Discretizing Unobserved Heterogeneity*

Stéphane Bonhomme† Thibaut Lamadon‡ Elena Manresa§

Revised draft: August 2019

Abstract

We study panel data estimators based on a discretization of unobserved heterogeneity when individual heterogeneity is not discrete in the population. We focus on *two-step grouped fixed-effects* (GFE) estimators, where individuals are classified into groups in a first step using *kmeans* clustering, and the model is estimated in a second step allowing for group-specific heterogeneity. In addition to reducing the number of parameters in estimation, GFE methods can allow for rich, time-varying forms of heterogeneity, under the assumption that individual heterogeneity is low-dimensional. We analyze the asymptotic properties of two-step GFE estimators as the number of groups grows with the sample size. We document their finite sample performance in a structural dynamic discrete choice model of migration, and in several specifications of a probit model with individual heterogeneity.

JEL codes: C23, C38.

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

*We thank the co-editor and four anonymous referees, Anna Simoni, Manuel Arellano, Jesus Carro, Gary Chamberlain, Tim Christensen, Alfred Galichon, Chris Hansen, Joe Hotz, Guido Imbens, Grégory Jolivet, Arthur Lewbel, Anna Mikusheva, Roger Moon, Whitney Newey, Juan Pantano, Philippe Rigollet, Martin Weidner, and seminar audiences at the 2016 Summer Meetings of the Econometric Society, the 2016 Panel Data Conference in Perth, the 2016 IO/Econometrics Cornell/Penn State Conference, and various other places for comments. The authors acknowledge support from the NSF grant number SES-1658920. The usual disclaimer applies.

†University of Chicago, sbonhomme@uchicago.edu

‡University of Chicago, lamadon@uchicago.edu

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

1 Introduction

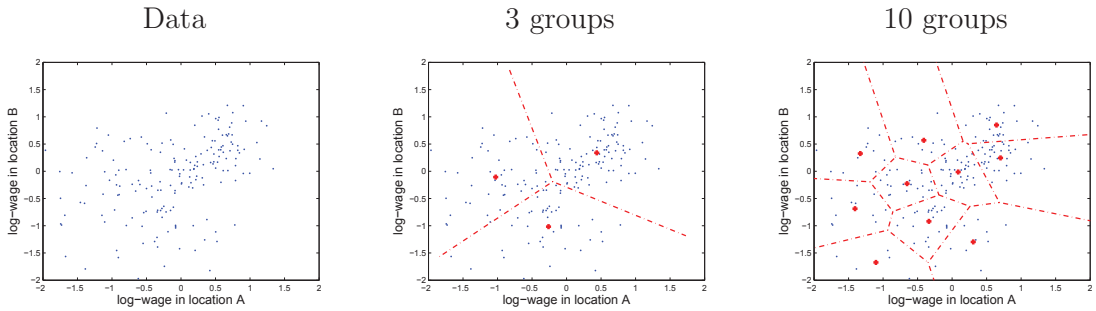
Unobserved heterogeneity is prevalent in modern economics, both in reduced-form and structural work, and accounting for it often makes large quantitative differences. In nonlinear panel data models, fixed-effects approaches are conceptually attractive as they do not require restricting the form of unobserved heterogeneity. These nonlinear fixed-effects methods are well understood from a theoretical perspective (e.g., Hahn and Newey, 2004, Arellano and Hahn, 2007). However, they involve large numbers of parameters, and face challenges in the presence of multiple individual unobservables such as time-varying heterogeneity.

Discrete approaches offer tractable alternatives. Consider as an example structural dynamic discrete choice models, which are popular in labor economics and industrial organizations. Starting with Keane and Wolpin (1997), numerous papers have modeled individual heterogeneity as a small number of unobserved types. In this context, discreteness is appealing for estimation since it leads to a finite number of unobserved state variables and reduces the number of parameters to estimate. However, the properties of discrete estimators have so far only been studied under particular restrictions on the form of heterogeneity, typically under discreteness. In this paper we consider a class of easy-to-implement discrete estimators, and we study their properties in nonlinear models without imposing that individual unobserved heterogeneity is discrete or independent of covariates in the population.

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. Then, in a second step, we estimate the model by allowing for group-specific heterogeneity. The aim of the kmeans classification is to group together individuals whose latent types are most similar. Kmeans is a popular tool that has been extensively used and studied in machine learning and computer science, and fast and reliable implementations are available. Classifying individuals into types using kmeans is related to the GFE estimators recently introduced by Hahn and Moon (2010) and Bonhomme and Manresa (2015). However, unlike those methods, and unlike random-effects methods such as finite mixtures, in our sequential approach we avoid jointly estimating the individual types and the model's parameters.

Figure 1 provides an illustration of the first step in a migration setting. According to the dynamic location choice model that we will describe in detail in Section 5, log-wages are informative about unobserved individual returns in locations A and B. We classify individuals into groups using location-specific means of log-wages. Depending on the number of groups K, the kmeans algorithm will deliver different partitions of individuals. Taking $K = 3$ will

Figure 1: Kmeans clustering



Notes: Source NLSY79. The sample is described in Section 5. The kmeans partitions are indicated in dashed.

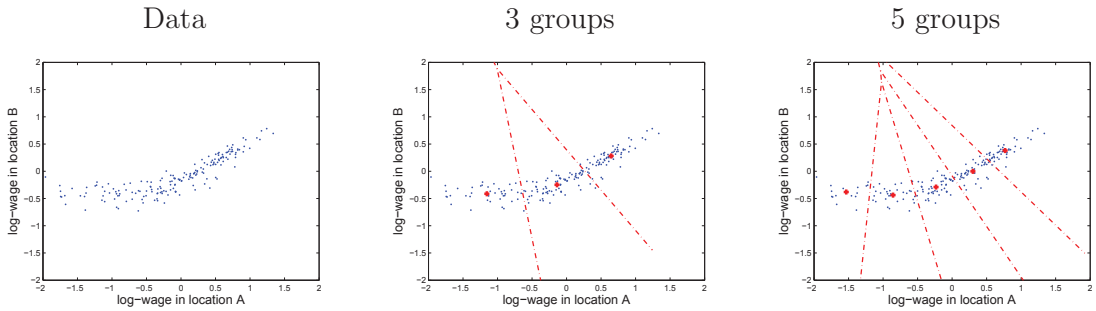
result in a drastic reduction in the number of parameters, however the approximation to the latent heterogeneity may be too coarse. Taking a larger K, such as $K = 10$, may reduce approximation error while still substantially reducing the number of parameters to estimate relative to fixed-effects.

We analyze the properties of two-step GFE estimators under two main assumptions. *First*, we assume that unobserved heterogeneity depends on a low-dimensional vector of latent types. The types can be continuous. Hence, we use discrete heterogeneity as a dimension reduction device, rather than viewing discreteness as a substantive assumption about population unobservables. This is an important difference relative to the literature.

Our setup covers models with rich, time-varying heterogeneity, in addition to more standard models with time-invariant individual effects. In time-varying settings, our first assumption implies that cross-sectional heterogeneity is low-dimensional. However, unlike methods that explicitly incorporate time effects such as linear factor models and models with interactive fixed-effects (e.g., Bai, 2009, Pesaran, 2006), here we do not specify the mapping between the underlying types and the heterogeneity in the model, and we do not impose a factor structure in estimation.

In many economic models, agents' heterogeneity in preferences and technology is driven by low-dimensional economic types, which manifest themselves in potentially complex ways in the data. Through the use of kmeans, and in contrast with fixed-effects methods, GFE provides a tool to exploit such nonlinear factor structures. To illustrate, consider Figure 2, where in this example log-wages in the two locations are closely related to each other and approximately lie on a curve. Such a structure could arise from the presence of a one-dimensional ability factor,

Figure 2: Kmeans in the presence of a low underlying dimension



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

for example. The kmeans-based partition efficiently adapts to the data structure in a way that guarantees a low error of approximation.

Our *second* main assumption is that the researcher has individual-specific moments from which the underlying types can be approximated. Moments may take the form of external measurements of the types, such as measures of individual skills or firm productivity. When such information is not available, we show that moments can be constructed from the panel data. Indeed, individual-specific averages of outcomes and covariates depend on the distribution of heterogeneity, and they will be functions of the latent types in large samples. For example, in structural models, moments that are internal to the model can be based on choices, state variables, or payoffs.

Recovering types from moments requires an injectivity condition. A population moment is a function of the latent types. Injectivity requires that any two individuals with the same population moments have the same type. Verifying that a given moment vector satisfies injectivity is conceptually related to verifying that a parameter is identified. Similarly to the case of identification, checking that moments satisfy injectivity may not be obvious. We discuss the choice of moments in various models. In static models, we show how injectivity can be guaranteed, given an identification assumption, by using individual-specific distributions of observables as moments.

In practice, moments are estimated with noise. The performance of GFE depends on the ability to estimate the moments with enough precision. In settings with internal moments that are constructed from the panel, this requires the number of periods T to be sufficiently large. Hence, similarly to nonlinear fixed-effects estimators, consistency of GFE requires that T tends

to infinity together with the number of individual units N.

Our main theoretical result is an expansion of the two-step GFE estimator, which highlights the presence of two kinds of biases. First, estimating group membership indicators and group-specific parameters contributes to an incidental parameter bias that is related to the bias affecting fixed-effects estimators. We use the half-panel jackknife method of Dhaene and Jochmans (2015) to reduce this source of bias. Second, since the discrete approximation of population heterogeneity may be imperfect, two-step GFE is affected by approximation error. As a result, consistency requires the number of groups K to grow with the sample size. We propose a simple data-driven rule for K that controls the size of the approximation error.

We provide simple guidelines to implement two-step GFE. Given moments such as means or other characteristics of individual data, we use the kmeans algorithm to estimate the number of groups and the partition of individuals into groups. Given those, we estimate the model's parameters while allowing for group-specific fixed-effects. We can then readily obtain bias-reduced estimates by repeating the same procedure on two halves of the sample. Finally, standard errors of bias-reduced estimators can be computed using formulas that we provide.

In order to characterize the size of the approximation error, we use results from the literature on vector quantization, see among others Gersho and Gray (1992), Gray and Neuhoff (1998), and Graf and Luschgy (2000, 2002). The approximation error increases with the dimension d of the latent types. When d is small, for instance when types are one- or two-dimensional, a moderate number of groups K will suffice to guarantee a low approximation error. In models with time-invariant heterogeneity, GFE then has an asymptotic performance similar to fixed-effects, while having a smaller number of parameters. Moreover, GFE can dominate fixed-effects in settings with richer heterogeneity, for example when unobservables vary over time.

When the dimension d is large, approximation error tends to deteriorate the performance of two-step GFE. A relevant situation that may lead to a large d is when there are multiple covariates whose distribution is individual-specific. We propose two methods to reduce approximation error. The first one is based on a conditional first step that explicitly accounts for the presence of conditioning covariates. The second one is an iteration that relies on the structure of the model. Both methods involve relatively low additional computational cost compared to two-step GFE, and we report simulations suggesting they may improve performance in settings where d is moderately large. In addition, while we do not provide a full statistical analysis of these extensions, we derive a rate of convergence for the conditional first step that demonstrates that it is subject to a smaller approximation error than the baseline two-step method.

When applying GFE to settings with time-varying unobservables, we group cross-sectional

heterogeneity. As a result, the number of time-specific parameters may be large, affecting the properties of the estimator. Hence we introduce a two-way GFE estimator where, in addition to the moments to classify individuals, we use aggregate moments to classify time periods. Notice that "time" is simply a label, and one can use two-way GFE in settings where, for example, unobservables vary across products and markets, and group both products and markets in the first step. We analyze the asymptotic properties of two-way GFE in static models, under the assumption that individual and time heterogeneity have a low dimension.

We illustrate the behavior of GFE estimators in a structural dynamic discrete choice model of migration. In this context, two-step methods provide an alternative to finite mixtures and related approaches, such as the ones developed in Arcidiacono and Jones (2003) and Arcidiacono and Miller (2011), for example. We set up a simulation exercise based on estimates from a simple dynamic model of location choice in the spirit of Kennan and Walker (2011), which we estimate on NLSY data.

We consider two data generating processes where unobserved heterogeneity is continuous, and assess the magnitude of the biases of GFE estimators and the performance of bias reduction. In the first specification, mobility costs are constant across individuals, so we can compare GFE to fixed-effects. We find that the two methods perform similarly. In addition, jackknife bias reduction and our model-based iteration tend to improve GFE performance. In the second specification, costs and returns are both heterogeneous but depend on a one-dimensional factor within each location. In this case, fixed-effects is no longer feasible since individual-specific costs can only be estimated for individuals who move during the sample period. In contrast, we find that GFE still performs well. This demonstrates the ability of GFE to take advantage of commonalities between different dimensions of heterogeneity, without having to rely on exact discreteness in the population.

We conclude with a simulation study based on a probit model, where the dimension d of heterogeneity increases with the number of covariates. In line with our theory, we find that two-step GFE performs well when d is small ($d = 1, 2$), but shows substantial bias when d is larger ($d = 4, 6$). Interestingly, iterated GFE and conditional first step methods tend to significantly improve performance in those cases. We find that GFE exhibits similar patterns in two specifications with richer heterogeneity: a model with random coefficients that depend on a scalar latent type, and a time-varying one-factor specification (that we do not exploit in estimation). By comparison, the performance of fixed-effects deteriorates in the random coefficients model, and fixed-effects is not feasible in the model with time-varying heterogeneity.

This paper is related to discrete random-effects approaches, which have been studied under

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

Previously to this paper, the properties of GFE estimators have been characterized under the assumption that unobserved heterogeneity is discrete in the population. Under suitable conditions, estimated types converge to the true population types as both dimensions of the panel increase; see Hahn and Moon (2010), Lin and Ng (2012), Saggio (2012), Bonhomme and Manresa (2015), Bai and Ando (2016), Su, Shi and Phillips (2016), and Vogt and Linton (2016). In the context of structural dynamic discrete choice estimation, also under a discrete population framework, Buchinsky, Hahn and Hotz (2005) propose to classify types based on kmeans clustering and perform the Hotz and Miller (1993) estimation strategy using the estimated types. Pantano and Zheng (2013) use a related approach based on subjective expectations data.

However, there has been little work studying properties of discrete estimators as the number of groups tends to infinity with the sample size. Important exceptions are Bester and Hansen (2016) and Frederiksen, Honoré and Hu (2007), who focus on setups with known groups, and Gao, Lu and Zhou (2015) and Wolfe and Ohlede (2014), who study stochastic blockmodels in networks.

The outline of the paper is as follows. We introduce the setup and two-step GFE estimators in Section 2. We study their asymptotic properties in Section 3. In Section 4 we present several extensions. In Section 5 we report simulations based on a structural dynamic discrete choice model, and on several specifications of a probit model. We conclude in Section 6. The proofs may be found in Appendix A, and a supplementary appendix contains additional results.

2 Two-step grouped fixed-effects

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 $i = 1, ..., N$. We denote the conditional density of Y_i given X_i, with respect to some measure, as $f_i(\alpha_{i0}, \theta_0)$, where the α_{i0} are individual-specific vectors and θ_0 is a vector of common parameters. We are interested in estimating θ_0, as well as average effects depending on the α_{i0}. In the asymptotic analysis we will let both N and T tend to infinity. We defer the formal presentation of regularity conditions

until the next section.

We focus on conditional densities that take the following form:

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

where α_{it0} may vary over time, and $\alpha_{i0} = (\alpha'_{i10}, ..., \alpha'_{iT0})'$. In models with first-order dependence we assume that Y_{i0} is observed and we condition on it, so the researcher has $T + 1$ periods of data. Higher-order dependence can be accommodated similarly. In dynamic settings, Y_{it} may contain sequentially exogenous covariates in addition to outcome variables.

We consider densities of exogenous covariates of the form:

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

where $\mu_{i0} = (\mu'_{i10}, ..., \mu'_{iT0})'$ are individual-specific. 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 dimension of μ_{i0} will play an important role when studying the properties of two-step GFE.

2.1 Main assumptions

Our first main assumption is that α_{it0} and μ_{it0}, $t = 1, ..., T$, depend on a low-dimensional vector ξ_{i0}. This assumption, which requires that individual unobservables have a low "underlying dimension", is key to the performance of GFE.

Assumption 1. *(underlying dimension) There exist vectors ξ_{i0} of dimension d, vectors λ_{t0} of dimension d_λ, and two functions α and μ, such that $\alpha_{it0} = \alpha(\xi_{i0}, \lambda_{t0})$ and $\mu_{it0} = \mu(\xi_{i0}, \lambda_{t0})$.*

In our theory we cover three cases. A first scenario is when $\alpha_{it0} = \alpha_{i0}$ and $\mu_{it0} = \mu_{i0}$ are vectors of time-invariant fixed-effects. In this case we can abstract from λ_{t0} without loss of generality. While the GFE approach can be used in such models, it may be particularly useful in settings with richer heterogeneity. Specifically, we cover two other scenarios: when α_{it0} and μ_{it0} may vary over time unrestrictedly, and when they are constant within sub-periods but can vary unrestrictedly between sub-periods. Variation in unobservables over time (e.g., over the business cycle) or age (over the life cycle) is of interest in many applications. Moreover, t could denote labor or product markets, for example. In such cases, it may be appealing to allow unobservables to vary between counties or MSA.

In models with time-varying unobserved heterogeneity, Assumption 1 requires that the unobservables follow a factor structure. This structure may be linear, as in the interactive specification $\alpha_{it0} = \xi_{i0}'\lambda_{t0}$ considered in Bai (2009) and Pesaran (2006), among many others. More generally, the link between α_{it0}, ξ_{i0} and λ_{t0} may be nonlinear, and we will not need to impose it in estimation. However, it is crucial for good performance of GFE that the factor loadings ξ_{i0} have a low dimension; say, $d = 1$ or $d = 2$. One can think of ξ_{i0} as a vector of individual "types". We will refer to d as the "underlying dimension", or simply "dimension", of heterogeneity. In contrast, the dimension d_λ of the vector λ_{t0} of latent time effects will not appear directly in the convergence rate of the estimator. Hence, this setup is one where cross-sectional heterogeneity is low-dimensional, and time heterogeneity can be general. Lastly, the researcher will not need knowledge of d in applications.

In GFE estimation we rely on individual-specific moment vectors h_i that are informative about the vectors of types ξ_{i0}. We now state our second main assumption, where S denotes the precision of h_i; that is, S is the number of observations that we use to construct h_i. Here and in the following, $\|\cdot\|$ denotes an Euclidean norm.

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

The injectivity condition in Assumption 2 requires the individual moment h_i to be informative about ξ_{i0}, in the sense that, for large S, ξ_{i0} can be uniquely recovered from h_i. Intuitively, injectivity will guarantee that one can separate the types of two individuals ξ_{i0} and $\xi_{i'0}$ simply by using their moments h_i and $h_{i'}$. Note that neither φ not ψ need to be known by the econometrician. Injectivity is a key requirement for consistency of two-step GFE estimators. More generally, the choice of moments h_i is important in order to ensure good performance.

Moments can be internal or external to the model. External moments are available when the researcher has a set of measures on some latent skills or traits of the individual, for example. Alternatively, and this is the leading case in this paper, when T is large enough one can construct internal moments using the outcomes and covariates in the panel, by taking $S = T$. Indeed, an average $h_i = \frac{1}{T}\sum_{t=1}^T h(Y_{it}, X_{it})$ of functions of outcomes and covariates will, under Assumption 1 and suitable regularity conditions, converge as T tends to infinity to a function $\varphi(\xi_{i0})$ of the type ξ_{i0}. The function φ will generally depend on θ_0. Moreover, the convergence rate in Assumption 2 will hold under appropriate conditions on h and the serial dependence of Y_{it} and X_{it}. In models with time-varying heterogeneity, φ will also depend on the distribution of λ_{t0}, and we will need suitable conditions on the dependence properties of λ_{t0}.

2.2 Estimator

Two-step GFE consists of a *classification* step and an *estimation* step, which we now describe.

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

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

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

The optimization problem in (2) is referred to as *kmeans* in machine learning and computer science. In (2) the minimum is taken with respect to all possible partitions $\{k_i\}$. Computing a global minimum may be challenging, yet fast and stable heuristic algorithms have been developed, such as iterative descent, genetic algorithms, or variable neighborhood search.[1] Lloyd's algorithm is often considered to be a simple and reliable benchmark when initialized using a randomly generated set of starting values.

