\alpha} \|\frac{\partial}{\partial\xi'}\big|_{\xi=\widetilde{\xi}} \mathbb{E}_{\xi_{i0}=\xi}(\operatorname{vec} \frac{\partial^2 \ell_{it}(\alpha,\theta_0)}{\partial\theta\partial\alpha'})\|$, $\sup_{\widetilde{\xi},\alpha} \|\frac{\partial}{\partial\xi'}\big|_{\xi=\widetilde{\xi}} \mathbb{E}_{\xi_{i0}=\xi}(\operatorname{vec} \frac{\partial^2 \ell_{it}(\alpha,\theta_0)}{\partial\alpha\partial\alpha'})\|$, 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

---

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

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 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:

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

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

Theorem 1. *Let the conditions of Lemmas 1 and 2 and 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 estima-

tors (e.g., Hahn and Newey, 2004).[11] 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 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:*

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

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 $T^{\frac{1}{2}}$).

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*

---

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

appendix holds. Then the $O_p(1/T)$ term in (11) takes the explicit form $C/T + o_p(1/T)$, where:

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

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 $\|V\|_\Omega^2 = V'\Omega V$.

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 $\overline{\alpha}^t(\theta,\xi)=\text{argmax}_\alpha\,\mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}(\ell_{it}(\alpha,\theta))$.[12]

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 space for α_{it0} are compact, and θ_0 belongs to the interior of Θ.*

(ii) *N,T,K 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, 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) *$\|\frac{\partial}{\partial\xi'}|_{\xi=\widetilde{\xi}}\mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}(\text{vec}\frac{\partial^2 \ell_{it}(\alpha,\theta_0)}{\partial\theta\partial\alpha'})\|$, $\|\frac{\partial}{\partial\xi'}|_{\xi=\widetilde{\xi}}\mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}(\text{vec}\frac{\partial^2 \ell_{it}(\alpha,\theta_0)}{\partial\alpha\partial\alpha'})\|$, and $\|\frac{\partial}{\partial\xi'}|_{\xi=\widetilde{\xi}}\mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}(\frac{\partial\ell_{it}(\overline{\alpha}^t(\theta,\widetilde{\xi}),\theta)}{\partial\alpha})\|$ 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 $\mathbb{E}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}(\text{vec}\frac{\partial}{\partial\theta'}|_{\theta_0}\frac{\partial\ell_{it}(\overline{\alpha}^t(\theta,\xi),\theta)}{\partial\alpha})$ 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 $\|\text{Var}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}(\frac{\partial\ell_{it}(\overline{\alpha}^t(\theta,\xi),\theta)}{\partial\alpha})\|$ and $\|\text{Var}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}(\text{vec}\frac{\partial}{\partial\theta'}|_{\theta_0}\frac{\partial\ell_{it}(\overline{\alpha}^t(\theta,\xi),\theta)}{\partial\alpha})\|$ 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.[13] 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).

---

[12]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 $(\lambda_{t0}, \lambda_{t-1,0}, ...)$.

[13]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.

Theorem 2. *Let the conditions of Lemmas 1 and 2 and 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.

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 $Y_{it} = \mathbf{1}\left\{ X_{it}'\theta_0 + \alpha(\psi(\varphi_{i0}), \lambda_{t0}) + U_{it} \geq 0 \right\}$, 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, 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 $T^{\frac{1}{2}}$, the O_p terms in (13) are $O_p(1/T + T^{\frac{1}{2}}/N)$.[14] 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.

---

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

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 $C/T + o_p(1/T)$ 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.[15] 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:

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

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 $\sum_{i=1}^{N} \sum_{t=1}^{T} \ln f(Y_{it} \mid X_{it}, \alpha(\widehat{k}_i, \widehat{l}_t), \theta)$, with respect to θ and the KL group-specific parameters $\alpha(k, l)$.

---

[15]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.

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 term in this expansion can be shown to be $O_p(1/T + 1/N)$. We leave to future work the formal study of the validity of bias reduction methods for inference, such as two-way split panel jackknife (Fernández-Val and Weidner, 2016), as N and T tend to infinity at the same rate.

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}.[16]

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 $f(Y_{it} \mid X_{it}, \alpha_{i0}, \theta_0)$ where X_{it} has finite support. In this case, we have under appropriate conditions:

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

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

---

[16]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.

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, 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 (2019): "A Distributional Framework for Matched Employer-Employee Data," *Econometrica*, 87(3), 699–739.

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

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

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

[8] 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.

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

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

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

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

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

[14] 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.

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

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

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

APPENDIX

<u>Proof of Lemma 1.</u> 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.

<u>Proofs of Theorems 1 and 2.</u> 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 $k \in \{1, ..., K\}$, $\widehat{\alpha}^j(k, \theta) = \text{argmax}_\alpha \sum_{i=1}^N \mathbf{1}\{\widehat{k}_i = k\}\ell_{ij}(\alpha, \theta)$. Likewise, denote $\overline{\alpha}^j(\theta, \xi) = \text{argmax}_\alpha \mathbb{E}_{\xi_{i0}=\xi, \lambda_0=\lambda}(\ell_{ij}(\alpha, \theta))$. 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.

<u>To show consistency</u> 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 <u>two key steps</u> 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_{i0}^j for conciseness:

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

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

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

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:

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

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

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:

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

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

where we have used Lemma A3 in the last equality.

Finally, <u>to show Theorems 1 and 2</u> 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 $\overline{\ell}_i(\theta)=\frac{1}{p}\sum_{j=1}^{p}\ell_{ij}\left(\overline{\alpha}^{j}(\theta,\xi_{i0}),\theta\right)$ denotes the individual target log-likelihood, then $s_i=\frac{\partial\overline{\ell}_i(\theta_0)}{\partial\theta}$ and $H=\mathrm{plim}_{N,T\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\xi_{i0},\lambda_0}(-\frac{\partial^2\overline{\ell}_i(\theta_0)}{\partial\theta\partial\theta'})$.

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

<u>Proof of Corollary 2.</u> 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 $\mathrm{Var}_{h_i=h}(\widehat{w}_i(\widehat{\alpha}_i(\theta_0)-\gamma(h_i)))=O(\frac{1}{T})$ and $\mathrm{Var}_{h_i=h}((\widehat{g}_i-\lambda(h_i))\widehat{w}_i)=O(\frac{1}{T})$, 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:*

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

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

SUPPLEMENTAL MATERIAL
"Discretizing Unobserved Heterogeneity"

S1 Proofs of technical lemmas

<u>Lemma A1.</u> From Assumption 3 (iii)-(iv) or 4 (iii)-(iv), both $\frac{\partial \overline{\alpha}^j(\theta,\xi)}{\partial \theta'}$ and $\frac{\partial \overline{\alpha}^j(\theta,\xi)}{\partial \xi'}$ are uniformly bounded (in probability in the time-varying case). Let $a^j(k,\theta) = \overline{\alpha}^j(\theta, \psi(\widehat{h}(k)))$. We thus have, using Lemmas 1 and 2:

$$\sup_{\theta \in \Theta} \frac{1}{Np} \sum_{i,j} \left\| a^j(\widehat{k}_i, \theta) - \overline{\alpha}^j(\theta, \xi_{i0}) \right\|^2 = \sup_{\theta \in \Theta} \frac{1}{Np} \sum_{i,j} \left\| \overline{\alpha}^j(\theta, \psi(\widehat{h}(\widehat{k}_i))) - \overline{\alpha}^j(\theta, \psi(\varphi(\xi_{i0}))) \right\|^2$$

$$= O_p \left(\frac{1}{N} \sum_i \| \widehat{h}(\widehat{k}_i) - \varphi(\xi_{i0}) \|^2 \right) = O_p(\delta). \tag{S1}$$

Let $\theta \in \Theta$. Expanding: $\sum_{i,j} \ell_{ij}(a^j(\widehat{k}_i, \theta), \theta) \leq \sum_{i,j} \ell_{ij}(\widehat{\alpha}^j(\widehat{k}_i, \theta), \theta)$ to second order around $\overline{\alpha}^j(\theta, \xi_{i0})$, and using:

$$\max_{i,j} \sup_{(\alpha,\theta)} \| v_{ij}^\alpha(\alpha, \theta) \| = O_p(1), \tag{S2}$$

we have, for some $a_{ij}(\theta)$ between $\widehat{\alpha}^j(\widehat{k}_i, \theta)$ and $\overline{\alpha}^j(\theta, \xi_{i0})$:

$$\frac{1}{2Np} \sum_{i,j} \left(\widehat{\alpha}^j(\widehat{k}_i, \theta) - \overline{\alpha}^j(\theta, \xi_{i0}) \right)' [-v_{ij}^\alpha(a_{ij}(\theta), \theta)] \left(\widehat{\alpha}^j(\widehat{k}_i, \theta) - \overline{\alpha}^j(\theta, \xi_{i0}) \right)$$

$$\leq \frac{1}{Np} \sum_{i,j} v_{ij}(\overline{\alpha}^j(\theta, \xi_{i0}), \theta)' \left(\widehat{\alpha}^j(\widehat{k}_i, \theta) - a^j(\widehat{k}_i, \theta) \right) + O_p(\delta)$$

$$= \frac{1}{Np} \sum_{i,j} \overline{v}_j(\widehat{k}_i, \theta)' \left(\widehat{\alpha}^j(\widehat{k}_i, \theta) - a^j(\widehat{k}_i, \theta) \right) + O_p(\delta), \tag{S3}$$

where $\overline{v}_j(k, \theta)$ denotes the mean over i of $v_{ij}(\overline{\alpha}^j(\theta, \xi_{i0}), \theta)$ in group $\widehat{k}_i = k$, and the $O_p(\delta)$ terms are uniform in θ by (S1).

Now, by Assumption 3 (ii) or 4 (ii) there exists a constant $\underline{c} > 0$ such that:

$$\min_{i,j} \inf_{(\alpha,\theta)} \text{mineig} \left[-v_{ij}^\alpha(\alpha, \theta) \right] \geq \underline{c} + o_p(1), \tag{S4}$$

where mineig(M) is the minimum eigenvalue of M. Let $A = \frac{1}{Np} \sum_{i,j} \| \widehat{\alpha}^j(\widehat{k}_i, \theta) -$

$\overline{\alpha}^j(\theta, \xi_{i0})\|^2$. By (S3) and the Cauchy Schwarz inequality, we have:

$$A \leq O_p\left[\left(\frac{1}{Np}\sum_{i,j}\left\|\overline{v}_j(\widehat{k}_i, \theta)\right\|^2\right)^{\frac{1}{2}}\left(\frac{1}{Np}\sum_{i,j}\left\|\widehat{\alpha}^j(\widehat{k}_i, \theta) - a^j(\widehat{k}_i, \theta)\right\|^2\right)^{\frac{1}{2}}\right] + O_p(\delta).$$

By (S1) and the triangle inequality: $(\frac{1}{Np}\sum_{i,j}\|\widehat{\alpha}^j(\widehat{k}_i, \theta) - a^j(\widehat{k}_i, \theta)\|^2)^{\frac{1}{2}} \leq A^{\frac{1}{2}} + O_p(\delta^{\frac{1}{2}})$. Hence: $A = O_p\left[\left(\frac{1}{Np}\sum_{i,j}\|\overline{v}_j(\widehat{k}_i, \theta)\|^2\right)^{\frac{1}{2}}\left(A^{\frac{1}{2}} + O_p(\delta^{\frac{1}{2}})\right)\right] + O_p(\delta)$, which implies:

$$A = O_p\left(\frac{1}{Np}\sum_{i,j}\|\overline{v}_j(\widehat{k}_i, \theta)\|^2\right) + O_p(\delta). \tag{S5}$$

We are now going to show that, for all $\theta \in \Theta$:

$$\frac{1}{Np}\sum_{i,j}\left\|\overline{v}_j(\widehat{k}_i, \theta)\right\|^2 = O_p(\delta). \tag{S6}$$

Using (S5) and (S6) will then imply (A1). Under the conditions of Theorem 1, it is easy to see that (S6) holds. We are now going to show (S6) under the conditions of Theorem 2. Let, for all $j, \theta, h, \xi, \lambda$: $\rho_j(h, \xi, \lambda, \theta) = \mathbb{E}_{h_i=h, \xi_{i0}=\xi, \lambda_0=\lambda}(v_{ij}(\overline{\alpha}^j(\theta, \xi), \theta))$, and, for all i, j, θ: $\zeta_{ij}(\theta) = v_{ij}(\overline{\alpha}^j(\theta, \xi_{i0}), \theta) - \rho_j(h_i, \xi_{i0}, \lambda_0, \theta)$. By Assumption 4 (v), and letting $h_i = \varphi(\xi_{i0}) + \varepsilon_i$, we can expand $\rho_j(h_i, \xi_{i0}, \lambda_0, \theta)$ twice around $\varphi(\xi_{i0})$ as: $\rho_j(\varphi(\xi_{i0}), \xi_{i0}, \lambda_0, \theta) + \frac{\partial\rho_j(\varphi(\xi_{i0}), \xi_{i0}, \lambda_0, \theta)}{\partial h'}\varepsilon_i + \frac{1}{2}\varepsilon_i'\frac{\partial^2\rho_j(a_{i\theta}^j, \xi_{i0}, \lambda_0, \theta)}{\partial h\partial h'}\varepsilon_i$, where $a_{i\theta}^j$ lies between h_i and $\varphi(\xi_{i0})$. Hence, taking expectations, using that $\mathbb{E}_{\xi_{i0}, \lambda_0}\left[\rho_j(h_i, \xi_{i0}, \lambda_0, \theta)\right] = 0$, and using Assumptions 2 and 4 (v), we have:

$$\frac{1}{Np}\sum_{i,j}\|\rho_j(\varphi(\xi_{i0}), \xi_{i0}, \lambda_0, \theta)\|^2 = \frac{1}{Np}\sum_{i,j}\left\|\frac{\partial\rho_j(\varphi(\xi_{i0}), \xi_{i0}, \lambda_0, \theta)}{\partial h'}\mathbb{E}_{\xi_{i0}, \lambda_0}\left[\varepsilon_i\right]\right\|^2 + o_p\left(\frac{1}{T}\right),$$

which is $O_p(\frac{1}{T})$. Hence: $\frac{1}{Np}\sum_{i,j}\|\rho_j(h_i, \xi_{i0}, \lambda_0, \theta)\|^2 = O_p(\frac{1}{T})$. It thus follows from the triangle inequality that:

$$\frac{1}{Np}\sum_{i,j}\|\overline{v}_j(\widehat{k}_i, \theta)\|^2 \leq O_p\left(\frac{1}{T}\right) + \frac{2}{Np}\sum_{i,j}\|\overline{\zeta}_j(\widehat{k}_i, \theta)\|^2, \tag{S7}$$

where $\overline{\zeta}_j(k, \theta)$ denotes the mean of $\zeta_{ij}(\theta)$ in group $\widehat{k}_i = k$. Now, using that

