Supplementary Appendix for "Discretizing Unobserved Heterogeneity"[1]

In this supplementary appendix we provide details on various aspects of the theory, and we report additional results.

B Conditional first step

In this section of the appendix we study some theoretical properties of GFE based on a conditional first-step, under a series regression specification. We first study the case where α_{i0} does not vary over time, and then turn to the time-varying case.

B.1 Convergence rate for the first step

We make the following assumption, where $g(x,\alpha) = \mathbb{E}_{X_{it}=x\,,\,\alpha_{i0}=\alpha}[h(Y_{it},X_{it})]$, and f_μ denotes the conditional density of X_{it} given $\mu_{i0}=\mu$.

Assumption B1.

(i) *There exists an integrable function ω_X such that $f_\mu(x) \leq \omega_X(x)$ for all x,μ.*

The maximum eigenvalue of $\int P_q(x)P_q(x)'\omega_X(x)dx$ is $O(1)$, and $\sup_\alpha \int \|g(x,\alpha)\|^2\omega_X(x)dx$ is bounded.

There exists a constant $\underline{c}>0$, independent of q and T, such that the minimum eigenvalue of $\frac{1}{T}\sum_{t=1}^T P_q(X_{it})P_q(X_{it})'$ is larger than $\underline{c}+o_p(1)$, uniformly in i.

(ii) *Let $\mathcal{A}$ denote the parameter space for α_i. For all $\alpha\in\mathcal{A}$, $g(\cdot,\alpha)\in\mathcal{G}\subset L^2(\omega_X)$.*

$\|g(x,\alpha')-g(x,\alpha)\| \leq C(x)\|\alpha'-\alpha\|$, where $\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T C(X_{it})^2 = O_p(1)$.

There exist a constant $a>0$ and a sequence of functions $\beta_q(\cdot)$ such that $\sup_{x,\alpha}\|g(x,\alpha) - P_q(x)'\beta_q(\alpha)\| = O(q^{-a})$.

(iii) *Let $\varepsilon_{it} = h(Y_{it},X_{it}) - g(X_{it},\alpha_{i0})$. Then $\frac{1}{N}\sum_{i=1}^N \left\|\frac{1}{T}\sum_{t=1}^T P_q(X_{it})\varepsilon_{it}\right\|^2 = O_p(q/T)$.*

(iv) *There exists a Lipschitz-continuous mapping $\psi:\mathcal{G}\to\mathcal{A}$ such that $\psi(g(\cdot,\alpha_{i0}))=\alpha_{i0}$.*

Parts (i) and (ii) in Assumption B1 will be satisfied when using a suitable family P_q. The condition on the maximum eigenvalue of $\int P_q(x)P_q(x)'\omega_X(x)dx$ in part (i) is without loss of generality. In part (ii), $a=s/r$ for splines and power series, where s is the number of continuous derivatives in x of

[1]We thank the IFAU for access to, and help with, the Swedish administrative data.

$g(x,\alpha)$, and r is the dimension of X_{it}. Part (iii) holds under standard conditions on moments and serial dependence. Part (iv) is an injectivity condition.

To illustrate these conditions, suppose X_{it} is scalar, and consider a family of polynomials that are orthonormal with respect to the uniform distribution on $[-1,1]$. If ω_X is bounded by a constant $B>0$, then $\int P_q(x)P_q(x)'\omega_X(x)dx$ is bounded by BI_q, where I_q is the $q\times q$ identity matrix, so its maximum eigenvalue is bounded. If in addition we assume that $f_\mu(x)\geq b$ for all x,μ, where $b>0$ is a constant, then $\mathbb{E}_{\mu_{i0}=\mu}\left(P_q(X_{it})P_q(X_{it})'\right)\geq bI_q$ for all μ and q. Newey (1997, Theorem 1) provides conditions under which, for given i, the minimum eigenvalue of $\frac{1}{T}\sum_{t=1}^T P_q(X_{it})P_q(X_{it})'$ is larger than $\underline{c}+o_p(1)$, for some constant $\underline{c}>0$. Specifically, Newey's conditions require that X_{it} be i.i.d. over time (here, conditional on μ_{i0}), and $\sup_x\|P_q(x)\|_F\leq\zeta_q$ with $q\zeta_q^2/T=o(1)$, where $\|A\|_F=(\operatorname{Tr}A'A)^{1/2}$ denotes the Fröbenius norm of A. From this he obtains that $\|\frac{1}{T}\sum_{t=1}^T Z_{it}\|_F=o_p(1)$, where $Z_{it}=P_q(X_{it})P_q(X_{it})'-\mathbb{E}_{\mu_{i0}}\left(P_q(X_{it})P_q(X_{it})'\right)$.

In the present context, to guarantee that the minimum eigenvalue of $\frac{1}{T}\sum_{t=1}^T P_q(X_{it})P_q(X_{it})'$ be larger than $\underline{c}+o_p(1)$ uniformly in i, we need the stronger uniform condition $\max_i\|\frac{1}{T}\sum_{t=1}^T Z_{it}\|_F=o_p(1)$. Results from random matrix theory are helpful to provide primitive conditions for this. For example, using the matrix Bernstein theorem (Theorem 1.4. in Tropp, 2012), under the assumption that X_{it} are independent over time conditional on μ_{i0}, we have for all $\epsilon>0$, μ, q, and T:

$$\Pr\left(\left\|\frac{1}{T}\sum_{t=1}^T Z_{it}\right\|\geq\epsilon\,\middle|\,\mu_{i0}=\mu\right)\leq q\exp\left(-T\frac{\epsilon^2}{2(\zeta_q^4+\frac{\zeta_q^2\epsilon}{3})}\right),\tag{B1}$$

where recall that $\|A\|$ denotes the spectral norm of A. A related inequality holds under a form of serial dependence (β-mixing), see Banna, Merlevède and Youssef (2016). Using the union bound and the fact that $\|A\|_F\leq\sqrt{q}\,\|A\|$, (B1) implies that $\max_i\|\frac{1}{T}\sum_{t=1}^T Z_{it}\|_F=o_p(1)$, provided that $q\zeta_q^4\ln(Nq)/T\to 0$.

Part (iv) is an injectivity condition that generalizes the one in the unconditional case. As an example, consider the probit model of Example 2: $Y_{it}=\mathbf{1}\{X_{it}'\theta_0+\alpha_{i0}+U_{it}\geq 0\}$, where U_{it} are i.i.d. standard normal, independent of all X_{it}'s. Taking $h(Y_{it},X_{it})=Y_{it}$, we have $g(X_{it},\alpha_{i0})=\Phi(X_{it}'\theta_0+\alpha_{i0})$. Provided that α_{i0} and X_{it} are such that g is bounded between $\epsilon>0$ and $1-\epsilon$, and denoting $c_0=\int\omega_X(x)dx>0$ and $c_1=c_0^{-1}\int x\omega_X(x)dx$, the following function satisfies part (iv):

$$\psi(g(\cdot,\alpha))=c_0^{-1}\int\Phi^{-1}(g(x,\alpha))\omega_X(x)dx-c_1'\theta_0.$$

We now state our main result on the convergence rate of the conditional kmeans estimator, where:

$$B_\alpha(K)=\min_{(\widetilde{\alpha},k_1,\ldots,k_N)}\frac{1}{N}\sum_{i=1}^N\|\alpha_{i0}-\widetilde{\alpha}(k_i)\|^2$$

denotes the approximation error associated with α_{i0} alone.

Theorem B1. *Suppose that α_{i0} is time-invariant, of fixed dimension. Let parts (i) to (iii) in Assumption B1 hold. Then, as N, T, K, q tend to infinity:*

$$\frac{1}{N}\sum_{i=1}^{N}\int \left\| P_q(x)'\widehat{\beta}_q(\widehat{k}_i) - g(x,\alpha_{i0})\right\|^2 \omega_X(x)dx = O_p(q/T) + O_p(q^{-2a}) + O_p(B_\alpha(K)).$$

If in addition part (iv) in Assumption B1 holds, then:

$$\frac{1}{N}\sum_{i=1}^{N}\left\| \psi\left(P_q(\cdot)'\widehat{\beta}_q(\widehat{k}_i)\right) - \alpha_{i0}\right\|^2 = O_p(q/T) + O_p(q^{-2a}) + O_p(B_\alpha(K)).$$

Note that the second part in Theorem B1 implies (19) in the main text. As a special case, Theorem B1 implies that, when $g(x,\alpha) = P_q(x)'\beta_q(\alpha)$ (i.e., there is no remainder term), then the rate is $O_p(1/T) + O_p(B_\alpha(K))$. As an example, this rate is achieved when X_{it} has finite support and we use indicator functions as P_q. More generally, when g is nonparametric and covariates are continuous, a larger number of covariates increases q^{-2a}, thus requiring choosing a larger q and incurring a larger q/T. Nevertheless, even in this nonparametric case, the resulting rate may improve relative to the one of our baseline two-step estimation.

To give a concrete example, assume that there are two covariates and g has four continuous derivatives, so $a = 2$. Suppose also that α_{i0} is scalar and μ_{i0} has dimension two, and that (α_{i0}, μ_{i0}') has an absolutely continuous density, so the underlying dimension of heterogeneity is $d = 3$. Taking $q \propto T^{1/5}$ we obtain a rate of $O_p(T^{-4/5}) + O_p(K^{-2})$ in the conditional case. In contrast, the corresponding rate in the unconditional case is $O_p(T^{-1}) + O_p(K^{-2/3})$. When N and T grow at the same rate, the best rates achievable are $O_p(T^{-4/5})$ and $O_p(T^{-2/3})$, respectively (since $K \leq N$). In addition, the required numbers of groups to achieve these rates are $K \propto T^{2/5}$ and T, respectively. This shows that the conditional method achieves a better rate, and that the number of groups needed to achieve that rate is smaller. As in other applications of series methods, smoothness assumptions are critical to ensure good performance. Under sufficient smoothness, the theoretical difference between the two methods increases as the dimension of X_{it} increases.