Algorithm 1. *(Lloyd's algorithm for kmeans)*

- *Given initial values for* $\widetilde{h}(1), ..., \widetilde{h}(K)$*, iterate between the following two steps until convergence:*
- *Given* $\widetilde{h}(1), ..., \widetilde{h}(K)$*, compute* $k_i = \operatorname{argmin}_{k=1,...,K} \|h_i - \widetilde{h}(k)\|$ *for all* i*.*
- *Given* $k_1, ..., k_N$*, compute* $\widetilde{h}(k) = \sum_{i=1}^{N} \mathbf{1}\{k_i = k\} h_i / \sum_{i=1}^{N} \mathbf{1}\{k_i = k\}$ *for all* k*.*

In the asymptotic analysis, following the statistical literature on kmeans clustering since Pollard (1981, 1982), we will focus on the properties of the global minimum in (2) and abstract from optimization error. Note that, while we focus on an unweighted version of kmeans, the quadratic loss function in (2) can accommodate different weights on different components of h_i, for example based on inverse variances. Our theoretical results will still apply in this case, provided that the (normalized) weights be bounded away from zero and one.

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

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

$$\left(\widehat{\theta}, \widehat{\alpha}\right) = \underset{(\theta,\alpha)}{\text{argmax}} \sum_{i=1}^{N} \ln f_i \left(\alpha\left(\widehat{k}_i\right), \theta\right), \tag{3}$$

where the maximization is with respect to θ and $\alpha = (\alpha(1)', ..., \alpha(K)')'$, with $\alpha(k)$ being parameter vectors. Note that, in contrast with fixed-effects maximum likelihood, this second step involves a maximization with respect to K vectors of individual effects instead of N. 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 alone. Let:

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

denote the value of the kmeans objective function corresponding to K groups, where for conciseness we have not indicated the dependence of $\widehat{h}$ and $\widehat{k}_i$ on K. We suggest setting:

$$\widehat{K} = \min_{K \geq 1} \left\{ K \,:\, \widehat{Q}(K) \leq \gamma \widehat{V}_h \right\}, \tag{4}$$

where $\widehat{V}_h = \mathbb{E}[\|h_i - \varphi(\xi_{i0})\|^2] + o_p(1/S)$, and $\gamma > 0$ is a user-specific parameter. We recommend setting $\gamma = 1$ as a default, and checking how GFE estimates vary when taking $\gamma < 1$; that is, when using a larger K.

The intuition for this choice is that, for $K = \widehat{K}$ given by (4) with $\gamma = 1$, the within-group variation in h_i is of the same order as the noise level $\widehat{V}_h$. When h_i is a mean of S independent measurements h_{is}, one may take $\widehat{V}_h = \frac{1}{NS^2} \sum_{i=1}^{N} \sum_{s=1}^{S} \|h_{is} - h_i\|^2$. When there is dependence among the measurements, trimming or bootstrap strategies may be used to construct a suitable $\widehat{V}_h$ (see Hahn and Kuersteiner, 2011, or Arellano and Hahn, 2016).

2.3 Examples

In the next section we will present asymptotic theory for two-step GFE estimators. Then, in Section 4 we will analyze several extensions of the baseline approach. Before turning to the theory, we first describe two examples of applications.

Example 1: dynamic discrete choice model. A prototypical structural dynamic discrete choice model features the following elements (see for example Aguirregabiria and Mira, 2010): choices $j_{it} \in \{1, ..., J\}$, payoff variables Y_{it}, and observed and unobserved state variables X_{it} and α_i, respectively. As an example, in the location choice model of Section 5, j_{it} is location at time t, and log-wages Y_{it} depend on latent location-specific returns $\alpha_i(j_{it})$. The individual log-likelihood function conditional on initial choices and state variables typically takes the form:

$$\ln f_i(\alpha_i, \theta) = \sum_{t=1}^{T} \underbrace{\ln f\left(j_{it} \,|\, X_{it}, \alpha_i, \theta\right)}_{\text{choices}} + \underbrace{\ln f\left(X_{it} \,|\, j_{i,t-1}, X_{i,t-1}, \alpha_i, \theta\right)}_{\text{state variables}} + \underbrace{\ln f\left(Y_{it} \,|\, j_{it}, X_{it}, \alpha_i, \theta\right)}_{\text{payoff variables}}. \tag{5}$$

Note that, in such models, the law of motion of observed covariates (the state variables X_{it}) is parametrically specified given α_i. Hence, the same "type" α_i governs the heterogeneity in choices, states, and payoffs.

Computing choice probabilities $f\left(j_{it} \,|\, X_{it}, \alpha_i, \theta\right)$ in (5) requires solving the dynamic optimization problem, which can be demanding. Using two-step GFE, we will estimate a partition $\{\widehat{k}_i\}$ in a first step that does not require solving the model. In the second step, we will take the partition $\{\widehat{k}_i\}$ as given, and maximize the log-likelihood in (5) with respect to θ and type-specific parameters $\alpha(k)$. Note that this two-step method contrasts with iterative methods for random-effects estimation, where parameter updates are iterated until convergence to a deterministic solution (as in the EM algorithm) or to a stationary regime (as in Markov Chain Monte Carlo).

In some economic settings, external moments h_i may be available. For example, in structural models of the labor market the researcher may have access to measures of academic ability or some dimensions of skills (e.g., cognitive or non-cognitive, as in Cunha *at al.*, 2010), such as test scores or psychometric measures taken before the individual entered the labor market. In the absence of such external moments, and provided the panel is sufficiently long, one can construct moments h_i based on payoff variables, observed state variables, and choices. By construction, in model (5), for large T all those moments will be functions of α_{i0}. The key requirement will then be that the injectivity condition in Assumption 2 be satisfied. In the application in Section 5 we will use means of log-wages in a first step for two-step estimation. We will also rely on a likelihood-based iteration that exploits the full model's structure, hence using information on choices.

In other settings, the researcher may wish to account for time-varying heterogeneity α_{it}. For example, unobserved ability or human capital may evolve over the life cycle (or stages of

the life-cycle), or unobserved product characteristics may vary between markets (or between some aggregate markets but not within). Such extensions are readily implemented using GFE.

Example 2: probit model. Consider the following probit model:

$$\left\{ \begin{array}{rcl} Y_{it} & = & \mathbf{1}\left\{ X_{it}'\theta_0 + \alpha_{i0} + U_{it} > 0 \right\}, \\ X_{it} & = & \mu_{i0} + V_{it}, \end{array} \right. \tag{6}$$

where U_{it} are i.i.d. standard normal, independent of $X_i, \alpha_{i0}, \mu_{i0}$, and V_{it} are i.i.d. independent of all U_{it}'s, X_i, α_{i0}, and μ_{i0}. Given a partition of $\{1, ..., N\}$ into K groups, in the second step we will run a probit regression of Y_{it} on X_{it} and group indicators.

Consider the moment vectors $h_i = (\overline{Y}_i, \overline{X}_i')'$, where $\overline{Z}_i = \frac{1}{T}\sum_{t=1}^{T} Z_{it}$ denote individual means of Z_{it}. We have:

$$h_i = \left(\begin{array}{c} G_{\theta_0}\left(\alpha_{i0} + \mu_{i0}'\theta_0\right) \\ \mu_{i0} \end{array} \right) + O_p\left(\frac{1}{\sqrt{T}}\right),$$

where G_{θ_0} is the cumulative distribution function of $-(U_{it} + V_{it}'\theta_0)$. It is easy to see that Assumption 2 holds, with $S = T$, provided G_{θ_0} is strictly increasing over the support of $\alpha_{i0} + \mu_{i0}'\theta_0$. However, injectivity would generally fail if one were to use only $\overline{Y}_i$ as moments. Note that, in this model with exogenous covariates, the underlying dimension d in Assumption 1 is that of the type governing both α_{i0} and μ_{i0}, so d will often increase with the number of covariates. We will return to this issue in Section 4.

Model (6) can be modified in several ways that can be implemented using GFE. One modification is to allow $\theta_{i0} = \theta(\xi_{i0})$ to be a function of the underlying type. Another modification is to allow $\alpha_{it0} = \alpha(\xi_{i0}, \lambda_{t0})$ and $\mu_{it0} = \mu(\xi_{i0}, \lambda_{t0})$ to vary over time. In Section 5 we will study the performance of GFE estimators on data simulated from different versions of the probit model.

3 Asymptotic properties of two-step GFE

In this section we study some asymptotic properties of two-step GFE. We consider the two steps in turn.

3.1 First step: classification

Our first result is a rate of convergence for the kmeans estimator $\widehat{h}(\widehat{k}_i)$ in (2). Let us define the following quantity:

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

where, similarly to (2), the minimum is taken with respect to all partitions $\{k_i\}$ and vectors $\widetilde{\xi}(k)$. The term $B_\xi(K)$ represents the approximation error one would make if one were to discretize the latent types ξ_{i0} directly. It is a non-increasing function of K. Moreover, as we will discuss below, the rate at which $B_\xi(K)$ tends to zero depends crucially on the dimension d of ξ_{i0}.

We have the following characterization of the convergence rate of kmeans. In the asymptotic analysis we let $S = S_N$ and $K = K_N$ tend to infinity jointly with N.

Lemma 1. *Let Assumption 2 hold, where h_i has fixed dimension and φ is Lipschitz-continuous. Then, as N, S, K tend to infinity we have:*

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

Lemma 1 provides an upper bound on the rate of convergence of the discrete estimator $\widehat{h}(\widehat{k}_i)$ of $\varphi(\xi_{i0})$. The bound has two terms: an $O_p(1/S)$ term that depends on the number of measures used to construct the moments h_i, and an $O_p\left(B_\xi(K)\right)$ term that reflects the presence of an approximation error. Lemma 1 will be instrumental in deriving the asymptotic properties of GFE estimators in the next subsection.

The approximation error in (7) has been extensively studied in the literature on vector quantization, where it is referred to as the "empirical quantization error". Graf and Luschgy (2002) provide explicit characterizations in the case where ξ_{i0} has compact support. Specifically, in their Theorem 5.3 they show the following result.[2]

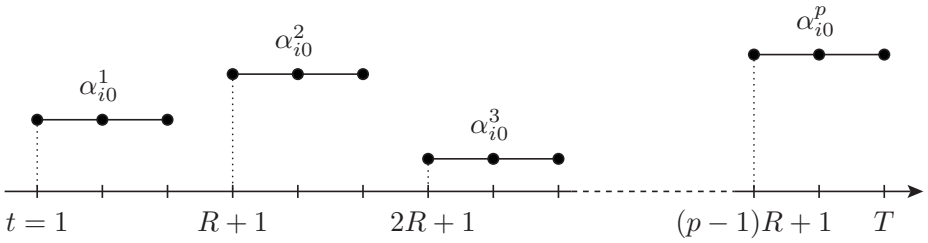
Lemma 2. *(Graf and Luschgy, 2002) Suppose that ξ_{i0} are random vectors with a distribution whose support is compact in $\mathbb{R}^d$. Then, as N, K tend to infinity we have:*

$$B_\xi(K) = O_p(K^{-\frac{2}{d}}).$$

Lemma 2 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, for example. Importantly, Lemmas 1 and 2 imply that the relevant dimension is that of the type ξ_{i0}, not the number of moments that we use in estimation. In other words, when φ is Lipschitz-continuous, the approximation error depends on

[2]While results on empirical quantization errors have been derived in the large-N limit under general conditions (see for example Theorem 6.2 in Graf and Luschgy, 2000), rates as N and K tend to infinity jointly are so far limited to distributions with compact support; see Graf and Luschgy (2002, p.875) for a discussion. We will not impose compact support in our simulations in Section 5.

Figure 3: Time-varying unobservables



the *underlying* dimension of $\varphi(\xi_{i0})$, not on its actual dimension. When the dimension of ξ_{i0} is low, the approximation error may still be relatively small for moderate K.

Finally, in the following corollary we establish a bound on the within-group mean squared error of the unobserved heterogeneity.

Corollary 1. *Let the conditions of Lemmas 1 and 2 hold, and Assumptions 1 and 2 hold, with ψ Lipschitz-continuous, and α and μ Lipschitz-continuous in their first argument. We have:*

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\alpha_{it0}-\overline{\alpha}_{t0}(\widehat{k}_i)\right\|^2 = O_p\left(\frac{1}{S}\right)+O_p(K^{-\frac{2}{d}}), \tag{8}$$

where $\overline{\alpha}_{t0}(k)$ denotes the mean over i of α_{it0} in group $\widehat{k}_i = k$. A similar rate holds for μ_{it0}.

3.2 Second step: estimation

To characterize the asymptotic properties of two-step GFE estimators we start by introducing some notation. Let p denote the number of values that α_{it0} can take over time, and let α_{i0}^j, $j = 1, ..., p$, denote those values. We assume that there are $T = pR$ periods, and that $\alpha_{it0} = \alpha_{i0}^1$ when $t \in \{1, ..., R\}$, $\alpha_{it0} = \alpha_{i0}^2$ when $t \in \{R+1, ..., 2R\}$, ..., and $\alpha_{it0} = \alpha_{i0}^p$ when $t \in \{(p-1)R+1, ..., pR\}$. We illustrate the timing in Figure 3.

For example, when $p = 1$ and $R = T$ the model allows for time-invariant unobserved heterogeneity, while when $p = T$ and $R = 1$ the model allows for unrestricted time variation in unobservables. With intermediate values of p and R, the model allows for unobservables that vary between sub-periods but not within. For simplicity, we consider a balanced structure where T is identical for all individuals, and all sub-periods have exactly R observations.

We define the average log-likelihood function for individual i on subperiod j, and the average

log-likelihood for i over sub-periods, respectively, as:

$$\ell_{ij}(\alpha_i^j, \theta) = \frac{1}{R} \sum_{t=(j-1)R+1}^{jR} \ln f(Y_{it} \,|\, Y_{i,t-1}, X_{it}, \alpha_i^j, \theta), \quad \text{and} \quad \ell_i(\alpha_i, \theta) = \frac{1}{p} \sum_{j=1}^{p} \ell_{ij}(\alpha_i^j, \theta).$$

Note that $\ell_i(\alpha_i, \theta) = (1/T) \ln f_i(\alpha_i, \theta)$, where $\ln f_i$ is given by (1).

Throughout the analysis, we use the shorthand notation $\mathbb{E}_{Z_i}(W_i)$ for the conditional expectation of W_i given Z_i, and $\mathbb{E}_{Z_i=z}(W_i)$ for the conditional expectation of W_i given $Z_i = z$. $\mathbb{E}(W_i)$ simply denotes the unconditional expectation of W_i. We denote as λ_0 the process $\{\lambda_{t0},\, t \in \mathbb{R}\}$, and we use the notation $\mathbb{E}_{\lambda_0=\lambda}(W_i)$ to denote the conditional expectation of W_i given a realization λ of λ_0. We use similar notations for variances. Moreover, unless there is some risk of confusion, for conciseness we omit "almost surely" statements throughout, and we simply write $\inf_\theta$ instead of $\inf_{\theta \in \Theta}$ and use a similar notation for suprema, minima and maximas. Finally, when A is a matrix, $\|A\|$ denotes the spectral norm of A.

We now state and discuss all the regularity conditions for our main theorem.

Assumption 3. *(regularity)*

(i) $(Y_i', X_i', \xi_{i0}', h_i')'$ *are independent and identically distributed across* i *conditional on* λ_0.

$(Y_{it}', X_{it}', \lambda_{t0}')'$ *are stationary over time, for all* i.

$\ell_{ij}(\alpha_i^j, \theta)$ *is three times differentiable in both its arguments, for all* i, j.[3]

The parameter space Θ *for* θ_0 *is compact, the space for* α_{i0}^j *is compact, and* θ_0 *belongs to the interior of* Θ.

(ii) N, S, K *tend jointly to infinity.*

$\max_j \sup_{\xi,\lambda,\alpha,\theta} |\mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}(\ell_{ij}(\alpha, \theta))| = O(1)$, *and similarly for the first three derivatives of* ℓ_{ij} *in both its arguments.*

Moreover, one of the two following sets of conditions holds:

(a) $R = T/p$ *tends to infinity.*

$\min_j \inf_{\xi,\lambda,\alpha,\theta} \mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}(-\frac{\partial^2 \ell_{ij}(\alpha,\theta)}{\partial\alpha\partial\alpha'})$ *is positive definite.*

$\max_{i,j} \sup_{\alpha,\theta} \left|\ell_{ij}(\alpha, \theta) - \mathbb{E}_{\xi_{i0},\lambda_0}\left(\ell_{ij}(\alpha, \theta)\right)\right| = o_p(1)$, *and similarly for the first three derivatives of* ℓ_{ij} *in both its arguments.*

[3]By this, we mean that the functions $\frac{1}{R}\sum_{t=(j-1)R+1}^{jR} \ln f(y_{it} \,|\, y_{i,t-1}, x_{it}, \alpha, \theta)$ are three times differentiable with respect to α and θ, for almost all y_i, x_i.

(b) *The minimum (respectively, maximum) eigenvalue of* $(-\frac{\partial^2 \ell_{ij}(\alpha,\theta)}{\partial\alpha\partial\alpha'})$ *is bounded away from zero (resp., infinity) with probability one, uniformly in* i, j, α, θ.

The third derivatives of $\ell_{ij}(\alpha,\theta)$ *are* $O_p(1)$*, uniformly in* i, j, α, θ.

$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}[\ell_{ij}(\alpha_{i0}^j,\theta_0) - \mathbb{E}_{\xi_{i0},\lambda_0}(\ell_{ij}(\alpha_{i0}^j,\theta_0))]^2 = O_p(1/R)$*, and similarly for the first three derivatives of* ℓ_{ij} *in both its arguments.*

(iii) For all θ, ξ, j*, let* $\overline{\alpha}^j(\theta,\xi) = \operatorname{argmax}_\alpha \mathbb{E}_{\xi_{i0}=\xi,\lambda_0}(\ell_{ij}(\alpha,\theta))$.

$\min_j \inf_{\xi,\lambda,\theta} \mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}(-\frac{\partial^2 \ell_{ij}(\overline{\alpha}^j(\theta,\xi),\theta)}{\partial\alpha\partial\alpha'})$ *is positive definite.*

$\mathbb{E}\left[\frac{1}{p}\sum_{j=1}^{p}\ell_{ij}(\overline{\alpha}^j(\theta,\xi_{i0}),\theta)\right]$ *has a unique maximum at* θ_0 *on* Θ*, and its second derivative* $-H$ *is negative definite.*

$\sup_\theta \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\|\frac{\partial^2 \ell_{ij}(\overline{\alpha}^j(\theta,\xi_{i0}),\theta)}{\partial\theta\partial\alpha'}\|^2 = O_p(1)$.

(iv) $\max_j \sup_{\widetilde{\xi},\lambda,\alpha} \|\frac{\partial}{\partial\xi'}\big|_{\xi=\widetilde{\xi}} \mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}(\operatorname{vec}\frac{\partial^2 \ell_{ij}(\alpha,\theta_0)}{\partial\theta\partial\alpha'})\| = O(1)$.

$\max_j \sup_{\widetilde{\xi},\lambda,\alpha} \|\frac{\partial}{\partial\xi'}\big|_{\xi=\widetilde{\xi}} \mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}(\operatorname{vec}\frac{\partial^2 \ell_{ij}(\alpha,\theta_0)}{\partial\alpha\partial\alpha'})\| = O(1)$.

$\max_j \sup_{\widetilde{\xi},\lambda,\theta} \|\frac{\partial}{\partial\xi'}\big|_{\xi=\widetilde{\xi}} \mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}(\frac{\partial \ell_{ij}(\overline{\alpha}^j(\theta,\widetilde{\xi}),\theta)}{\partial\alpha})\| = O(1)$.

(v) $\mathbb{E}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}(\frac{\partial \ell_{ij}(\overline{\alpha}^j(\theta,\xi),\theta)}{\partial\alpha})$ *and* $\mathbb{E}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}(\operatorname{vec}\frac{\partial}{\partial\theta'}\big|_{\theta_0}\frac{\partial \ell_{ij}(\overline{\alpha}^j(\theta,\xi),\theta)}{\partial\alpha})$ *are twice differentiable with respect to* h*, with first and second derivatives that are uniformly bounded in* j*,* ξ*,* λ*,* $h \in \mathcal{H}$*, and* $\theta \in \Theta$*, where* $\mathcal{H}$ *denotes the support of* h_i.

$\|\operatorname{Var}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}(\frac{\partial \ell_{ij}(\overline{\alpha}^j(\theta,\xi),\theta)}{\partial\alpha})\| = O(1/R)$*, uniformly in* j*,* ξ*,* λ*,* h*, and* θ.