$\widehat{k}_1, ..., \widehat{k}_N$ are functions of $h_1, ..., h_N$, we have:

$$\mathbb{E}\left[\frac{1}{Np}\sum_{i,j}\|\overline{\zeta}_j(\widehat{k}_i, \theta)\|^2\right]$$

$$= \frac{1}{Np}\sum_{k,j}\mathbb{E}\left[\frac{\sum_i\sum_{i'}\mathbf{1}\{\widehat{k}_i=k\}\mathbf{1}\{\widehat{k}_{i'}=k\}\mathbb{E}_{h_1,...,h_N,\xi_{10},...,\xi_{N0},\lambda_0}\left(\zeta_{ij}(\theta)'\zeta_{i'j}(\theta)\right)}{\sum_i\mathbf{1}\{\widehat{k}_i = k\}}\right].$$

Furthermore, since observations are independent across i given λ_0:

$$\mathbb{E}_{h_1,...,h_N,\xi_{10},...,\xi_{N0},\lambda_0}\left(\zeta_{i_1,j}(\theta)'\zeta_{i_2,j}(\theta)\right)$$

$$= \mathbb{E}_{h_{i_1},\xi_{i_1,0},\lambda_0}\left(\zeta_{i_1,j}(\theta)\right)'\mathbb{E}_{h_{i_2},\xi_{i_2,0},\lambda_0}\left(\zeta_{i_2,j}(\theta)\right) = 0 \quad \text{for all } i_1 \neq i_2 \text{ and } j.$$

Hence:

$$\mathbb{E}\left[\frac{1}{Np}\sum_{i,j}\|\overline{\zeta}_j(\widehat{k}_i, \theta)\|^2\right] = \frac{1}{Np}\sum_{k,j}\mathbb{E}\left[\frac{\sum_i\mathbf{1}\{\widehat{k}_i=k\}\mathbb{E}_{h_i,\xi_{i0},\lambda_0}\left(\zeta_{ij}(\theta)'\zeta_{ij}(\theta)\right)}{\sum_i\mathbf{1}\{\widehat{k}_i = k\}}\right].$$

Finally, using that $\mathbb{E}_{h_i,\xi_{i0},\lambda_0}\left(\zeta_{ij}(\theta)\right) = 0$, and using part (v) in Assumption 4:

$$\mathbb{E}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}\left(\zeta_{ij}(\theta)'\zeta_{ij}(\theta)\right) = \text{Tr}\left[\text{Var}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}(v_{ij}(\overline{\alpha}^j(\theta, \xi_{i0}), \theta))\right] = O\left(1\right),$$

uniformly in h, ξ, λ.[1] This implies that $\mathbb{E}\left[\frac{1}{Np}\sum_{i,j}\|\overline{\zeta}_j(\widehat{k}_i, \theta)\|^2\right] = O\left(\frac{K}{N}\right)$, and shows (S6) and (A1).

We are now going to show:

$$\sup_{\theta\in\Theta}\frac{1}{Np}\sum_{i,j}\|\overline{v}_j(\widehat{k}_i, \theta)\|^2 = o_p\left(1\right). \tag{S8}$$

Using a bounding argument similar to the one we used to show (A1), (A2) will then follow. To see that (S8) holds, let $Z(\theta) = \frac{1}{Np}\sum_{i,j}\|\overline{v}_j(\widehat{k}_i, \theta)\|^2$. By (S6), $Z(\theta) = O_p(\delta)$ for all $\theta \in \Theta$. Moreover: $\frac{\partial Z(\theta)}{\partial \theta} = \frac{2}{Np}\sum_{i,j}\overline{v}_j^\theta(\widehat{k}_i, \theta)\overline{v}_j(\widehat{k}_i, \theta) = O_p\left(\sqrt{\sup_{\theta\in\Theta}Z(\theta)}\right)$ uniformly in θ, using the Cauchy Schwarz inequality with either Assumption 3 (ii) or 4 (ii), where $\overline{v}_j^\theta(k, \widetilde{\theta})$ is the mean of $\frac{\partial}{\partial\theta}\big|_{\theta=\widetilde{\theta}}v_{ij}(\overline{\alpha}^j(\theta, \xi_{i0}), \theta)'$

---

[1]Note that the dimension of v_{ij} is fixed throughout, independent of the sample size.

in group $\widehat{k}_i = k$. Since Θ is compact, it follows that $\sup_{\theta \in \Theta} Z(\theta) = o_p(1)$.[2]

<u>Lemma A2.</u> Let us omit references to θ_0 and α_{i0}^j throughout, and let:

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

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

<u>We first bound A</u>. Expanding: $\sum_i \mathbf{1}\{\widehat{k}_i = k\} v_{ij}(\widehat{\alpha}^j(k)) = 0$ for all k, j, we have, for a_{ij} between α_{i0}^j and $\widehat{\alpha}^j(\widehat{k}_i)$:

$$
\sum_i \mathbf{1}\{\widehat{k}_i = k\} v_{ij}(\alpha_{i0}^j) + \sum_i \mathbf{1}\{\widehat{k}_i = k\} v_{ij}^\alpha(\alpha_{i0}^j)(\widehat{\alpha}^j(\widehat{k}_i) - \alpha_{i0}^j)
$$

$$
+ \frac{1}{2} \sum_i \mathbf{1}\{\widehat{k}_i = k\} v_{ij}^{\alpha\alpha}(a_{ij}) \left(\widehat{\alpha}^j(\widehat{k}_i) - \alpha_{i0}^j \right) \otimes \left(\widehat{\alpha}^j(\widehat{k}_i) - \alpha_{i0}^j \right) = 0.
$$

It follows that $\widehat{\alpha}^j(\widehat{k}_i) = \widetilde{\alpha}_j(\widehat{k}_i) + \widetilde{v}_j(\widehat{k}_i) + \widetilde{w}_j(\widehat{k}_i)$, where:

$$
\widetilde{\alpha}_j(k) = \left(\sum_i \mathbf{1}\{\widehat{k}_i = k\}(-v_{ij}^\alpha) \right)^{-1} \left(\sum_i \mathbf{1}\{\widehat{k}_i = k\}(-v_{ij}^\alpha)\alpha_{i0}^j \right),
$$

$$
\widetilde{v}_j(k) = \left(\sum_i \mathbf{1}\{\widehat{k}_i = k\}(-v_{ij}^\alpha) \right)^{-1} \left(\sum_i \mathbf{1}\{\widehat{k}_i = k\} v_{ij} \right),
$$

$$
\widetilde{w}_j(k) = \frac{1}{2} \left(\sum_i \mathbf{1}\{\widehat{k}_i = k\}(-v_{ij}^\alpha) \right)^{-1} \left(\sum_i \mathbf{1}\{\widehat{k}_i = k\} v_{ij}^{\alpha\alpha}(a_{ij}) \left(\widehat{\alpha}^j(\widehat{k}_i) - \alpha_{i0}^j \right)^{\otimes 2} \right),
$$

where $a^{\otimes 2} = a \otimes a$. Hence, we have:

$$
A = \frac{1}{Np} \sum_{i,j} \mathbb{E}_{\xi_{i0},\lambda_0} \left(v_{ij}^\theta \right) \left[\mathbb{E}_{\xi_{i0},\lambda_0} \left(v_{ij}^\alpha \right) \right]^{-1} v_{ij}^\alpha \left(\widetilde{w}_j(\widehat{k}_i) + \widetilde{\alpha}_j(\widehat{k}_i) - \alpha_{i0}^j + \widetilde{v}_j(\widehat{k}_i) + (v_{ij}^\alpha)^{-1} v_{ij} \right).
$$

---

[2]Let $\upsilon > 0, \epsilon > 0$. There is $M > 0$ 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 diameters $\operatorname{diam} B_r \leq \frac{1}{2M}\sqrt{\upsilon}$. Since: $\sup_{\theta \in \Theta} Z(\theta) \leq \max_r Z(\theta_r) + \sup_\theta \left\| \frac{\partial Z(\theta)}{\partial \theta} \right\| \frac{1}{2M}\sqrt{\upsilon}$, and since: $a > \upsilon \Rightarrow a - \sqrt{a}\frac{1}{2}\sqrt{\upsilon} > \frac{\upsilon}{2}$, we have: $\Pr\left(\sup_{\theta \in \Theta} Z(\theta) > \upsilon \right) \leq \frac{\epsilon}{2} + \Pr \left(\max_r Z(\theta_r) > \frac{\upsilon}{2} \right)$, which, by (S6), is smaller than ϵ for N, T, K large enough.

Note first that:

$$\frac{1}{Np}\sum_{i,j}\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\theta}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\alpha}\right)\right]^{-1}v_{ij}^{\alpha}\widetilde{w}_j(\widehat{k}_i)=O_p(\frac{1}{Np}\sum_{i,j}\|\widehat{\alpha}^j(\widehat{k}_i)-\alpha_{i0}^j\|^2)=O_p(\delta),$$

where we have used (S2), (A1), and either Assumption 3 (ii) or Assumption 4 (ii).

Next, let $z_j(\xi_{i0})' = \mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\theta}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\alpha}\right)\right]^{-1}$. We have:

$$\frac{1}{Np}\sum_{i,j}\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\theta}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\alpha}\right)\right]^{-1}v_{ij}^{\alpha}\left(\widetilde{\alpha}_j(\widehat{k}_i)-\alpha_{i0}^j\right)$$

$$=\frac{1}{Np}\sum_{i,j}\left(z_j(\xi_{i0})'-\widetilde{z}_j\left(\widehat{k}_i\right)'\right)v_{ij}^{\alpha}\left(\widetilde{\alpha}_j(\widehat{k}_i)-\alpha_{i0}^j\right), \tag{S9}$$

where, for all k, j:

$$\widetilde{z}_j(k)=\left(\sum_i\mathbf{1}\{\widehat{k}_i=k\}(-v_{ij}^{\alpha})\right)^{-1}\left(\sum_i\mathbf{1}\{\widehat{k}_i=k\}(-v_{ij}^{\alpha})z_j(\xi_{i0})\right). \tag{S10}$$

Now we have, using that: $\alpha^j\mapsto\sum_i\left(\alpha^j(\widehat{k}_i)-\alpha_{i0}^j\right)'\left(-v_{ij}^{\alpha}\right)\left(\alpha^j(\widehat{k}_i)-\alpha_{i0}^j\right)$ is minimized at $\alpha^j=\widetilde{\alpha}_j$, and using (S2) and (S4):

$$\frac{1}{Np}\sum_{i,j}\left\|\widetilde{\alpha}_j(\widehat{k}_i)-\alpha_{i0}^j\right\|^2=O_p\left(\frac{1}{Np}\sum_{i,j}\left(\widetilde{\alpha}_j(\widehat{k}_i)-\alpha_{i0}^j\right)'\left(-v_{ij}^{\alpha}\right)\left(\widetilde{\alpha}_j(\widehat{k}_i)-\alpha_{i0}^j\right)\right)$$

$$=O_p\left(\frac{1}{Np}\sum_{i,j}\left(\widehat{\alpha}^j(\widehat{k}_i)-\alpha_{i0}^j\right)'\left(-v_{ij}^{\alpha}\right)\left(\widehat{\alpha}^j(\widehat{k}_i)-\alpha_{i0}^j\right)\right)=O_p\left(\frac{1}{Np}\sum_{i,j}\left\|\widehat{\alpha}^j(\widehat{k}_i)-\alpha_{i0}^j\right\|^2\right),$$

where the last expression is $O_p(\delta)$ by (A1). Likewise, since by Assumption 3 (iv) or 4 (iv) $\frac{\partial\,\mathrm{vec}\,z_j(\xi)}{\partial\xi'}$ is bounded (in probability) uniformly in j and ξ, we have:

$$\frac{1}{Np}\sum_{i,j}\left\|\widetilde{z}_j(\widehat{k}_i)-z_j(\xi_{i0})\right\|^2=O_p\left(\frac{1}{Np}\sum_{i,j}\left(\widetilde{z}_j(\widehat{k}_i)-z_j(\xi_{i0})\right)'\left(-v_{ij}^{\alpha}\right)\left(\widetilde{z}_j(\widehat{k}_i)-z_j(\xi_{i0})\right)\right)$$

$$=O_p\left(\frac{1}{Np}\sum_{i,j}\left(z_j\left(\psi\left(\widehat{h}(\widehat{k}_i)\right)\right)-z_j(\xi_{i0})\right)'\left(-v_{ij}^{\alpha}\right)\left(z_j\left(\psi\left(\widehat{h}(\widehat{k}_i)\right)\right)-z_j(\xi_{i0})\right)\right)$$

$$=O_p\left(\frac{1}{Np}\sum_{i,j}\left\|\widehat{h}(\widehat{k}_i)-\varphi(\xi_{i0})\right\|^2\right)=O_p(\delta), \tag{S11}$$

where we have used (S2), (S4), Lemmas 1 and 2, and that ψ is Lipschitz-continuous. Combining results, and using the Cauchy Schwarz inequality in (S9), we obtain:

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

The last term in A is:

$$A_3 = \frac{1}{Np} \sum_{i,j} \mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^\theta\right) \left[\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^\alpha\right)\right]^{-1} \left(-v_{ij}^\alpha\right) \left(\left(-v_{ij}^\alpha\right)^{-1} v_{ij} - \widetilde{v}_j(\widehat{k}_i)\right).$$

Since $\widetilde{v}_j(k) = (\sum_i \mathbf{1}\{\widehat{k}_i = k\}(-v_{ij}^\alpha))^{-1}(\sum_i \mathbf{1}\{\widehat{k}_i = k\}(-v_{ij}^\alpha)(-v_{ij}^\alpha)^{-1} v_{ij})$, we have:

$$A_3 = \frac{1}{Np} \sum_{i,j} \left(z_j(\xi_{i0})' - \widetilde{z}_j\left(\widehat{k}_i\right)'\right) \left(-v_{ij}^\alpha\right)\left(-v_{ij}^\alpha\right)^{-1} v_{ij} = \frac{1}{Np} \sum_{i,j} \left(z_j(\xi_{i0})' - \widetilde{z}_j\left(\widehat{k}_i\right)'\right) v_{ij}$$

$$= \frac{1}{Np} \sum_{i,j} \left(z_j(\xi_{i0})' - z_j^*\left(\widehat{k}_i\right)'\right) v_{ij} + \frac{1}{Np} \sum_{i,j} \left(z_j^*\left(\widehat{k}_i\right)' - \widetilde{z}_j\left(\widehat{k}_i\right)'\right) v_{ij}, \qquad \text{(S12)}$$

where $\widetilde{z}_j(k)$ is given by (S10), and:

$$z_j^*(k) = \left(\sum_i \mathbf{1}\{\widehat{k}_i = k\} \mathbb{E}_{\xi_{i0},\lambda_0}\left(-v_{ij}^\alpha\right)\right)^{-1} \left(\sum_i \mathbf{1}\{\widehat{k}_i = k\} \mathbb{E}_{\xi_{i0},\lambda_0}\left(-v_{ij}^\alpha\right) z_j(\xi_{i0})\right).$$

$$\text{(S13)}$$

Under the conditions of Theorem 1, it is easy to see that $A_3 = O_p(\delta)$. We are now going to show that $A_3 = O_p(\delta)$ under the conditions of Theorem 2. To see that the first term on the right-hand-side of (S12) is $O_p(\delta)$, we use an argument similar to the one we used to show (S6). Let $\zeta_{ij} = v_{ij} - \mathbb{E}_{h_i,\xi_{i0},\lambda_0}(v_{ij})$. Following the same steps as the ones leading to (S7), we obtain:

$$\frac{1}{Np} \sum_{i,j} \left\|\mathbb{E}_{h_i,\xi_{i0},\lambda_0}(v_{ij})\right\|^2 = O_p\left(\frac{1}{T}\right). \qquad \text{(S14)}$$

Moreover, by an argument similar to (S11), since $\mathbb{E}_{\xi_{i0},\lambda_0}(-v_{ij}^\alpha)$ is bounded away

from zero with probability one, we have:

$$\frac{1}{Np} \sum_{i,j} \left\| z_j(\xi_{i0}) - z_j^*(\widehat{k}_i) \right\|^2 = O_p(\delta). \tag{S15}$$

Let $z' = (z_1', ..., z_p')$, and $z^*(k)' = (z_1^*(k)', ..., z_p^*(k)')$. Since ζ_{ij} are independent across i, with zero mean, conditional on $h_1, ..., h_N, \xi_{10}, ..., \xi_{N0}, \lambda_0$, we thus have, denoting $\zeta_i = (\zeta_{i1}', ..., \zeta_{ip}')'$:

$$\mathbb{E} \left[\left\| \frac{1}{Np} \sum_{i,j} \left(z_j(\xi_{i0})' - z_j^*\left(\widehat{k}_i\right)' \right) v_{ij} \right\|^2 \right]$$

$$\leq 2O\left(\frac{1}{T}\right) \mathbb{E} \left[\frac{1}{Np} \sum_{i,j} \left\| z_j(\xi_{i0}) - z_j^*\left(\widehat{k}_i\right) \right\|^2 \right]$$

$$+ 2\mathbb{E} \left[\left\| \frac{1}{Np} \sum_{i,j} \left(z_j(\xi_{i0})' - z_j^*\left(\widehat{k}_i\right)' \right) \zeta_{ij} \right\|^2 \right]$$

$$= O\left(\frac{\delta}{T}\right) + \frac{2}{N^2 p^2} \sum_i \mathbb{E} \left[\left(z_i' - z^*\left(\widehat{k}_i\right)' \right) \mathbb{E}_{h_i, \xi_{i0}, \lambda_0} \left[\zeta_i \zeta_i' \right] \left(z_i - z^*\left(\widehat{k}_i\right) \right) \right]$$

$$= O\left(\frac{\delta}{T}\right) + O\left(\frac{\delta p}{NT}\right) = O(\delta^2),$$

where we have used, in turn, the triangle and Cauchy Schwarz inequalities, (S14), (S15), conditional independence of the ζ_i across i, part (v) in Assumption 4, and (S15) one more time. Note that, by part (v) in Assumption 4, $\left\| \mathbb{E}_{h_i, \xi_{i0}, \lambda_0} \left[\zeta_i \zeta_i' \right] \right\| \leq \text{Tr}\, \mathbb{E}_{h_i, \xi_{i0}, \lambda_0} \left[\zeta_i \zeta_i' \right] \leq p \max_j \text{Tr}\, \mathbb{E}_{h_i, \xi_{i0}, \lambda_0} \left[\zeta_{ij} \zeta_{ij}' \right] = O_p(p^2/T)$.

Turning to the second term in (S12), we have:

$$\frac{1}{Np} \sum_{i,j} \left(z_j^*\left(\widehat{k}_i\right)' - \widetilde{z}_j\left(\widehat{k}_i\right)' \right) v_{ij} = \frac{1}{Np} \sum_{i,j} \left(z_j^*\left(\widehat{k}_i\right)' - \widetilde{z}_j\left(\widehat{k}_i\right)' \right) \overline{v}_j\left(\widehat{k}_i\right),$$

where by (S6) we have: $\frac{1}{Np} \sum_{i,j} \|\overline{v}_j(\widehat{k}_i)\|^2 = O_p(\delta)$. Moreover:

$$\frac{1}{Np} \sum_{i,j} \left\| z_j^*\left(\widehat{k}_i\right) - \widetilde{z}_j\left(\widehat{k}_i\right) \right\|^2 \leq \frac{2}{Np} \sum_{i,j} \left\| z_j(\xi_{i0}) - z_j^*\left(\widehat{k}_i\right) \right\|^2$$

$$+ \frac{2}{Np} \sum_{i,j} \left\| z_j(\xi_{i0}) - \widetilde{z}_j\left(\widehat{k}_i\right) \right\|^2,$$

where the second term on the right-hand side is $O_p(\delta)$ due to (S11), and the first term is $O_p(\delta)$ due to (S15). This shows that $A_3 = O_p(\delta)$, hence that $A = O_p(\delta)$.

Let us now turn to B. Letting: $\pi'_{ij} = v^\theta_{ij} \left(v^\alpha_{ij}\right)^{-1} - \mathbb{E}_{\xi_{i0},\lambda_0} \left(v^\theta_{ij}\right) \left[\mathbb{E}_{\xi_{i0},\lambda_0} \left(v^\alpha_{ij}\right)\right]^{-1}$, we have:

$$ B = \frac{1}{Np} \sum_{i,j} \pi'_{ij} v^\alpha_{ij} \left(\widetilde{w}_j(\widehat{k}_i) + \widetilde{v}_j(\widehat{k}_i) + \widetilde{\alpha}_j(\widehat{k}_i) - \alpha^j_{i0} \right). $$

First, we have: $\frac{1}{Np} \sum_{i,j} \pi'_{ij} v^\alpha_{ij} \widetilde{w}_j(\widehat{k}_i) = O_p(\delta)$. Next, we have: $\frac{1}{Np} \sum_{i,j} \pi'_{ij} v^\alpha_{ij} \widetilde{v}_j(\widehat{k}_i) = \frac{1}{Np} \sum_{i,j} \widetilde{\pi}_j(\widehat{k}_i)' v^\alpha_{ij} \widetilde{v}_j(\widehat{k}_i)$, where $\widetilde{\pi}_j(k)$ is defined similarly to $\widetilde{\alpha}_j(k)$. To see that this quantity is $O_p(\delta)$, note that, by the definition of $\widetilde{v}_j(k)$ and using (S4) and (S6):

$$ \frac{1}{Np} \sum_{i,j} \left\| \widetilde{v}_j(\widehat{k}_i) \right\|^2 = O_p \left(\frac{1}{Np} \sum_{i,j} \left\| \overline{v}_j(\widehat{k}_i) \right\|^2 \right) = O_p(\delta). $$

Moreover, letting $\tau_{ij} = \pi'_{ij} v^\alpha_{ij}$, we have:

$$ \frac{1}{Np} \sum_{i,j} \left\| \widetilde{\pi}_j(\widehat{k}_i) \right\|^2 = O_p \left(\frac{1}{Np} \sum_{i,j} \left\| \overline{\tau}_j(\widehat{k}_i) \right\|^2 \right). $$

Now, the τ_{ij} are independent across i, with zero conditional mean given ξ_{i0}, λ_0:

$$ \mathbb{E}_{\xi_{i0},\lambda_0} \left(\pi'_{ij} v^\alpha_{ij} \right) = \mathbb{E}_{\xi_{i0},\lambda_0} \left(\left(v^\theta_{ij} \left(v^\alpha_{ij}\right)^{-1} - \mathbb{E}_{\xi_{i0},\lambda_0} \left(v^\theta_{ij}\right) \left[\mathbb{E}_{\xi_{i0},\lambda_0} \left(v^\alpha_{ij}\right)\right]^{-1} \right) v^\alpha_{ij} \right) = 0. $$

Using an argument similar to the one we used to show (S6), and using Assumption 4 (v) in the time-varying case, it thus follows that $\frac{1}{Np} \sum_{i,j} \|\widetilde{\pi}_j(\widehat{k}_i)\|^2 = O_p(\delta)$. Hence, by the Cauchy Schwarz inequality: $\frac{1}{Np} \sum_{i,j} \pi'_{ij} v^\alpha_{ij} \widetilde{v}_j(\widehat{k}_i) = O_p(\delta)$.

We lastly bound the third term B_3 in B:

$$ \frac{1}{Np} \sum_{i,j} \pi'_{ij} v^\alpha_{ij} \left(\widetilde{\alpha}_j(\widehat{k}_i) - \alpha^j_{i0} \right) = \frac{1}{Np} \sum_{i,j} \pi'_{ij} v^\alpha_{ij} \left[\left(\alpha^*_j(\widehat{k}_i) - \alpha^j_{i0} \right) + \left(\widetilde{\alpha}_j(\widehat{k}_i) - \alpha^*_j(\widehat{k}_i) \right) \right], $$

where $\widetilde{\alpha}_j(k)$ and $\alpha^*_j(k)$ are given by expressions similar to (S10) and (S13), with α^j_{i0} in place of $z_j(\xi_{i0})$ in those formulas. The first term is $O_p(\delta)$ since, similarly to (S15): $\frac{1}{Np} \sum_{i,j} \|\alpha^*_j(\widehat{k}_i) - \alpha^j_{i0}\|^2 = O_p(\delta)$, and the $\tau_{ij} = \pi'_{ij} v^\alpha_{ij}$ are conditionally

independent across i with zero mean given ξ_{i0} and λ_0 (using a similar argument to the first term in (S12)). The second term is:

$$\frac{1}{Np}\sum_{i,j}\pi'_{ij}v^{\alpha}_{ij}\left(\widetilde{\alpha}_j(\widehat{k}_i)-\alpha^*_j(\widehat{k}_i)\right) = \frac{1}{Np}\sum_{i,j}\widetilde{\pi}_j(\widehat{k}_i)'v^{\alpha}_{ij}\left(\widetilde{\alpha}_j(\widehat{k}_i)-\alpha^*_j(\widehat{k}_i)\right).$$

We have already shown that: $\frac{1}{Np}\sum_{i,j}\|\widetilde{\pi}_j(\widehat{k}_i)\|^2=O_p(\delta)$. Moreover, using similar arguments to the ones we used to bound $\frac{1}{Np}\sum_{i,j}\|z^*_j(\widehat{k}_i)-\widetilde{z}_j(\widehat{k}_i)\|^2$ above, we have: $\frac{1}{Np}\sum_{i,j}\|\widetilde{\alpha}_j(\widehat{k}_i)-\alpha^*_j(\widehat{k}_i)\|^2=O_p(\delta)$. This shows that $B_3=O_p(\delta)$, hence that $B=O_p(\delta)$.