Proof of Theorem B1. Let:

$$(\underline{\beta}_q, \{\underline{k}_i\}) = \underset{(b,\{k_i\})}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} \left\| g(X_{it},\alpha_{i0}) - P_q(X_{it})'b(k_i)\right\|^2.$$

Let also:

$$(\underline{\alpha}, \{\underline{k}_i^*\}) = \underset{(\widetilde{\alpha},\{k_i\})}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| \alpha_{i0} - \widetilde{\alpha}(k_i)\right\|^2.$$

By Assumption B1 (ii), we have, for a and $\beta_q(\cdot)$ defined in that assumption:

$$\begin{aligned}
&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|g(X_{it},\alpha_{i0})-P_q(X_{it})'\underline{\beta}_q(\underline{k}_i)\right\|^2 \\
&\leq \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|g(X_{it},\alpha_{i0})-P_q(X_{it})'\beta_q(\underline{\alpha}(\underline{k}_i^*))\right\|^2 \\
&\leq \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|g(X_{it},\alpha_{i0})-g(X_{it},\underline{\alpha}(\underline{k}_i^*))\right\|^2+O_p(q^{-2a}) \\
&= O_p(B_\alpha(K))+O_p(q^{-2a}). \qquad \text{(B2)}
\end{aligned}$$

Moreover, since:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|h(Y_{it},X_{it})-P_q(X_{it})'\widehat{\beta}_q(\widehat{k}_i)\right\|^2 \leq \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|h(Y_{it},X_{it})-P_q(X_{it})'\underline{\beta}_q(\underline{k}_i)\right\|^2,$$

we have, using that $h(Y_{it},X_{it})=g(X_{it},\alpha_{i0})+\varepsilon_{it}$:

$$\begin{aligned}
&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|g(X_{it},\alpha_{i0})-P_q(X_{it})'\widehat{\beta}_q(\widehat{k}_i)\right\|^2 \\
&\leq \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|g(X_{it},\alpha_{i0})-P_q(X_{it})'\underline{\beta}_q(\underline{k}_i)\right\|^2 \\
&\quad+\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\varepsilon_{it}'\left(P_q(X_{it})'\widehat{\beta}_q(\widehat{k}_i)-P_q(X_{it})'\underline{\beta}_q(\underline{k}_i)\right). \qquad \text{(B3)}
\end{aligned}$$

Hence, using parts (ii) and (iii) in Assumption B1, equation (B2), and the Cauchy Schwarz and triangle inequalities, we obtain:

$$\begin{aligned}
&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|P_q(X_{it})'\beta_q(\alpha_{i0})-P_q(X_{it})'\widehat{\beta}_q(\widehat{k}_i)\right\|^2 \\
&\leq O_p(1)\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{T}\sum_{t=1}^{T}P_q(X_{it})'\varepsilon_{it}\right\|^2\right)^{\frac{1}{2}}\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\beta}_q(\widehat{k}_i)-\underline{\beta}_q(\underline{k}_i)\right\|^2\right)^{\frac{1}{2}} \\
&\quad+O_p(q^{-2a})+O_p(B_\alpha(K)) \\
&\leq O_p\left(\sqrt{\frac{q}{T}}\right)\left(\frac{2}{N}\sum_{i=1}^{N}\left\|\widehat{\beta}_q(\widehat{k}_i)-\beta_q(\alpha_{i0})\right\|^2+\frac{2}{N}\sum_{i=1}^{N}\left\|\underline{\beta}_q(\underline{k}_i)-\beta_q(\alpha_{i0})\right\|^2\right)^{\frac{1}{2}} \\
&\quad+O_p(q^{-2a})+O_p(B_\alpha(K)).
\end{aligned}$$

It thus follows from part (i) in Assumption B1 that:

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\beta}_q(\widehat{k}_i)-\beta_q(\alpha_{i0})\right\|^2 \\
&\leq O_p\left(\sqrt{\frac{q}{T}}\right)\left(\frac{2}{N}\sum_{i=1}^{N}\left\|\widehat{\beta}_q(\widehat{k}_i)-\beta_q(\alpha_{i0})\right\|^2+\frac{2}{N}\sum_{i=1}^{N}\left\|\underline{\beta}_q(\underline{k}_i)-\beta_q(\alpha_{i0})\right\|^2\right)^{\frac{1}{2}} \\
&\quad+O_p(q^{-2a})+O_p(B_\alpha(K)),
\end{aligned}$$

hence that:

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\beta}_q(\widehat{k}_i)-\beta_q(\alpha_{i0})\right\|^2 \\
&\leq O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\underline{\beta}_q(\underline{k}_i)-\beta_q(\alpha_{i0})\right\|^2\right)+O_p\left(\frac{q}{T}\right)+O_p(q^{-2a})+O_p(B_\alpha(K)) \\
&=O_p\left(\frac{q}{T}\right)+O_p(q^{-2a})+O_p(B_\alpha(K)),
\end{aligned}$$

where the last identity comes from the fact that, by (B2) and parts (i) and (ii) in Assumption B1:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\underline{\beta}_q(\underline{k}_i)-\beta_q(\alpha_{i0})\right\|^2=O_p(q^{-2a})+O_p(B_\alpha(K)).$$

Hence, using parts (i) and (ii) in Assumption B1 we have:

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\int\left\|P_q(x)'\widehat{\beta}_q(\widehat{k}_i)-g(x,\alpha_{i0})\right\|^2\omega_X(x)dx \\
&\leq\frac{1}{N}\sum_{i=1}^{N}\int\left\|P_q(x)'\widehat{\beta}_q(\widehat{k}_i)-P_q(x)'\beta_q(\alpha_{i0})\right\|^2\omega_X(x)dx+O_p(q^{-2a}) \\
&\leq O_p(1)\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\beta}_q(\widehat{k}_i)-\beta_q(\alpha_{i0})\right\|^2+O_p(q^{-2a}) \\
&=O_p\left(\frac{q}{T}\right)+O_p(q^{-2a})+O_p(B_\alpha(K)).
\end{aligned}$$

Lastly, we prove the second part in Theorem B1. We have, using the first part and part (iv) in Assumption B1:

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\left\|\psi\left(P_q(\cdot)'\widehat{\beta}_q(\widehat{k}_i)\right)-\alpha_{i0}\right\|^2 \\
&=\frac{1}{N}\sum_{i=1}^{N}\left\|\psi\left(P_q(\cdot)'\widehat{\beta}_q(\widehat{k}_i)\right)-\psi(g(\cdot,\alpha_{i0}))\right\|^2 \\
&\leq O_p(1)\frac{1}{N}\sum_{i=1}^{N}\int\left\|P_q(x)'\widehat{\beta}_q(\widehat{k}_i)-g(x,\alpha_{i0})\right\|^2\omega_X(x)dx \\
&=O_p\left(\frac{q}{T}\right)+O_p(q^{-2a})+O_p(B_\alpha(K)).
\end{aligned}$$

B.2 Choice of K

In the conditional case, we define:

$$\widehat{Q}_X(K) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|P_q(X_{it})'\widehat{\beta}(\widehat{k}_i) - P_q(X_{it})'\widehat{\beta}_{qi}\right\|^2,$$

where $\widehat{\beta}_{qi} = \text{argmin}_b \sum_{t=1}^{T}\|h(Y_{it}, X_{it}) - P_q(X_{it})'b\|^2$. We let:

$$\widehat{K} = \min_{K\geq 1}\left\{K \,:\, \widehat{Q}_X(K) \leq \gamma \widehat{V}_X\right\}, \tag{B4}$$

where $\widehat{V}_X = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|P_q(X_{it})'\widehat{\beta}_{qi} - P_q(X_{it})'\beta_q(\alpha_{i0})\right\|^2 + o_p(q/T)$.

For example, when ε_{it} are independent over time and homoskedastic, one can take:

$$\widehat{V}_X = \frac{q}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\|h(Y_{it}, X_{it}) - P_q(X_{it})'\widehat{\beta}_{qi}\|^2.$$

Corollary B1. *Let Assumption B1 hold. Suppose that* $(1+\gamma)\widehat{V}_X = O_p(q/T)$. *Take* $K \geq \widehat{K}$. *Then:*

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\psi\left(P_q(\cdot)'\widehat{\beta}_q(\widehat{k}_i)\right) - \alpha_{i0}\right\|^2 = O_p(q/T) + O_p(q^{-2a}).$$

Note that, to implement the formula for $\widehat{K}$, we need to compute $\widehat{\beta}_{qi}$. An alternative approach, which we use in the simulations, is to replace $\widehat{\beta}_{qi}$ by a conditional kmeans estimate $\widehat{\beta}(\widehat{k}_i)$ obtained using a large number of groups $K_{\max}$. In all the models that we consider in the simulation section (Section G), we use this approach with $K_{\max} = 30$.

Proof of Corollary B1. Using (B4) and the triangle inequality, we have:

$$\begin{aligned}
&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|P_q(X_{it})'\beta_q(\alpha_{i0}) - P_q(X_{it})'\widehat{\beta}_q(\widehat{k}_i)\right\|^2 \\
&\leq \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|P_q(X_{it})'\beta_q(\alpha_{i0}) - P_q(X_{it})'\widehat{\beta}_{qi}\right\|^2 \\
&\quad + \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|P_q(X_{it})'\widehat{\beta}_{qi} - P_q(X_{it})'\widehat{\beta}_q(\widehat{k}_i)\right\|^2 \\
&= O_p\left(\widehat{V}_X\right) + o_p(q/T) + O_p\left(\gamma\widehat{V}_X\right) = O_p\left(\frac{q}{T}\right).
\end{aligned}$$

The result then follows from the same arguments as in the proof of Theorem B1.

B.3 Time-varying heterogeneity

The conditional kmeans first step can be adapted to models where α_{it0} varies over time. Let us consider two cases. A first situation is when $\alpha_{it0} = \alpha(\xi_{i0}, \lambda_t^{(o)})$, where $\lambda_t^{(o)}$ is a vector of observed time-varying factors. In our theory we require time-stationarity, so $\lambda_t^{(o)}$ needs to be a stationary factor. In this case the baseline conditional kmeans method is unchanged, subject to including $\lambda_t^{(o)}$ in the set of covariates X_{it}. Moreover, Theorem B1 holds exactly as in the time-invariant case.

A more general situation is when $\alpha_{it0} = \alpha(\xi_{i0}, \lambda_{t0})$, where λ_{t0} is a vector of unobserved factors. Here we focus on the case where the factors may depend unrestrictedly on time (i.e., we take $p = T$), although it would be easy to modify the setup to allow for heterogeneity varying only between subperiods but not within, as in Theorem 1. We modify the first step algorithm in the following way.