$\|\operatorname{Var}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}(\operatorname{vec}\frac{\partial}{\partial\theta'}\big|_{\theta_0}\frac{\partial \ell_{ij}(\overline{\alpha}^j(\theta,\xi),\theta)}{\partial\alpha})\| = O(1/R)$*, uniformly in* j*,* ξ*,* λ*, and* h.

In part (i) in Assumption 3 we treat heterogeneity as random in order to use Lemma 2, which requires that ξ_{i0} be i.i.d. draws from a distribution. Nevertheless, since we do not restrict how α_{i0}, μ_{i0} and λ_0 depend on one another, our analysis is close to a "fixed-effects" setup that would condition on the realizations of the heterogeneity. We also assume that observations are stationary over time. While our results require asymptotic stationarity of the time-series processes, one might allow for non-stationary initial conditions at the cost of complicating some of the assumptions.

In part (ii) we require strict concavity of the log-likelihood as a function of α. We distinguish two cases, where the stronger condition in part (iib) allows us to cover settings with time-varying heterogeneity where p grows to infinity at the same rate as T. This regime is interesting, since a fixed-effects estimator treating all the α_i^j as parameters is inconsistent in this case, and we

will see that GFE remains consistent under suitable conditions. Concavity holds in a number of nonlinear panel data models such as probit and logit models, tobit, Poisson, or multinomial logit; see Chen, Fernández-Val and Weidner (2014) and Fernández-Val and Weidner (2016), for example. However, concavity may be difficult to check, and may not hold, in complex nonlinear models such as structural economic models. In Corollary E1 in Appendix E we show that, in models with time-invariant heterogeneity where p remains fixed as the sample size increases, Theorem 1 continues to hold without concavity, under a standard identification assumption and an assumption bounding the derivatives of the empirical GFE objective function. In future work it will be important to generalize the analysis to models with time-varying heterogeneity and non-concave log-likelihoods.

In part (iii) we work with the following "target" log-likelihood function:

$$\overline{\ell}(\theta) = \frac{1}{Np} \sum_{i=1}^{N} \sum_{j=1}^{p} \ell_{ij}(\overline{\alpha}^j(\theta, \xi_{i0}), \theta).$$

Under our assumptions, $\overline{\ell}(\theta)$ approximates the GFE log-likelihood in large samples. It is the least-favorable log-likelihood (Severini and Wong, 1992, Arellano and Hahn, 2007), and it enjoys several properties of an actual log-likelihood function. In particular, we have $\overline{\alpha}^j(\theta_0, \xi_{i0}) = \alpha_{i0}^j$. Note that, in models with time-varying heterogeneity, $\overline{\alpha}^j(\theta, \xi_{i0})$ depends on the process λ_0 in addition to the type ξ_{i0}, even though we leave the dependence on λ_0 implicit in the notation. For example, in a static model with $p = T$, $\overline{\alpha}^t(\theta, \xi_{i0})$ is a function of ξ_{i0} and λ_{t0}, while in a dynamic model it will generally depend in addition on the history of the time effects $\lambda_{t0}, \lambda_{t-1,0}, ...$

In parts (iv) and (v) we impose additional regularity conditions specific to the GFE estimation problem. In part (iv) we require that some moments be bounded asymptotically. In part (v) we impose regularity of certain conditional expectations and variances given the moments h_i. Note that $\frac{\partial \ell_{ij}}{\partial \alpha}$ is an average of R zero-mean random variables, where $R = T/p$ is the number of observations in every sub-period, so the conditional variances in part (v) will be $O(1/R)$ under suitable conditions on serial dependence within sub-periods. In addition, one can verify that part (v) is not needed when heterogeneity is time-invariant and T tends to infinity.

In Appendix F we verify all the conditions needed to apply our main theorem, including all the different parts of Assumption 3, in two dynamic linear models under a set of simple primitive conditions. We consider a model with time-invariant heterogeneity, and a model where time-varying heterogeneity follows a factor structure.

We are now in position to state our main theoretical result. Let us denote:

$$s_i = \frac{1}{p} \sum_{j=1}^{p} \left\{ \frac{\partial \ell_{ij}(\alpha_{i0}^j, \theta_0)}{\partial \theta} + \mathbb{E}_{\xi_{i0}, \lambda_0} \left(\frac{\partial^2 \ell_{ij}(\alpha_{i0}^j, \theta_0)}{\partial \theta \partial \alpha'} \right) \left[\mathbb{E}_{\xi_{i0}, \lambda_0} \left(-\frac{\partial^2 \ell_{ij}(\alpha_{i0}^j, \theta_0)}{\partial \alpha \partial \alpha'} \right) \right]^{-1} \frac{\partial \ell_{ij}(\alpha_{i0}^j, \theta_0)}{\partial \alpha} \right\},$$

and:

$$H = \mathbb{E}\left[\frac{1}{p}\sum_{j=1}^{p}\left\{\mathbb{E}_{\xi_{i0},\lambda_0}\left(-\frac{\partial^2 \ell_{ij}(\alpha_{i0}^j,\theta_0)}{\partial\theta\partial\theta'}\right)\right.\right.$$
$$\left.\left. - \mathbb{E}_{\xi_{i0},\lambda_0}\left(\frac{\partial^2 \ell_{ij}(\alpha_{i0}^j,\theta_0)}{\partial\theta\partial\alpha'}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_0}\left(-\frac{\partial^2 \ell_{ij}(\alpha_{i0}^j,\theta_0)}{\partial\alpha\partial\alpha'}\right)\right]^{-1}\mathbb{E}_{\xi_{i0},\lambda_0}\left(\frac{\partial^2 \ell_{ij}(\alpha_{i0}^j,\theta_0)}{\partial\alpha\partial\theta'}\right)\right\}\right].$$

The score s_i is the derivative of $\frac{1}{p}\sum_{j=1}^{p}\ell_{ij}(\overline{\alpha}^j(\theta,\xi_{i0}),\theta)$ with respect to θ evaluated at θ_0, and the Hessian H is the same matrix as in Assumption 3 part (iii).

Theorem 1. *Let the conditions in Corollary 1 hold.*

(a) Let Assumption 3 with part (iia) hold. Then, as N, S, K and $R = T/p$ tend to infinity such that $Kp/(NT)$ tends to zero, we have:

$$\widehat{\theta} = \theta_0 + H^{-1}\frac{1}{N}\sum_{i=1}^{N} s_i + O_p\left(\frac{1}{S}\right) + O_p\left(\frac{Kp}{NT}\right) + O_p\left(K^{-\frac{2}{d}}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right). \tag{9}$$

If in addition $\frac{1}{N}\sum_{i=1}^{N} s_i = O_p(1/\sqrt{NT})$, then we have:

$$\frac{1}{Np}\sum_{i=1}^{N}\left\|\widehat{\alpha}(\widehat{k}_i) - \alpha_{i0}\right\|^2 = O_p\left(\frac{1}{S}\right) + O_p\left(\frac{Kp}{NT}\right) + O_p\left(K^{-\frac{2}{d}}\right). \tag{10}$$

(b) If in Assumption 3 part (iia) is replaced with part (iib), then (9) and (10) hold as N, S, K tend to infinity such that $Kp/(NT)$ tends to zero, irrespective of $R = T/p$ being fixed or growing.

There are several similarities and differences between the expansion of $\widehat{\theta}$ in Theorem 1 and standard large-N,T expansions of fixed-effects estimators (e.g., Hahn and Newey, 2004).[4] Under time-invariant heterogeneity, the score s_i and Hessian H are the same as in the expansion of the fixed-effects estimator. Moreover, similarly to fixed-effects, GFE is subject to bias. However, Theorem 1 continues to hold in the presence of time-varying heterogeneity, in settings where fixed-effects is severely biased or inconsistent. Indeed, when estimating p parameters per individual, the incidental parameter bias of fixed-effects is of the order of p/T. In particular, when p is proportional to T the bias does not vanish asymptotically, and fixed-effects is inconsistent. In contrast, two-step GFE is consistent provided K/N tends to zero.

[4] Although we formulate Theorem 1 in a likelihood setup, it holds more generally for M-estimators, interpreting $\sum_{i=1}^{N}\sum_{j=1}^{p}\ell_{ij}(\alpha_i^j,\theta)$ as the objective function in the M-estimation. In addition, a similar result holds for partial likelihood estimators such as the ones we use in our empirical illustration.

The expansion in (9) highlights the presence of three terms. The $1/S$ term reflects the noise in h_i. In a panel data setting with $S = T$, this term is akin to incidental parameter bias. The $Kp/(NT)$ term arises from the estimation of Kp group-specific parameters using NT observations, and may thus also be interpreted as reflecting incidental parameter bias. Lastly, the term $K^{-\frac{2}{d}}$ reflects the approximation error associated with estimating the heterogeneity using K groups. In Appendix A, we provide a similar expansion as in (9) for GFE estimators of average effects $M_0 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} m\left(X_{it}, \alpha_{it0}, \theta_0\right)$ that are functions of both common parameters and individual heterogeneity.

Theorem 1 stands in contrast with the previous literature on GFE estimation under discrete heterogeneity (e.g., Hahn and Moon, 2010, Bonhomme and Manresa, 2015). In an environment where heterogeneity is *not* restricted to have a small number of points of support, classification noise affects the properties of second-step estimators in general. Our framework, which leads to different properties compared to previous results obtained under discrete heterogeneity, motivates choosing K appropriately in order to control the approximation error, and combining discrete estimation with bias reduction.

Example 2 (continued). In the probit model of Example 2, let us first consider the case where α_{i0} and μ_{i0} are time-invariant, so $p = 1$. Using $h_i = (\overline{Y}_i, \overline{X}_i')'$ as moments, we obtain, using Theorem 1 and the fact that $Kp/(NT) \leq 1/T$:

$$\widehat{\theta} = \theta_0 + H^{-1}\frac{1}{N}\sum_{i=1}^{N} s_i + O_p\left(\frac{1}{T}\right) + O_p\left(K^{-\frac{2}{d}}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right), \tag{11}$$

where s_i and H coincide with the score and Hessian in fixed-effects probit maximum likelihood. Next, consider the case where $\alpha_{it0} = \alpha(\xi_{i0}, \lambda_{t0})$ and $\mu_{it0} = \alpha(\xi_{i0}, \lambda_{t0})$ vary over time, and $p = T$. Using the same moments in the first step, the expansion in Theorem 1 then becomes:

$$\widehat{\theta} = \theta_0 + \widetilde{H}^{-1}\frac{1}{N}\sum_{i=1}^{N} \widetilde{s}_i + O_p\left(\frac{1}{T}\right) + O_p\left(\frac{K}{N}\right) + O_p\left(K^{-\frac{2}{d}}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right). \tag{12}$$

We provide the expressions for s_i, H, $\widetilde{s}_i$, and $\widetilde{H}$ in Appendix A.

In Example 2, in the large-T limit, h_i is a function of the type ξ_{i0} that is driving both α_{i0} and μ_{i0} (or α_{it0} and μ_{it0} in the time-varying case). For example, if X_{it} is a scalar covariate and (α_{i0}, μ_{i0}) follows a non-degenerate bivariate probability distribution, then the type has dimension $d = 2$. More generally, d will often increase with the number of covariates, thus contributing to a slower convergence rate of two-step GFE. The extensions that we will introduce in Section 4 aim at improving performance in such cases.

Remark on external moments. When external moments are available, Theorem 1 holds even when T is fixed as N, S, K tend to infinity, provided the moments h_i be independent of the data $(Y_i', X_i')'$ conditional on the latent types ξ_{i0} and the process of time effects λ_0. Part (v) in Assumption 3 may not be satisfied absent independence. Notice that, in contrast, the fixed-effects estimator of θ_0 is generally inconsistent for fixed T.

In addition, the availability of external moments can expand the applicability of two-step GFE beyond the situations covered by Theorem 1. First, it may be that external moments capture additional heterogeneity beyond the type ξ_{i0}. It is easy to adapt Theorem 1 to such cases. Second, the researcher may have access to external measurements of α_{i0}, as opposed to measurements of a type that drives both α_{i0} and μ_{i0} as in Assumption 1. In this case, an expansion of the form (9) can be derived under similar assumptions, with three differences: the relevant "target likelihood" and the corresponding score and Hessian are then based on $\overline{\alpha}^j(\theta, \varphi) = \text{argmax}_\alpha \, \mathbb{E}_{\varphi_{i0}=\varphi, \lambda_0}(\ell_{ij}(\alpha, \theta))$ where $\varphi_{i0} = \text{plim}_{S\to\infty} h_i$, the expansion features an additional $O_p(1/N)$ term compared to (9), and the approximation error depends on the underlying dimension of φ_{i0}.

3.3 Implications for inference

We now discuss some implications of Theorem 1 for inference. The following corollary provides an expression of the expansion of GFE estimators when we set the number of groups according to the rule given by (4).

Corollary 2. *Let the conditions in Theorem 1 hold. Let K be such that (i) $S\widehat{Q}(K) = O_p(1)$ and (ii) $SKp/(NT) = O(1)$. Then we have:*

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

Condition (i) in Corollary 2 guarantees that the impact of approximation error, which decreases as a function of K, is of the same or lower order as $1/S$. This requires setting a large enough K. This condition is automatically satisfied when using our rule to select the number of groups, provided γ in (4) is bounded. In turn, Condition (ii) requires that K be not too large, so that the term $Kp/(NT)$ does not dominate $1/S$. We further discuss this condition below.

The following result is then a direct consequence of Corollary 2.

Corollary 3. *Let the conditions in Corollary 2 hold, and suppose that the two following conditions hold:*

(i) $\frac{1}{N}\sum_{i=1}^{N} s_i = \frac{1}{\sqrt{NT}} Z + o_p\left(\frac{1}{\sqrt{NT}}\right)$, *where* $Z \sim \mathcal{N}(0, H)$.
(ii) $\sqrt{NT}/S = o(1)$.
Then we have:

$$\sqrt{NT}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, H^{-1}). \tag{14}$$

While Condition (i) in Corollary 3 is standard, Condition (ii) is more restrictive since it requires that $\sqrt{NT}/S$ tends to zero. In panel data models where $S = T$, this requires T to tend to infinity faster than N. In this asymptotic regime, the $1/S = 1/T$ incidental parameter bias term does not affect the asymptotic distribution to first order. However, when T and N grow at a similar rate, incidental parameter bias has a first-order effect on inference. To reduce this bias, we use the half-panel jackknife method of Dhaene and Jochmans (2015).

Specifically, we first compute the two-step GFE estimator $\widehat{\theta}$ on the full sample. Then, we compute $\widehat{\theta}_1$ and $\widehat{\theta}_2$ on the first $T/2$ periods and the last $T/2$ periods, respectively (considering T even for simplicity). The bias-reduced estimator is then:

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

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

Corollary 4. *Let the conditions in Corollary 2 hold, with* $S = T$. *In addition, suppose that Condition* (i) *in Corollary 3 holds, and that (13) takes the form:*

$$\widehat{\theta} = \theta_0 + H^{-1}\frac{1}{N}\sum_{i=1}^{N} s_i + \frac{C}{T} + o_p\left(\frac{1}{T}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right), \tag{15}$$

for some constant C. *Then, as* N *and* T *tend to infinity jointly such that* N/T *tends to a positive constant, we have:*

$$\sqrt{NT}(\widehat{\theta}^{\mathrm{BR}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, H^{-1}).$$

We provide conditions that ensure (15) in Appendix E, in a setup where p does not tend to infinity. In that case, we provide an explicit expression for the constant C as a function of the model's parameters. The conditions we rely on require the number of groups K to grow relatively fast with the sample size, which corresponds to taking $\gamma = o(1)$ in (4). It would be interesting to derive primitive conditions for (15) under our recommended choice $\gamma = 1$, and in models with time-varying heterogeneity where p grows with the sample size, although this does not follow directly from our methods of proof.

To apply either Corollary 3 or Corollary 4, we need Condition (ii) in Corollary 2 to hold. To provide intuition about this condition, note that it follows from results in vector quantization (see Graf and Luschgy, 2000, 2002) that, when setting K as in (4) with $\gamma = 1$, K typically grows proportionally to $S^{\frac{d}{2}}$, so Condition (ii) in Corollary 2 requires that $S^{1+\frac{d}{2}}p/(NT) = O(1)$.

As an example, consider the case $S = T$ with N and T growing at the same rate, as in Corollary 4. The condition then requires:

$$T^{\frac{d}{2}-1}p = O(1). \tag{16}$$

When $d = 1$, (16) allows for time-varying heterogeneity, although p cannot grow faster than $\sqrt{T}$. Hence, unobservables need to be varying relatively slowly over time. In Subsection 4.4 we will present a two-way extension of GFE that relaxes this condition, in static models and under an assumption about the dimensionality of time heterogeneity.

Moreover, our theory imposes a tight bound on the dimension d. When $d = 2$, (16) only allows for time-invariant heterogeneity, while when $d \geq 3$ the conditions of Corollary 4 cannot be satisfied when N and T grow at the same rate. This discussion highlights the fact that two-step GFE is designed for settings where individual heterogeneity has a low underlying dimension.

Finally, for feasible inference based on Corollaries 3 or 4, one can replace H by the following empirical counterpart:

$$\begin{aligned}\widehat{H} =& \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\{\widehat{\mathbb{E}}_{\widehat{k}_i}\left(-\frac{\partial^2 \ell_{ij}\left(\widehat{\alpha}^j(\widehat{k}_i),\widehat{\theta}\right)}{\partial\theta\partial\theta'}\right)\right. \\ &\left. - \widehat{\mathbb{E}}_{\widehat{k}_i}\left(\frac{\partial^2 \ell_{ij}(\widehat{\alpha}^j(\widehat{k}_i),\widehat{\theta})}{\partial\theta\partial\alpha'}\right)\left[\widehat{\mathbb{E}}_{\widehat{k}_i}\left(-\frac{\partial^2 \ell_{ij}(\widehat{\alpha}^j(\widehat{k}_i),\widehat{\theta})}{\partial\alpha\partial\alpha'}\right)\right]^{-1}\widehat{\mathbb{E}}_{\widehat{k}_i}\left(\frac{\partial^2 \ell_{ij}(\widehat{\alpha}^j(\widehat{k}_i),\widehat{\theta})}{\partial\alpha\partial\theta'}\right)\right\},\end{aligned} \tag{17}$$

where $\widehat{\mathbb{E}}_k(Z_{ij})$ denotes the mean over i of Z_{ij} in group $\widehat{k}_i = k$, and $\widehat{\alpha}^j(k)$ are GFE estimates defined similarly as α^j_{i0}. Under a slight modification of the conditions of Theorem 1, $\widehat{H}$ is consistent for H (see Appendix A). Lastly, Corollaries 3 and 4 hold under correct specification of the likelihood model. Under misspecification, the asymptotic covariance matrix and its empirical counterpart need to be modified .

4 Extensions of the two-step approach

The theory presented in the previous section has highlighted several aspects of GFE that are needed for good performance. GFE relies on injective moments, as formalized in Assumption

2. It also requires the dimension of the types ξ_{i0} to be sufficiently small in order to control the approximation error. Lastly, while GFE can allow for time-varying unobservables, (16) shows that, when N and T tend to infinity at the same rate, the amount of time variation that can be allowed for is limited. In this section we present several extensions of two-step GFE that aim to alleviate these issues. We focus on standard panel data settings where no external information is available and $S = T$.

4.1 Choice of moments and injectivity

The choice of moments is a key input, since it determines the quality of the approximation to the unobserved heterogeneity. Specific models may suggest particular individual summary statistics to be used in the classification step. For example, in linear models, individual averages of outcomes and covariates are natural choices. However, in nonlinear models, choosing suitable moments faces several challenges.

The moments h_i should, asymptotically, be injective functions of the heterogeneity, as stated in Assumption 2. Intuitively, the moments should be rich enough in order to capture all relevant sources of heterogeneity. One way to ensure injectivity is to use the entire empirical distribution of the data, thereby capturing all the relevant heterogeneity in the classification step. In Appendix I, we describe this approach in the context of a static model, and we use it to decompose the variance of log-wages into terms that reflect worker and firm heterogeneity as well as sorting. We use an estimator proposed in Bonhomme, Lamadon and Manresa (2019), where firms are classified based on their empirical wage distributions, in data generating processes where heterogeneity is continuous.