Lemma A3. For given k,j, θ-differentiating: $\sum_i\mathbf{1}\{\widehat{k}_i=k\}v_{ij}(\widehat{\alpha}^j(k,\theta),\theta)=0$, and using (S4), we obtain:

$$\frac{\partial\widehat{\alpha}^j(k,\theta)}{\partial\theta'}=\left(\sum_i\mathbf{1}\{\widehat{k}_i=k\}\left(-v^{\alpha}_{ij}\left(\widehat{\alpha}^j(\widehat{k}_i,\theta),\theta\right)\right)\right)^{-1}\sum_i\mathbf{1}\{\widehat{k}_i=k\}v^{\theta}_{ij}\left(\widehat{\alpha}^j(\widehat{k}_i,\theta),\theta\right)'.$$
(S16)

Let us define, at $\theta=\theta_0$ (and omitting θ_0 and α^j_{i0} from the notation):

$$\frac{\partial\widetilde{\alpha}^j(k)}{\partial\theta'}=\left(\sum_i\mathbf{1}\{\widehat{k}_i=k\}(-v^{\alpha}_{ij})\right)^{-1}\sum_i\mathbf{1}\{\widehat{k}_i=k\}(v^{\theta}_{ij})',$$

$$\frac{\partial\widetilde{\alpha}^j_*(k)}{\partial\theta'}=\left(\sum_i\mathbf{1}\{\widehat{k}_i=k\}(-v^{\alpha}_{ij})\right)^{-1}\sum_i\mathbf{1}\{\widehat{k}_i=k\}(-v^{\alpha}_{ij})\underbrace{\left[\mathbb{E}_{\xi_{i0},\lambda_0}(-v^{\alpha}_{ij})\right]^{-1}\mathbb{E}_{\xi_{i0},\lambda_0}(v^{\theta}_{ij})'}_{=\frac{\partial\overline{\alpha}^j(\xi_{i0})}{\partial\theta'}}.$$

Using (A1) and (S4), we have: $\frac{1}{Np}\sum_{i,j}\|\frac{\partial\widehat{\alpha}^j(\widehat{k}_i)}{\partial\theta'}-\frac{\partial\widetilde{\alpha}^j(\widehat{k}_i)}{\partial\theta'}\|^2=o_p(1)$. Moreover:

$$\frac{\partial\widetilde{\alpha}^j(k)}{\partial\theta'}-\frac{\partial\widetilde{\alpha}^j_*(k)}{\partial\theta'}=\left(\frac{\sum_i\mathbf{1}\{\widehat{k}_i=k\}(-v^{\alpha}_{ij})}{\sum_i\mathbf{1}\{\widehat{k}_i=k\}}\right)^{-1}\left(\frac{\sum_i\mathbf{1}\{\widehat{k}_i=k\}\tau'_{ij}}{\sum_i\mathbf{1}\{\widehat{k}_i=k\}}\right),$$

where $\tau'_{ij}=(v^{\theta}_{ij})'-(-v^{\alpha}_{ij})\left[\mathbb{E}_{\xi_{i0},\lambda_0}(-v^{\alpha}_{ij})\right]^{-1}\mathbb{E}_{\xi_{i0},\lambda_0}(v^{\theta}_{ij})'$ are conditionally independent across i, with zero mean given ξ_{i0} and λ_0. Hence, using (S4), and a similar argument to the one we used to show (S6), we have: $\frac{1}{Np}\sum_{i,j}\|\frac{\partial\widetilde{\alpha}^j(\widehat{k}_i)}{\partial\theta'}-\frac{\partial\widetilde{\alpha}^j_*(\widehat{k}_i)}{\partial\theta'}\|^2=o_p(1)$. Lastly, using (S4) we have, as in (S11): $\frac{1}{Np}\sum_{i,j}\|\frac{\partial\widetilde{\alpha}^j_*(\widehat{k}_i)}{\partial\theta'}-\frac{\partial\overline{\alpha}^j(\xi_{i0})}{\partial\theta'}\|^2=o_p(1)$. Combining results shows (A5).

34

Lemma A4. In the following we again evaluate all functions at θ_0, and omit θ_0 for the notation. In particular, $\widehat{\alpha}_i$ is a shorthand for $\widehat{\alpha}_i(\theta_0)$. We will use the notation $\widehat{w}_i = -v_i^\alpha(\widehat{\alpha}_i)$. The choice of $K = \widehat{K}$ with $\gamma = o(1)$ implies that:

$$\frac{1}{N} \sum_i \|h_i - \widehat{h}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right). \tag{S17}$$

We also have: $\frac{1}{N} \sum_i \|\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\|^2 = O_p\left(\frac{1}{T}\right)$. Let, for all k:

$$\widetilde{\alpha}(k) = \left(\sum_i \mathbf{1}\{\widehat{k}_i=k\}\widehat{w}_i\right)^{-1} \sum_i \mathbf{1}\{\widehat{k}_i=k\}\widehat{w}_i\widehat{\alpha}_i. \tag{S18}$$

Expanding: $\sum_i \mathbf{1}\{\widehat{k}_i=k\}v_i(\widehat{\alpha}(k)) = 0$ around $\widehat{\alpha}_i$, using that $v_i(\widehat{\alpha}_i) = 0$, we obtain:

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

where $a_i(k)$ lies between $\widehat{\alpha}_i$ and $\widehat{\alpha}(k)$, and $v_i^{\alpha\alpha}(a_i(k))$ is a matrix of third derivatives with $(\dim \alpha_{i0})^2$ columns.

To see that (A6) holds, we rely on the following decomposition:

$$\left.\frac{\partial}{\partial\theta}\right|_{\theta_0} \Delta L(\theta) = \frac{1}{N}\sum_i \frac{\partial \ell_i(\widehat{\alpha}(\widehat{k}_i))}{\partial\theta} - \frac{1}{N}\sum_i \frac{\partial \ell_i(\widehat{\alpha}_i)}{\partial\theta}$$

$$= \frac{1}{N}\sum_i v_i^\theta(\widehat{\alpha}_i)\left(\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right) + \frac{1}{2N}\sum_i v_i^{\theta\alpha}(a_i)\left(\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right)^{\otimes 2}$$

$$= \frac{1}{N}\sum_i v_i^\theta(\widehat{\alpha}_i)\left(\widetilde{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right) + \frac{1}{2N}\sum_i v_i^{\theta\alpha}(a_i)\left(\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right)^{\otimes 2}$$

$$+ \frac{1}{2N}\sum_i v_i^\theta(\widehat{\alpha}_i)\left(\mathbb{E}_{\widehat{k}_i}[\widehat{w}_{i'}]\right)^{-1}\mathbb{E}_{\widehat{k}_i}\left[v_{i'}^{\alpha\alpha}\left(a_{i'}(\widehat{k}_{i'})\right)\left(\widehat{\alpha}(\widehat{k}_{i'}) - \widehat{\alpha}_{i'}\right)^{\otimes 2}\right]$$

$$= \underbrace{\frac{1}{N}\sum_i v_i^\theta(\widehat{\alpha}_i)\left(\widetilde{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right)}_{=A_1} + \underbrace{\frac{1}{2N}\sum_i v_i^{\theta\alpha}(a_i)\left(\widetilde{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right)^{\otimes 2}}_{=A_2} + o_p\left(\frac{1}{T}\right)$$

$$+ \underbrace{\frac{1}{2N}\sum_i v_i^\theta(\widehat{\alpha}_i)\left(\mathbb{E}_{\widehat{k}_i}[\widehat{w}_{i'}]\right)^{-1}\mathbb{E}_{\widehat{k}_i}\left[v_{i'}^{\alpha\alpha}\left(a_{i'}(\widehat{k}_{i'})\right)\left(\widetilde{\alpha}(\widehat{k}_{i'}) - \widehat{\alpha}_{i'}\right)^{\otimes 2}\right]}_{=A_3},$$

35

where a_i and $a_i(\widehat{k}_i)$ lie between $\widehat{\alpha}_i$ and $\widehat{\alpha}(\widehat{k}_i)$, and $\mathbb{E}_k$ denotes a mean in group $\widehat{k}_i = k$. Let $\gamma(h) = \{\mathbb{E}_{h_i=h}(\widehat{w}_i)\}^{-1}\mathbb{E}_{h_i=h}(\widehat{w}_i\widehat{\alpha}_i)$, and $\nu_i = \widehat{\alpha}_i - \gamma(h_i)$. Let $\widehat{g}_i = v_i^\theta(\widehat{\alpha}_i)(\widehat{w}_i)^{-1}$, $\lambda(h) = \mathbb{E}_{h_i=h}(\widehat{g}_i\widehat{w}_i)\{\mathbb{E}_{h_i=h}(\widehat{w}_i)\}^{-1}$, and $\tau_i = \widehat{g}_i' - \lambda(h_i)'$. Using (S17) we can show, using that γ is Lipschitz-continuous, that $\frac{1}{N}\sum_i\|\gamma(h_i) - \widetilde{\gamma}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right)$. Moreover, we have: $\mathbb{E}[\widehat{w}_i\nu_i \mid h_1,...,h_N] = \mathbb{E}_{h_i}[\widehat{w}_i\widehat{\alpha}_i] - \mathbb{E}_{h_i}[\widehat{w}_i\widehat{\alpha}_i] = 0$. Similar arguments to the proof of Lemma A1 give: $\frac{1}{N}\sum_i\|\mathbb{E}_{\widehat{k}_i}[\widehat{w}_{i'}\nu_{i'}]\|^2 = O_p(\frac{K}{NT}) = o_p\left(\frac{1}{T}\right)$. Hence: $\frac{1}{N}\sum_i\|\widetilde{\nu}(\widehat{k}_i)\|^2 = \frac{1}{N}\sum_i\|\left(\mathbb{E}_{\widehat{k}_i}[\widehat{w}_{i'}]\right)^{-1}\mathbb{E}_{\widehat{k}_i}[\widehat{w}_{i'}\nu_{i'}]\|^2 = o_p\left(\frac{1}{T}\right)$. Likewise, we have: $\frac{1}{N}\sum_i\|\lambda(h_i) - \widetilde{\lambda}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right)$, and: $\frac{1}{N}\sum_i\|\widetilde{\tau}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right)$.[3]

Let us now expand the three terms A_1, A_2, A_3 in the above decomposition:

$$A_1 = \frac{1}{N}\sum_i \widehat{g}_i\widehat{w}_i\left(\widetilde{\alpha}(\widehat{k}_i)-\widehat{\alpha}_i\right) = \frac{1}{N}\sum_i\left(\widehat{g}_i-\widetilde{g}(\widehat{k}_i)\right)\widehat{w}_i\left(\widetilde{\alpha}(\widehat{k}_i)-\widehat{\alpha}_i\right)$$

$$= -\frac{1}{N}\sum_i\left(\lambda(h_i) - \widetilde{\lambda}(\widehat{k}_i) + \tau_i' - \widetilde{\tau}(\widehat{k}_i)'\right)\widehat{w}_i\left(\gamma(h_i) - \widetilde{\gamma}(\widehat{k}_i) + \nu_i - \widetilde{\nu}(\widehat{k}_i)\right)$$

$$= -\frac{1}{N}\sum_i\tau_i'\widehat{w}_i\nu_i + o_p\left(\frac{1}{T}\right) = -\frac{1}{N}\sum_i\tau_i'\mathbb{E}_{\xi_{i0}}(-v_i^\alpha(\alpha_{i0}))\nu_i + o_p\left(\frac{1}{T}\right),$$

$$A_2 = \frac{1}{2N}\sum_i\mathbb{E}_{\xi_{i0}}\left(v_i^{\theta\alpha}(\alpha_{i0})\right)\left(\widetilde{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right)^{\otimes 2} + o_p\left(\frac{1}{T}\right)$$

$$= \frac{1}{2N}\sum_i\mathbb{E}_{\xi_{i0}}\left(v_i^{\theta\alpha}(\alpha_{i0})\right)\left(\widetilde{\gamma}(\widehat{k}_i) - \gamma(h_i) + \widetilde{\nu}(\widehat{k}_i) - \nu_i\right)^{\otimes 2} + o_p\left(\frac{1}{T}\right)$$

$$= \frac{1}{2N}\sum_i\mathbb{E}_{\xi_{i0}}\left(v_i^{\theta\alpha}(\alpha_{i0})\right)\nu_i^{\otimes 2} + o_p\left(\frac{1}{T}\right),$$

$$A_3 = \frac{1}{2N}\sum_i\mathbb{E}_{\xi_{i0}}\left(v_i^\theta(\alpha_{i0})\right)\left[\mathbb{E}_{\xi_{i0}}(-v_i^\alpha(\alpha_{i0}))\right]^{-1}\mathbb{E}_{\xi_{i0}}\left[v_i^{\alpha\alpha}(\alpha_{i0})\right]\nu_i^{\otimes 2} + o_p\left(\frac{1}{T}\right).$$

Combining, we get:

$$\left.\frac{\partial}{\partial\theta}\right|_{\theta_0}\Delta L(\theta) = -\frac{1}{N}\sum_i\tau_i'\mathbb{E}_{\xi_{i0}}(-v_i^\alpha(\alpha_{i0}))\nu_i + o_p\left(\frac{1}{T}\right)$$

$$+ \frac{1}{2N}\sum_i\left[\mathbb{E}_{\xi_{i0}}\left(v_i^{\theta\alpha}(\alpha_{i0})\right) + \mathbb{E}_{\xi_{i0}}\left(v_i^\theta(\alpha_{i0})\right)\left[\mathbb{E}_{\xi_{i0}}(-v_i^\alpha(\alpha_{i0}))\right]^{-1}\mathbb{E}_{\xi_{i0}}\left[v_i^{\alpha\alpha}(\alpha_{i0})\right]\right]\nu_i^{\otimes 2}.$$

---

[3] Here $\widetilde{\gamma}(k)$, $\widetilde{\lambda}(k)$, $\widetilde{\nu}(k)$, and $\widetilde{\tau}(k)$ are defined similarly to $\widetilde{\alpha}(k)$ in (S18), with $\gamma(h_i)$, $\lambda(h_i)$, ν_i, and τ_i, respectively, replacing $\widehat{\alpha}_i$ in that formula.