Algorithm 3. *(conditional kmeans, time-varying heterogeneity)*

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

Adapting the theory requires additional assumptions and different arguments. We start with a definition.

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

We make the following assumption, where $g(x, \alpha) = \mathbb{E}_{X_{it}=x\,, \alpha_{it0}=\alpha}[h(Y_{it}, X_{it})]$.

Assumption B2.

(i) Let $\varepsilon_{it} = h(Y_{it}, X_{it}) - g(X_{it}, \alpha_{it0})$*. The* $NTq \times 1$ *random vector* $(P_q(X_{it})\varepsilon_{it})_{i=1,...,N, t=1,...,T}$ *satisfies Definition 1.*

(ii) There exists a Lipschitz-continuous mapping $\psi : \mathcal{G} \to \mathcal{A}$ *such that* $\psi(g(\cdot, \alpha_{it0})) = \alpha_{it0}$*.*

Part (i) in Assumption B2 requires $P_q(X_{it})\varepsilon_{it}$ to be sub-Gaussian (e.g., Vershynin, 2010). This is stronger than the corresponding requirements in Theorem 1. For example, i.i.d. Gaussian random variables and i.i.d. bounded random variables are sub-Gaussian. More generally, this assumption

allows for dependence across observations. As an example, a random vector $W \sim \mathcal{N}(0, \Sigma)$ is sub-Gaussian when the maximal eigenvalue of Σ is bounded from above by 2λ. This allows for weak dependence, across both individual units and time periods.

We have the following result.

Corollary B2. *Let Assumption 1 hold, for time-varying α_{it0} and α Lipschitz-continuous in its first argument. Let parts (i) and (ii) in Assumption B1 hold, and let part (i) in Assumption B2 hold. Then, as N, T, K, q tend to infinity:*

$$\begin{aligned}&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\int \left\|P_q(x)'\widehat{\beta}_q(\widehat{k}_i, t) - g(x, \alpha_{it0})\right\|^2 \omega_X(x)dx \\ &= O_p((\ln K)/T) + O_p(q^{-2a}) + O_p(qK/N) + O_p(B_\alpha(K)).\end{aligned}$$

If in addition part (ii) in Assumption B2 holds, then:

$$\begin{aligned}&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\psi\left(P_q(\cdot)'\widehat{\beta}_q(\widehat{k}_i, t)\right) - \alpha_{it0}\right\|^2 \\ &= O_p((\ln K)/T) + O_p(q^{-2a}) + O_p(qK/N) + O_p(B_\alpha(K)).\end{aligned}$$

From Corollary B2 we obtain a convergence rate for conditional kmeans that depends on the underlying dimension d_α of α_{it0}, instead of the dimension d that appears in Theorem 1. Also, note that, in the time-invariant case, under the assumptions of Corollary B2 one can show that the conditional kmeans estimator satisfies the following convergence rate: $O_p(\min(q, \ln K)/T) + O_p(qK/(NT)) + O_p(q^{-2a}) + O_p(B_\alpha(K))$, which is as least as fast as the rate in Theorem B1.

Proof of Corollary B2. Let:

$$(\underline{\beta}_q, \{\underline{k}_i\}) = \underset{(b,\{k_i\})}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left\|g(X_{it}, \alpha_{it0}) - P_q(X_{it})'b(k_i, t)\right\|^2,$$

and let:

$$(\underline{\alpha}, \{\underline{k}_i^*\}) = \underset{(\widetilde{\alpha},\{k_i\})}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\alpha_{it0} - \widetilde{\alpha}(k_i, t)\right\|^2.$$

We have, as in the proof of Theorem B1:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|g(X_{it}, \alpha_{it0}) - P_q(X_{it})'\underline{\beta}_q(\underline{k}_i, t)\right\|^2 = O_p(B_\alpha(K)) + O_p(q^{-2a}). \tag{B5}$$

Moreover, we have:

$$\begin{aligned}
&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|g(X_{it},\alpha_{it0})-P_q(X_{it})'\widehat{\beta}_q(\widehat{k}_i,t)\right\|^2 \\
&\leq \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|g(X_{it},\alpha_{it0})-P_q(X_{it})'\underline{\beta}_q(\underline{k}_i,t)\right\|^2 \\
&\quad+\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(P_q(X_{it})\varepsilon_{it})'\left(\widehat{\beta}_q(\widehat{k}_i,t)-\underline{\beta}_q(\underline{k}_i,t)\right). \qquad \text{(B6)}
\end{aligned}$$

The main difference with the proof of Theorem B1 is how to bound the cross-product term in (B6). To do so, we apply a version the Hanson-Wright tail inequality for quadratic forms, due to Hsu, Kakade and Zhang (2012, Theorem 2.1), which allows for dependent data.

Lemma B1. *(Hsu, Kadade and Zhang, 2012) Let Z be a random vector such that, for some $\lambda > 0$, $\mathbb{E}\left[\exp(\tau' Z)\right] \leq \exp(\lambda \left\|\tau\right\|^2)$ for all $\tau \in \mathbb{R}^{\dim Z}$. Let Q be a positive semi-definite matrix. Then, for all $s > 0$:*

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

Let $V_{it} = P_q(X_{it})\varepsilon_{it}$. We will bound:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(P_q(X_{it})\varepsilon_{it})'\left(\widehat{\beta}_q(\widehat{k}_i,t)-\underline{\beta}_q(\underline{k}_i,t)\right)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\overline{V}_t(\widehat{k}_i,\underline{k}_i)'\left(\widehat{\beta}_q(\widehat{k}_i,t)-\underline{\beta}_q(\underline{k}_i,t)\right),$$

where $\overline{V}_t(k,k')$ denotes the t-specific linear projection of V_{it} on group indicators $\mathbf{1}\{\widehat{k}_i = k\}$ and $\mathbf{1}\{\underline{k}_i = k'\}$. For this, we will use the Cauchy Schwarz inequality and bound:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\overline{V}_t(\widehat{k}_i,\underline{k}_i)\right\|^2.$$

Let, for given partitions $\{k_{i1}\},\{k_{i2}\}$: $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\overline{V}_t(k_{1i},k_{2i})\right\|^2 = \frac{v'Qv}{NT}$, where we have defined $v = (V_{11}',...,V_{NT}')'$, Q is an $NTq \times NTq$ projection matrix with $\operatorname{tr} Q \leq 2KTq$, $Q^2 = Q$, and $\left\|Q\right\| = 1$. By Lemma B1 and part (i) in Assumption B2 we have, for all s:

$$\Pr\left[v'Qv > 4\lambda KTq + 4\lambda\sqrt{2KTqs} + 4\lambda s\right] \leq \exp(-s),$$

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

$$\Pr\left[v'Qv > 8\lambda KTq + 6\lambda s\right] \leq \exp(-s),$$

hence, for all $b > 0$:

$$\Pr\left[\frac{v'Qv}{NT} > b\right] \leq \exp\left[-\left(\frac{bNT}{6\lambda} - \frac{4KTq}{3}\right)\right].$$

Lastly, by the union bound, given that the number of partitions $\{k_{i1}\} \cap \{k_{i2}\}$ is bounded by K^{2N}:

$$\begin{aligned}\Pr\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\overline{V}_t(\widehat{k}_i,\underline{k}_i)\|^2 > b\right] &\leq K^{2N}\max_{(\{k_{i1}\},\{k_{i2}\})}\Pr\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\overline{V}_t(k_{i1},k_{i2})\|^2 > b\right] \\ &\leq \exp\left[2N\ln K + \frac{4KTq}{3} - \frac{bNT}{6\lambda}\right].\end{aligned}$$

This implies that:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\overline{V}_t(\widehat{k}_i,\underline{k}_i)\|^2 = O_p((\ln K)/T) + O_p(qK/N).$$

The rest of the proof is as in Theorem B1.

C Model-based iteration and one-step GFE estimator

In this section we provide details on model-based iterated GFE, and we outline a continuously updated counterpart, the "one-step" GFE estimator.

C.1 Model-based iteration

We are going to derive a bound on the within-group mean squared error of α_{i0}, similarly to Corollary 1, for the case of iterated GFE.

Let $(\widehat{\theta},\widehat{\alpha})$ denote the two-step GFE estimator, with partition $\{\widehat{k}_i\}$, and let $\{\widehat{k}_i^{(2)}\}$ denote the partition after one iteration. We use the same notation as in the proof of Theorem 1. In particular, we define δ in the same way. We have:

$$\sum_{i=1}^{N}\sum_{j=1}^{p}\ell_{ij}\left(\widehat{\alpha}^j(\widehat{k}_i,\widehat{\theta}),\widehat{\theta}\right) \leq \sum_{i=1}^{N}\sum_{j=1}^{p}\ell_{ij}\left(\widehat{\alpha}^j(\widehat{k}_i^{(2)},\widehat{\theta}),\widehat{\theta}\right).$$

Hence, expanding, we obtain:

$$\begin{aligned}&-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{p}\left(\widehat{\alpha}^j(\widehat{k}_i^{(2)},\widehat{\theta}) - \widehat{\alpha}^j(\widehat{k}_i,\widehat{\theta})\right)' v_{ij}^{\alpha}\left(a_{ij},\widehat{\theta}\right)\left(\widehat{\alpha}^j(\widehat{k}_i^{(2)},\widehat{\theta}) - \widehat{\alpha}^j(\widehat{k}_i,\widehat{\theta})\right) \\ &\leq \sum_{i=1}^{N}\sum_{j=1}^{p}v_{ij}\left(\widehat{\alpha}^j(\widehat{k}_i,\widehat{\theta}),\widehat{\theta}\right)\left(\widehat{\alpha}^j(\widehat{k}_i^{(2)},\widehat{\theta}) - \widehat{\alpha}^j(\widehat{k}_i,\widehat{\theta})\right),\end{aligned}$$

where a_{ij} is between $\widehat{\alpha}^j(\widehat{k}_i,\widehat{\theta})$ and $\widehat{\alpha}^j(\widehat{k}_i^{(2)},\widehat{\theta})$.