While including a large number of measurements in the first step helps making injectivity more plausible, this may come at a cost. Consider the probit model in Example 2. Only including $\overline{Y}_i$ as a moment is not enough to ensure injectivity in general, since the mean μ_{i0} of X_{it} is heterogeneous. On the other hand, adding $\overline{X}_i$ to the vector of moments, while guaranteeing injectivity, is likely to increase the dimension d of heterogeneity and the approximation error, thus contributing to a slower convergence rate of the kmeans estimator. More generally, adding moments to inform the classification may add other dimensions of heterogeneity, and lead to a lower signal-to-noise ratio.

The presence of an approximation error is an important feature of GFE. It is particularly salient in models with conditioning covariates, when covariates and outcomes are not driven by a common low-dimensional type. In order to broaden the scope of applications of GFE, we next describe two methods that exploit additional information on the model to improve performance:

nonparametric smoothness conditions in the first case, and parametric assumptions from the panel model in the second case.

4.2 GFE with a conditional first step

A first strategy to improve performance is to incorporate conditioning covariates in the first step using a conditional approach. To proceed, suppose that the moments are individual-specific averages of the form $h_i = \frac{1}{T}\sum_{t=1}^{T} h(Y_{it}, X_{it})$. To account for covariates in the first step, we now propose a variation of two-step GFE that relies on clusterwise regression (Späth, 1979) instead of kmeans, under a linear series specification.

Consider a vector of approximating functions $P_q(x) = (P_{1q}(x), ..., P_{qq}(x))'$, such as orthogonal polynomials, splines, or wavelets, of degree q. Let us first focus on a setting with time-invariant heterogeneity. We define the conditional kmeans estimator as:

$$\left(\widehat{\beta}_q, \widehat{k}_1, ..., \widehat{k}_N\right) = \min_{(b,k_1,...,k_N)} \sum_{i=1}^{N}\sum_{t=1}^{T} \left\|h(Y_{it}, X_{it}) - P_q(X_{it})'b(k_i)\right\|^2. \tag{18}$$

This step replaces (2) in the two-step method.

We perform the minimization using the following algorithm, which extends Lloyd's algorithm to a conditional classification step.

Algorithm 2. *(conditional kmeans)*

- *Given initial values for* $b_q(1), ..., b_q(K)$*, iterate between the following two steps until convergence:*
- *Given* $b_q(1), ..., b_q(K)$*, compute* $k_i = \text{argmin}_{k=1,...,K} \sum_{t=1}^{T} \left\|h(Y_{it}, X_{it}) - P_q(X_{it})'b_q(k)\right\|^2$ *for all* i*.*
- *Given* $k_1, ..., k_N$*, compute* $b_q(k) = \text{argmin}_b \sum_{i=1}^{N} \mathbf{1}\{k_i = k\} \sum_{t=1}^{T} \left\|h(Y_{it}, X_{it}) - P_q(X_{it})'b\right\|^2$ *for all* k*.*

The main difference with Lloyd's algorithm is the update step, since the parameters $b_q(k)$ are now updated using within-group regressions. While here we describe the method based on series linear regression, other linear or nonlinear regression techniques can be used, such as the Lasso, regression trees, or neural networks for example. In Section 5 we will compare the performance of several specifications on simulated data. Also, note that estimates from kmeans may provide good starting values for its conditional counterpart. We use this approach in our simulations.

To illustrate the conditional first step, consider our two main examples. In Example 1, when using discrete choices as outcomes, Algorithm 2 effectively groups individuals together when estimates of their conditional choice probabilities are most similar. In Example 2, conditional kmeans based on linear regressions consists in iterating between classifying individual units, and running linear probability regressions within groups.

As a counterpart to Corollary 1, we show that, when using conditional kmeans, the within-group mean squared error of α_{i0} admits the following bound, as N, T, K and q tend to infinity:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\alpha_{i0}-\overline{\alpha}_0(\widehat{k}_i)\right\|^2 = O_p\left(\frac{q}{T}\right)+O_p\left(q^{-2a}\right)+O_p(K^{-\frac{2}{d_\alpha}}), \tag{19}$$

where $a>0$ is a constant, and $d_\alpha>0$ is the underlying dimension of α_{i0}. In Theorem B1 in Appendix B we derive (19) under appropriate regularity conditions.

The first and second terms in (19) are standard contributions to the rate of convergence of series estimators (as in Newey, 1997, for example). In particular, the second term depends on the smoothness in X_{it} of the conditional expectation function of $h(Y_{it}, X_{it})$. Importantly, the dimension d_α that appears in the third term in (19) is the underlying dimension of α_{i0} alone, and it does not depend on μ_{i0}. For instance, in Example 2, $d_\alpha = 1$ irrespective of the dimension of X_{it}. In contrast, the d that appears in Corollary 1 and Theorem 1 depends on the number of covariates in general. When d_α is small relative to d, the convergence rate in the conditional kmeans first step can improve relative to the rate of unconditional kmeans. At the same time, for improvements to occur we need the term q^{-2a} to be sufficiently small, where a depends on smoothness conditions and the dimensionality of X_{it}.

The conditional kmeans estimator in (18) is designed for models where α_{i0} is time-invariant. In models with time-varying heterogeneity where $\alpha_{it0} = \alpha(\xi_{i0}, \lambda_t^{(o)})$, for $\lambda_t^{(o)}$ a vector of observed factors, the same estimator can be applied by incorporating $\lambda_t^{(o)}$ to the set of covariates X_{it}. More generally, when $\alpha_{it0} = \alpha(\xi_{i0}, \lambda_{t0})$ with λ_{t0} a vector of unobserved factors, one can interact the groups with time indicators, and estimate $b_q(k,t)$ coefficients in Algorithm 2. We provide details on this approach, and a convergence rate, in Appendix B.

Lastly, using the rate in (19), it is not immediate to derive the properties of second-step GFE estimators in the spirit of Theorem 1 and Corollary 4. In particular, the presence of the q/T term, where q tends to infinity, complicates the derivation of asymptotic distributions. We leave the development of an inference theory for GFE estimators with a conditional first step to future work.

4.3 Using the model: iterated GFE

A second strategy to improve the performance of two-step GFE is to exploit the parametric structure of the model using a model-based iteration. To describe the method, suppose that we have computed two-step estimates $\widehat{\theta}$ and $\widehat{\alpha}$ from (3), based on an unconditional or conditional first step. We propose to compute a new partition of individual units according to the following model-based classification rule:

$$\widehat{k}_i^{(2)} = \underset{k=1,...,K}{\text{argmax}} \ \ln f_i\left(\widehat{\alpha}\left(k\right), \widehat{\theta}\right), \quad \text{for all } i = 1, ..., N. \tag{20}$$

We then update second-step estimates as:

$$\left(\widehat{\theta}^{(2)}, \widehat{\alpha}^{(2)}\right) = \underset{(\theta,\alpha)}{\text{argmax}} \ \sum_{i=1}^{N} \ln f_i\left(\alpha\left(\widehat{k}_i^{(2)}\right), \theta\right). \tag{21}$$

The method may be iterated further. Notice that the value of the log-likelihood is weakly increasing in each iteration.

Similarly to the conditional kmeans approach, model-based iterations can provide substantial improvements in finite sample performance. However, it seems difficult to provide a theoretical basis for these improvements. In Appendix C we give details on the model-based iteration, and we discuss a "one-step" counterpart where the likelihood function is jointly optimized with respect to every group partition and parameter values.

4.4 Aggregate moments: two-way kmeans clustering

In applications, it may be appealing to group not only i, as in our baseline GFE approach, but also t. As an example, consider a model of demand for differentiated products with unobserved product characteristics that vary across markets t as well as products i (as in Berry, Levinsohn and Pakes, 1995). In this context, fixed-effects methods are popular alternatives to instrumental variables strategies. As an example, Moon, Shum and Weidner (2018) model unobserved product characteristics through a factor-analytic "interactive fixed-effects" specification in the spirit of Bai (2009). In this subsection, we describe a two-way GFE method that is able to approximate general unobservables with low underlying dimension, through a data-based classification of both products i and markets t.

To describe two-way GFE, let us suppose that the likelihood function takes the static form:

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

with:

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

where we assume that observations are i.i.d. across i and t. This setup is therefore more restrictive than the one we considered in Theorem 1, where we allowed for serial dependence. In addition, here we suppose that both ξ_{i0} and λ_{t0} in $\alpha_{it0} = \alpha(\xi_{i0}, \lambda_{t0})$ and $\mu_{it0} = \mu(\xi_{i0}, \lambda_{t0})$ are low-dimensional. Hence, in the setup we focus on in this subsection, both cross-sectional and time heterogeneity have a small underlying dimension.

We propose augmenting the baseline two-step GFE approach by adding to (2) a second kmeans classification step based on *aggregate* moments $w_t = \frac{1}{N}\sum_{i=1}^{N} w(Y_{it}, X_{it})$; that is, we propose to compute:

$$\left(\widehat{w}, \widehat{l}_1, ..., \widehat{l}_T\right) = \underset{(\widetilde{w}, l_1, ..., l_T)}{\operatorname{argmin}} \sum_{t=1}^{T} \left\| w_t - \widetilde{w}(l_t) \right\|^2, \tag{22}$$

where $\{l_t\}$ are partitions of $\{1, ..., T\}$ into at most p groups. Given the group indicators $\widehat{k}_i$ and $\widehat{l}_t$, we then compute the second-step estimator:

$$\left(\widehat{\theta}, \widehat{\alpha}\right) = \underset{(\theta, \alpha)}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{t=1}^{T} \ln f(Y_{it} \,|\, X_{it}, \alpha(\widehat{k}_i, \widehat{l}_t), \theta). \tag{23}$$

In Theorem D1 in Appendix D, we provide conditions under which $\widehat{\theta}$ admits the following expansion:

$$\begin{aligned} \widehat{\theta} \;=\; \theta_0 + \widetilde{H}^{-1} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{s}_{it} + O_p\left(\frac{1}{T}\right) + O_p\left(\frac{1}{N}\right) + O_p\left(\frac{Kp}{NT}\right) \\ + O_p\left(K^{-\frac{2}{d}}\right) + O_p\left(p^{-\frac{2}{d_\lambda}}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right), \end{aligned} \tag{24}$$

where $\widetilde{s}_{it}$ and $\widetilde{H}$ are given by (D1) and (D2) in Appendix D, and the dimensions d and d_λ were introduced in Assumption 1. Note that, while p was part of the model in Theorem 1, here it is set by the researcher, and plays a similar role as K.

There are several differences between (24) and the corresponding expansion in Theorem 1. First, there are now two kinds of approximation error, since we are approximating not only ξ_{i0} but also λ_{t0}. As a result, here the dimension of λ_{t0} matters. Second, the estimation of the partition $\{\widehat{l}_t\}$ contributes an additional $O_p(1/N)$ term. Third, and this is a key advantage of the two-way approach relative to the baseline two-step one, here p can be chosen small relative to T, which results in a smaller number of group-specific parameters to estimate, and hence a smaller $Kp/(NT)$ term in the expansion.

As an example, suppose that ξ_{i0} and λ_{t0} are scalar. Then, taking K and p proportional to $\sqrt{T}$ and $\sqrt{N}$, respectively, we obtain:

$$\widehat{\theta} \;=\; \theta_0 + \widetilde{H}^{-1}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{s}_{it} + O_p\left(\frac{1}{T}\right) + O_p\left(\frac{1}{N}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right). \tag{25}$$

The rates for K and p are implied by our selection rule, see (4) for the choice of K, with a similar rule for p. Under the additional assumption that, as N and T tend to infinity, the term $O_p(T^{-1})+O_p(N^{-1})$ in (25) is equal to $C_1/T+C_2/N+o_p(1/T)+o_p(1/N)$ for some constants C_1 and C_2, one can apply the two-way split panel jackknife method of Fernández-Val and Weidner (2016) for bias correction, and perform inference as N and T tend to infinity at the same rate.

5 Illustrations

In this section we apply GFE estimators to a dynamic structural model of location choice, and to several specifications of a probit model with individual-specific heterogeneity.

5.1 A dynamic model of location choice

In dynamic structural discrete choice models, modeling unobserved heterogeneity as discrete types has two main advantages: unobserved state variables have a small number of points of support, and the number of parameters is relatively small. For these reasons, discrete type methods are widely used. Here we report simulation results for two-step GFE and iterated GFE estimators in the dynamic structural migration model of Example 1, in an environment where heterogeneity is continuous.

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

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

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

$$\Pr(j_{it} = j \mid S_{i,t-1}) = \frac{\exp V_t(j, S_{i,t-1})}{\sum_{j'=1}^{J} \exp V_t(j', S_{i,t-1})}. \tag{26}$$

Estimation and simulation. We estimate the model on the NLSY79, using observations on males who were at least 22 years old in 1979. We keep observations until 1994. We regress log-wages on indicators of years of education and race and a full set of age indicators, and then compute log-wage residuals $\ln W_{it}$. We focus on a stylized setup with $J = 2$ large regions: North-East and South (region A), and North-Central and West (B). There are $N = 1889$ workers, who are observed on average for 12.3 years with a maximum of $T = 16$ years. The probability of moving between the two regions is low in the data: 1.5% per year, and only 10.5% of workers move at all during the observation period. Mean log-wage residuals are 0.09 higher in region A compared to B.

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

$$(\widehat{\alpha}, \widehat{k}_1, ..., \widehat{k}_N) = \underset{(\widetilde{\alpha}, k_1, ..., k_N)}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\ln W_{it} - \widetilde{\alpha}(k_i, j_{it})\right)^2, \tag{27}$$

which amounts to classifying individuals according to location-specific means of log-wages. Here, for a location j that the agent has visited, $\alpha_i(j)$ coincides with the average log-wage in that location, so the j-specific mean of log-wages satisfies the injectivity condition in Assumption 2.

Expected returns in non-visited locations, which enter the rational beliefs of agents, are pinned down from job movers.

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

$$(\widehat{\theta},\widehat{c}) = \underset{(\theta,c)}{\operatorname{argmax}} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{J} \mathbf{1}\{j_{it}=j\} \ln \Pr\left(j_{it}=j \,|\, j_{i,t-1}, \mathcal{J}_{i,t-1}, \widehat{\alpha}(\widehat{k}_i, \mathcal{J}_{i,t-1}), c(\widehat{k}_i, \mathcal{J}_{i,t-1}), \theta\right), \tag{28}$$

where θ contains all structural parameters (including the utility parameter ρ) except for the mobility costs. We set the discount factor to $\beta = 0.95$, and do not estimate it. The likelihood is conditional on the initial location and location-specific return of the agent in period 0. We use a steepest ascent algorithm to maximize the objective in (28), as in the nested fixed point method of Rust (1994). In this model, other estimation methods could be used (e.g., Hotz and Miller, 1993, Aguirregabiria and Mira, 2002, Su and Judd, 2012). Given parameter estimates $\widehat{\alpha}(k,j)$, $\widehat{c}(k,j)$, $\widehat{\sigma}^2$, and $\widehat{\theta}$, we update the estimated partition of individuals using the full model's structure, as in (20) and (21); see Appendix H.

Agents in the model face uncertainty about future values of realized types $\alpha_i(j)$. Hence, here we only discretize the decision problem for estimation purposes. This approach contrasts with a model with discrete types in the population, where after observing her type in one location the agent would generically be able to infer her type in all other possible locations. In the present setup, uncertainty about future types diminishes over time as the agent visits more locations, but it only disappears when all locations have been visited.

To construct the data generating process (DGP) we first estimate the model with homogeneous costs using GFE with $K = 10$ groups, using an iteration starting from the two steps (27) and (28). Following Kennan and Walker (2011), each agent is a "stayer type" with some probability (which depends on the initial $\alpha_i(j_{i1})$ through a logistic specification), in which case they never move. Hence, while the main parameters of interest are ρ and c, the model also features the intercept and slope coefficients (a and b) in the probability of being a mover type; that is, of not being a stayer type. The estimates we obtain are $\widehat{\rho} = 0.34$, $\widehat{c} = 2.13$, $\widehat{a} = -1.94$, and $\widehat{b} = -0.61$. According to the DGP, the probability of being a mover type is low and depends negatively on the initial location-specific return $\alpha_i(j_{i1})$. In addition, the estimated mobility cost is high. The effect of wages on utility is positive, although we will see below that it is quantitatively small.

We then estimate a second specification allowing heterogeneity in mobility costs. For this, we use the same groups as in the specification with homogeneous costs, so the first GFE step is unchanged, and we simply modify the second GFE step. The resulting estimates are $\widehat{\rho} = 0.23$,

$\widehat{a} = -1.61$, and $\widehat{b} = -0.47$, and the mean of the estimated mobility costs is $\widehat{\mathbb{E}}(\widehat{c}_i(j)) = 2.37$, indicating somewhat larger mobility costs and a lower effect of wages on mobility than in the homogeneous specification.

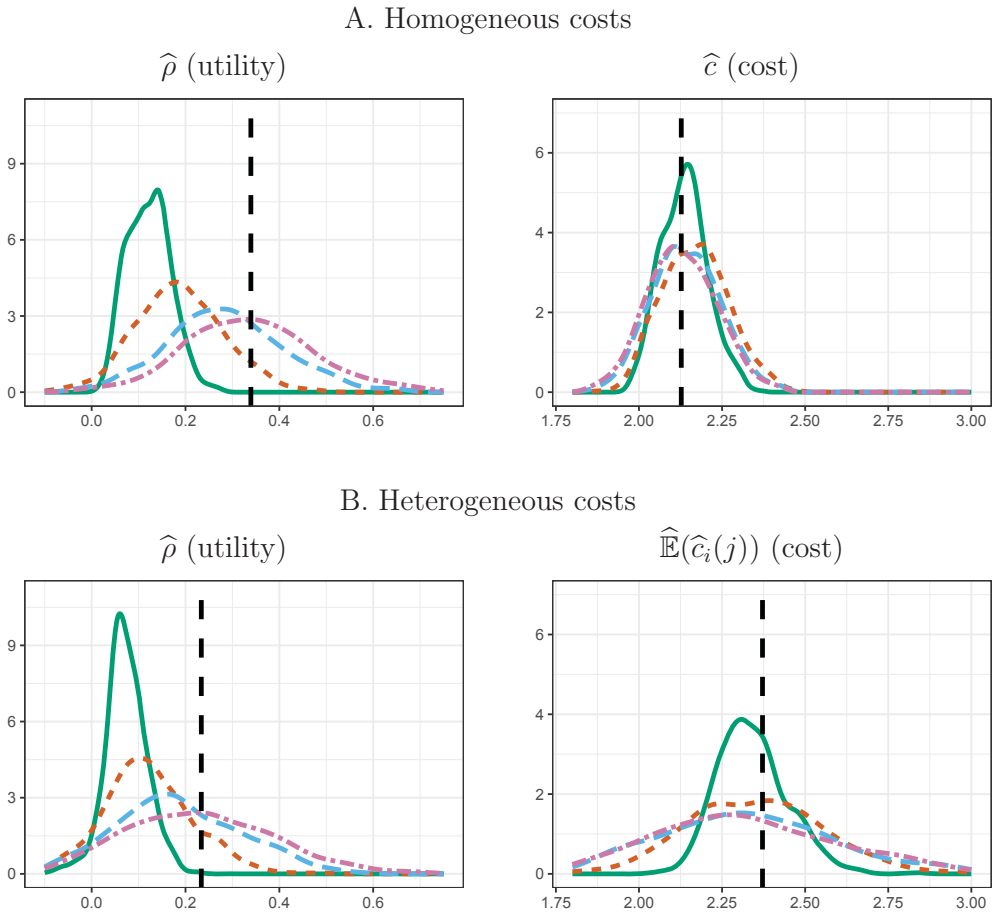
We next solve and simulate the model (as described in Appendix H) based on these parameter values, together with i.i.d. normal specifications for the shocks to log wages and $(\alpha_i(A), \alpha_i(B))$, with means and variances calibrated to our estimates. We simulate the model for $T = 16$ periods, the α's being drawn independently of the initial location. There are two DGP, depending on whether costs are homogeneous or heterogeneous. In the DGP with cost heterogeneity, we specify each location-specific log-cost as a linear function of the corresponding location-specific return. Hence, with $J = 2$ locations, the dimension of heterogeneity is $d = 2$. Note that neither returns nor costs are discrete in the DGP, although we use a discrete approach in estimation.

We report the results of 500 Monte Carlo simulations for the two specifications of mobility costs. We use a kmeans routine with 100 randomly generated starting values, and checked that varying this number had little impact on the results. We estimate the number of groups based on (4) with $\gamma = 1$, in every simulation (and in every subsample when using half-panel jackknife for bias correction).