Now, $\frac{\partial \widehat{\alpha}_i(\theta_0)}{\partial \theta'} = \widehat{g}_i'$, and:

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

Let $\omega_i = \{\mathbb{E}_{h_i}(\widehat{w}_i)\}^{-1}\widehat{w}_i$, and $\widetilde{\nu}_i(\theta) = \widehat{\alpha}_i(\theta) - \mathbb{E}_{h_i}(\omega_i\widehat{\alpha}_i(\theta))$. Combining the above with the expression of the bias of the FE score, we obtain:

$$\frac{\partial}{\partial \theta}\Big|_{\theta_0} \Delta L(\theta) = -\frac{\partial}{\partial \theta}\Big|_{\theta_0} \frac{1}{2N} \sum_i \widetilde{\nu}_i(\theta)' \mathbb{E}_{\xi_{i0}} \left[-v_i^\alpha\left(\overline{\alpha}(\theta, \xi_{i0}), \theta\right)\right] \widetilde{\nu}_i(\theta) + o_p\left(\frac{1}{T}\right). \quad \text{(S19)}$$

Lastly, let $\widehat{\alpha}_i(\theta) = \mathbb{E}_{h_i}(\widehat{\alpha}_i(\theta)) + \nu_i(\theta)$, and $\omega_i = \mathbb{E}_{h_i}(\omega_i) + \eta_i = 1 + \eta_i$. We have: $\widetilde{\nu}_i(\theta) = \nu_i(\theta) - \mathbb{E}_{h_i}(\eta_i\nu_i(\theta))$, from which it follows that: $\frac{1}{N}\sum_i \|\widetilde{\nu}_i(\theta_0) - \nu_i(\theta_0)\|^2 = o_p(1/T)$. Likewise: $\frac{1}{N}\sum_i \|\frac{\partial\widetilde{\nu}_i(\theta_0)}{\partial\theta'} - \frac{\partial\nu_i(\theta_0)}{\partial\theta'}\|^2 = o_p(1/T)$. Hence, (S19) implies (A6).

S2 Complements and extensions

S2.1 Average effects

Let $m_i(\alpha_i, \theta) = \frac{1}{T}\sum_{t=1}^T m(X_{it}, \alpha_i, \theta)$ in the time-invariant case, and $m_i(\alpha_i, \theta) = \frac{1}{T}\sum_{t=1}^T m(X_{it}, \alpha_{it}, \theta)$ in the time-varying case. Let $\widehat{M} = \frac{1}{N}\sum_i m_i\left(\widehat{\alpha}(\widehat{k}_i), \widehat{\theta}\right)$ be the GFE estimator of $M_0 = \frac{1}{N}\sum_i m_i(\alpha_{i0}, \theta_0)$. We use a common notation as in the proofs of Theorems 1 and 2, and denote $m_{ij}(\alpha_i^j, \theta) = m_i(\alpha_i, \theta)$ in the time-invariant case, and $m_{ij}(\alpha_i^j, \theta) = m(X_{it}, \alpha_{it}, \theta)$ in the time-varying case.

Assumption S1. *(average effects)*

(i) $m_{ij}(\alpha, \theta)$ *is twice differentiable in both its arguments, for all i, j.*

(ii) $\max_{i,j} \sup_{\alpha,\theta} \|m_{ij}(\alpha, \theta)\| = O_p(1)$, *and similarly for the first two derivatives of m_{ij};* $\max_j \sup_{\widetilde{\xi},\lambda} \|\frac{\partial}{\partial\xi'}\big|_{\xi=\widetilde{\xi}} \mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}(\frac{\partial m_{ij}(\alpha_{i0}^j,\theta_0)}{\partial\alpha})\| = O(1)$; *and, letting* $\tau_{ij}^m = \frac{\partial m_{ij}(\alpha_{i0}^j,\theta_0)}{\partial\alpha'} - \mathbb{E}_{\xi_{i0},\lambda_0}[\frac{\partial m_{ij}(\alpha_{i0}^j,\theta_0)}{\partial\alpha'}]\mathbb{E}_{\xi_{i0},\lambda_0}[v_{ij}^\alpha(\alpha_{i0}^j,\theta_0)]v_{ij}^\alpha(\alpha_{i0}^j,\theta_0)$, *the function* $\mathbb{E}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}(\text{vec}\,\tau_{ij}^m)$ *is twice differentiable with respect to h, with first and second derivatives that are uniformly bounded in j, ξ, λ, and h, and* $\|\text{Var}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}(\text{vec}\,\tau_{ij}^m)\| = O(\frac{p}{T})$, *uniformly in j, ξ, λ, and h.*

Let s_i and H as in Theorem 1 or 2, and let $\overline{s} = \frac{1}{N} \sum_i s_i$. Define:

$$s_i^m = \frac{1}{p} \sum_j \left\{ \mathbb{E}_{\xi_{i0},\lambda_0} \left(\frac{\partial m_{ij}}{\partial \alpha'} \right) \left[\mathbb{E}_{\xi_{i0},\lambda_0} \left(-\frac{\partial^2 \ell_{ij}}{\partial \alpha \partial \alpha'} \right) \right]^{-1} \frac{\partial \ell_{ij}}{\partial \alpha} + \mathbb{E}_{\xi_{i0},\lambda_0} \left(\frac{\partial m_{ij}}{\partial \theta'} \right) H^{-1} \overline{s} \right.$$
$$\left. + \mathbb{E}_{\xi_{i0},\lambda_0} \left(\frac{\partial m_{ij}}{\partial \alpha'} \right) \left[\mathbb{E}_{\xi_{i0},\lambda_0} \left(-\frac{\partial^2 \ell_{ij}}{\partial \alpha \partial \alpha'} \right) \right]^{-1} \mathbb{E}_{\xi_{i0},\lambda_0} \left(\frac{\partial^2 \ell_{ij}}{\partial \alpha \partial \theta'} \right) H^{-1} \overline{s} \right\}.$$

Corollary S1. *Let the conditions of Theorem 1 or 2 hold, and let Assumption S1 hold. Then, as N, T, K tend to infinity such that $Kp/(NT)$ tends to zero:*

$$\widehat{M} \;=\; M_0 + \frac{1}{N} \sum_i s_i^m + O_p \left(\frac{1}{T} \right) + O_p \left(\frac{Kp}{NT} \right) + O_p \left(K^{-\frac{2}{d}} \right) + o_p \left(\frac{1}{\sqrt{NT}} \right).$$

Proof. We have, by a Taylor expansion:

$$\widehat{M} - M_0 = \frac{1}{Np} \sum_{i,j} m_{ij} \left(\widehat{\alpha}^j(\widehat{k}_i, \widehat{\theta}), \widehat{\theta} \right) - \frac{1}{Np} \sum_{i,j} m_{ij} \left(\alpha_{i0}^j, \theta_0 \right)$$
$$= \frac{1}{Np} \sum_{i,j} \frac{\partial m_{ij} \left(\alpha_{i0}^j, \theta_0 \right)}{\partial \alpha'} \left(\widehat{\alpha}^j(\widehat{k}_i, \widehat{\theta}) - \alpha_{i0}^j \right) + \frac{1}{Np} \sum_{i,j} \frac{\partial m_{ij} \left(\alpha_{i0}^j, \theta_0 \right)}{\partial \theta'} \left(\widehat{\theta} - \theta_0 \right) + O_p(\delta),$$

where δ is defined as in the proofs of Theorems 1 and 2.

Using similar arguments to the ones we used to establish Lemma A2, under Assumption S1 we have (recall that $\overline{\alpha}^j(\theta_0, \xi_{i0}) = \alpha_{i0}^j$):

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

Moreover, using (A5) and Assumption S1 we obtain:

$$\frac{1}{Np} \sum_{i,j} \frac{\partial m_{ij} \left(\alpha_{i0}^j, \theta_0 \right)}{\partial \alpha'} \left\{ \left(\widehat{\alpha}^j(\widehat{k}_i, \widehat{\theta}) - \overline{\alpha}^j(\widehat{\theta}, \xi_{i0}) \right) - \left(\widehat{\alpha}^j(\widehat{k}_i, \theta_0) - \overline{\alpha}^j(\theta_0, \xi_{i0}) \right) \right\}$$
$$= o_p \left(\| \widehat{\theta} - \theta_0 \| \right) + O_p(\delta) = o_p \left(\frac{1}{\sqrt{NT}} \right) + O_p(\delta).$$

Combining, we obtain:

$$
\widehat{M} - M_0 = \frac{1}{Np} \sum_{i,j} \frac{\partial m_{ij}\left(\alpha_{i0}^j, \theta_0\right)}{\partial \alpha'} \left(\widehat{\alpha}^j(\widehat{k}_i, \widehat{\theta}) - \widehat{\alpha}^j(\widehat{k}_i, \theta_0)\right)
$$

$$
+ \frac{1}{Np} \sum_{i,j} \frac{\partial m_{ij}\left(\alpha_{i0}^j, \theta_0\right)}{\partial \alpha'} \left(\widehat{\alpha}^j(\widehat{k}_i, \theta_0) - \alpha_{i0}^j\right) + \frac{1}{Np} \sum_{i,j} \frac{\partial m_{ij}\left(\alpha_{i0}^j, \theta_0\right)}{\partial \theta'} \left(\widehat{\theta} - \theta_0\right) + O_p\left(\delta\right)
$$

$$
= \frac{1}{Np} \sum_{i,j} \frac{\partial m_{ij}\left(\alpha_{i0}^j, \theta_0\right)}{\partial \alpha'} \left(\overline{\alpha}^j(\widehat{\theta}, \xi_{i0}) - \overline{\alpha}^j(\theta_0, \xi_{i0})\right)
$$

$$
+ \frac{1}{Np} \sum_{i,j} \mathbb{E}_{\xi_{i0}, \lambda_0} \left[\frac{\partial m_{ij}\left(\alpha_{i0}^j, \theta_0\right)}{\partial \alpha'}\right] \mathbb{E}_{\xi_{i0}, \lambda_0} \left[-v_{ij}^\alpha(\alpha_{i0}^j, \theta_0)\right]^{-1} v_{ij}(\alpha_{i0}^j, \theta_0)
$$

$$
+ \frac{1}{Np} \sum_{i,j} \frac{\partial m_{ij}\left(\alpha_{i0}^j, \theta_0\right)}{\partial \theta'} \left(\widehat{\theta} - \theta_0\right) + O_p\left(\delta\right) + o_p\left(\frac{1}{\sqrt{NT}}\right).
$$

The result comes from expanding $\overline{\alpha}^j(\widehat{\theta}, \xi_{i0})$ around θ_0, and then substituting $\widehat{\theta} - \theta_0$ by its influence function. ∎

S2.2   Two-way GFE

We have the following lemma, whose proof is analogous to that of Lemma 1.

Lemma S1. *Suppose that there exist random vectors $h_i = \frac{1}{T} \sum_t h(Y_{it}, X_{it})$ and $w_t = \frac{1}{N} \sum_i w(Y_{it}, X_{it})$, with fixed dimensions, and Lipschitz-continuous functions φ and ϕ, such that $h_i = \varphi(\xi_{i0}) + o_p(1)$, $\frac{1}{N} \sum_i \|h_i - \varphi(\xi_{i0})\|^2 = O_p\left(1/T\right)$, $w_t = \phi(\lambda_{t0}) + o_p(1)$, and $\frac{1}{T} \sum_t \|w_t - \phi(\lambda_{t0})\|^2 = O_p\left(1/N\right)$ as N, T tend to infinity. Then we have, as N, T, K tend to infinity: $\frac{1}{N} \sum_i \|\widehat{h}(\widehat{k}_i) - \varphi(\xi_{i0})\|^2 = O_p\left(\frac{1}{T}\right) + O_p\left(B_\xi(K)\right)$, and, as N, T, L tend to infinity: $\frac{1}{T} \sum_t \|\widehat{w}(\widehat{l}_t) - \phi(\lambda_{t0})\|^2 = O_p\left(\frac{1}{N}\right) + O_p\left(B_\lambda(L)\right)$, where $B_\lambda(L)$ is defined analogously to $B_\xi(K)$.*

For all θ, ξ, and λ, let $\overline{\alpha}(\theta, \xi, \lambda) = \text{argmax}_\alpha \, \mathbb{E}_{\xi_{i0}=\xi, \lambda_{t0}=\lambda}(\ell_{it}(\alpha, \theta))$. In addition, let $\xi_0 = (\xi_{10}', ..., \xi_{N0}')'$.