Let us first assume that p does not grow with the sample size. Using similar arguments as in the proof of Theorem 1 we then have:

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

Using Theorem 1, we then have:

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

Moreover, since p does not grow with the sample size:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|v_{ij}\left(\alpha_{i0}^j,\theta_0\right)\right\|^2=O_p\left(\frac{1}{T}\right),$$

so:

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

Using again Theorem 1, we obtain:

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

It follows that:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\overline{\alpha}_0^j(\widehat{k}_i^{(2)})-\alpha_{i0}^j\right\|^2=O_p(\delta). \tag{C1}$$

Let us now turn to the case where p grows with the sample size. In that case, one can bound:

$$\begin{aligned}
&\left\|\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}v_{ij}\left(\widehat{\alpha}^j(\widehat{k}_i,\widehat{\theta}),\widehat{\theta}\right)\left(\widehat{\alpha}^j(\widehat{k}_i^{(2)},\widehat{\theta})-\widehat{\alpha}^j(\widehat{k}_i,\widehat{\theta})\right)\right\|^2\\
&\leq\left\|\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}v_{ij}\left(\alpha_{i0}^j,\theta_0\right)\left(\widehat{\alpha}^j(\widehat{k}_i^{(2)},\widehat{\theta})-\widehat{\alpha}^j(\widehat{k}_i,\widehat{\theta})\right)\right\|^2+O_p(\delta)\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\widehat{\alpha}^j(\widehat{k}_i^{(2)},\widehat{\theta})-\widehat{\alpha}^j(\widehat{k}_i,\widehat{\theta})\right\|^2\\
&=\left\|\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\overline{v}_j\left(\widehat{k}_i,\widehat{k}_i^{(2)}\right)\left(\widehat{\alpha}^j(\widehat{k}_i^{(2)},\widehat{\theta})-\widehat{\alpha}^j(\widehat{k}_i,\widehat{\theta})\right)\right\|^2+O_p(\delta)\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\widehat{\alpha}^j(\widehat{k}_i^{(2)},\widehat{\theta})-\widehat{\alpha}^j(\widehat{k}_i,\widehat{\theta})\right\|^2\\
&\leq\left[\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\overline{v}_j\left(\widehat{k}_i,\widehat{k}_i^{(2)}\right)\right\|^2+O_p(\delta)\right]\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\widehat{\alpha}^j(\widehat{k}_i^{(2)},\widehat{\theta})-\widehat{\alpha}^j(\widehat{k}_i,\widehat{\theta})\right\|^2,
\end{aligned}$$

where $\overline{v}_j(k,k')$ denotes the j-specific linear projection of $v_{ij}\left(\alpha_{i0}^j,\theta_0\right)$ on group indicators $\mathbf{1}\{\widehat{k}_i=k\}$ and $\mathbf{1}\{\widehat{k}_i^{(2)}=k'\}$.

Moreover, under a sub-Gaussianity assumption, and using a union bound argument together with Lemma B1, we will obtain the following bound:

$$\frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}\left\|\overline{v}_j\left(\widehat{k}_i,\widehat{k}_i^{(2)}\right)\right\|^2=O_p\left(\frac{\ln K}{T}\right)+O_p\left(\frac{Kp}{NT}\right).$$

Combining results, and using Theorem 1, we obtain a bound that, similarly to (C1), depends on $K^{-2/d}$, where d is the underlying dimension of (α'_{i0},μ'_{i0}). Hence we obtain a convergence rate, but no improvement, for iterated GFE relative to the baseline two-step approach.

C.2 One-step GFE estimator

A continuously updated counterpart to the iterated estimator is the *one-step* GFE estimator, which is defined as follows:

$$\left(\widehat{\theta}^{\text{1step}}, \widehat{\alpha}^{\text{1step}}, \{\widehat{k}_i^{\text{1step}}\}\right) = \underset{(\theta,\alpha,\{k_i\})}{\text{argmax}} \sum_{i=1}^{N} \ln f_i\left(\alpha\left(k_i\right), \theta\right), \tag{C2}$$

where the maximum is taken with respect to all possible parameter values (θ, α) and all possible partitions $\{k_i\}$ of $\{1, ..., N\}$ into at most K groups. This corresponds to the classification maximum likelihood estimator of Bryant and Williamson (1978); see also Hahn and Moon (2010) and Bonhomme and Manresa (2015). Unlike in two-step GFE, (C2) requires optimizing the likelihood function with respect to every partition and parameter value. This poses two difficulties. First, the estimator may be substantially more computationally intensive than two-step methods. Second, this complicates the statistical analysis since the discrete classification depends on parameter values and the objective function of the one-step estimator is therefore not smooth. In the case of the kmeans estimator, Pollard (1981, 1982) derived asymptotic properties for fixed K and T, as N tends to infinity. Deriving the properties of one-step estimators in (C2) as N, T, K tend jointly to infinity is an interesting avenue for future work.

D Two-way grouped fixed-effects

In this section of the appendix we consider two-way GFE estimators, and we derive an expansion in the spirit of Theorem 1.

We have the following lemma, the proof of which is analogous to that of Lemma 1, and omitted for brevity. Here we consider the general case where the precision of h_i is S, and the precision of w_t is J, while in the main text we focus on the case where $S = T$ and $J = N$.

Lemma D1. *Suppose that there exist random vectors h_i and w_t, with fixed dimensions, and Lipschitz-continuous functions φ and ϕ, such that $h_i = \varphi(\xi_{i0}) + o_p(1)$, $\frac{1}{N}\sum_{i=1}^{N} \|h_i - \varphi(\xi_{i0})\|^2 = O_p\left(1/S\right)$, $w_t = \phi(\lambda_{t0}) + o_p(1)$, and $\frac{1}{T}\sum_{t=1}^{T} \|w_t - \phi(\lambda_{t0})\|^2 = O_p\left(1/J\right)$ as N, T, S, J tend to infinity. Then we have, as N, S, K tend to infinity:*

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

and, as T, J, p tend to infinity:

$$\frac{1}{T}\sum_{t=1}^{T} \left\|\widehat{w}(\widehat{l}_t) - \phi(\lambda_{t0})\right\|^2 = O_p\left(\frac{1}{J}\right) + O_p\left(B_\lambda(p)\right),$$

where $B_\lambda(p)$ is defined analogously as $B_\xi(K)$.

Assumption D1. *(regularity, two-way)*

(i) $(Y_{it}', X_{it}', \xi_{i0}', \lambda_{t0}')'$, $i = 1, .., N$, $t = 1, ..., T$, *are i.i.d.*

$\ell_{it}(\alpha_{it}, \theta)$ *is three times differentiable in both its arguments, for all* i, t.

The parameter space Θ *for* θ_0 *is compact, the spaces for* ξ_{i0} *and* λ_{t0} *are compact, and* θ_0 *belongs to the interior of* Θ.

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

$\sup_{\xi,\lambda,\alpha,\theta} |\mathbb{E}_{\xi_{i0}=\xi,\lambda_{t0}=\lambda}(\ell_{it}(\alpha, \theta))| = O(1)$, *and similarly for the first three derivatives of* ℓ_{it} *in both its arguments.*

The minimum (respectively, maximum) eigenvalue of $(-\frac{\partial^2 \ell_{it}(\alpha,\theta)}{\partial\alpha\partial\alpha'})$ *is bounded away from zero (resp., infinity) with probability one, uniformly in* i, t, α, θ.

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

$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}[\ell_{it}(\alpha_{it0}, \theta_0) - \mathbb{E}_{\xi_{i0},\lambda_{t0}}(\ell_{it}(\alpha_{it0}, \theta_0))]^2 = O_p(1)$, *and similarly for the first three derivatives of* ℓ_{it} *in both its arguments.*

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

$\inf_{\xi,\lambda,\theta} \mathbb{E}_{\xi_{i0}=\xi,\, \lambda_{t0}=\lambda}(-\frac{\partial^2 \ell_{it}(\overline{\alpha}(\theta,\xi,\lambda),\theta)}{\partial\alpha\partial\alpha'})$ *is positive definite.*

$\mathbb{E}\left[\ell_{it}(\overline{\alpha}(\theta, \xi_{i0}, \lambda_{t0}), \theta)\right]$ *has a unique maximum at* θ_0 *on* Θ, *and its second derivative* $-H$ *is negative definite.*

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

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

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

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

$\sup_{\widetilde{\xi},\lambda,\theta} \| \frac{\partial}{\partial\xi'}\big|_{\xi=\widetilde{\xi}} \mathbb{E}_{\xi_{i0}=\xi,\lambda_{t0}=\lambda}(\frac{\partial \ell_{it}(\overline{\alpha}(\theta,\xi,\lambda),\theta)}{\partial\alpha})\| = O(1)$.

$\sup_{\xi,\widetilde{\lambda},\theta} \| \frac{\partial}{\partial\lambda'}\big|_{\lambda=\widetilde{\lambda}} \mathbb{E}_{\xi_{i0}=\xi,\lambda_{t0}=\lambda}(\frac{\partial \ell_{it}(\overline{\alpha}(\theta,\xi,\lambda),\theta)}{\partial\alpha})\| = O(1)$.

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

$\|\text{Var}_{h_i=h,\xi_{i0}=\xi,w_t=w,\lambda_{t0}=\lambda}(\frac{\partial \ell_{it}(\overline{\alpha}(\theta,\xi,\lambda),\theta)}{\partial\alpha})\| = O(1)$, *uniformly in* h, w, ξ, λ, θ.

$\|\text{Var}_{h_i=h,\xi_{i0}=\xi,w_t=w,\lambda_{t0}=\lambda}(\text{vec}\, \frac{\partial}{\partial\theta'}\big|_{\theta_0} \frac{\partial \ell_{it}(\overline{\alpha}(\theta,\xi,\lambda),\theta)}{\partial\alpha})\| = O(1)$, *uniformly in* h, w, ξ, λ.