Parameter estimates. In Panels A and B of Figure 4 we report the densities of parameter estimates across Monte Carlo replications under homogeneous and heterogeneous costs, respectively. We show four types of estimates: two-step GFE (solid curve), bias-corrected two-step GFE (dotted), a single iteration and bias-corrected (dashed), and iterated three times and bias-corrected (dashed-dotted). We focus on the wage coefficient in utility ρ, and the mobility cost c (or the average of costs $c_i(j)$ across individuals and locations), and report additional results in Appendix H. The results for the two specifications of mobility costs agree quite well. Two-step GFE has moderate bias for the cost parameter, but it is biased downwards for the wage coefficient in utility. The Monte Carlo average of the estimated number of groups $\widehat{K}$ is 6.4. Using both bias reduction and an iteration improves the performance of the estimator of ρ substantially. Note that, when combined with half-panel jackknife, a single iteration seems sufficient to correct for most of the bias. At the same time, bias reduction tends to be associated with a variance increase.

In Figure 5 we show the results of two alternative estimators for the model with homogeneous costs: a fixed-effects estimator with its bias-corrected counterpart, and a random-effects estimator with a fixed number of types ($K = 2, 4, 8$), computed using the EM algorithm. We

Figure 4: GFE estimates of structural parameters across simulations

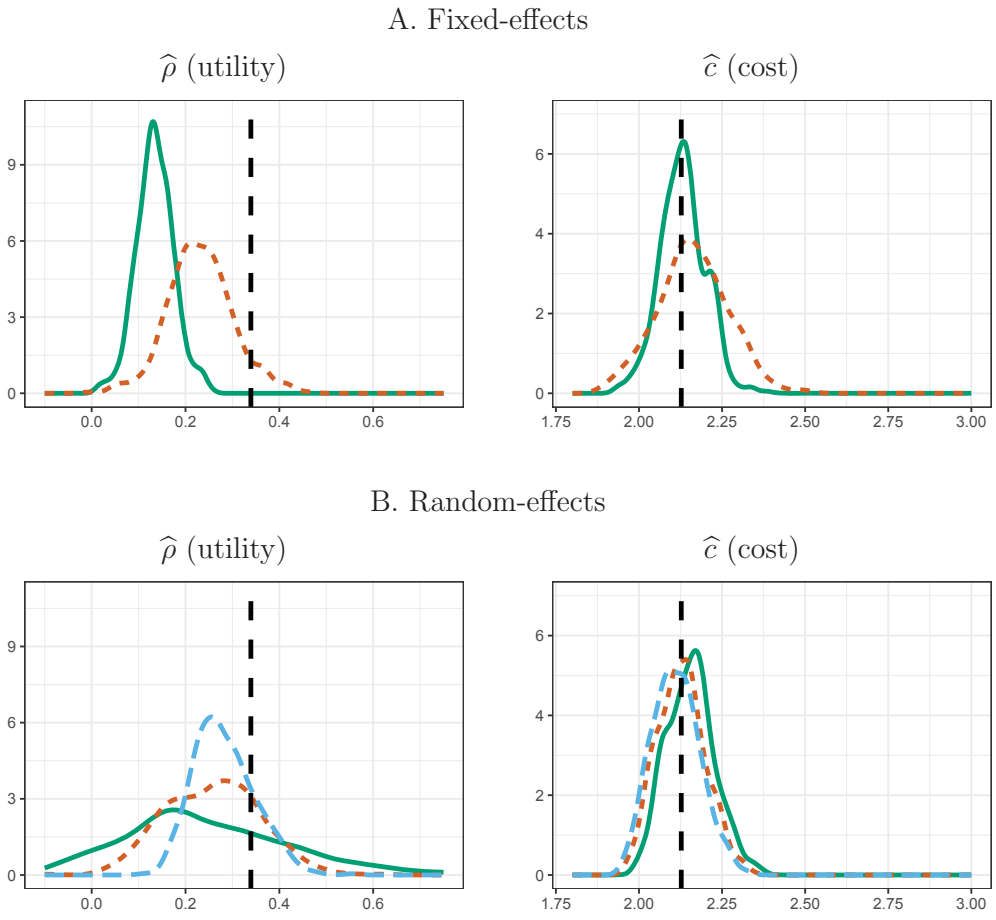


Notes: Solid is two-step GFE, dotted is bias-corrected, dashed is iterated once and bias-corrected, dashed-dotted is iterated three times and bias-corrected. The vertical line indicates the true parameter value. $N = 1889$, $T = 16$. *Unobserved heterogeneity is continuously distributed in the DGP. Costs are homogeneous in Panel A, and heterogeneous in Panel B. We estimate the number of groups* K *in every replication.* 500 *replications.*

provide details and report additional results in Appendix H. The random-effects estimator performs quite well, especially when $K = 8$. Yet, computing a random-effects estimator in this model is more challenging than computing fixed-effects or GFE estimators. In addition, to our knowledge there are no theoretical guarantees for discrete random-effects estimators when population heterogeneity is not discrete.

The fixed-effects estimator and its bias-corrected counterpart perform similarly to two-step

Figure 5: Fixed-effects and random-effects estimates of structural parameters across simulations, model with homogeneous costs



Notes: See the notes to Figure 4. On the top panel we show results for fixed-effects (solid) and bias-corrected fixed-effects (dotted). On the bottom panel we report discrete random-effects estimates with $K = 2$ *(solid),* $K = 4$ *(dotted), and* $K = 8$ *(dashed) groups.* 500 *replications.*

GFE and bias-corrected GFE in the model with homogeneous costs. However, since many individuals do not move between regions during the sample period, fixed-effects estimation is infeasible in the model with cost heterogeneity. Discrete estimation is often used in such contexts, and it is typically justified under the assumption that the types driving returns and costs have a small number of points of support. It is interesting to see that GFE recovers the wage effect ρ and the average of mobility costs $c_i(j)$ well in this case, especially when combined with iteration and bias reduction. This illustrates the ability of GFE to take advantage of the

presence of common features between different dimensions of (continuous) heterogeneity.

Counterfactual. As an example of a counterfactual experiment that the model can be used for, we next compute the average steady-state probability of working in region A when varying the mean log-wage difference between A and B. In Figure 6 we show the log-wage difference on the x-axis, and the probability of working in A on the y-axis. The dashed curves on the graphs show the values in the DGP, while the solid and dotted curves are means and 95% pointwise bands across simulations for the two-step estimator, two-step bias-corrected, iterated once and bias-corrected, and iterated three times and bias-corrected, respectively.

Panel A in Figure 6 shows that the model with homogeneous costs predicts small effects of wages on mobility on average. When increasing the wage in A relative to B by 30% the probability of working in A increases by less than 2 percentage points, from 56.8% to 58.4%. When focusing on workers who are of a "mover type" (in the second row), whose mobility may be affected by the change in wages, we see a more substantial effect, as increasing the wage in region A by 30% increases the long-run probability of working in A from 53.5% to 67.0%. In both cases the two-step estimators are biased downward. In contrast the bias-corrected and bias-corrected iterated estimators are close to unbiased. However they are less precisely estimated, reflecting the estimates of the wage coefficient ρ in Figure 4. The results for the specification with heterogeneous costs in Panel B are similar, with a slightly smaller long-run effect of wages.

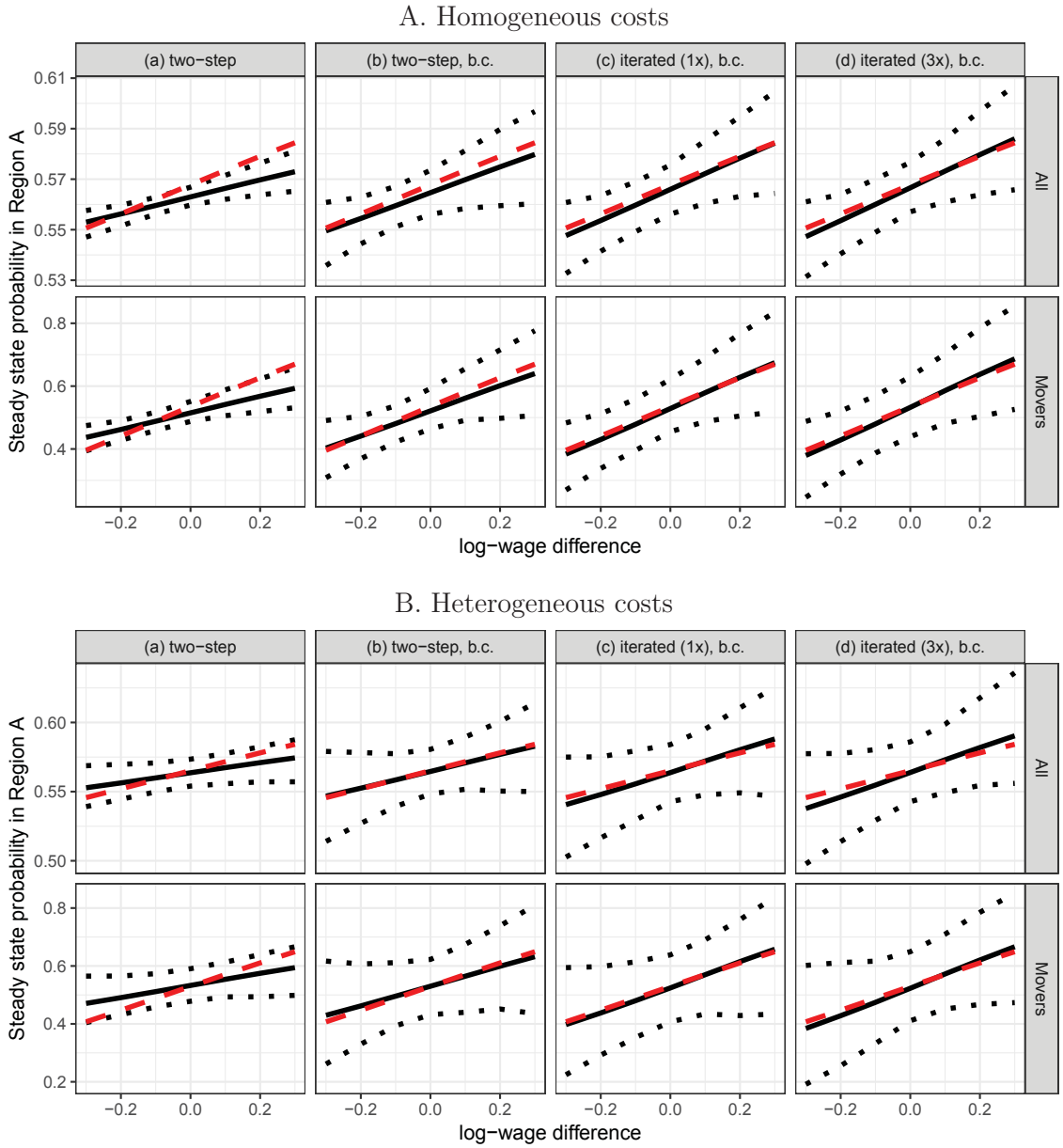
This application illustrates a potential use of discrete GFE estimators in the presence of continuous unobserved heterogeneity. Computation of two-step estimators is easy, and the jackknife bias reduction and iteration provide finite-sample improvements at moderate additional computational cost. This stylized illustration thus suggests that the methods we propose could be useful in structural models.

5.2 Simulations in probit models

In this subsection we study the performance of GFE methods in various specifications of the probit model of Example 2. The purpose of this exercise is twofold: to assess the relevance of our asymptotic results for finite-sample performance in settings where heterogeneity is low-dimensional, and to investigate the ability of the two-step method and its extensions to handle multi-dimensional heterogeneity in models with conditioning covariates.

We start by focusing on a probit model with a time-invariant heterogeneous intercept, as in (6). We consider four DGP that vary in terms of the dimension of heterogeneity. Both DGP 1

Figure 6: Long-run effects of wages



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

and 2 have one covariate, DGP 3 has three covariates, and DGP 4 has five covariates. In all the DGP, μ_{i0} has as many components as covariates, and those components are independent. In addition, we set $\alpha_{i0} = \mu_{i0}$ in DGP 1, while in all the other DGP α_{i0} is not perfectly correlated with the components of μ_{i0}. Hence, the dimension d of heterogeneity in the four DGP is 1, 2, 4, and 6, respectively. In Appendix G we provide details about the DGP.

In Table 1 we report estimates using two-step GFE, iterated GFE, and three versions with a conditional first step based on linear, cubic, and neural network specifications. For two-step GFE, we use the moments $h_i = (\overline{Y}_i, \overline{X}_i')'$. We initialize the iteration and conditional first-step algorithm at two-step estimates, and we iterate ten times. We checked that varying the number of iterations had little impact on the estimates. In all cases, $N = 1000$ and $T = 20$, and we report the results of 500 simulations. We set $K = 10$ for all DGP, and for DGP 1 and 2 we also report results where we estimate K using (4) with $\gamma = 1$ (in rows indicated with a star). In DGP 3 and 4, the dimension of heterogeneity is large, and as a result our rule gives impractical values of K. In Appendix G, for these two DGP we report results based on an alternative rule for K that is tailored to models with conditioning covariates.

In Table 1, we show the mean estimates across simulations of the first component of θ_0 (whose true value is equal to 1), and we report standard deviations in parentheses. The bias of two-step GFE varies substantially with the dimension of heterogeneity. In DGP 1, the dimension is $d = 1$ and two-step GFE is almost unbiased, both when $K = 10$ and when we estimate K using our rule (in DGP 1^*). In DGP 2, where heterogeneity is bi-dimensional, two-step GFE is more biased when $K = 10$, but the bias is small when K is estimated (in DGP 2^*). In that case, our rule, which adapts to the dimension, gives $K \approx 20$ on average. In both DGP 1 and 2, the jackknife tends to reduce the bias of the two-step estimator.

In contrast, in DGP 3 and 4, heterogeneity has dimension $d = 4$ and 6, respectively, and two-step GFE is severely biased. This is in line with our theory, since for such values of d we expect a large approximation error. In addition, given the size of the approximation error, the jackknife does not improve performance in those cases. By comparison, although the fixed-effects estimator suffers from a small-T bias, the half-panel jackknifed fixed-effects estimator is close to unbiased. Hence, fixed-effects and bias-reduced fixed-effects perform better than two-step GFE in DGP 3 and 4.

Interestingly, both the model-based iteration and the GFE method based on a conditional first step outperform the two-step estimator, especially in DGP 3 and 4. The performance of iterated GFE and its jackknifed counterpart is comparable to that of uncorrected and bias-corrected fixed-effects. In the bottom panel of Table 1, we report the results of GFE methods

Table 1: Parameter estimates in a fixed-effects probit model

	Two-step		Iterated		Fixed-effects	
DGP	UC	BC	UC	BC	UC	BC
1	0.985 (0.018)	0.995 (0.020)	1.085 (0.021)	0.993 (0.020)	1.087 (0.021)	0.990 (0.021)
1*	0.985 (0.019)	1.002 (0.019)	1.087 (0.021)	0.999 (0.021)	1.087 (0.021)	0.990 (0.021)
2	0.861 (0.019)	0.841 (0.027)	1.080 (0.021)	0.990 (0.022)	1.089 (0.022)	0.991 (0.021)
2*	0.940 (0.018)	0.970 (0.024)	1.088 (0.022)	0.998 (0.022)	1.089 (0.022)	0.991 (0.021)
3	0.745 (0.019)	0.754 (0.025)	1.112 (0.027)	0.976 (0.028)	1.124 (0.027)	0.980 (0.026)
4	0.747 (0.021)	0.752 (0.027)	1.137 (0.032)	0.966 (0.034)	1.152 (0.031)	0.968 (0.031)
	Conditional first step					
	Linear		Cubic		Neural network	
DGP	UC	BC	UC	BC	UC	BC
1	1.039 (0.019)	0.983 (0.020)	1.057 (0.019)	1.011 (0.021)	1.078 (0.020)	1.027 (0.024)
1*	1.043 (0.019)	1.002 (0.021)	1.059 (0.020)	1.014 (0.022)	1.079 (0.020)	1.027 (0.024)
2	0.971 (0.019)	0.923 (0.026)	1.009 (0.021)	0.978 (0.026)	1.023 (0.021)	1.001 (0.028)
2*	1.001 (0.020)	0.971 (0.024)	1.017 (0.021)	0.991 (0.026)	1.035 (0.023)	1.016 (0.029)
3	0.924 (0.023)	0.917 (0.034)	0.926 (0.024)	0.933 (0.037)	0.986 (0.026)	0.987 (0.042)
4	0.897 (0.026)	0.898 (0.038)	0.882 (0.027)	0.893 (0.042)	0.964 (0.030)	0.978 (0.049)

Notes: Model of Example 2. Means over 500 simulations, and standard deviations in parentheses. The true value of the first component of θ_0 *is* 1. $N = 1000$ *and* $T = 20$. *DGP 1 and 2 have one covariate, DGP 3 has three covariates, DGP 4 has five covariates. In the rows indicated DGP 1–2–3–4,* $K = 10$. *In the rows indicated DGP* 1*–2*, *we estimate* K *using (4) with* $\gamma = 1$. *UC denotes uncorrected, BC denotes bias-corrected. See Appendix G for details.*

based on a conditional first step. We consider three specifications: linear in covariates, cubic, and based on a simple neural network model. In Appendix G we give details on the estimators. The estimators using a conditional first step perform better than baseline two-step GFE in terms of bias, especially in the case of the neural network specification.

Next, we show results for two probit models with richer heterogeneity. In Panel A of Table 2, we show results for a model where the intercept α_{it0} is time-varying, specified as a one-factor model with a time trend as the time-varying factor. Notice that we do not assume knowledge of the factor structure in estimation. In particular, there is no feasible fixed-effects estimator in this DGP. Although the stationary assumption of our theory is not met, time trends are often used in applied work, and it is of interest to see how GFE performs in this context. We use the moments $h_i = (\overline{Y}_i, \overline{X}_i')'$. Moreover, to implement the conditional first step we simply add the time trend to the vector of covariates.

Panel A of Table 2 shows that, as in the specification with time-invariant heterogeneity, two-step GFE performs well when the dimension d is small, especially when using our rule to estimate K and half-panel jackknife. However, as in the model with time-invariant fixed-effects, two-step GFE is severely biased in DGP 3 and 4, due to the presence of a large dimension of heterogeneity. In those two DGP, iterated GFE and the methods based on a conditional first step tend to improve performance relative to the two-step approach. In Tables S3 and S4 in Appendix G, we report results using a rule for K based on a conditional first step.

In Panel B of Table 2, we consider a random coefficients model where the coefficient of the first covariate is constant equal to one, and the other coefficients are now heterogeneous and perfectly collinear with the intercept α_{i0}. Hence, while there are three random coefficients in DGP 3 and five in DGP 4 including the intercept, the dimension d of heterogeneity is the same as in the fixed-effects probit model. For two-step GFE we use the moments $h_i = (\overline{Y}_i, \overline{X}_i', \overline{YX}_i')'$, where $\overline{YX}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}X_{it}$. We verified analytically that the injectivity requirement in Assumption 2 is satisfied.

Panel B of Table 2 shows that, in this setting, fixed-effects does not perform well since it is substantially biased even after jackknife bias reduction. Indeed, fixed-effects does not exploit the presence of a low underlying dimension of heterogeneity, and the number of fixed-effects to estimate is too large given the length of the panel. Notice also the very large standard deviation of the fixed-effects estimator. By comparison, although two-step GFE is biased in these DGP where heterogeneity has a large dimension due to the presence of multiple covariates, both iterated GFE and conditional first step methods perform only slightly worse than in the model without random coefficients.