Assumption S2. *(regularity, two-way)*

(i) *$(Y_{it}', X_{it}')'$, $i = 1, .., N$, $t = 1, ..., T$, are i.i.d. given ξ_0 and λ_0, ξ_{i0} are i.i.d., and λ_{t0} are i.i.d.; $\ell_{it}(\alpha, \theta)$ is three times differentiable in (θ, α); Θ is compact, the spaces for ξ_{i0} and λ_{t0} are compact, and θ_0 belongs to the interior of Θ.*

(ii) N, T, K, L tend jointly to infinity; $\sup_{\xi,\lambda,\alpha,\theta} |\mathbb{E}_{\xi_{i0}=\xi,\lambda_{t0}=\lambda}(\ell_{it}(\alpha,\theta))| = O(1)$, and similarly for the first three derivatives of ℓ_{it}; the minimum (resp., 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, and the third derivatives of $\ell_{it}(\alpha,\theta)$ are $O_p(1)$, uniformly in i, t, α, θ; $\frac{1}{NT}\sum_{i,t}[\ell_{it}(\alpha_{it0},\theta_0) - \mathbb{E}_{\xi_{i0},\lambda_{t0}}(\ell_{it}(\alpha_{it0},\theta_0))]^2 = O_p(1)$, and similarly for the first three derivatives of ℓ_{it}.

(iii) $\inf_{\xi,\lambda,\theta} \mathbb{E}_{\xi_{i0}=\xi,\lambda_{t0}=\lambda}(-\frac{\partial^2\ell_{it}(\overline{\alpha}(\theta,\xi,\lambda),\theta)}{\partial\alpha\partial\alpha'}) > 0$; $\mathbb{E}\left[\frac{1}{NT}\sum_{i,t}\ell_{it}(\overline{\alpha}(\theta,\xi_{i0},\lambda_{t0}),\theta)\right]$ has a unique maximum at θ_0 on Θ, and its second derivative is $-H < 0$.

(iv) $\frac{\partial}{\partial\xi'}\big|_{\widetilde{\xi}}\mathbb{E}_{\xi_{i0}=\xi,\lambda_{t0}=\lambda}(\text{vec }\frac{\partial^2\ell_{it}(\alpha,\theta_0)}{\partial\theta\partial\alpha'})=O(1)$; $\frac{\partial}{\partial\lambda'}\big|_{\widetilde{\lambda}}\mathbb{E}_{\xi_{i0}=\xi,\lambda_{t0}=\lambda}(\text{vec }\frac{\partial^2\ell_{it}(\alpha,\theta_0)}{\partial\theta\partial\alpha'})=O(1)$; $\frac{\partial}{\partial\xi'}\big|_{\widetilde{\xi}}\mathbb{E}_{\xi_{i0}=\xi,\lambda_{t0}=\lambda}(\text{vec }\frac{\partial^2\ell_{it}(\alpha,\theta_0)}{\partial\alpha\partial\alpha'})=O(1)$; $\frac{\partial}{\partial\lambda'}\big|_{\widetilde{\lambda}}\mathbb{E}_{\xi_{i0}=\xi,\lambda_{t0}=\lambda}(\text{vec }\frac{\partial^2\ell_{it}(\alpha,\theta_0)}{\partial\alpha\partial\alpha'})=O(1)$; $\frac{\partial}{\partial\xi'}\big|_{\widetilde{\xi}}\mathbb{E}_{\xi_{i0}=\xi,\lambda_{t0}=\lambda}(\frac{\partial\ell_{it}(\overline{\alpha}(\theta,\xi,\lambda),\theta)}{\partial\alpha})=O(1)$; $\frac{\partial}{\partial\lambda'}\big|_{\widetilde{\lambda}}\mathbb{E}_{\xi_{i0}=\xi,\lambda_{t0}=\lambda}(\frac{\partial\ell_{it}(\overline{\alpha}(\theta,\xi,\lambda),\theta)}{\partial\alpha})=O(1)$, uniformly in $\xi, \widetilde{\xi}, \lambda, \widetilde{\lambda}, \alpha, \theta$.

(v) $\mathbb{E}_{h_i=h,\xi_{i0}=\xi,w_t=w,\lambda_{t0}=\lambda}(\frac{\partial\ell_{it}(\overline{\alpha}(\theta,\xi,\lambda),\theta)}{\partial\alpha})$, $\mathbb{E}_{h_i=h,\xi_{i0}=\xi,w_t=w,\lambda_{t0}=\lambda}(\text{vec }\frac{\partial}{\partial\theta'}\big|_{\theta_0}\frac{\partial\ell_{it}(\overline{\alpha}(\theta,\xi,\lambda),\theta)}{\partial\alpha})$ are twice differentiable with respect to h and w, with first and second derivatives that are uniformly bounded in $h \in \mathcal{H}$, $w \in \mathcal{W}$, ξ, λ, and $\theta \in \Theta$, where $\mathcal{H}$ and $\mathcal{W}$ are the supports of h_i and w_t; $\|\text{Var}_{h_i=h,\xi_{i0}=\xi,w_t=w,\lambda_{t0}=\lambda}(\frac{\partial\ell_{it}(\overline{\alpha}(\theta,\xi,\lambda),\theta)}{\partial\alpha})\|$ and $\|\text{Var}_{h_i=h,\xi_{i0}=\xi,w_t=w,\lambda_{t0}=\lambda}(\text{vec }\frac{\partial}{\partial\theta'}\big|_{\theta_0}\frac{\partial\ell_{it}(\overline{\alpha}(\theta,\xi,\lambda),\theta)}{\partial\alpha})\|$ are $O(1)$, uniformly in h, w, ξ, λ, θ.

Theorem S1. *Let the conditions in Lemma S1 hold. Suppose that $B_\xi(K) = O_p(K^{-\frac{2}{d}})$ and $B_\lambda(L) = O_p(L^{-\frac{2}{d_\lambda}})$. Suppose that α and μ are Lipschitz-continuous in both arguments, and that there exist two Lipschitz-continuous functions ψ and Ψ such that $\xi_{i0} = \psi(\varphi(\xi_{i0}))$ and $\lambda_{t0} = \Psi(\phi(\lambda_{t0}))$. Lastly, let Assumption S2 hold. Then, as N, T, K, L tend to infinity such that $KL/(NT)$ tends to zero, we have:*

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

Proof. The proof closely follows the steps of that of Theorem 2. Here we simply highlight the main differences. Let $\delta = \frac{1}{T} + \frac{1}{N} + \frac{KL}{NT} + K^{-\frac{2}{d}} + L^{-\frac{2}{d_\lambda}}$. To show

consistency, a key step is to show, for all $\theta \in \Theta$:

$$\frac{1}{NT}\sum_{i,t}\left\|\overline{v}(\widehat{k}_i,\widehat{l}_t,\theta)\right\|^2 = O_p(\delta), \tag{S20}$$

where $\overline{v}(k,l,\theta)$ denotes the mean of $v_{it}(\overline{\alpha}(\theta,\xi_{i0},\lambda_{t0}),\theta)$ in the intersection of groups $\widehat{k}_i=k$ and $\widehat{l}_t=l$. Let: $\rho(h,\xi,w,\lambda,\theta) = \mathbb{E}_{h_i=h,\xi_{i0}=\xi,w_t=w,\lambda_{t0}=\lambda}(v_{it}(\overline{\alpha}(\theta,\xi,\lambda),\theta))$, and let, for all i,t,θ: $\zeta_{it}(\theta) = v_{it}(\overline{\alpha}(\theta,\xi_{i0},\lambda_{t0}),\theta) - \rho(h_i,\xi_{i0},w_t,\lambda_{t0},\theta)$. Proceeding as in the proof of Lemma A1 we have:

$$\frac{1}{NT}\sum_{i,t}\|\rho(h_i,\xi_{i0},w_t,\lambda_{t0},\theta)\|^2 = O_p\left(\frac{1}{T}\right) + O_p\left(\frac{1}{N}\right).$$

We thus only need to bound:

$$\mathbb{E}\left[\frac{1}{NT}\sum_{i,t}\|\overline{\zeta}(\widehat{k}_i,\widehat{l}_t,\theta)\|^2\right] = \frac{1}{NT}\sum_{k,\ell}\mathbb{E}_{k\ell}\left[\mathbb{E}_{h_i,\xi_{i0},w_t,\lambda_{t0}}\left(\zeta_{it}(\theta)'\zeta_{it}(\theta)\right)\right],$$

where we have used that observations are independent across i and t given ξ_0 and λ_0, and $\mathbb{E}_{k\ell}$ denotes a mean in groups $\widehat{k}_i = k$ and $\widehat{l}_t = l$. To bound this quantity, we use part (v) in Assumption S2. We thus obtain (S20).

Similarly to the proof of Lemma A2, we then show:

$$\frac{1}{NT}\sum_{i,t}\left\{v_{it}^\theta\left(\widehat{\alpha}(\widehat{k}_i,\widehat{l}_t)-\alpha_{it0}\right) + \mathbb{E}_{\xi_{i0},\lambda_{t0}}\left(v_{it}^\theta\right)\left[\mathbb{E}_{\xi_{i0},\lambda_{t0}}\left(v_{it}^\alpha\right)\right]^{-1}v_{it}\right\} = O_p(\delta), \tag{S21}$$

where we omit references to θ_0 and α_{it0}. The first key term is:

$$A_3 = \frac{1}{NT}\sum_{i,t}\mathbb{E}_{\xi_{i0},\lambda_{t0}}\left(v_{it}^\theta\right)\left[\mathbb{E}_{\xi_{i0},\lambda_{t0}}\left(v_{it}^\alpha\right)\right]^{-1}\left(-v_{it}^\alpha\right)\left(\left(-v_{it}^\alpha\right)^{-1}v_{it} - \widetilde{v}(\widehat{k}_i,\widehat{l}_t)\right),$$

where $\widetilde{v}$ is defined analogously to the proof of Lemma A2. To show that $A_3 = O_p(\delta)$, we use that the $\zeta_{it}(\theta_0)$ are independent across i and t, with zero mean conditional on $h_1, ..., h_N$, $w_1, ..., w_T$, ξ_0, and λ_0.

Let $\pi'_{it} = v^\theta_{it} (v^\alpha_{it})^{-1} - \mathbb{E}_{\xi_{i0},\lambda_{t0}} (v^\theta_{it}) [\mathbb{E}_{\xi_{i0},\lambda_{t0}} (v^\alpha_{it})]^{-1}$. <u>The second key term</u> is:

$$
\begin{aligned}
B_3 &= \frac{1}{NT} \sum_{i,t} \pi'_{it} v^\alpha_{it} \left(\widetilde{\alpha}(\widehat{k}_i, \widehat{l}_t) - \alpha_{it0} \right) \\
&= \frac{1}{NT} \sum_{i,t} \pi'_{it} v^\alpha_{it} \left(\alpha^*(\widehat{k}_i, \widehat{l}_t) - \alpha_{it0} \right) + \frac{1}{NT} \sum_{i,t} \pi'_{it} v^\alpha_{it} \left(\widetilde{\alpha}(\widehat{k}_i, \widehat{l}_t) - \alpha^*(\widehat{k}_i, \widehat{l}_t) \right),
\end{aligned}
$$

where $\widetilde{\alpha}(k,l)$ and $\alpha^*(k,l)$ are defined analogously to the proof of Lemma A2. To show that $B_3 = O_p(\delta)$, we use that $\tau_{it} = \pi'_{it} v^\alpha_{it}$ are independent across i and t with zero mean given ξ_0, λ_0.

<u>The final step</u>, as in the proof of Lemma A3, is to show that:

$$
\frac{1}{NT} \sum_{i,t} \left\| \frac{\partial \widehat{\alpha}(\widehat{k}_i, \widehat{l}_t, \theta_0)}{\partial \theta'} - \frac{\partial \overline{\alpha}(\theta_0, \xi_{i0}, \lambda_{t0})}{\partial \theta'} \right\|^2 = o_p(1). \tag{S22}
$$

The proof of (S22) follows similar arguments to the proof of Lemma A3. ∎

S2.3 GFE based on conditional moments

Assumption S3. *(heterogeneity, conditional case)*

There exist vectors ξ_{i0} of fixed dimension d, and ν_{i0} of dimension d_ν, and functions α and μ Lipschitz-continuous in ξ, such that $\alpha_{i0} = \alpha(\xi_{i0})$ and $\mu_{i0} = \mu(\xi_{i0}, \nu_{i0})$.

Differently from Assumption 1, here μ_{i0} depends on an additional heterogeneity component ν_{i0}, and by Assumption 2 the moment h_i is only injective for ξ_{i0}.

Assumption S4. *(regularity, conditional case)*

   (i) *$(Y'_i, X'_i, \xi'_{i0}, \nu'_{i0}, h'_i)'$ are i.i.d.; $(Y'_{it}, X'_{it})'$ are stationary for all i; $\ell_{it}(\alpha, \theta)$ is three times differentiable in both its arguments for all i, t; and Θ is compact, the space for α_{i0} is compact, and θ_0 belongs to the interior of Θ.*

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

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

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

(v) $\mathbb{E}_{h_i=h,\xi_{i0}=\xi}(\frac{\partial\ell_{it}(\overline{\alpha}(\theta,\xi),\theta)}{\partial\alpha})$ *is twice differentiable with respect to* h *and* ξ, *with first and second derivatives that are uniformly bounded in* ξ, $h \in \mathcal{H}$, *and* $\theta \in \Theta$; *and* $\|\mathrm{Var}_{h_i=h,\xi_{i0}=\xi}(\frac{\partial\ell_i(\overline{\alpha}(\theta,\xi),\theta)}{\partial\alpha})\| = O(1)$, *uniformly in* ξ, h *and* θ.