Let us denote:

$$\widetilde{s}_{it} = \frac{\partial \ell_{it}(\alpha_{it0}, \theta_0)}{\partial \theta} + \mathbb{E}_{\xi_{i0},\lambda_{t0}}\left(\frac{\partial^2 \ell_{it}(\alpha_{it0}, \theta_0)}{\partial \theta \partial \alpha'}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_{t0}}\left(-\frac{\partial^2 \ell_{it}(\alpha_{it0}, \theta_0)}{\partial \alpha \partial \alpha'}\right)\right]^{-1}\frac{\partial \ell_{it}(\alpha_{it0}, \theta_0)}{\partial \alpha}, \quad \text{(D1)}$$

and:

$$\begin{aligned}\widetilde{H} =& \mathbb{E}\Bigg[\mathbb{E}_{\xi_{i0},\lambda_{t0}}\left(-\frac{\partial^2 \ell_{it}(\alpha_{it0}, \theta_0)}{\partial \theta \partial \theta'}\right) \\ &- \mathbb{E}_{\xi_{i0},\lambda_{t0}}\left(\frac{\partial^2 \ell_{it}(\alpha_{it0}, \theta_0)}{\partial \theta \partial \alpha'}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_{t0}}\left(-\frac{\partial^2 \ell_{it}(\alpha_{it0}, \theta_0)}{\partial \alpha \partial \alpha'}\right)\right]^{-1}\mathbb{E}_{\xi_{i0},\lambda_{t0}}\left(\frac{\partial^2 \ell_{it}(\alpha_{it0}, \theta_0)}{\partial \alpha \partial \theta'}\right)\Bigg]. \quad \text{(D2)}\end{aligned}$$

Theorem D1. *Let the conditions in Lemma D1 hold. Suppose that $B_\xi(K) = O_p(K^{-\frac{2}{d}})$ and $B_\lambda(p) = O_p(p^{-\frac{2}{d_\lambda}})$. Suppose that α and μ are Lipschitz-continuous in both arguments. Suppose that there exist two Lipschitz-continuous functions ψ and Ψ such that $\xi_{i0} = \psi(\varphi(\xi_{i0}))$ and $\lambda_{t0} = \Psi(\phi(\lambda_{t0}))$. Lastly, let Assumption D1 hold. Then, as N, T, S, J, K, p tend to infinity such that $Kp/(NT)$ tends to zero, we have:*

$$\begin{aligned}\widehat{\theta} \;=\; \theta_0 + \widetilde{H}^{-1}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{s}_{it} + O_p\left(\frac{1}{S}\right) + O_p\left(\frac{1}{J}\right) + O_p\left(\frac{Kp}{NT}\right) \\ + O_p\left(K^{-\frac{2}{d}}\right) + O_p\left(p^{-\frac{2}{d_\lambda}}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right). \quad \text{(D3)}\end{aligned}$$

Proof of Theorem D1

In the proof we closely follow the steps of the proof of Theorem 1. Let, for all $\theta \in \Theta$, $k \in \{1, ..., K\}$, and $l \in \{1, ..., p\}$:

$$\widehat{\alpha}(k, l, \theta) = \text{argmax}_\alpha \sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{1}\{\widehat{k}_i = k\}\mathbf{1}\{\widehat{l}_t = l\}\ell_{it}(\alpha, \theta), \quad \text{(D4)}$$

and define $\overline{\alpha}(\theta, \xi, \lambda)$ according to Assumption D1 (*iii*). Let also $\delta = \frac{1}{S} + \frac{1}{J} + \frac{Kp}{NT} + K^{-\frac{2}{d}} + p^{-\frac{2}{d_\lambda}}$.

A key step in the proof is to establish the following main intermediate results:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial \ell_{it}(\widehat{\alpha}(\widehat{k}_i, \widehat{l}_t, \theta_0), \theta_0)}{\partial \theta} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial}{\partial \theta}\Big|_{\theta_0}\ell_{it}(\overline{\alpha}(\theta, \xi_{i0}, \lambda_{t0}), \theta) + O_p(\delta), \quad \text{(D5)}$$

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial^2}{\partial \theta \partial \theta'}\Big|_{\theta_0}\left(\ell_{it}\left(\widehat{\alpha}(\widehat{k}_i, \widehat{l}_t, \theta), \theta\right) - \ell_{it}(\overline{\alpha}(\theta, \xi_{i0}, \lambda_{t0}), \theta)\right) = o_p(1). \quad \text{(D6)}$$

Consistency of $\widehat{\theta}$. The consistency proof follows Theorem 1 closely. First, we show that $\overline{\alpha}(\theta, \xi, \lambda)$ is Lipschitz-continuous, with coefficients that are uniformly bounded with respect to θ, ξ and λ.

Next, we define $a(k, l, \theta) = \overline{\alpha}(\theta, \psi(\widehat{h}(k)), \Psi(\widehat{w}(l)))$, and we use Lemma D1 to show that:

$$\sup_{\theta \in \Theta}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|a(\widehat{k}_i, \widehat{l}_t, \theta) - \overline{\alpha}(\theta, \xi_{i0}, \lambda_{t0})\right\|^2 = O_p(\delta). \quad \text{(D7)}$$

Next, we use that, for all θ:

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\ell_{it}\left(a(\widehat{k}_i,\widehat{l}_t,\theta),\theta\right)\leq\sum_{i=1}^{N}\sum_{t=1}^{T}\ell_{it}\left(\widehat{\alpha}(\widehat{k}_i,\widehat{l}_t,\theta),\theta\right), \tag{D8}$$

and we expand both sides of this inequality as in the proof of Theorem 1 to obtain, for all $\theta\in\Theta$:

$$\begin{aligned}&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\widehat{\alpha}(\widehat{k}_i,\widehat{l}_t,\theta)-\overline{\alpha}(\theta,\xi_{i0},\lambda_{t0})\right\|^2\\&\leq O_p\left[\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\overline{v}(\widehat{k}_i,\widehat{l}_t,\theta)\|^2\right)^{\frac{1}{2}}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\widehat{\alpha}(\widehat{k}_i,\widehat{l}_t,\theta)-a(\widehat{k}_i,\widehat{l}_t,\theta)\right\|^2\right)^{\frac{1}{2}}\right]+O_p(\delta),\end{aligned}$$

where $\overline{v}(k,l,\theta)$ denotes the mean of $v_{it}(\overline{\alpha}(\theta,\xi_{i0},\lambda_{t0}),\theta)$ in the intersection of groups $\widehat{k}_i=k$ and $\widehat{l}_t=l$.

As in the proof of Theorem 1, the key step is to show:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\overline{v}(\widehat{k}_i,\widehat{l}_t,\theta)\right\|^2=O_p\left(\delta\right). \tag{D9}$$

Let, for all θ,h,w,ξ,λ:

$$\rho(h,\xi,w,\lambda,\theta)=\mathbb{E}_{h_i=h,\xi_{i0}=\xi,w_t=w,\lambda_{t0}=\lambda}(v_{it}(\overline{\alpha}(\theta,\xi,\lambda),\theta)),$$

and let, for all i,t,θ:

$$\zeta_{it}(\theta)=v_{it}(\overline{\alpha}(\theta,\xi_{i0},\lambda_{t0}),\theta)-\rho(h_i,\xi_{i0},w_t,\lambda_{t0},\theta).$$

Expanding $\rho(h_i,\xi_{i0},w_t,\lambda_{t0},\theta)$ around $h_i=\varphi(\xi_{i0})$ and $w_t=\phi(\lambda_{t0})$, we have:

$$\begin{aligned}\rho(h_i,\xi_{i0},w_t,\lambda_{t0},\theta)=&\rho(\varphi(\xi_{i0}),\xi_{i0},\phi(\lambda_{t0}),\lambda_{t0},\theta)+\frac{\partial\rho(\varphi(\xi_{i0}),\xi_{i0},\phi(\lambda_{t0}),\lambda_{t0},\theta)}{\partial h'}(h_i-\varphi(\xi_{i0}))\\&+\frac{\partial\rho(\varphi(\xi_{i0}),\xi_{i0},\phi(\lambda_{t0}),\lambda_{t0},\theta)}{\partial w'}(w_t-\phi(\lambda_{t0}))+O_p\left(\frac{1}{S}\right)+O_p\left(\frac{1}{J}\right).\end{aligned}$$

Hence, taking expectations:

$$\begin{aligned}0=&\mathbb{E}_{\xi_{i0},\lambda_{t0}}(v_{it}(\overline{\alpha}(\theta,\xi_{i0},\lambda_{t0}),\theta)))\\=&\mathbb{E}_{\xi_{i0},\lambda_{t0}}\left[\rho(h_i,\xi_{i0},w_t,\lambda_{t0},\theta)\right]\\=&\rho(\varphi(\xi_{i0}),\xi_{i0},\phi(\lambda_{t0}),\lambda_{t0},\theta)+\frac{\partial\rho(\varphi(\xi_{i0}),\xi_{i0},\phi(\lambda_{t0}),\lambda_{t0},\theta)}{\partial h'}\mathbb{E}_{\xi_{i0},\lambda_{t0}}(h_i-\varphi(\xi_{i0}))\\&+\frac{\partial\rho(\varphi(\xi_{i0}),\xi_{i0},\phi(\lambda_{t0}),\lambda_{t0},\theta)}{\partial w'}\mathbb{E}_{\xi_{i0},\lambda_{t0}}(w_t-\phi(\lambda_{t0}))+O_p\left(\frac{1}{S}\right)+O_p\left(\frac{1}{J}\right).\end{aligned}$$

It follows that:

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

We thus only need to bound:

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

where we have used that observations are independent across i and t. To bound this quantity, we use part (v) in Assumption D1. We thus obtain (D9).

The rest of the consistency part is as in the proof of Theorem 1.

Proof of (D5). To show (D5), we are now going to show that:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{v_{it}^{\theta}\left(\widehat{\alpha}(\widehat{k}_i,\widehat{l}_t)-\alpha_{it0}\right)+\mathbb{E}_{\xi_{i0},\lambda_{t0}}\left(v_{it}^{\theta}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_{t0}}\left(v_{it}^{\alpha}\right)\right]^{-1}v_{it}\right\}=O_p\left(\delta\right), \tag{D10}$$

where we omit references to θ_0 and α_{it0}.