Table 2: Parameter estimates in probit models with richer heterogeneity

A. Probit with time-varying unobservables

	Two-step		Iterated		Fixed-effects	
DGP	UC	BC	UC	BC	UC	BC
1	1.013 (0.021)	1.010 (0.026)	1.284 (0.037)	1.021 (0.064)	–	–
1*	1.015 (0.021)	1.028 (0.025)	1.319 (0.039)	1.093 (0.070)	–	–
2	0.736 (0.024)	0.744 (0.039)	1.116 (0.035)	1.066 (0.061)	–	–
2*	0.889 (0.023)	0.935 (0.037)	1.278 (0.044)	1.196 (0.080)	–	–
3	0.563 (0.027)	0.548 (0.040)	1.245 (0.046)	1.090 (0.079)	–	–
4	0.570 (0.027)	0.555 (0.040)	1.310 (0.053)	1.057 (0.089)	–	–

	Conditional first step					
	Linear		Cubic		Neural network	
DGP	UC	BC	UC	BC	UC	BC
1	1.095 (0.023)	0.997 (0.031)	1.169 (0.028)	1.057 (0.041)	1.181 (0.029)	1.117 (0.043)
1*	1.110 (0.023)	1.027 (0.032)	1.186 (0.027)	1.087 (0.040)	1.195 (0.029)	1.143 (0.045)
2	0.948 (0.024)	0.936 (0.041)	0.994 (0.026)	0.970 (0.051)	1.003 (0.032)	1.027 (0.057)
2*	1.041 (0.028)	1.056 (0.046)	1.096 (0.035)	1.098 (0.060)	1.085 (0.035)	1.144 (0.062)
3	0.890 (0.035)	0.897 (0.062)	0.778 (0.039)	0.788 (0.071)	0.976 (0.044)	1.024 (0.087)
4	0.852 (0.041)	0.877 (0.077)	0.722 (0.036)	0.738 (0.064)	0.938 (0.052)	0.996 (0.103)

B. Probit with random coefficients

	Two-step		Iterated		Fixed-effects	
DGP	UC	BC	UC	BC	UC	BC
3	0.707 (0.023)	0.716 (0.029)	1.117 (0.028)	0.949 (0.037)	1.438 (0.047)	0.854 (0.077)
4	0.615 (0.026)	0.622 (0.034)	1.110 (0.033)	0.946 (0.048)	2.173 (0.142)	0.601 (0.815)

	Conditional first step					
	Linear		Cubic		Neural network	
DGP	UC	BC	UC	BC	UC	BC
3	0.909 (0.027)	0.854 (0.041)	0.931 (0.029)	0.900 (0.044)	1.019 (0.029)	0.947 (0.051)
4	0.841 (0.038)	0.790 (0.069)	0.837 (0.036)	0.841 (0.063)	1.007 (0.043)	0.907 (0.082)

Notes: See notes to Table 1. Panel A corresponds to an extension of the model of Example 2 with a time-varying intercept α_{it0} that has a one-factor structure. Panel B corresponds to a specification with heterogeneous θ_{i0}, whose components are driven by a single scalar type. See Appendix G for details.

Overall, these simulations and the ones we report in Appendix G suggest that, while two-step GFE may perform well when the dimension of heterogeneity in outcomes and covariates is small as formalized in Assumption 1, iterated GFE and methods with a conditional first step can improve performance in settings with multiple heterogeneous covariates.

6 Conclusion

Two-step grouped fixed-effects (GFE) methods based on an initial data-driven classification are effective dimension reduction devices. In this paper we analyze their properties in settings where population heterogeneity is not discrete. Our framework relies on two main assumptions: low-dimensional individual heterogeneity, and the availability of moments to approximate the latent types. In many economic models, individual types are low-dimensional. By taking advantage of this feature, GFE can allow for rich, time-varying forms of heterogeneity. Moments can be constructed from the panel data, and they can be combined with external measurements of the types when available.

Unlike in environments where population heterogeneity is discrete and groups can be estimated consistently, in our asymptotic framework we acknowledge that incidental parameter bias and approximation error play a key role. We show theoretically and through simulations that combining GFE with bias reduction techniques may be beneficial. We propose two alternatives to two-step GFE, a conditional method and a model-based iteration, which we show in simulations can reduce approximation error when the dimension of heterogeneity in outcomes and covariates is not small. Developing the theory for such extensions is an important avenue for future work. Another useful direction for the theory will be to account for first-step numerical optimization error in the analysis of the statistical properties of GFE estimators.

Finally, GFE methods could be of interest beyond our illustrations. In the spirit of the dynamic migration model that we have analyzed, we think GFE is a useful tool for structural estimation. Another example is given by our illustration on matched employer employee data in Appendix I, where we evaluate the performance of the two-step GFE method of Bonhomme, Lamadon and Manresa (2019) in the presence of continuous firm heterogeneity. Other potential applications include models with multi-sided heterogeneity, nonlinear factor models, nonparametric or semi-parametric panel data models such as quantile regression with individual effects, and network models.

References

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

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

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

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

[5] 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.

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

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

[8] Bai, J., and T. Ando (2016):"Panel Data Models with Grouped Factor Structure Under Unknown Group Membership," *Journal of Applied Econometrics*, 31(1), 163–191.

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

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

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

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

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

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

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

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

[17] 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.

[18] Frederiksen, A., B. E. Honoré, and L. Hu (2007): "Discrete Time Duration Models with Group-Level Heterogeneity," *Journal of Econometrics*, 141(2), 1014–1043.

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

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

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

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

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

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

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

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

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

[28] 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.

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

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

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

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

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

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

[35] Moon, H. R., M. Shum, and M. Weidner (2018): "Estimation of Random Coefficients Logit Demand Models with Interactive Fixed Effects," *Journal of Econometrics*, 206(2), 613–644.

[36] Newey, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79(1), 147–168.

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

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

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

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

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

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

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

[44] Severini, T. A., and W. H. Wong (1992): "Profile Likelihood and Conditionally Parametric Models," *The Annals of Statistics*, 20(4), 1768–1802.

[45] Späth, H. (1979): "Algorithm 39: Clusterwise Linear Regression," *Computing*, 22(4), 367–373.

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

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

[48] Su, L., Z. Shi, and P. C. B. Phillips (2016): "Identifying Latent Structures in Panel Data," *Econometrica*, 84(6), 2215–2264.

[49] Vogt and O. Linton (2016): "Classification of Nonparametric Regression Functions in Longitudinal Data Models," *Journal of the Royal Statistical Society: Series B*, 79, 5–27.

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

APPENDIX

A Proofs

A.1 Proof of Lemma 1

Let us define, similarly to (7):

$$B_{\varphi(\xi)}(K) = \min_{(\widetilde{h},\{k_i\})} \frac{1}{N}\sum_{i=1}^{N} \left\|\varphi(\xi_{i0}) - \widetilde{h}(k_i)\right\|^2,$$

and let us denote:

$$(\underline{h},\{\underline{k}_i\}) = \operatorname*{argmin}_{(\widetilde{h},\{k_i\})} \sum_{i=1}^{N} \left\|\varphi(\xi_{i0}) - \widetilde{h}(k_i)\right\|^2, \tag{A1}$$

where we refer to Chapter 1 in Graf and Luschgy (2000) for general results on existence of optimal empirical quantizers; that is, solutions to (A1).

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

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

Letting $\varepsilon_i = h_i - \varphi(\xi_{i0})$, we thus have, using the triangle inequality twice:

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

Now, by Assumption 2, $\frac{1}{N}\sum_{i=1}^{N} \|\varepsilon_i\|^2 = O_p(1/S)$. 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'). We thus have:

$$B_{\varphi(\xi)}(K) = \min_{(\widetilde{h},\{k_i\})} \frac{1}{N}\sum_{i=1}^{N} \left\|\varphi(\xi_{i0}) - \widetilde{h}(k_i)\right\|^2 \leq \tau^2 \min_{(\widetilde{\xi},\{k_i\})} \frac{1}{N}\sum_{i=1}^{N} \left\|\xi_{i0} - \widetilde{\xi}(k_i)\right\|^2 = \tau^2 B_{\xi}(K).$$

Hence $B_{\varphi(\xi)}(K) = O_p(B_{\xi}(K))$, and Lemma 1 follows.

A.2 Proof of Corollary 1

Using Lemmas 1 and 2, and that $\alpha_{it0} = \alpha(\psi(\varphi(\xi_{i0})), \lambda_{t0})$ with α and ψ Lipschitz-continuous, we have:

$$\begin{aligned}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\alpha_{it0} - \alpha(\psi(\widehat{h}(\widehat{k}_i)), \lambda_{t0})\|^2 &= O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\varphi(\xi_{i0}) - \widehat{h}(\widehat{k}_i)\right\|^2\right) \\ &= O_p(1/S) + O_p(K^{-\frac{2}{d}}).\end{aligned}$$

Hence:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\alpha_{it0} - \overline{\alpha}_{t0}(\widehat{k}_i)\right\|^2 \leq \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\alpha_{it0} - \alpha(\psi(\widehat{h}(\widehat{k}_i)), \lambda_{t0})\|^2 = O_p(1/S) + O_p(K^{-\frac{2}{d}}).$$

This shows Corollary 1.

A.3 Proof of Theorem 1

We will use the following notation: $v_{ij} = \frac{\partial \ell_{ij}}{\partial \alpha}$, $v_{ij}^{\alpha} = \frac{\partial^2 \ell_{ij}}{\partial \alpha \partial \alpha'}$, $v_{ij}^{\theta} = \frac{\partial^2 \ell_{ij}}{\partial \theta \partial \alpha'}$, 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 $k \in \{1, ..., K\}$:

$$\widehat{\alpha}^j(k, \theta) = \text{argmax}_{\alpha} \sum_{i=1}^{N} \mathbf{1}\{\widehat{k}_i = k\}\ell_{ij}(\alpha, \theta), \tag{A2}$$

and define $\overline{\alpha}^j(\theta, \xi)$ according to Assumption 3 (*iii*). Let also $\delta = \frac{1}{S} + \frac{Kp}{NT} + K^{-\frac{2}{d}}$.

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

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

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

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

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

Next, since:

$$s_i = \frac{1}{p}\sum_{j=1}^{p}\frac{\partial}{\partial \theta}\Big|_{\theta_0}\ell_{ij}\left(\overline{\alpha}^j(\theta, \xi_{i0}), \theta\right), \quad H = -\mathbb{E}\left[\frac{1}{p}\sum_{j=1}^{p}\frac{\partial^2}{\partial \theta \partial \theta'}\Big|_{\theta_0}\ell_{ij}\left(\overline{\alpha}^j(\theta, \xi_{i0}), \theta\right)\right], \tag{A5}$$

equation (9) will then follow from approximating $\frac{\partial}{\partial\theta'}\big|_{\widetilde{\theta}}\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\frac{\partial\ell_{ij}(\widehat{\alpha}^j(\widehat{k}_i,\theta),\theta)}{\partial\theta}$ by its value at θ_0, using that $\widetilde{\theta}$ is consistent, and using either part (*iia*) (in part (a) of the theorem) or part (*iib*) (in part (b) of the theorem) in Assumption 3. (10) will then follow.

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

$$\sup_{\theta\in\Theta}\left|\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\ell_{ij}\left(\widehat{\alpha}^j(\widehat{k}_i,\theta),\theta\right)-\underbrace{\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\ell_{ij}\left(\overline{\alpha}^j\left(\theta,\xi_{i0}\right),\theta\right)}_{=\overline{\ell}(\theta)}\right|=o_p(1). \tag{A6}$$

Compactness of the parameter space, continuity of the target likelihood $\overline{\ell}(\theta)$, and identification of θ_0 (from part (*iii*) in Assumption 3), will then imply that $\widehat{\theta}$ is consistent for θ_0 (e.g., Theorem 2.1 in Newey and McFadden, 1994).

From Assumption 3 (*iii*) and (*iv*) we have that both:

$$\frac{\partial\overline{\alpha}^j\left(\theta,\widetilde{\xi}\right)}{\partial\theta'}=\mathbb{E}_{\xi_{i0}=\widetilde{\xi},\lambda_0}\left[-v_{ij}^{\alpha}\left(\overline{\alpha}^j(\theta,\widetilde{\xi}),\theta\right)\right]^{-1}\mathbb{E}_{\xi_{i0}=\widetilde{\xi},\lambda_0}\left[v_{ij}^{\theta}\left(\overline{\alpha}^j(\theta,\widetilde{\xi}),\theta\right)\right]'$$

and:

$$\frac{\partial\overline{\alpha}^j\left(\theta,\widetilde{\xi}\right)}{\partial\xi'}=\mathbb{E}_{\xi_{i0}=\widetilde{\xi},\lambda_0}\left[-v_{ij}^{\alpha}\left(\overline{\alpha}^j(\theta,\widetilde{\xi}),\theta\right)\right]^{-1}\frac{\partial}{\partial\xi'}\bigg|_{\xi=\widetilde{\xi}}\mathbb{E}_{\xi_{i0}=\xi,\lambda_0}\left[v_{ij}\left(\overline{\alpha}^j(\theta,\widetilde{\xi}),\theta\right)\right] \tag{A7}$$

are uniformly $O_p(1)$ ($O(1)$ in models with time-invariant heterogeneity).

Let $a^j(k,\theta)=\overline{\alpha}^j(\theta,\psi(\widehat{h}(k)))$. We have:

$$\begin{aligned}
&\sup_{\theta\in\Theta}\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|a^j(\widehat{k}_i,\theta)-\overline{\alpha}^j\left(\theta,\xi_{i0}\right)\right\|^2\\
&=\sup_{\theta\in\Theta}\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\overline{\alpha}^j(\theta,\psi(\widehat{h}(\widehat{k}_i)))-\overline{\alpha}^j\left(\theta,\psi(\varphi(\xi_{i0}))\right)\right\|^2\\
&=O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{h}(\widehat{k}_i)-\varphi(\xi_{i0})\right\|^2\right)=O_p(\delta),
\end{aligned} \tag{A8}$$

where we have used Lemmas 1 and 2, Assumption 2, and that $\frac{\partial\overline{\alpha}^j(\theta,\xi)}{\partial\xi'}$ is uniformly $O_p(1)$ and ψ is Lipschitz-continuous.

Now, we have, for all θ:

$$\sum_{i=1}^{N}\sum_{j=1}^{p}\ell_{ij}\left(a^j(\widehat{k}_i,\theta),\theta\right)\leq\sum_{i=1}^{N}\sum_{j=1}^{p}\ell_{ij}\left(\widehat{\alpha}^j(\widehat{k}_i,\theta),\theta\right). \tag{A9}$$

Moreover, expanding, we have:

$$\begin{aligned}
&\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\ell_{ij}(\widehat{\alpha}^j(\widehat{k}_i,\theta),\theta)-\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\ell_{ij}(\overline{\alpha}^j(\theta,\xi_{i0}),\theta)\\
&=\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}v_{ij}(\overline{\alpha}^j(\theta,\xi_{i0}),\theta)'\left(\widehat{\alpha}^j(\widehat{k}_i,\theta)-\overline{\alpha}^j(\theta,\xi_{i0})\right)\\
&\quad+\frac{1}{2Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left(\widehat{\alpha}_j(\widehat{k}_i,\theta)-\overline{\alpha}^j(\theta,\xi_{i0})\right)'v_{ij}^{\alpha}(a_{ij}(\theta),\theta)\left(\widehat{\alpha}^j(\widehat{k}_i,\theta)-\overline{\alpha}^j(\theta,\xi_{i0})\right),
\end{aligned}\tag{A10}$$

and:

$$\begin{aligned}
&\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\ell_{ij}\left(a^j\left(\widehat{k}_i,\theta\right),\theta\right)-\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\ell_{ij}(\overline{\alpha}^j(\theta,\xi_{i0}),\theta)\\
&=\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}v_{ij}(\overline{\alpha}^j(\theta,\xi_{i0}),\theta)'\left(a^j\left(\widehat{k}_i,\theta\right)-\overline{\alpha}^j(\theta,\xi_{i0})\right)\\
&\quad+\frac{1}{2Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left(a^j\left(\widehat{k}_i,\theta\right)-\overline{\alpha}^j(\theta,\xi_{i0})\right)'v_{ij}^{\alpha}(b_{ij}(\theta),\theta)\left(a^j\left(\widehat{k}_i,\theta\right)-\overline{\alpha}^j(\theta,\xi_{i0})\right)\\
&=\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}v_{ij}(\overline{\alpha}^j(\theta,\xi_{i0}),\theta)'\left(a^j\left(\widehat{k}_i,\theta\right)-\overline{\alpha}^j(\theta,\xi_{i0})\right)+O_p\left(\delta\right),
\end{aligned}\tag{A11}$$

where $a_{ij}(\theta)$ lies between $\widehat{\alpha}^j(\widehat{k}_i,\theta)$ and $\overline{\alpha}^j(\theta,\xi_{i0})$, and $b_{ij}(\theta)$ lies between $a^j(\widehat{k}_i,\theta)$ and $\overline{\alpha}^j(\theta,\xi_{i0})$. Here we have used (A8), and that, either by Assumption 3 (*iia*) (part (a) of the theorem), or by Assumption 3 (*iib*) (part (b) of the theorem):

$$\max_{i,j}\sup_{(\alpha,\theta)}\|v_{ij}^{\alpha}(\alpha,\theta)\|=O_p(1).\tag{A12}$$

Hence, using (A9), (A10), and (A11):

$$\begin{aligned}
&\frac{1}{2Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left(\widehat{\alpha}^j(\widehat{k}_i,\theta)-\overline{\alpha}^j(\theta,\xi_{i0})\right)'\left[-v_{ij}^{\alpha}(a_{ij}(\theta),\theta)\right]\left(\widehat{\alpha}^j(\widehat{k}_i,\theta)-\overline{\alpha}^j(\theta,\xi_{i0})\right)\\
&\leq\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}v_{ij}(\overline{\alpha}^j(\theta,\xi_{i0}),\theta)'\left(\widehat{\alpha}^j(\widehat{k}_i,\theta)-a^j(\widehat{k}_i,\theta)\right)+O_p(\delta)\\
&=\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\overline{v}_j(\widehat{k}_i,\theta)'\left(\widehat{\alpha}^j(\widehat{k}_i,\theta)-a^j(\widehat{k}_i,\theta)\right)+O_p(\delta),
\end{aligned}$$

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.

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

$$\min_{i,j}\inf_{(\alpha,\theta)}\text{mineig}\left[-v_{ij}^{\alpha}(\alpha,\theta)\right]\geq\underline{c}+o_p(1),\tag{A13}$$

where mineig(A) is the minimum eigenvalue of A.