Corollary S2. *Let the conditions of Lemmas 1 and 2 hold. Let Assumptions 2, S3, and S4 hold. Let K be given by (6), with $\gamma = O(1)$. Then, as N, T, K tend to infinity such that $T^{1+\frac{d}{2}} = O(N)$ we have:*

$$\widehat{\theta} = \theta_0 + O_p\left(\frac{1}{T}\right) + O_p\left(\frac{1}{\sqrt{NT}}\right). \tag{S23}$$

Proof. Let $\delta = \frac{1}{T} + \frac{K}{N} + K^{-\frac{2}{d}}$.[4] To show <u>consistency</u>, the key step is to show:

$$\frac{1}{N}\sum_i \left\|\overline{v}(\widehat{k}_i,\theta)\right\|^2 = O_p(\delta), \quad \forall\theta \in \Theta. \tag{S24}$$

Let, for all θ, h, ξ: $\rho(h,\xi,\theta) = \mathbb{E}_{h_i=h,\xi_{i0}=\xi}(v_i(\overline{\alpha}(\theta,\xi),\theta))$, and let, for all i,θ: $\zeta_i(\theta) = v_i(\overline{\alpha}(\theta,\xi_{i0}),\theta) - \rho(h_i,\xi_{i0},\theta)$. One can show, using similar techniques to the proof of Lemma A1, that: $\frac{1}{N}\sum_i \|\overline{\zeta}(\widehat{k}_i,\theta)\|^2 = O_p(\frac{K}{N})$, and that this implies (S24).[5]

<u>We then show</u>: $\frac{1}{N}\sum_i \frac{\partial\ell_i(\widehat{\alpha}(\widehat{k}_i,\theta_0),\theta_0)}{\partial\theta} = O_p(\delta)$, which will follow from:

$$\frac{1}{N}\sum_i v_i^\theta\left(\widehat{\alpha}(\widehat{k}_i) - \alpha_{i0}\right) = O_p(\delta), \tag{S25}$$

where from now on we omit references to θ_0 and α_{i0}. We have:

$$\frac{1}{N}\sum_i v_i^\theta\left(\widehat{\alpha}(\widehat{k}_i) - \alpha_{i0}\right) = \frac{1}{N}\sum_i v_i^\theta\left(\widetilde{\alpha}(\widehat{k}_i) - \alpha_{i0} + \widetilde{v}(\widehat{k}_i)\right) + O_p(\delta),$$

---

[4]Note that if $K = \widehat{K}$ is given by (6) with $\gamma = O(1)$, then $K = O(T^{\frac{d}{2}})$ and $\delta = O(\frac{1}{T} + \frac{T^{\frac{d}{2}}}{N})$, so if $T^{1+\frac{d}{2}} = O(N)$ then $\delta = O(\frac{1}{T})$.

[5]Note that, in the case of Theorem 1 (i.e., in the absence of additional heterogeneity ν_{i0}), the left-hand side in (S24) is $O_p(\frac{1}{T})$.

where $\widetilde{\alpha}(k)$ and $\widetilde{v}(k)$ are as in the proof of Lemma A2; that is, denoting $w_i = (-v_i^\alpha)$, we have $\widetilde{\alpha}(k) = \overline{w}(k)^{-1}\overline{w\alpha}(k)$ and $\widetilde{v}(k) = \overline{w}(k)^{-1}\overline{v}(k)$.

Let $\gamma_v(h_i) = \mathbb{E}_{h_i}(v_i)$, $\zeta_i^v = v_i - \gamma_v(h_i)$, $\gamma_w(h_i) = \mathbb{E}_{h_i}(w_i)$, $\zeta_i^w = w_i - \gamma_w(h_i)$, $\gamma_{v^\theta}(h_i) = \mathbb{E}_{h_i}(v_i^\theta)$, and $\zeta_i^{v^\theta} = v_i^\theta - \gamma_{v^\theta}(h_i)$. <u>First</u>, we have:

$$\frac{1}{N}\sum_i v_i^\theta \widetilde{v}(\widehat{k}_i) = \frac{1}{N}\sum_i v_i^\theta \overline{w}(\widehat{k}_i)^{-1}\overline{v}(\widehat{k}_i) = \frac{1}{N}\sum_i \overline{v}^\theta(\widehat{k}_i)\overline{w}(\widehat{k}_i)^{-1}v_i$$
$$=\frac{1}{N}\sum_i(\overline{\gamma}_{v^\theta}(\widehat{k}_i) + \overline{\zeta}^{v^\theta}(\widehat{k}_i))(\overline{\gamma}_w(\widehat{k}_i) + \overline{\zeta}^w(\widehat{k}_i))^{-1}v_i=\frac{1}{N}\sum_i \overline{\gamma}_{v^\theta}(\widehat{k}_i)\overline{\gamma}_w(\widehat{k}_i)^{-1}v_i + O_p(\delta),$$

where for example $\overline{\gamma}_w(k)$ is the mean of $\gamma_w(h_i)$ in group $\widehat{k}_i = k$, and we have used that $\frac{1}{N}\sum_i \|\overline{\zeta}^{v^\theta}(\widehat{k}_i)\|^2 = O_p(K/N)$, $\frac{1}{N}\sum_i \|\overline{\zeta}^w(\widehat{k}_i)\|^2 = O_p(K/N)$, and $\frac{1}{N}\sum_i \|v_i\|^2 = O_p(1/T)$. Moreover:

$$\frac{1}{N}\sum_i \overline{\gamma}_{v^\theta}(\widehat{k}_i)\overline{\gamma}_w(\widehat{k}_i)^{-1}v_i = \frac{1}{N}\sum_i \overline{\gamma}_{v^\theta}(\widehat{k}_i)\overline{\gamma}_w(\widehat{k}_i)^{-1}\gamma_v(h_i) + O_p(\delta),$$

where we have used that $\frac{1}{N}\sum_i \|\overline{\zeta}^v(\widehat{k}_i)\|^2 = O_p(K/(NT))$. Lastly, we have:

$$\frac{1}{N}\sum_i \overline{\gamma}_{v^\theta}(\widehat{k}_i)\overline{\gamma}_w(\widehat{k}_i)^{-1}\gamma_v(h_i) = \frac{1}{N}\sum_i \gamma_{v^\theta}(h_i)\gamma_w(h_i)^{-1}\gamma_v(h_i)$$
$$+ \frac{1}{N}\sum_i \left[\overline{\gamma}_{v^\theta}(\widehat{k}_i)\overline{\gamma}_w(\widehat{k}_i)^{-1} - \gamma_{v^\theta}(h_i)\gamma_w(h_i)^{-1}\right]\gamma_v(h_i),$$

where the first term is $O_p(\delta)$ since it is a mean of i.i.d. terms with mean $O(1/T)$ and variance $O(1/T)$, and the second term is $O_p(\delta)$ since $\frac{1}{N}\sum_i \|h_i - \overline{h}(\widehat{k}_i)\|^2 = O_p(\delta)$ and the γ functions are Lipschitz-continuous.

<u>Second</u>, let $v_i^\theta w_i^{-1} = \eta(h_i, \xi_{i0}) + e_i$, where $\mathbb{E}_{h_i=h,\xi_{i0}=\xi}(e_i w_i) = 0$. We have:

$$\frac{1}{N}\sum_i v_i^\theta\left(\widetilde{\alpha}(\widehat{k}_i)-\alpha_{i0}\right)=\frac{1}{N}\sum_i \eta(h_i,\xi_{i0})w_i\left(\widetilde{\alpha}(\widehat{k}_i)-\alpha_{i0}\right) + \frac{1}{N}\sum_i e_i w_i\left(\widetilde{\alpha}(\widehat{k}_i)-\alpha_{i0}\right),$$

where the first term is $O_p(\delta)$ since $\frac{1}{N}\sum_i \|h_i - \overline{h}(\widehat{k}_i)\|^2 = O_p(\delta)$, $\frac{1}{N}\sum_i \|\xi_{i0} - \overline{\xi}(\widehat{k}_i)\|^2 = O_p(\delta)$, $\frac{1}{N}\sum_i \|\widetilde{\alpha}(\widehat{k}_i) - \alpha_{i0}\|^2 = O_p(\delta)$, η is Lipschitz-continuous, and

w_i is uniformly bounded (as in the proof of Lemma A2), and the second term is:

$$\frac{1}{N} \sum_i e_i w_i \left(\widetilde{\alpha}(\widehat{k}_i) - \alpha_{i0} \right) = \frac{1}{N} \sum_i e_i w_i \left(\widetilde{\alpha}(\widehat{k}_i) - \overline{\alpha}(\widehat{k}_i) \right) + \frac{1}{N} \sum_i e_i w_i \left(\overline{\alpha}(\widehat{k}_i) - \alpha_{i0} \right)$$

$$= \frac{1}{N} \sum_i \overline{ew}(\widehat{k}_i) \left(\widetilde{\alpha}(\widehat{k}_i) - \overline{\alpha}(\widehat{k}_i) \right) + O_p(\delta) = O_p(\delta),$$

where we have used that the $(e_i w_i)$'s have zero mean given $h_1, ..., h_N, \xi_{10}, ..., \xi_{N0}$ with bounded conditional variance, and $\frac{1}{N} \sum_i \|\overline{ew}(\widehat{k}_i)\|^2 = O_p(K/N) = O_p(\delta)$.

  <u>Finally, to show:</u> $\frac{1}{N} \sum_i \frac{\partial^2}{\partial\theta\partial\theta'}\big|_{\theta_0} (\ell_i(\widehat{\alpha}(\widehat{k}_i, \theta), \theta) - \ell_i(\overline{\alpha}(\theta, \xi_{i0}), \theta)) = o_p(1)$, we use similar arguments to the proof of Lemma A3.[6] ∎

Example: a linear homoskedastic model. Consider the model $Y_{it} = X_{it}\theta_0 + \alpha_{i0} + U_{it}$, where X_{it} are scalar and U_{it} are i.i.d. with mean zero and variance σ^2 given $X_{i1}, ..., X_{iT}, \alpha_{i0}$. Let $\widehat{\theta}$ be the GFE estimator based on a moment $h_i = \varphi(\alpha_{i0}) + \varepsilon_i$ that satisfies Assumptions 1 and 2 for $\xi_{i0} = \alpha_{i0}$; that is, h_i is only informative about α_{i0}, but not about the heterogeneity in X_{it}. Let $\zeta_i^X = \overline{X}_i - \mathbb{E}_{h_i}(\overline{X}_i)$, $\zeta_i^\alpha = \alpha_{i0} - \mathbb{E}_{h_i}(\alpha_{i0})$, and $\zeta_i^U = \overline{U}_i - \mathbb{E}_{h_i}(\overline{U}_i)$. We assume that K is large enough for the approximation error to be of smaller order, and that K/N tends to zero, as in Corollary 2. Under appropriate conditions in the regression model, using similar arguments to the proof of Corollary S2 (though with no need for any restriction on the relative rates of N and T), one can show that $\widehat{\theta}$ admits the following expansion:

$$\widehat{\theta} = \theta_0 + \frac{\frac{1}{N} \sum_i \zeta_i^X (\zeta_i^\alpha + \zeta_i^U) + \frac{1}{NT} \sum_{i,t} (X_{it} - \overline{X}_i)(U_{it} - \overline{U}_i)}{\mathbb{E}[(X_{it} - \overline{X}_i)^2] + \mathrm{Var}(\zeta_i^X)} + o_p\left(\frac{1}{T}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right). \tag{S26}$$

Notice two differences between (S26) and the expansion of the FE estimator: the presence of $\mathrm{Var}(\zeta_i^X)$ in the denominator, and the presence of $\frac{1}{N} \sum_i \zeta_i^X (\zeta_i^\alpha + \zeta_i^U)$ in the numerator. In addition, notice that (S26) simplifies to the expression in Corollary 2 in the absence of additional heterogeneity ν_{i0}.

---

[6]Although the arguments are as in the proof of Lemma A3, the target log-likelihood is different since here $\overline{\alpha}(\theta, \xi_{i0})$ only depends on ξ_{i0}, not on $(\xi'_{i0}, \nu'_{i0})'$. In particular, the matrix H^{cond} in Assumption S4 differs from the matrix H in Assumption 3; see (S26) for an example.

S3   Simulations

Model of wages and participation (see (2)).   We model the initial condition as: $Y_{i0} = \mathbf{1}\{u(\alpha_{i0}) \geq c(1; \theta_0) + U_{i0}\}$, with U_{i0} standard normal, independent of α_{i0}. We set $c(0; \theta_0) = 0$ and $c(1; \theta_0) = -1$. We set α_{i0} and V_{it} to be independent standard normals. In the simulations based on models (2) and (3) we weight the moments by the share of between-i variance to total variance.[7] To compute the variance $\widehat{V}_h$ to set the number of groups in this dynamic model, we use a Newey-West expression with one lag. Lastly, for kmeans computation we use Lloyd's algorithm with 100 random starting values. Table S1 shows additional simulation results for this model.