We will bound, in turn:

$$A \equiv \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}_{\xi_{i0},\lambda_{t0}}\left(v_{it}^{\theta}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_{t0}}\left(v_{it}^{\alpha}\right)\right]^{-1}v_{it}^{\alpha}\left(\widehat{\alpha}(\widehat{k}_i,\widehat{l}_t)-\alpha_{it0}+(v_{it}^{\alpha})^{-1}v_{it}\right),$$
$$B \equiv \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(v_{it}^{\theta}\left(v_{it}^{\alpha}\right)^{-1}-\mathbb{E}_{\xi_{i0},\lambda_{t0}}\left(v_{it}^{\theta}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_{t0}}\left(v_{it}^{\alpha}\right)\right]^{-1}\right)v_{it}^{\alpha}\left(\widehat{\alpha}(\widehat{k}_i,\widehat{l}_t)-\alpha_{it0}\right).$$

To bound A, the key term is:

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

where $\widetilde{v}$ is defined analogously as in the proof of Theorem 1.

Let $z(\xi,\lambda)'=\mathbb{E}_{\xi_{i0}=\xi,\lambda_{t0}=\lambda}\left(v_{it}^{\theta}\right)\left[\mathbb{E}_{\xi_{i0}=\xi,\lambda_{t0}=\lambda}\left(v_{it}^{\alpha}\right)\right]^{-1}$. We have:

$$A_3=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(z(\xi_{i0},\lambda_{t0})'-z^*\left(\widehat{k}_i,\widehat{l}_t\right)'\right)v_{it}+\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(z^*\left(\widehat{k}_i,\widehat{l}_t\right)'-\widetilde{z}\left(\widehat{k}_i,\widehat{l}_t\right)'\right)v_{it},$$

where $\widetilde{z}\left(k,l\right)$ and $z^*\left(k,l\right)$ are defined analogously as in the proof of Theorem 1.

To see that the first term in A_3 is $O_p(\delta)$, we use an argument similar to the one we used in the proof of Theorem 1. Let: $\zeta_{it}=v_{it}-\mathbb{E}_{h_i,\xi_{i0},w_t,\lambda_{t0}}(v_{it})$. We have, as in the proof of Theorem 1:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\mathbb{E}_{h_i,\xi_{i0},w_t,\lambda_{t0}}(v_{it})\right\|^2=O_p\left(\frac{1}{S}\right)+O_p\left(\frac{1}{J}\right). \tag{D11}$$

Moreover, we have:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|z(\xi_{i0},\lambda_{t0})-z^*(\widehat{k}_i,\widehat{l}_t)\right\|^2=O_p(\delta). \tag{D12}$$

Since ζ_{it} are independent across i and t, with zero mean conditional on $h_1, ..., h_N$, $g_1, ..., g_T$, $\xi_{10}, ..., \xi_{N0}$, and $\lambda_{10}, ..., \lambda_{T0}$, we thus have:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(z(\xi_{i0},\lambda_{t0})' - z^*\left(\widehat{k}_i,\widehat{l}_t\right)'\right)v_{it}\right\|^2\right] \\
&\leq 2\left(O\left(\frac{1}{S}\right)+O\left(\frac{1}{J}\right)\right)\mathbb{E}\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|z(\xi_{i0},\lambda_{t0})' - z^*\left(\widehat{k}_i,\widehat{l}_t\right)'\right\|^2\right] \\
&\quad + 2\mathbb{E}\left[\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(z(\xi_{i0},\lambda_{t0})' - z^*\left(\widehat{k}_i,\widehat{l}_t\right)'\right)\zeta_{it}\right\|^2\right] \\
&= O\left(\frac{\delta}{S}\right)+O\left(\frac{\delta}{J}\right) \\
&\quad + \frac{1}{N^2T^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[\left(z(\xi_{i0},\lambda_{t0})' - z^*\left(\widehat{k}_i,\widehat{l}_t\right)'\right)\mathbb{E}_{h_i,\xi_{i0},w_t,\lambda_{t0}}\left[\zeta_{it}\zeta_{it}'\right]\left(z(\xi_{i0},\lambda_{t0})' - z^*\left(\widehat{k}_i,\widehat{l}_t\right)'\right)\right] \\
&= O\left(\frac{\delta}{S}\right)+O\left(\frac{\delta}{J}\right)+O\left(\frac{\delta}{NT}\right) = O(\delta^2).
\end{aligned}$$

The second term in A_3 is also $O_p(\delta)$, using similar arguments as in the proof of Theorem 1 and equation (D9). This shows that $A = O_p(\delta)$.

Let us now turn to B. Let $\pi_{it}' = v_{it}^{\theta}\left(v_{it}^{\alpha}\right)^{-1} - \mathbb{E}_{\xi_{i0},\lambda_{t0}}\left(v_{it}^{\theta}\right)\left[\mathbb{E}_{\xi_{i0},\lambda_{t0}}\left(v_{it}^{\alpha}\right)\right]^{-1}$. As in the proof of Theorem 1, the key step is to bound:

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

where $\widetilde{\alpha}(k,l)$ and $\alpha^*(k,l)$ are defined analogously as in the proof of Theorem 1.

The first term is $O_p(\delta)$, since the $\tau_{it} = \pi_{it}'v_{it}^{\alpha}$ are independent across i and t with zero mean given ξ_{i0}, λ_{t0}. The second term is:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\pi_{it}'v_{it}^{\alpha}\left(\widetilde{\alpha}(\widehat{k}_i,\widehat{l}_t) - \alpha^*(\widehat{k}_i,\widehat{l}_t)\right) \quad = \quad \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{\pi}(\widehat{k}_i,\widehat{l}_t)'v_{it}^{\alpha}\left(\widetilde{\alpha}(\widehat{k}_i,\widehat{l}_t) - \alpha^*(\widehat{k}_i,\widehat{l}_t)\right).$$

As in the proof of Theorem 1, $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\widetilde{\pi}(\widehat{k}_i,\widehat{l}_t)\|^2 = O_p(\delta)$. Moreover, we have:

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

This shows that $B_3 = O_p(\delta)$, hence that $B = O_p(\delta)$. This implies that (D5) holds.

Proof of (D6). The proof of (D6) follows the same steps as the corresponding part in the proof of Theorem 1. The key step is to show that:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\frac{\partial\widehat{\alpha}(\widehat{k}_i,\widehat{l}_t,\theta_0)}{\partial\theta'}-\frac{\partial\overline{\alpha}^j(\theta_0,\xi_{i0},\lambda_{t0})}{\partial\theta'}\right\|^2=o_p\left(1\right). \tag{D13}$$

The proof of (D13) follows similar arguments as in the proof of Theorem 1.

E Further results for fixed p

In this section we derive two additional properties of GFE estimators, in the case where p is fixed; that is, when heterogeneity is time-invariant and its dimension does not grow with the sample size. Without loss of generality we set $p=1$. For simplicity, in this section we assume that $S=T$, and we remove p and j from the notation. Finally, we denote $\eta_{i0}=(\alpha_{i0}',\mu_{i0}')'$ the full vector of heterogeneity.

E.1 Expansion of two-step GFE

We start with a result that shows that the expansion in Theorem 1 continues to hold under suitable conditions, when the concavity in α of Assumption 3 part (iia) fails to hold.

Corollary E1. *Let the conditions in Theorem 1 hold, with $\xi_{i0}=\eta_{i0}$, except that Assumption 3 part (iia) is replaced by the following:*

(i) For all $\epsilon>0$, $\inf_{\theta,\eta=(\alpha',\mu')'}\inf_{\|(\widetilde{\alpha},\widetilde{\theta})-(\alpha,\theta)\|>\epsilon}\mathbb{E}_{\eta_{i0}=\eta}[\ell_i(\alpha,\theta)-\ell_i(\widetilde{\alpha},\widetilde{\theta})]>0$.

(ii) The function $\widehat{\ell}_i(\theta)=\ell_i(\widehat{\alpha}(\widehat{k}_i,\theta),\theta)$ is three times differentiable on a neighborhood of θ_0 (that is, the function is three times differentiable for almost all sample realizations), where $\widehat{\alpha}(k,\theta)$ for all k is the solution of (3) given any $\theta\in\Theta$. Moreover, $\frac{1}{N}\sum_{i=1}^{N}\|\frac{\partial^2\widehat{\ell}_i(\theta)}{\partial\theta\partial\theta'}\|^2=O_p(1)$ uniformly in a neighborhood of θ_0, and similarly for the third derivative of $\widehat{\ell}_i$.

Then, for p fixed as N,T,K tend to infinity, we have:

$$\widehat{\theta} \;=\; \theta_0+H^{-1}\frac{1}{N}\sum_{i=1}^{N}s_i+O_p\left(\frac{1}{T}\right)+O_p\left(K^{-\frac{2}{d}}\right)+o_p\left(\frac{1}{\sqrt{NT}}\right), \tag{E1}$$

and:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\right\|^2=O_p\left(\frac{1}{T}\right)+O_p\left(K^{-\frac{2}{d}}\right). \tag{E2}$$

Condition (i) in Corollary E1 is an identification condition. Condition (ii) is a differentiability condition on the sample GFE objective function. Such an assumption is not needed in order to characterize the first-order properties of fixed-effects estimators, since, under suitable conditions, fixed-effects estimators of individual effects are uniformly consistent in the sense that $\max_{i=1,\dots,N}\|\widehat{\alpha}_i-$

$\alpha_{i0}\| = o_p(1)$; see, e.g., Hahn and Kuersteiner (2011). In contrast, our characterization of GFE estimators is based on establishing a rate of convergence for the average $\frac{1}{N}\sum_{i=1}^{N}\|\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\|^2$. Lastly, the score s_i and Hessian H in Corollary E1 are the same as in Theorem 1.

Proof of Corollary E1. The proof follows closely the proof of Theorem 1. The beginning of the consistency proof is identical. Here, $\delta = 1/T + K^{-2/d}$. One difference is that we introduce the fixed-effects estimator of α_i, for given θ, which is defined as $\widehat{\alpha}_i(\theta) = \text{argmax}_\alpha\ \ell_i(\alpha,\theta)$. We have, for all θ:

$$\sum_{i=1}^{N}\ell_i\left(a(\widehat{k}_i,\theta),\theta\right) \leq \sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}(\widehat{k}_i,\theta),\theta\right) \leq \sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}_i(\theta),\theta\right). \tag{E3}$$

Moreover, expanding the fixed-effects log-likelihood around $\overline{\alpha}_i(\theta) \equiv \overline{\alpha}(\theta,\eta_{i0})$, we have (as in Arellano and Hahn, 2007, for example):

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

uniformly in θ.