Using the Cauchy Schwarz inequality, we thus have:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\widehat{\alpha}^j(\widehat{k}_i,\theta)-\overline{\alpha}^j(\theta,\xi_{i0})\right\|^2$$
$$\leq O_p\left[\left(\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\|\overline{v}_j(\widehat{k}_i,\theta)\|^2\right)^{\frac{1}{2}}\left(\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\widehat{\alpha}^j(\widehat{k}_i,\theta)-a^j(\widehat{k}_i,\theta)\right\|^2\right)^{\frac{1}{2}}\right]+O_p(\delta).$$

Let: $A\equiv\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\|\widehat{\alpha}^j(\widehat{k}_i,\theta)-\overline{\alpha}^j(\theta,\xi_{i0})\|^2$. By (A8) and the triangle inequality we have:

$$\left(\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\widehat{\alpha}^j(\widehat{k}_i,\theta)-a^j(\widehat{k}_i,\theta)\right\|^2\right)^{\frac{1}{2}}\leq A^{\frac{1}{2}}+O_p(\delta^{\frac{1}{2}}).$$

Hence:

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

which implies:

$$A=O_p\left(\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\|\overline{v}_j(\widehat{k}_i,\theta)\|^2\right)+O_p(\delta). \tag{A14}$$

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

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\overline{v}_j(\widehat{k}_i,\theta)\right\|^2=O_p\left(\delta\right), \tag{A15}$$

which, using (A14), will imply:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\widehat{\alpha}^j(\widehat{k}_i,\theta)-\overline{\alpha}^j(\theta,\xi_{i0})\right\|^2=O_p(\delta). \tag{A16}$$

To show (A15), let, for all j,θ,h,ξ,λ:

$$\rho_j(h,\xi,\lambda,\theta)=\mathbb{E}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}(v_{ij}(\overline{\alpha}^j(\theta,\xi),\theta)),$$

and let, 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 3 (v), expanding $\rho_j(h_i,\xi_{i0},\lambda_0,\theta)$ around $h_i=\varphi(\xi_{i0})$, and using that $h_i=\varphi(\xi_{i0})+\varepsilon_i$, we have:

$$\rho_j(h_i,\xi_{i0},\lambda_0,\theta)=\rho_j(\varphi(\xi_{i0}),\xi_{i0},\lambda_0,\theta)+\frac{\partial\rho_j(\varphi(\xi_{i0}),\xi_{i0},\lambda_0,\theta)}{\partial h'}\varepsilon_i+O_p\left(\frac{1}{S}\right),$$

where the $O_p(1/S)$ is uniform in i, j, θ. Hence, taking expectations:

$$\begin{aligned}0 =&\mathbb{E}_{\xi_{i0},\lambda_0}(v_{ij}(\overline{\alpha}^j(\theta,\xi_{i0}),\theta))) = \mathbb{E}_{\xi_{i0},\lambda_0}\left[\rho_j(h_i,\xi_{i0},\lambda_0,\theta)\right]\\ =&\rho_j(\varphi(\xi_{i0}),\xi_{i0},\lambda_0,\theta) + \frac{\partial\rho_j(\varphi(\xi_{i0}),\xi_{i0},\lambda_0,\theta)}{\partial h'}\mathbb{E}_{\xi_{i0},\lambda_0}(\varepsilon_i) + O_p\left(\frac{1}{S}\right).\end{aligned}$$

Hence, using that, by Assumption 3 (v), $\frac{\partial\rho_j(\varphi(\xi_{i0}),\xi_{i0},\lambda_0,\theta)}{\partial h'} = O_p(1)$ uniformly in i, j, and using Assumption 2, we obtain:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\|\rho_j(h_i,\xi_{i0},\lambda_0,\theta)\|^2 = O_p\left(\frac{1}{S}\right).$$

It thus follows that:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\|\overline{v}_j(\widehat{k}_i,\theta)\|^2 \le O_p\left(\frac{1}{S}\right) + \frac{2}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\|\overline{\zeta}_j(\widehat{k}_i,\theta)\|^2, \tag{A17}$$

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

Now:

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\|\overline{\zeta}_j(\widehat{k}_i,\theta)\|^2\right]\\ &= \frac{1}{Np}\sum_{k=1}^{K}\sum_{j=1}^{p}\mathbb{E}\left[\frac{\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\mathbf{1}\{\widehat{k}_{i_1}=k\}\mathbf{1}\{\widehat{k}_{i_2}=k\}\mathbb{E}_{h_1,...,h_N,\xi_{10},...,\xi_{N0},\lambda_0}\left(\zeta_{i_1,j}(\theta)'\zeta_{i_2,j}(\theta)\right)}{\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}}\right],\end{aligned}$$

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

Furthermore, since observations are independent across i given λ_0:

$$\begin{aligned}&\mathbb{E}_{h_1,...,h_N,\xi_{10},...,\xi_{N0},\lambda_0}\left(\zeta_{i_1,j}(\theta)'\zeta_{i_2,j}(\theta)\right)\\ &= \mathbb{E}_{h_{i_1},\xi_{i_1,0},\lambda_0}\left(\zeta_{i_1,j}(\theta)\right)'\mathbb{E}_{h_{i_2},\xi_{i_2,0},\lambda_0}\left(\zeta_{i_2,j}(\theta)\right) = 0 \quad \text{for all } i_1 \neq i_2 \text{ and } j.\end{aligned}$$

Hence:

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

Finally, using that $\mathbb{E}_{h_i,\xi_{i0},\lambda_0}\left(\zeta_{ij}(\theta)\right) = 0$, and using part (v) in Assumption 3:

$$\mathbb{E}_{h_i,\xi_{i0},\lambda_0}\left(\zeta_{ij}(\theta)'\zeta_{ij}(\theta)\right) = \mathrm{Tr}\left[\mathrm{Var}_{h_i,\xi_{i0},\lambda_0}(v_{ij}(\overline{\alpha}^j(\theta,\xi_{i0}),\theta))\right] = O_p\left(\frac{1}{R}\right),$$

uniformly in i, j. Note that the dimension of v_{ij} is fixed, independent of the sample size.

Hence, since $T = pR$:

$$\mathbb{E}\left[\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\|\overline{\zeta}_j(\widehat{k}_i,\theta)\|^2\right] = O\left(\frac{Kp}{NpR}\right) = O\left(\frac{Kp}{NT}\right).$$

This implies (A15), and shows (A16).

We are now going to show that:

$$\sup_{\theta\in\Theta}\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\|\overline{v}_j(\widehat{k}_i,\theta)\|^2=o_p\left(1\right). \tag{A18}$$

Using a bounding argument similar to the one we used to show (A16), we will then obtain that:

$$\sup_{\theta\in\Theta}\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\widehat{\alpha}^j(\widehat{k}_i,\theta)-\overline{\alpha}^j(\theta,\xi_{i0})\right\|^2=o_p\left(1\right). \tag{A19}$$

To see that (A18) holds, let $Z(\theta)=\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\|\overline{v}_j(\widehat{k}_i,\theta)\|^2$. By (A15), $Z(\theta)=O_p(Kp/NT)$ for all $\theta\in\Theta$. Moreover:

$$\frac{\partial Z(\theta)}{\partial\theta}=\frac{2}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\overline{v}_j^{\theta}(\widehat{k}_i,\theta)\overline{v}_j(\widehat{k}_i,\theta)=O_p\left(\sqrt{\sup_{\theta\in\Theta}Z(\theta)}\right),$$

uniformly in θ, using the Cauchy Schwarz inequality with either part (iia) or part (iib) in Assumption 3. Here $\overline{v}_j^{\theta}(k,\theta)$ denotes the mean of $v_{ij}^{\theta}(\overline{\alpha}^j(\theta,\xi_{i0}),\theta)$ in group $\widehat{k}_i=k$. Since Θ is compact it follows that $\sup_{\theta\in\Theta}Z(\theta)=o_p(1)$.[5]

Taylor expanding one more time, and using (A12), (A18), (A19), and the Cauchy Schwarz inequality, we obtain that (A6) holds. Consistency of $\widehat{\theta}$ follows.

Proof of (A3). We are now going to show (A3). First, evaluating (A16) at θ_0 we obtain:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\widehat{\alpha}^j(\widehat{k}_i,\theta_0)-\alpha_{i0}^j\right\|^2=O_p(\delta). \tag{A20}$$

To show (A3), we are now going to show that:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\{v_{ij}^{\theta}\left(\widehat{\alpha}^j(\widehat{k}_i)-\alpha_{i0}^j\right)+\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\theta}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\alpha}\right)\right]^{-1}v_{ij}\right\}=O_p\left(\delta\right), \tag{A21}$$

where from now on we omit references to θ_0 and α_{i0}^j for conciseness.

[5]Let $\upsilon>0,\epsilon>0$. There is a constant $M>0$ such that $\Pr\left(\sup_{\theta\in\Theta}\left\|\frac{\partial Z(\theta)}{\partial\theta}\right\|>M\sqrt{\sup_{\theta\in\Theta}Z(\theta)}\right)<\frac{\epsilon}{2}$. Take a finite cover of $\Theta=B_1\cup...\cup B_R$, where B_r are balls with centers θ_r and 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 (A15), is smaller than ϵ for N,S,K large enough.

Note that (A20) and (A21) will indeed imply (A3), since:

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

where we have expanded $v^\theta_{ij}(\widehat{\alpha}_j(\widehat{k}_i))$ around $\overline{\alpha}^j(\theta_0,\xi_{i0}) = \alpha^j_{i0}$, and used (A20) together with (A12) and either part (iia) or part (iib) in Assumption 3.

To show (A21) we will bound, in turn:

$$\begin{aligned}
A &\equiv \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\mathbb{E}_{\xi_{i0},\lambda_0}\left(v^\theta_{ij}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_0}\left(v^\alpha_{ij}\right)\right]^{-1}v^\alpha_{ij}\left(\widehat{\alpha}^j(\widehat{k}_i)-\alpha^j_{i0}+(v^\alpha_{ij})^{-1}v_{ij}\right),\\
B &\equiv \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left(v^\theta_{ij}\left(v^\alpha_{ij}\right)^{-1} - \mathbb{E}_{\xi_{i0},\lambda_0}\left(v^\theta_{ij}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_0}\left(v^\alpha_{ij}\right)\right]^{-1}\right)v^\alpha_{ij}\left(\widehat{\alpha}^j(\widehat{k}_i)-\alpha^j_{i0}\right).
\end{aligned}$$

Let us start with bounding A. We have, for all k, j:

$$\begin{aligned}
0 &= \sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i = k\}v_{ij}(\widehat{\alpha}^j(k))\\
&= \sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i = k\}v_{ij}(\alpha^j_{i0}) + \sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i = k\}v^\alpha_{ij}(\alpha^j_{i0})(\widehat{\alpha}^j(\widehat{k}_i)-\alpha^j_{i0})\\
&\quad + \frac{1}{2}\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i = k\}v^{\alpha\alpha}_{ij}(a_{ij})\left(\widehat{\alpha}^j(\widehat{k}_i)-\alpha^j_{i0}\right)\otimes\left(\widehat{\alpha}^j(\widehat{k}_i)-\alpha^j_{i0}\right),
\end{aligned}$$

for a_{ij} between α^j_{i0} and $\widehat{\alpha}^j(\widehat{k}_i)$.

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:

$$\begin{aligned}
\widetilde{\alpha}_j(k) &= \left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i = k\}(-v^\alpha_{ij})\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i = k\}(-v^\alpha_{ij})\alpha^j_{i0}\right),\\
\widetilde{v}_j(k) &= \left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i = k\}(-v^\alpha_{ij})\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i = k\}v_{ij}\right),
\end{aligned}$$

and:

$$\widetilde{w}_j(k) = \frac{1}{2}\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i = k\}(-v^\alpha_{ij})\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i = k\}v^{\alpha\alpha}_{ij}(a_{ij})\left(\widehat{\alpha}^j(\widehat{k}_i)-\alpha^j_{i0}\right)\otimes\left(\widehat{\alpha}^j(\widehat{k}_i)-\alpha^j_{i0}\right)\right). \tag{A22}$$

Hence. we obtain:

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

Note first that:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\theta}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\alpha}\right)\right]^{-1}v_{ij}^{\alpha}\widetilde{w}_j(\widehat{k}_i) = O_p\left(\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\widehat{\alpha}^{j}(\widehat{k}_i)-\alpha_{i0}^{j}\right\|^2\right)=O_p(\delta), \tag{A23}$$

where we have used (A12), (A20), (A22), and either part (iia) or part (iib) in Assumption 3.

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

$$\begin{aligned}&\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\theta}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\alpha}\right)\right]^{-1}v_{ij}^{\alpha}\left(\widetilde{\alpha}_j(\widehat{k}_i)-\alpha_{i0}^{j}\right)\\ &=\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left(z_j(\xi_{i0})'-\widetilde{z}_j\left(\widehat{k}_i\right)'\right)v_{ij}^{\alpha}\left(\widetilde{\alpha}_j(\widehat{k}_i)-\alpha_{i0}^{j}\right),\end{aligned} \tag{A24}$$

where, for all k, j:

$$\widetilde{z}_j(k)=\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}(-v_{ij}^{\alpha})\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}(-v_{ij}^{\alpha})z_j(\xi_{i0})\right). \tag{A25}$$

Now we have, using that $\widetilde{\alpha}_j = \operatorname{argmin}_{(\alpha_j(1),\ldots,\alpha_j(K))}\sum_{i=1}^{N}\left(\alpha^j(\widehat{k}_i)-\alpha_{i0}^{j}\right)'(-v_{ij}^{\alpha})\left(\alpha^j(\widehat{k}_i)-\alpha_{i0}^{j}\right)$, and using (A12), (A13), and (A20):

$$\begin{aligned}&\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\widetilde{\alpha}_j(\widehat{k}_i)-\alpha_{i0}^{j}\right\|^2=O_p\left(\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left(\widetilde{\alpha}_j(\widehat{k}_i)-\alpha_{i0}^{j}\right)'(-v_{ij}^{\alpha})\left(\widetilde{\alpha}_j(\widehat{k}_i)-\alpha_{i0}^{j}\right)\right)\\ &=O_p\left(\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left(\widehat{\alpha}^{j}(\widehat{k}_i)-\alpha_{i0}^{j}\right)'(-v_{ij}^{\alpha})\left(\widehat{\alpha}^{j}(\widehat{k}_i)-\alpha_{i0}^{j}\right)\right)\\ &=O_p\left(\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\widehat{\alpha}^{j}(\widehat{k}_i)-\alpha_{i0}^{j}\right\|^2\right)=O_p(\delta).\end{aligned}$$

Likewise, since by Assumption 3 (iv) $\frac{\partial \operatorname{vec} z_j(\xi)}{\partial \xi'}$ is $O_p(1)$ uniformly in j and ξ ($O(1)$ in models with

time-invariant heterogeneity), we have:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\widetilde{z}_j(\widehat{k}_i)-z_j(\xi_{i0})\right\|^2 = O_p\left(\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left(\widetilde{z}_j(\widehat{k}_i)-z_j(\xi_{i0})\right)'(-v_{ij}^{\alpha})\left(\widetilde{z}_j(\widehat{k}_i)-z_j(\xi_{i0})\right)\right)$$
$$= O_p\left(\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left(z_j\left(\psi\left(\widehat{h}(\widehat{k}_i)\right)\right)-z_j(\xi_{i0})\right)'(-v_{ij}^{\alpha})\left(z_j\left(\psi\left(\widehat{h}(\widehat{k}_i)\right)\right)-z_j(\xi_{i0})\right)\right)$$
$$= O_p\left(\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\widehat{h}(\widehat{k}_i)-\varphi(\xi_{i0})\right\|^2\right) = O_p(\delta), \tag{A26}$$

where we have used (A12), (A13), Lemmas 1 and 2, and that ψ is Lipschitz-continuous.

Combining results, using the Cauchy Schwarz inequality in (A24), we obtain:

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

The last term in A is:

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

Note that:

$$\widetilde{v}_j(k) = \left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}(-v_{ij}^{\alpha})\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}(-v_{ij}^{\alpha})(-v_{ij}^{\alpha})^{-1}v_{ij}\right). \tag{A27}$$

We thus have:

$$A_3 = \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left(z_j(\xi_{i0})'-\widetilde{z}_j\left(\widehat{k}_i\right)'\right)(-v_{ij}^{\alpha})(-v_{ij}^{\alpha})^{-1}v_{ij} = \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left(z_j(\xi_{i0})'-\widetilde{z}_j\left(\widehat{k}_i\right)'\right)v_{ij}$$
$$= \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left(z_j(\xi_{i0})'-z_j^{*}\left(\widehat{k}_i\right)'\right)v_{ij}+\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left(z_j^{*}\left(\widehat{k}_i\right)'-\widetilde{z}_j\left(\widehat{k}_i\right)'\right)v_{ij}, \tag{A28}$$

where $\widetilde{z}_j(k)$ is given by (A25), and:

$$z_j^{*}(k) = \left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}\mathbb{E}_{\xi_{i0},\lambda_0}\left(-v_{ij}^{\alpha}\right)\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}\mathbb{E}_{\xi_{i0},\lambda_0}\left(-v_{ij}^{\alpha}\right)z_j(\xi_{i0})\right). \tag{A29}$$

To see that the first term on the right-hand-side of (A28) is $O_p(\delta)$, we use an argument similar to the one we used to show (A15). Let: $\zeta_{ij} = v_{ij} - \mathbb{E}_{h_i,\xi_{i0},\lambda_0}(v_{ij})$. Following the same steps as the ones leading to (A17), we obtain:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\mathbb{E}_{h_i,\xi_{i0},\lambda_0}(v_{ij})\right\|^2 = O_p\left(\frac{1}{S}\right). \tag{A30}$$

Moreover, by a similar argument as (A26), since $\mathbb{E}_{\xi_{i0},\lambda_0}(-v_{ij}^{\alpha})$ is bounded away from zero with probability one, we have:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|z_j(\xi_{i0})-z_j^*(\widehat{k}_i)\right\|^2=O_p(\delta). \tag{A31}$$

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 $\zeta_i=(\zeta_{i1}',...,\zeta_{ip}')'$:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left(z_j(\xi_{i0})'-z_j^*\left(\widehat{k}_i\right)'\right)v_{ij}\right\|^2\right]\\
&\leq 2O\left(\frac{1}{S}\right)\mathbb{E}\left[\left\|\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left(z_j(\xi_{i0})'-z_j^*\left(\widehat{k}_i\right)'\right)\right\|^2\right]\\
&\quad+2\mathbb{E}\left[\left\|\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left(z_j(\xi_{i0})'-z_j^*\left(\widehat{k}_i\right)'\right)\zeta_{ij}\right\|^2\right]\\
&=O\left(\frac{\delta}{S}\right)+\frac{1}{N^2p^2}\sum_{i=1}^{N}\mathbb{E}\left[\left(z(\xi_{i0})'-z^*\left(\widehat{k}_i\right)'\right)\mathbb{E}_{h_i,\xi_{i0},\lambda_0}\left[\zeta_i\zeta_i'\right]\left(z(\xi_{i0})-z^*\left(\widehat{k}_i\right)\right)\right]\\
&=O\left(\frac{\delta}{S}\right)+O\left(\frac{\delta}{NR}\right)=O(\delta^2)+O\left(\frac{1}{S^2}\right)+O\left(\frac{1}{N^2R^2}\right)=O(\delta^2),
\end{aligned}$$

where we have used, in turn, the triangle and Cauchy Schwarz inequalities, (A30), (A31), conditional independence of the ζ_i and h_i across i, part (v) in Assumption 3, (A31) one more time, and the fact that $T=pR$. Note that $\|\mathbb{E}_{h_i,\xi_{i0},\lambda_0}\left[\zeta_i\zeta_i'\right]\|\leq \operatorname{Tr}\mathbb{E}_{h_i,\xi_{i0},\lambda_0}\left[\zeta_i\zeta_i'\right]\leq p\max_j \operatorname{Tr}\mathbb{E}_{h_i,\xi_{i0},\lambda_0}\left[\zeta_{ij}\zeta_{ij}'\right]=O_p(p/R)$ by part (v) in Assumption 3. Moreover, note that $1/(NR)=p/(NT)\leq\delta$.

Turning to the second term in (A28), we have:

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

where by (A15) we have: $\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\|\overline{v}_j(\widehat{k}_i)\|^2=O_p(\delta)$.

Moreover:

$$\begin{aligned}
&\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|z_j^*\left(\widehat{k}_i\right)-\widetilde{z}_j\left(\widehat{k}_i\right)\right\|^2\\
&\leq\frac{2}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|z_j(\xi_{i0})-z_j^*\left(\widehat{k}_i\right)\right\|^2+\frac{2}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|z_j(\xi_{i0})-\widetilde{z}_j\left(\widehat{k}_i\right)\right\|^2,
\end{aligned}$$

where the second term on the right-hand side is $O_p(\delta)$ due to (A26), and the first term is $O_p(\delta)$ due to (A31).

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:

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

First, similarly to (A23), we have:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\pi'_{ij}v^{\alpha}_{ij}\widetilde{w}_j(\widehat{k}_i) = O_p(\delta).$$

Next, we have:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\pi'_{ij}v^{\alpha}_{ij}\widetilde{v}_j(\widehat{k}_i) = \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\widetilde{\pi}_j(\widehat{k}_i)'v^{\alpha}_{ij}\widetilde{v}_j(\widehat{k}_i),$$

where $\widetilde{\pi}_j(k)$ is defined similarly to $\widetilde{\alpha}_j(k)$. To see that the right-hand side is $O_p(\delta)$, first note that, by (A27) and (A13):

$$\begin{aligned}\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\widetilde{v}_j(\widehat{k}_i)\right\|^2 &= \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\left(\frac{\sum_{i'=1}^{N}\mathbf{1}\{\widehat{k}_{i'}=\widehat{k}_i\}(-v^{\alpha}_{i'j})}{\sum_{i'=1}^{N}\mathbf{1}\{\widehat{k}_{i'}=\widehat{k}_i\}}\right)^{-1}\overline{v}_j(\widehat{k}_i)\right\|^2 \\ &= O_p\left(\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\overline{v}_j(\widehat{k}_i)\right\|^2\right).\end{aligned}$$

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

$$\begin{aligned}\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\widetilde{\pi}_j(\widehat{k}_i)\right\|^2 &= \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\left(\frac{\sum_{i'=1}^{N}\mathbf{1}\{\widehat{k}_{i'}=\widehat{k}_i\}(-v^{\alpha}_{i'j})}{\sum_{i'=1}^{N}\mathbf{1}\{\widehat{k}_{i'}=\widehat{k}_i\}}\right)^{-1}\overline{\tau}_j(\widehat{k}_i)\right\|^2 \\ &= O_p\left(\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\overline{\tau}_j(\widehat{k}_i)\right\|^2\right).\end{aligned}$$

Now, the τ_{ij} are independent across i, with conditional mean given ξ_{i0} and 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 (A15), and using part (v) in Assumption 3, it thus follows that $\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\|\widetilde{\pi}_j(\widehat{k}_i)\|^2 = O_p(\delta)$.