Probit model with time-varying heterogeneity (see (3)).   The U_{it}'s are standard normal independent of the X_{it}'s and the α_{it0}'s. The data generating process (DGP) for the scalar covariate is: $X_{it} = \mu_{it0} + V_{it}$, where V_{it} are i.i.d. standard normal independent of the U_{it}'s, α_{it0}'s, and μ_{it0}'s, and $\mu_{it0} = \alpha_{it0}$. We set $\theta_0 = 1$, and set ξ_{i0} and λ_{t0} to be i.i.d. Gamma(1,1) draws, independent of each other. Table S2 shows additional simulation results for this model, including for the two-way GFE estimator based on both the cross-sectional moments $(\frac{1}{N}\sum_i Y_{it}, \frac{1}{N}\sum_i X_{it})'$, and the individual-specific moments $(\overline{Y}_i, \overline{X}_i)'$.

Conditional moments: an example.   Consider the following probit model: $Y_{it} = \mathbf{1}\{X_{it}'\theta_0 + \alpha_{i0} + U_{it} \geq 0\}$, where the U_{it} are i.i.d. standard normal independent of the X_{it}'s and α_{i0}, and θ_0 is a vector of ones. The DGP for the k-th covariate is: $X_{itk} = \mathbf{1}\{\mu_{i0k} + V_{itk} > 0\}$, where V_{itk} are i.i.d. standard normal independent of the U_{it}'s, α_{i0}, and the μ_{i0k}'s, and α_{i0} and the μ_{i0k}'s follow independent standard normals. We vary the number of covariates between 1 and 3, so the total dimension of heterogeneity varies between 2 and 4. In this model, we expect the bias of FE to be moderate given the time horizon we consider ($T = 20$), since α_{i0} is scalar

---

[7]Specifically, we demean and rescale h_i so that all its components $h_{i\ell}$ have zero mean and unit variance, and multiply each component $h_{i\ell}$ by: $\max\left(\frac{\sum_i h_{i\ell}^2 - \frac{1}{T^2}\sum_{i,t}(h_{it\ell} - h_{i\ell})^2}{\sum_i h_{i\ell}^2}, 0\right)$. Using equal weights instead has small effects in these simulations, however we observed that this particular weighting can improve performance when some moments are substantially less informative about the heterogeneity than others.

and FE is a conditional approach. The question we ask here is how much the use of conditional moments can help reduce the bias of GFE due to the presence of additional heterogeneity in the covariates and the increased dimensionality of heterogeneity (see Subsection 4.2).

<u>Consider first</u> using $h_i = (\overline{Y}_i, \overline{X}_i')'$ as moments. In Table S3 we show the biases, standard deviations, and root mean squared errors of FE and GFE among 1000 simulations, for N=1000 and T=20. In the top panel we report GFE estimates as a function of the number of groups K. We see that, while the bias of GFE remains moderate with one covariate, the bias increases substantially with the dimension of heterogeneity, in agreement with our theory. By comparison, the bias of FE in the bottom panel is indeed quite small, and it only increases moderately with the number of covariates.

The situation is rather different when using <u>conditional moments</u> in GFE. In the middle panel in Table S3 we show simulation results for GFE based on covariates-specific conditional means $\overline{Y}_i(x) = \sum_{t=1}^{T} \mathbf{1}\{X_{it} = x\}Y_{it}/\sum_{t=1}^{T} \mathbf{1}\{X_{it} = x\}$. Importantly, in large samples these moments are only informative about α_{i0}, not μ_{i0}. We see that the bias of GFE with conditional moments increases only moderately with the number of covariates, and that FE and GFE with conditional moments have comparable — and quite small — biases.

Regarding implementation, note that, for a given i, all moments $\overline{Y}_i(x)$ may not be available since i's covariates may never take the value x in the sample. In Table S3, whenever $\overline{Y}_i(x)$ is not available, we set the moment to an imputed value, the overall conditional mean $\overline{Y}(x) = \sum_{i,t} \mathbf{1}\{X_{it} = x\}Y_{it}/\sum_{i,t} \mathbf{1}\{X_{it} = x\}$. The imputation does not affect the theory, provided the event that any of the $\overline{Y}_i(x)$'s is not available tends to zero with probability approaching one in large samples.[8] Moreover, we have obtained similar results using an alternative conditional first step implementation that does not rely on imputations.[9]

---

[8]To provide intuition in a simple case, suppose that X_{it} are binary, i.i.d. over time given μ_{i0}, with $\Pr(X_{it} = 1 \mid \mu_{i0} = \mu) \in (\epsilon, 1 - \epsilon)$ for all μ, for some $\epsilon > 0$. Then $\Pr(\exists i : X_{i1} = ... = X_{iT} = 0) \leq N(1 - \epsilon)^T$, which tends to zero whenever $(\ln N)/T \to 0$.

[9]This implementation is as follows. Let $I_i(x)$ be the indicator that there exists a t such that $X_{it} = x$, and let $x_1, ..., x_M$ denote the points of support of X_{it}. In the first step, we use a Lloyd's-like algorithm to minimize the function $\sum_{i=1}^{N} \sum_{m=1}^{M} I_i(x_m) \left(\overline{Y}_i(x_m) - g(x_m, k_i)\right)^2$, with

Table S1: Model (2) of wages and participation

T	Bias	std	RMSE	se/std	Bias	std	RMSE	se/std
	GFE, $\eta = 1$				FE, $\eta = 1$			
5	-0.570	0.058	0.573	1.082	-0.835	0.064	0.837	1.066
10	-0.207	0.040	0.211	1.003	-0.418	0.040	0.420	1.041
20	-0.088	0.027	0.092	0.993	-0.209	0.026	0.211	1.064
30	-0.055	0.023	0.060	0.960	-0.140	0.023	0.142	0.991
40	-0.040	0.019	0.044	1.000	-0.105	0.019	0.106	1.034
50	-0.031	0.017	0.036	0.982	-0.084	0.017	0.086	1.022
	GFE, $\eta = 2$				FE, $\eta = 2$			
5	-0.519	0.063	0.523	1.052	-0.876	0.068	0.879	1.063
10	-0.163	0.043	0.169	0.985	-0.442	0.041	0.444	1.070
20	-0.049	0.031	0.058	0.929	-0.225	0.028	0.227	1.042
30	-0.032	0.024	0.040	0.964	-0.153	0.022	0.154	1.068
40	-0.019	0.020	0.028	0.981	-0.113	0.019	0.115	1.045
50	-0.015	0.019	0.024	0.944	-0.091	0.018	0.093	1.000

Notes: 1000 *simulations,* $N = 1000$*. "RMSE" is root mean squared error, "se" is the average of standard error estimates across simulations, "std" is the standard deviation of the estimator across simulations.* η *is the risk aversion parameter.*

respect to $k_1, ..., k_N$ and $g(x_1, 1), ..., g(x_M, K)$.

Table S2: Probit model (3) with time-varying heterogeneity

σ = −10

T	2-way GFE Bias	std	RMSE	se/std	GFE Bias	std	RMSE	se/std	FE Bias	std	RMSE	se/std	IFE Bias	std	RMSE	se/std
5	−0.045	0.035	0.057	0.927	−0.044	0.035	0.056	0.926	0.442	0.071	0.448	0.706	0.116	0.064	0.133	0.473
10	−0.016	0.024	0.028	0.939	−0.014	0.024	0.028	0.939	0.198	0.036	0.201	0.762	0.100	0.036	0.107	0.488
20	−0.003	0.016	0.016	1.014	−0.000	0.016	0.016	1.014	0.098	0.019	0.100	0.911	0.087	0.020	0.089	0.596
30	−0.000	0.013	0.013	1.009	0.003	0.013	0.013	1.013	0.069	0.014	0.070	0.966	0.059	0.014	0.061	0.675
40	0.001	0.011	0.011	1.021	0.005	0.011	0.012	1.016	0.055	0.012	0.057	0.949	0.044	0.012	0.046	0.676
50	0.001	0.010	0.010	0.995	0.006	0.010	0.012	0.994	0.048	0.011	0.049	0.947	0.036	0.011	0.037	0.677

σ = 0

T	2-way GFE Bias	std	RMSE	se/std	GFE Bias	std	RMSE	se/std	FE Bias	std	RMSE	se/std	IFE Bias	std	RMSE	se/std
5	−0.044	0.042	0.060	0.883	−0.043	0.042	0.060	0.882	0.488	0.091	0.497	0.654	0.152	0.089	0.176	0.398
10	−0.022	0.026	0.035	0.951	−0.021	0.026	0.034	0.949	0.226	0.045	0.231	0.710	0.118	0.040	0.125	0.494
20	−0.009	0.018	0.020	0.969	−0.006	0.018	0.019	0.964	0.108	0.023	0.110	0.843	0.099	0.023	0.101	0.571
30	−0.004	0.014	0.015	1.001	0.001	0.014	0.014	1.000	0.072	0.017	0.074	0.918	0.068	0.017	0.070	0.615
40	−0.002	0.013	0.013	0.989	0.004	0.013	0.013	0.985	0.056	0.014	0.058	0.922	0.051	0.014	0.052	0.639
50	−0.001	0.011	0.012	0.965	0.005	0.012	0.013	0.961	0.046	0.012	0.047	0.926	0.040	0.012	0.042	0.643

σ = 1

T	2-way GFE Bias	std	RMSE	se/std	GFE Bias	std	RMSE	se/std	FE Bias	std	RMSE	se/std	IFE Bias	std	RMSE	se/std
5	−0.049	0.056	0.074	0.754	−0.048	0.056	0.074	0.754	0.565	0.146	0.583	0.506	0.207	0.117	0.238	0.359
10	−0.032	0.029	0.043	0.981	−0.030	0.029	0.042	0.979	0.251	0.062	0.258	0.603	0.141	0.043	0.147	0.513
20	−0.014	0.020	0.024	0.986	−0.010	0.020	0.022	0.983	0.114	0.027	0.117	0.825	0.125	0.029	0.128	0.514
30	−0.007	0.016	0.017	0.996	−0.001	0.016	0.016	0.992	0.074	0.019	0.077	0.889	0.085	0.021	0.088	0.561
40	−0.005	0.014	0.015	0.985	0.001	0.014	0.014	0.980	0.055	0.016	0.057	0.915	0.063	0.016	0.065	0.611
50	−0.003	0.012	0.012	1.027	0.004	0.012	0.013	1.025	0.044	0.014	0.046	0.951	0.050	0.014	0.052	0.632

σ = 10

T	2-way GFE Bias	std	RMSE	se/std	GFE Bias	std	RMSE	se/std	FE Bias	std	RMSE	se/std	IFE Bias	std	RMSE	se/std
5	−0.016	0.075	0.076	0.691	−0.015	0.075	0.076	0.692	0.706	0.262	0.753	0.386	0.300	0.255	0.394	0.218
10	−0.013	0.035	0.037	0.946	−0.010	0.035	0.036	0.947	0.323	0.097	0.337	0.486	0.183	0.060	0.192	0.458
20	−0.002	0.024	0.024	0.975	0.003	0.024	0.024	0.967	0.150	0.037	0.154	0.719	0.168	0.036	0.172	0.474
30	0.002	0.019	0.019	0.991	0.008	0.019	0.021	0.983	0.100	0.025	0.104	0.814	0.121	0.030	0.125	0.456
40	0.003	0.016	0.016	0.989	0.010	0.016	0.019	0.985	0.076	0.020	0.079	0.851	0.091	0.021	0.093	0.536
50	0.002	0.014	0.015	0.995	0.010	0.014	0.018	0.996	0.061	0.017	0.063	0.911	0.073	0.017	0.075	0.592

Notes: See notes to Table S1. IFE is interacted fixed-effects with one factor. σ is the substitution parameter.

Table S3: Probit model with binary covariates

K	Bias	std	RMSE	Bias	std	RMSE	Bias	std	RMSE
	GFE, 1 covariate			GFE, 2 covariates			GFE, 3 covariates		
5	-0.189	0.029	0.191	-0.293	0.031	0.295	-0.362	0.042	0.365
10	-0.083	0.027	0.088	-0.205	0.032	0.207	-0.275	0.035	0.278
20	-0.017	0.029	0.033	-0.118	0.030	0.122	-0.206	0.033	0.209
30	0.006	0.029	0.030	-0.081	0.030	0.086	-0.166	0.032	0.169
40	0.018	0.029	0.035	-0.056	0.030	0.064	-0.136	0.033	0.140
50	0.026	0.030	0.039	-0.040	0.031	0.051	-0.116	0.033	0.120
	Cond. GFE, 1 covariate			Cond. GFE, 2 covariates			Cond. GFE, 3 covariates		
5	-0.060	0.035	0.069	-0.085	0.037	0.093	-0.111	0.039	0.117
10	-0.045	0.033	0.056	-0.073	0.043	0.085	-0.100	0.044	0.109
20	-0.015	0.034	0.038	-0.046	0.037	0.059	-0.075	0.045	0.087
30	0.008	0.036	0.036	-0.031	0.036	0.047	-0.061	0.043	0.075
40	0.025	0.035	0.043	-0.020	0.037	0.041	-0.050	0.042	0.065
50	0.034	0.035	0.049	-0.012	0.036	0.038	-0.040	0.041	0.057
	FE, 1 covariate			FE, 2 covariates			FE, 3 covariates		
-	0.062	0.031	0.069	0.074	0.034	0.081	0.088	0.039	0.097

Notes: 1000 simulations, $N = 1000$, $T = 20$. In the top panel we show GFE estimates based on unconditional moments for different K values, in the middle panel we show GFE estimates based on conditional moments for different K values, in the bottom row we show FE estimates.