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

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

We then obtain, using similar arguments as in the proof of Theorem 1, that the following equality holds pointwise in θ:

$$\left|\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}(\widehat{k}_i,\theta),\theta\right) - \frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}_i\left(\theta\right),\theta\right)\right| = O_p(\delta). \tag{E4}$$

In turn, from the uniform convergence of the fixed-effects log-likelihood:

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

we obtain, using similar arguments as above:

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

Consistency of $\widehat{\theta}$ then follows, as in the proof of Theorem 1.

Let us now show equation (A3). From (E4) evaluated at θ_0 we have, for some a_i between $\widehat{\alpha}(\widehat{k}_i, \theta_0)$ and $\widehat{\alpha}_i(\theta_0)$, and omitting from now on the reference to θ_0 for conciseness:

$$O_p(\delta) = \frac{1}{N}\sum_{i=1}^{N} \ell_i(\widehat{\alpha}_i) - \frac{1}{N}\sum_{i=1}^{N} \ell_i(\widehat{\alpha}(\widehat{k}_i)) = \frac{1}{2N}\sum_{i=1}^{N} \left(\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right)' \left(-v_i^{\alpha}(a_i)\right) \left(\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right) \geq 0. \quad \text{(E5)}$$

Now, by the assumptions of Theorem 1 part (a), there exists a constant $\epsilon > 0$ and a positive definite matrix $\underline{\Sigma}$ such that:

$$\inf_{\eta=(\alpha',\mu')'} \inf_{\|\widetilde{\alpha}-\alpha\|\leq\epsilon} \mathbb{E}_{\eta_{i0}=\eta}\left(-v_i^{\alpha}(\widetilde{\alpha})\right) \geq \underline{\Sigma}.$$

For this ϵ we will first show that:

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left\{\|\widehat{\alpha}(\widehat{k}_i) - \alpha_{i0}\| > \epsilon\right\} = O_p(\delta). \quad \text{(E6)}$$

Showing (E6) will allow us to control the difference $\widehat{\alpha}(\widehat{k}_i) - \alpha_{i0}$ in an average sense.

To see that (E6) holds, let $\iota_i = \mathbf{1}\left\{\|\widehat{\alpha}(\widehat{k}_i) - \alpha_{i0}\| \leq \epsilon\right\}$, and note that by (E5) we have, since $\ell_i(\widehat{\alpha}_i) \geq \ell_i(\widehat{\alpha}(\widehat{k}_i))$ for all i:

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

Now, by part (i) in the assumptions of the corollary, and using that $\max_{i=1,\ldots,N}\|\widehat{\alpha}_i - \alpha_{i0}\| = o_p(1)$ and Assumption 3 part (iia), we have:

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

where $\zeta > 0$ is a constant. Hence $\frac{1}{N}\sum_{i=1}^{N}(1-\iota_i)(\zeta + o_p(1)) = O_p(\delta)$, from which (E6) follows.

Next, by Assumption 3 part (iia), $\sup_{\eta=(\alpha',\mu')'}\sup_{\widetilde{\alpha}}\left\|v_i^{\alpha}(\widetilde{\alpha}) - \mathbb{E}_{\eta_{i0}=\eta}\left(v_i^{\alpha}(\widetilde{\alpha})\right)\right\| = o_p(1)$. We thus have:

$$\min_{i=1,\ldots,N} \inf_{\|\alpha_i-\alpha_{i0}\|\leq\epsilon} \left(-v_i^{\alpha}(\alpha_i)\right) \geq \underline{\Sigma} + o_p(1). \quad \text{(E7)}$$

Using (E5) this implies that: $\frac{1}{N}\sum_{i=1}^{N}\iota_i\left\|\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right\|^2 = O_p(\delta)$. Hence, using in addition (E6) and the fact that the parameter space for α_i is compact, we have:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right\|^2 = \frac{1}{N}\sum_{i=1}^{N}\iota_i\left\|\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right\|^2 + \frac{1}{N}\sum_{i=1}^{N}(1-\iota_i)\left\|\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\right\|^2 = O_p(\delta). \quad \text{(E8)}$$

Given (E8), the rest of the proof of (A3) is as in the proof of Theorem 1, with some simplifications due to the fact that here p does not grow with the sample size.

However, the proof of (A4) is different from the one in Theorem 1. To proceed, let:

$$\widetilde{\iota}(k) = \mathbf{1}\left\{\sum_{i'=1}^{N}\mathbf{1}\{\widehat{k}_{i'} = k\}\left(-v_{i'}^{\alpha}\left(\widehat{\alpha}(\widehat{k}_{i'})\right)\right) \geq \frac{1}{2}\underline{\Sigma}\right\}.$$

We are first going to show that:

$$\frac{1}{N}\sum_{i=1}^{N}\left(1-\widetilde{\iota}\left(\widehat{k}_i\right)\right)=O_p(\delta). \tag{E9}$$

Let $\epsilon>0$ as in (E6), and define ι_i accordingly. From (E6) it suffices to show that:

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

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

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

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

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

This shows (E9).

We are now going to show (A4). By part (ii) in the assumptions of the corollary, the Cauchy

Schwarz inequality, and (E9), we have:

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

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

$$\frac{\partial\widehat{\alpha}(k)}{\partial\theta'}=\left(\sum_{i'=1}^{N}\mathbf{1}\{\widehat{k}_{i'}=k\}\left(-v_{i'}^{\alpha}\left(\widehat{\alpha}(\widehat{k}_{i'})\right)\right)\right)^{-1}\sum_{i'=1}^{N}\mathbf{1}\{\widehat{k}_{i'}=k\}\left(v_{i'}^{\theta}\left(\widehat{\alpha}(\widehat{k}_{i'})\right)\right)', \tag{E10}$$

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

Let now:

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

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

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

where we have used that $\mathbb{E}_{\eta_{i0}}(v_i)=0$, and that $\frac{\partial\overline{\alpha}_i}{\partial\theta'}=\left[\mathbb{E}_{\eta_{i0}}\left(-v_i^{\alpha}\right)\right]^{-1}\mathbb{E}_{\eta_{i0}}\left(v_i^{\theta}\right)'$. Hence, using (E8) and (E9):

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

so:

$$D = \frac{1}{N}\sum_{i=1}^{N}\widetilde{\iota}(\widehat{k}_i)\mathbb{E}_{\eta_{i0}}\left(v_i^{\theta}\right)\left(\frac{\partial\widehat{\alpha}(\widehat{k}_i,\theta_0)}{\partial\theta'}-\frac{\partial\overline{\alpha}_i(\theta_0)}{\partial\theta'}\right)+o_p(1).$$

Next, defining $z(\eta)' = \mathbb{E}_{\eta_{i0}=\eta}\left(v_i^\theta\right)\left[\mathbb{E}_{\eta_{i0}=\eta}\left(-v_i^\alpha\right)\right]^{-1}$ and $\widetilde{z}(k)$ as in the proof of Theorem 1 we have:

$$
\begin{aligned}
D &= \frac{1}{N}\sum_{i=1}^N \widetilde{\iota}(\widehat{k}_i) z(\eta_{i0})' \mathbb{E}_{\eta_{i0}}\left(-v_i^\alpha\right)\left(\frac{\partial \widehat{\alpha}(\widehat{k}_i)}{\partial \theta'} - \frac{\partial \overline{\alpha}_i}{\partial \theta'}\right) + o_p(1) \\
&= \frac{1}{N}\sum_{i=1}^N \widetilde{\iota}(\widehat{k}_i) z(\eta_{i0})' \left(-v_i^\alpha\left(\widehat{\alpha}(\widehat{k}_i)\right)\right)\left(\frac{\partial \widehat{\alpha}(\widehat{k}_i)}{\partial \theta'} - \frac{\partial \overline{\alpha}_i}{\partial \theta'}\right) + o_p(1) \\
&= \frac{1}{N}\sum_{i=1}^N \widetilde{\iota}(\widehat{k}_i) \widetilde{z}(\widehat{k}_i)' \left(-v_i^\alpha\left(\widehat{\alpha}(\widehat{k}_i)\right)\right)\left(\frac{\partial \widehat{\alpha}(\widehat{k}_i)}{\partial \theta'} - \frac{\partial \overline{\alpha}_i}{\partial \theta'}\right) + o_p(1) \\
&= \frac{1}{N}\sum_{i=1}^N \widetilde{\iota}(\widehat{k}_i) \widetilde{z}(\widehat{k}_i)' \left(\left(v_i^\theta\left(\widehat{\alpha}(\widehat{k}_i)\right)\right)' - \left(-v_i^\alpha\left(\widehat{\alpha}(\widehat{k}_i)\right)\right)\frac{\partial \overline{\alpha}_i}{\partial \theta'}\right) + o_p(1) \\
&= \frac{1}{N}\sum_{i=1}^N \widetilde{\iota}(\widehat{k}_i) \widetilde{z}(\widehat{k}_i)' \left(\underbrace{\left(\mathbb{E}_{\eta_{i0}}\left(v_i^\theta\right)\right)' - \left(\mathbb{E}_{\eta_{i0}}\left(-v_i^\alpha\right)\right)\frac{\partial \overline{\alpha}_i}{\partial \theta'}}_{=0}\right) + o_p(1) = o_p(1),
\end{aligned}
$$

where we have used (E8) in the third equality, (A33) and the expression of $\partial\widehat{\alpha}(k)/\partial\theta'$ in the fifth one, and we have expanded around α_{i0} and used (E8) in the last equality.

Finally, to show (E2), let us define, analogously to the beginning of the proof of (A3):

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

where ϵ is such that: $\inf_{\theta,\eta=(\alpha',\mu')'} \inf_{\|(\widetilde{\alpha},\widetilde{\theta})-(\alpha,\theta)\|\leq 2\epsilon} \mathbb{E}_{\eta_{i0}=\eta}\left(-v_i^\alpha(\widetilde{\alpha},\widetilde{\theta})\right) \geq \underline{\Sigma}$. Using that $\widehat{\theta}$ is consistent it is easy to verify that:

$$
\left|\frac{1}{N}\sum_{i=1}^N \ell_i\left(\widehat{\alpha}(\widehat{k}_i,\widehat{\theta}),\widehat{\theta}\right) - \frac{1}{N}\sum_{i=1}^N \ell_i\left(\widehat{\alpha}_i(\widehat{\theta}),\widehat{\theta}\right)\right| \leq \left|\frac{1}{N}\sum_{i=1}^N \ell_i\left(a(\widehat{k}_i,\widehat{\theta}),\widehat{\theta}\right) - \frac{1}{N}\sum_{i=1}^N \ell_i\left(\widehat{\alpha}_i(\widehat{\theta}),\widehat{\theta}\right)\right| = O_p(\delta). \tag{E11}
$$

Using similar arguments as at the beginning of the proof of (A3), but now at $\widehat{\theta}$, we obtain that:

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

hence that:

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

(E2) then comes from the fact that $\frac{1}{N}\sum_{i=1}^N \|\widehat{\alpha}_i(\widehat{\theta}) - \alpha_{i0}\|^2 = O_p(\delta)$.