Hence, by the Cauchy Schwarz inequality:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\pi'_{ij}v^{\alpha}_{ij}\widetilde{v}_j(\widehat{k}_i) = O_p(\delta).$$

We lastly bound the third term in B:

$$\begin{aligned} B_3 &\equiv \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\pi'_{ij}v^{\alpha}_{ij}\left(\widetilde{\alpha}_j(\widehat{k}_i)-\alpha^j_{i0}\right) \\ &= \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\pi'_{ij}v^{\alpha}_{ij}\left(\alpha^*_j(\widehat{k}_i)-\alpha^j_{i0}\right)+\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\pi'_{ij}v^{\alpha}_{ij}\left(\widetilde{\alpha}_j(\widehat{k}_i)-\alpha^*_j(\widehat{k}_i)\right), \end{aligned}$$

where $\widetilde{\alpha}_j(k)$ and $\alpha^*_j(k)$ are given by similar expressions as (A25) and (A29), with α^j_{i0} in place of $z_j(\xi_{i0})$ in those formulas.

The first term is $O_p(\delta)$ since, similarly to (A31): $\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\|\alpha^*_j(\widehat{k}_i)-\alpha^j_{i0}\|^2 = O_p(\delta)$, the $\tau_{ij}=\pi'_{ij}v^{\alpha}_{ij}$ and h_i are conditionally independent across i with zero mean given ξ_{i0} and λ_0, and part (v) in Assumption 3 holds (using a similar argument as for the first term in (A28)). The second term is:

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

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

This shows that $B_3 = O_p(\delta)$, hence that $B = O_p(\delta)$. This implies that (A3) holds.

Proof of (A4). Let:

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

We have:

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

where we have used that $\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p} v_{ij} = \mathbb{E}\left[\frac{1}{p}\sum_{j=1}^{p} v_{ij}\right] + o_p(1) = o_p(1)$, and we have used the identity: $\frac{\partial \overline{\alpha}^j(\theta,\xi_{i0})}{\partial \theta'} = \left[\mathbb{E}_{\xi_{i0},\lambda_0}\left(-v_{ij}^{\alpha}\right)\right]^{-1}\mathbb{E}_{\xi_{i0},\lambda_0}\left(v_{ij}^{\theta}\right)'$. Hence, using (A13) and (A20):

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

We will now show that:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\frac{\partial \widehat{\alpha}^j(\widehat{k}_i,\theta_0)}{\partial \theta'} - \frac{\partial \overline{\alpha}^j(\theta_0,\xi_{i0})}{\partial \theta'}\right\|^2 = o_p(1). \tag{A32}$$

Differentiating with respect to θ the following identity, for all k, j:

$$\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i = k\}v_{ij}(\widehat{\alpha}^j(k,\theta),\theta) = 0,$$

and using (A13), we obtain:

$$\frac{\partial \widehat{\alpha}^j(k,\theta)}{\partial \theta'} = \left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i = k\}\left(-v_{ij}^{\alpha}\left(\widehat{\alpha}^j(\widehat{k}_i,\theta),\theta\right)\right)\right)^{-1}\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i = k\}\left(v_{ij}^{\theta}\left(\widehat{\alpha}^j(\widehat{k}_i,\theta),\theta\right)\right)'. \tag{A33}$$

Let us define, at $\theta = \theta_0$ (and omitting the reference to θ_0 and α_{i0}^j from the notation):

$$\frac{\partial \widetilde{\alpha}^j(k)}{\partial \theta'} = \left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i = k\}(-v_{ij}^{\alpha})\right)^{-1}\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i = k\}(v_{ij}^{\theta})',$$

and:

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

We have:

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

where a_{ij} lies between α_{i0}^j and $\widehat{\alpha}^j(\widehat{k}_i)$. By (A13), (A16), and either part (iia) or part (iib) in Assumption 3, we thus have:

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

Moreover:

$$\begin{aligned}\frac{\partial\widetilde{\alpha}^j(k,\theta_0)}{\partial\theta'}-\frac{\partial\widetilde{\alpha}^j_*(k,\theta_0)}{\partial\theta'} &= \left(\sum_{i=1}^N \mathbf{1}\{\widehat{k}_i=k\}(-v^\alpha_{ij})\right)^{-1}\sum_{i=1}^N \mathbf{1}\{\widehat{k}_i=k\}\tau'_{ij}\\ &= \left(\frac{\sum_{i=1}^N \mathbf{1}\{\widehat{k}_i=k\}(-v^\alpha_{ij})}{\sum_{i=1}^N \mathbf{1}\{\widehat{k}_i=k\}}\right)^{-1}\left(\frac{\sum_{i=1}^N \mathbf{1}\{\widehat{k}_i=k\}\tau'_{ij}}{\sum_{j=1}^N \mathbf{1}\{\widehat{k}_i=k\}}\right),\end{aligned}$$

where the $\tau'_{ij}=(v^\theta_{ij})'-(-v^\alpha_{ij})\left[\mathbb{E}_{\xi_{i0},\lambda_0}(-v^\alpha_{ij})\right]^{-1}\mathbb{E}_{\xi_{i0},\lambda_0}(v^\theta_{ij})'$ are conditionally independent across i, with zero mean given ξ_{i0} and λ_0. Hence, using (A13), and using part (v) in Assumption 3 we have (using a similar argument to the one we used to show (A15)):

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

Lastly, using (A13) we have, as in (A26):

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

Combining results shows (A32). Finally, using the Cauchy Schwarz inequality, and either part (iia) or part (iib) in Assumption 3, implies (A4).

Proof of (10). We have:

$$\begin{aligned}\frac{1}{Np}\sum_{i=1}^N\left\|\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\right\|^2 &\le \frac{2}{Np}\sum_{i=1}^N\left\|\widehat{\alpha}(\widehat{k}_i,\widehat{\theta})-\overline{\alpha}^j(\widehat{\theta},\xi_{i0})\right\|^2+\frac{2}{Np}\sum_{i=1}^N\left\|\overline{\alpha}^j(\widehat{\theta},\xi_{i0})-\overline{\alpha}^j(\theta_0,\xi_{i0})\right\|^2\\ &\le \frac{2}{Np}\sum_{i=1}^N\left\|\widehat{\alpha}(\widehat{k}_i,\theta_0)-\overline{\alpha}^j(\theta_0,\xi_{i0})\right\|^2+O_p\left(\|\widehat{\theta}-\theta_0\|^2\right)+O_p\left(\|\widehat{\theta}-\theta_0\|^2\right)\\ &= O_p\left(\frac{1}{S}\right)+O_p\left(\frac{Kp}{NT}\right)+O_p\left(K^{-\frac{2}{d}}\right)+O_p\left(\frac{1}{NT}\right)=O_p(\delta),\end{aligned}$$

where we have used (A16), and that, by (A33), the definition of $\overline{\alpha}^j(\theta,\xi)$, and either part (iia) or part (iib) in Assumption 3, both $\frac{1}{Np}\sum_{i=1}^N\sum_{j=1}^p\|\frac{\partial\widehat{\alpha}^j(\widehat{k}_i,\theta)}{\partial\theta'}\|^2$ and $\frac{1}{Np}\sum_{i=1}^N\sum_{j=1}^p\|\frac{\partial\overline{\alpha}^j(\theta,\xi_{i0})}{\partial\theta'}\|^2$ are $O_p(1)$ uniformly in θ.

This shows (10) and ends the proof of Theorem 1.

A.4 Proof of Corollary 2

By the triangle inequality we have:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N\left\|\widehat{h}(\widehat{k}_i)-\varphi(\xi_{i0})\right\|^2 &\le \frac{2}{N}\sum_{i=1}^N\left\|\widehat{h}(\widehat{k}_i)-h_i\right\|^2+\frac{2}{N}\sum_{i=1}^N\|h_i-\varphi(\xi_{i0})\|^2\\ &= 2\widehat{Q}(K)+O_p\left(\frac{1}{S}\right)=O_p\left(\frac{1}{S}\right),\end{aligned}$$

where we have used that $S\widehat{Q}(K) = O_p(1)$, and Assumption 2.

Following the steps of the proof of Theorem 1 then gives:

$$\widehat{\theta} \;=\; \theta_0 + H^{-1}\frac{1}{N}\sum_{i=1}^{N} s_i + O_p\left(\frac{1}{S}\right) + O_p\left(\frac{Kp}{NT}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right),$$

which proves Corollary 2 since $SKp/(NT) = O(1)$.

A.5 Proof of Corollary 3

The result in Corollary 3 is immediate.

A.6 Proof of Corollary 4

First note that, using (15) and the fact that N and T grow at the same rate:

$$\widehat{\theta}^{\mathrm{BR}} = \theta_0 + H^{-1}\frac{1}{N}\sum_{i=1}^{N}\left(2s_i - \frac{s_{i1}+s_{i2}}{2}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right), \tag{A34}$$

where s_{i1} and s_{i2} are computed using the first and last $T/2$ periods of the panel, respectively. Moreover, $\frac{1}{N}\sum_{i=1}^{N}\left(2s_i - \frac{s_{i1}+s_{i2}}{2}\right) = \frac{1}{\sqrt{NT}}Z + o_p\left(\frac{1}{\sqrt{NT}}\right)$, where $Z \sim \mathcal{N}(0, H)$. This implies $\sqrt{NT}(\widehat{\theta}^{\mathrm{BR}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, H^{-1})$.

A.7 Consistency of Hessian estimate

Using Theorem 1, and parts (i) and (ii) in Assumption 3, we have:

$$\begin{aligned}\widehat{H} =& \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\{\widehat{\mathbb{E}}_{\widehat{k}_i}\left(-\frac{\partial^2 \ell_{ij}\left(\alpha_{i0}^{j}, \theta_0\right)}{\partial\theta\partial\theta'}\right)\right.\\ &\left. - \widehat{\mathbb{E}}_{\widehat{k}_i}\left(\frac{\partial^2 \ell_{ij}(\alpha_{i0}^{j}, \theta_0)}{\partial\theta\partial\alpha'}\right)\left[\widehat{\mathbb{E}}_{\widehat{k}_i}\left(-\frac{\partial^2 \ell_{ij}(\alpha_{i0}^{j}, \theta_0)}{\partial\alpha\partial\alpha'}\right)\right]^{-1}\widehat{\mathbb{E}}_{\widehat{k}_i}\left(\frac{\partial^2 \ell_{ij}(\alpha_{i0}^{j}, \theta_0)}{\partial\alpha\partial\theta'}\right)\right\} + o_p(1).\end{aligned}$$

Let $z_{ij} = \frac{\partial^2 \ell_{ij}(\alpha_{i0}^{j}, \theta_0)}{\partial\theta\partial\alpha'}$, and let $\zeta_{ij} = z_{ij} - \mathbb{E}_{\xi_{i0},\lambda_0}(z_{ij})$. We have:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\overline{z}_j(\widehat{k}_i) - \mathbb{E}_{\xi_{i0},\lambda_0}(z_{ij})\right\|^2 \leq \frac{2}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\overline{\zeta}_j(\widehat{k}_i)\right\|^2 + o_p(1),$$

where we have used Lemma 1 and the triangle inequality, and here we require that $\frac{\partial}{\partial\xi'}\mathbb{E}_{\xi_{i0}=\xi,\lambda_0}(\mathrm{vec}\, z_{ij})$ be uniformly $O_p(1)$.

Now, using the same arguments as the ones we used to show (A15), we obtain that:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\overline{\zeta}_j(\widehat{k}_i)\right\|^2 = o_p(1).$$

Note that in this step we need a slightly modified condition compared to Assumption 3 part (v), applied to vec $\frac{\partial^2 \ell_{ij}\left(\alpha_{i0}^j,\theta_0\right)}{\partial\theta\partial\alpha'}$.

Applying the same arguments to all terms in $\widehat{H}$, it follows that:

$$\widehat{H} = H + o_p(1).$$

A.8 Average effects

Average effects are of interest in many economic settings. For example, effects of counterfactual policies can often be written as averages over the cross-sectional agent heterogeneity. Here we characterize the asymptotic behavior of GFE estimators of such quantities. Let $m_i(\alpha_i,\theta) = \frac{1}{T}\sum_{t=1}^{T} m\left(X_{it},\alpha_{it},\theta\right)$. A GFE estimator of the population average $M_0 = \frac{1}{N}\sum_{i=1}^{N} m_i\left(\alpha_{i0},\theta_0\right)$ is:

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

Note that M_0 depends on X_i and α_{i0}. An alternative target parameter would be $\widetilde{M}_0 = \mathbb{E}(m_i\left(\alpha_{i0},\theta_0\right))$. Given that $M_0 - \widetilde{M}_0 = O_p(1/\sqrt{N})$ under standard conditions, it is straightforward to adapt Corollary A1 below to this case.

It is convenient to denote, for all i, j:

$$m_{ij}(\alpha^j,\theta) = \frac{1}{R}\sum_{t=(j-1)R+1}^{jR} m_{it}\left(\alpha_{it},\theta\right).$$

We make the following assumption.

Assumption A1. *(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} *in both its arguments.*

$\max_j \sup_{\widetilde{\xi},\lambda} \|\frac{\partial}{\partial\xi'}\big|_{\xi=\widetilde{\xi}} \mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}(\frac{\partial m_{ij}(\alpha_{i0}^j,\theta_0)}{\partial\alpha})\| = O(1)$.

Let:

$$s_i^m = \frac{1}{p}\sum_{j=1}^{p}\left\{\mathbb{E}_{\xi_{i0},\lambda_0}\left(\frac{\partial m_{ij}\left(\alpha_{i0}^j,\theta_0\right)}{\partial\alpha'}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_0}\left(-\frac{\partial^2\ell_{ij}\left(\alpha_{i0}^j,\theta_0\right)}{\partial\alpha\partial\alpha'}\right)\right]^{-1}\frac{\partial\ell_{ij}(\alpha_{i0}^j,\theta_0)}{\partial\alpha}\right.$$
$$+\mathbb{E}_{\xi_{i0},\lambda_0}\left(\frac{\partial m_{ij}\left(\alpha_{i0}^j,\theta_0\right)}{\partial\theta'}\right)H^{-1}\frac{1}{N}\sum_{i'=1}^{N}s_{i'}$$
$$\left.+\mathbb{E}_{\xi_{i0},\lambda_0}\left(\frac{\partial m_{ij}\left(\alpha_{i0}^j,\theta_0\right)}{\partial\alpha'}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_0}\left(-\frac{\partial^2\ell_{ij}\left(\alpha_{i0}^j,\theta_0\right)}{\partial\alpha\partial\alpha'}\right)\right]^{-1}\mathbb{E}_{\xi_{i0},\lambda_0}\left(\frac{\partial^2\ell_{ij}\left(\alpha_{i0}^j,\theta_0\right)}{\partial\alpha\partial\theta'}\right)H^{-1}\frac{1}{N}\sum_{i'=1}^{N}s_{i'}\right\},$$

where s_i and H appear in Theorem 1.

We have the following corollary to Theorem 1. Analogous results to Corollaries 2, 3, and 4 hold for average effects, although we omit them for brevity.

Corollary A1. *Let the conditions of Theorem 1 hold. In addition, let Assumption A1 hold. Then, as N, S, K tend to infinity such that $Kp/(NT)$ tends to zero, we have:*

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

Proof of Corollary A1. We have, by a Taylor expansion, and using Theorem 1 and parts (i) and (ii) in Assumption A1:

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

where δ is defined as in the proof of Theorem 1.

Using similar arguments to the ones we used to establish (A21) in the proof of Theorem 1, under Assumption A1 we have:

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

Moreover, using (A32) and Assumption A1 we obtain:

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

Combining results, we obtain:

$$\begin{aligned}
\widehat{M} - M_0 =& \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\frac{\partial m_{ij}\left(\alpha_{i0}^{j},\theta_0\right)}{\partial\alpha'}\left(\widehat{\alpha}^{j}(\widehat{k}_i,\widehat{\theta})-\widehat{\alpha}^{j}(\widehat{k}_i,\theta_0)\right)\\
&+\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\frac{\partial m_{ij}\left(\alpha_{i0}^{j},\theta_0\right)}{\partial\alpha'}\left(\widehat{\alpha}^{j}(\widehat{k}_i,\theta_0)-\alpha_{i0}^{j}\right)\\
&+\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\frac{\partial m_{ij}\left(\alpha_{i0}^{j},\theta_0\right)}{\partial\theta'}\left(\widehat{\theta}-\theta_0\right)+O_p\left(\delta\right)\\
=&\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\frac{\partial m_{ij}\left(\alpha_{i0}^{j},\theta_0\right)}{\partial\alpha'}\left(\overline{\alpha}^{j}(\widehat{\theta},\xi_{i0})-\overline{\alpha}^{j}(\theta_0,\xi_{i0})\right)\\
&+\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\mathbb{E}_{\xi_{i0},\lambda_0}\left[\frac{\partial m_{ij}\left(\alpha_{i0}^{j},\theta_0\right)}{\partial\alpha'}\right]\mathbb{E}_{\xi_{i0},\lambda_0}\left[-v_{ij}^{\alpha}(\alpha_{i0}^{j},\theta_0)\right]^{-1}v_{ij}(\alpha_{i0}^{j},\theta_0)\\
&+\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\frac{\partial m_{ij}\left(\alpha_{i0}^{j},\theta_0\right)}{\partial\theta'}\left(\widehat{\theta}-\theta_0\right)+O_p\left(\delta\right)+o_p\left(\frac{1}{\sqrt{NT}}\right).
\end{aligned}$$

The result comes from expanding $\overline{\alpha}^{j}(\widehat{\theta},\xi_{i0})$ around θ_0 (using that $\mathbb{E}_{\xi_{i0},\lambda_0}[v_{ij}(\overline{\alpha}^{j}(\theta,\xi_{i0}),\theta)]=0$ for all θ), and then substituting $\widehat{\theta}-\theta_0$ by its influence function.

A.9 Expansions in Example 2

Consider Example 2, with stationary observations. In the case where heterogeneity is time-invariant, we have:

$$\ell_{ij}(\alpha_i^j,\theta)\equiv\ell_i(\alpha_i,\theta)=\frac{1}{T}\sum_{t=1}^{T}Y_{it}\ln\left[\Phi\left(X_{it}'\theta+\alpha_i\right)\right]+(1-Y_{it})\ln\left[1-\Phi\left(X_{it}'\theta+\alpha_i\right)\right],$$

where Φ is the standard normal cumulative distribution function. Hence (11) holds, with:

$$s_i=\frac{1}{T}\sum_{t=1}^{T}W_{it}^{(1)}\left(X_{it}-\frac{\mathbb{E}_{\xi_{i0}}(W_{it}^{(2)}X_{it})}{\mathbb{E}_{\xi_{i0}}(W_{it}^{(2)})}\right),$$

and

$$H=\mathbb{E}\left[\mathbb{E}_{\xi_{i0}}(W_{it}^{(2)}X_{it}X_{it}')-\frac{\mathbb{E}_{\xi_{i0}}(W_{it}^{(2)}X_{it})\mathbb{E}_{\xi_{i0}}(W_{it}^{(2)}X_{it}')}{\mathbb{E}_{\xi_{i0}}(W_{it}^{(2)})}\right].$$

Here we have defined $W_{it}^{(k)}=W_k(Y_{it},X_{it},\alpha_{i0})$ with:

$$W_1(Y,X,\alpha)=\frac{\phi(X'\theta_0+\alpha)(Y-\Phi(X'\theta_0+\alpha))}{\Phi(X'\theta_0+\alpha)(1-\Phi(X'\theta_0+\alpha))},\quad W_2(Y,X,\alpha)=\frac{\phi(X'\theta_0+\alpha)^2}{\Phi(X'\theta_0+\alpha)(1-\Phi(X'\theta_0+\alpha))},$$

where ϕ denotes the standard normal density.

In the case where heterogeneity is time-varying:

$$\ell_{ij}(\alpha_i^j, \theta) \equiv \ell_{it}(\alpha_{it}, \theta) = Y_{it} \ln \left[\Phi\left(X_{it}'\theta + \alpha_{it}\right)\right] + (1 - Y_{it}) \ln \left[1 - \Phi\left(X_{it}'\theta + \alpha_{it}\right)\right].$$

Hence (12) holds, where $\widetilde{s}_i$ and $\widetilde{H}$ are identical to s_i and H above, except that $W_{it}^{(k)}$ is replaced by $\widetilde{W}_{it}^{(k)} = W_k(Y_{it}, X_{it}, \alpha_{it0})$, and the expectations are now conditional on both ξ_{i0} and λ_{t0}.