This ends the proof of Corollary E1.

E.2 Analytical expression of the asymptotic bias

The next result provides an explicit characterization of the $O_p(1/T)$ bias in (9), in cases where K grows relatively fast relative to T. We recall that here p is fixed.

Corollary E2. *Suppose that the conditions of Theorem 1 are satisfied. Let* $\widehat{\alpha}_i(\theta) = \text{argmax}_{\alpha_i} \ell_i(\alpha_i, \theta)$, *and let* $\widehat{g}_i(\theta) = \frac{\partial^2 \ell_i(\widehat{\alpha}_i(\theta),\theta)}{\partial\theta\partial\alpha'}(\frac{\partial^2 \ell_i(\widehat{\alpha}_i(\theta),\theta)}{\partial\alpha\partial\alpha'})^{-1}$. *Suppose in addition:*

(i) There is an $\upsilon > 0$ *such that* $T\, B_\xi(K^{1-\upsilon}) \xrightarrow{p} 0$ *as* N, T, K *tend to infinity. Moreover, for any diverging* $K_{N,T}$ *sequence,* $T\, B_\varepsilon(K_{N,T}) \xrightarrow{p} 0$ *as* N, T *tend to infinity.*

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

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

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

$$\widehat{\theta} \;=\; \theta_0 + H^{-1}\frac{1}{N}\sum_{i=1}^{N} s_i + \frac{C}{T} + o_p\left(\frac{1}{T}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right), \tag{E12}$$

where the expression of the constant C *is given by (E24)-(E25) in the proof.*

Under the conditions of Corollary E2, the kmeans objective is:

$$\frac{1}{N}\sum_{i=1}^{N} \|h_i - \widehat{h}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right). \tag{E13}$$

This happens when K grows sufficiently fast relative to T. As an example, when ξ_{i0} scalar and $B_\xi(K) = O_p(K^{-2})$, the condition requires TK^{-2} to tend to zero. The condition on B_ε is satisfied when ε_i is normal with zero mean and variance Σ/T for some $\Sigma > 0$, for example.

Corollary E2 shows that, when K is sufficiently large so that the approximation error $B_\xi(K)$ is small relative to $1/T$, and when in addition K/N tends to zero, the GFE estimator of θ_0 satisfies an expansion similar to that of the fixed-effects estimator, with a different first-order bias term; see, e.g., Hahn and Newey (2004, p.1302) for an expression of the bias of fixed-effects.

Proof of Corollary E2. We first show (E13). Let $\{k_i\} = \{k_{i1}\} \cap \{k_{i2}\}$ be the intersection of two partitions of $\{1, ..., N\}$: a first partition with (the integer part of) $K_1 = K^{1-\upsilon}$ groups, and a

second partition with (the integer part of) $K_2 = K^{\upsilon}$ groups. Since $(\widehat{h}, \{\widehat{k}_i\})$ solves (2), we have, using Condition (i) in the corollary:

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

To show that (E12) holds, we will follow a likelihood approach as in Arellano and Hahn (2007, 2016). Consider the difference between the GFE and fixed-effects concentrated likelihoods:

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

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

Specifically, we are going to show that:

$$\frac{\partial}{\partial\theta}\Big|_{\theta_0}\Delta L(\theta) \;=\; -\frac{\partial}{\partial\theta}\Big|_{\theta_0}\frac{1}{2N}\sum_{i=1}^{N}\nu_i(\theta)'\mathbb{E}_{\eta_{i0}}\left[-v_i^{\alpha}\left(\overline{\alpha}_i(\theta),\theta\right)\right]\nu_i(\theta) + o_p\left(\frac{1}{T}\right), \tag{E14}$$

where $\nu_i(\theta) = \widehat{\alpha}_i(\theta) - \mathbb{E}_{h_i}\left(\widehat{\alpha}_i(\theta)\right)$, and we denote $\overline{\alpha}_i\left(\theta\right) \equiv \overline{\alpha}(\theta,\eta_{i0})$. In this case also, we can define $\xi_{i0} = \eta_{i0}$.

To show (E14) we are first going to establish several preliminary results. Together with fourth-order differentiability in Condition (ii) of the corollary, those will allow us to derive the required expansions. In the following we will evaluate all functions at θ_0, and omit θ_0 for the notation. In particular, $\widehat{\alpha}_i$ will be a shorthand for $\widehat{\alpha}_i(\theta_0)$.

First, note that, from the proof of Theorem 1, and using (E13), we have:

$$\frac{1}{N}\sum_{i=1}^{N}\|\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\|^2 = O_p\left(\frac{1}{T}\right). \tag{E15}$$

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

$$\widehat{\alpha}(k) = \left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i = k\}(-v_i^{\alpha}(a_i(k)))\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i = k\}(-v_i^{\alpha}(a_i(k)))\widehat{\alpha}_i\right), \tag{E16}$$

for some $a_i(k)$ between $\widehat{\alpha}_i$ and $\widehat{\alpha}(k)$. Note that, by the conditions of Theorem 1, $(-v_i^{\alpha}(\alpha))$ is uniformly bounded away from zero with probability approaching one. Let $\widehat{\gamma}(k)$ and $\widehat{\nu}(k)$ denote the weighted

means of $\gamma(h_i)$ and ν_i in group $\widehat{k}_i = k$, respectively, where the weight is $(-v_i^{\alpha}(a_i(k)))$. Note that $\widehat{\alpha}(k) = \widehat{\gamma}(k) + \widehat{\nu}(k)$. By (E13), and since γ is Lipschitz-continuous, we have:

$$\frac{1}{N}\sum_{i=1}^{N} \|\gamma(h_i) - \widehat{\gamma}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right). \tag{E17}$$

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

$$\frac{1}{N}\sum_{i=1}^{N} \|\widehat{\nu}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right). \tag{E18}$$

Let $\widehat{g}_i = v_i^{\theta}(\widehat{\alpha}_i)(-v_i^{\alpha}(\widehat{\alpha}_i))^{-1} = \lambda(h_i) + \tau_i$, where $\lambda(h) = \mathbb{E}_{h_i=h}(\widehat{g}_i)$. Similarly to (E18), we have, using analogous notations for weighted group means:

$$\frac{1}{N}\sum_{i=1}^{N} \|\lambda(h_i) - \widehat{\lambda}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right), \quad \frac{1}{N}\sum_{i=1}^{N} \|\widehat{\tau}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right). \tag{E19}$$

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

$$\frac{1}{N}\sum_{i=1}^{N} \|\gamma(h_i) - \widetilde{\gamma}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right), \quad \frac{1}{N}\sum_{i=1}^{N} \|\widetilde{\nu}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right), \tag{E20}$$

$$\frac{1}{N}\sum_{i=1}^{N} \|\lambda(h_i) - \widetilde{\lambda}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right), \quad \frac{1}{N}\sum_{i=1}^{N} \|\widetilde{\tau}(\widehat{k}_i)\|^2 = o_p\left(\frac{1}{T}\right). \tag{E21}$$

Next, using (E16), (E17), and (E18), in addition to the parameter space for α_{i0} being compact and γ being bounded, we have that: $\frac{1}{N}\sum_{i=1}^{N} \|\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\|^3 = -\frac{1}{N}\sum_{i=1}^{N} \|\nu_i\|^3 + o_p(1/T)$. Hence, by Condition (iii) in the corollary:

$$\frac{1}{N}\sum_{i=1}^{N} \|\widehat{\alpha}(\widehat{k}_i) - \widehat{\alpha}_i\|^3 = o_p\left(\frac{1}{T}\right). \tag{E22}$$

To see that (E14) holds, first note that, denoting $a^{\otimes 2} = a \otimes a$:

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

where a_i lies between $\widehat{\alpha}_i$ and $\widehat{\alpha}(\widehat{k}_i)$ and so does $a_i(\widehat{k}_i)$, $v_i^{\theta\alpha}(a_i)$ is a matrix of third derivatives with $(\dim \alpha_{i0})^2$ columns, and we have defined, for all k:

$$\widetilde{\alpha}(k) = \left(\sum_{i=1}^{N} \mathbf{1}\{\widehat{k}_i = k\}(-v_i^{\alpha}(\widehat{\alpha}_i))\right)^{-1} \sum_{i=1}^{N} \mathbf{1}\{\widehat{k}_i = k\}(-v_i^{\alpha}(\widehat{\alpha}_i))\widehat{\alpha}_i, \tag{E23}$$

where we note that $(-v_i^{\alpha}(\widehat{\alpha}_i))$ is uniformly bounded away from zero with probability approaching one.

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

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

where we have used (E16), (E20), and (E21).

Next, we have, using in addition (E22):

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

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

$$\begin{aligned}
A_3 &= \frac{1}{2N}\sum_{i=1}^{N}\widehat{g}_i(-v_i^{\alpha}(\widehat{\alpha}_i))\left(\sum_{i'=1}^{N}\mathbf{1}\{\widehat{k}_{i'}=\widehat{k}_i\}(-v_{i'}^{\alpha}(\widehat{\alpha}_{i'}))\right)^{-1}\\
&\qquad\qquad\times\sum_{i'=1}^{N}\mathbf{1}\{\widehat{k}_{i'}=\widehat{k}_i\}v_{i'}^{\alpha\alpha}\left(a_{i'}(\widehat{k}_{i'})\right)\left(\widehat{\alpha}(\widehat{k}_{i'})-\widehat{\alpha}_{i'}\right)^{\otimes 2}\\
&= \frac{1}{2N}\sum_{i=1}^{N}\widetilde{g}(\widehat{k}_i)(-v_i^{\alpha}(\widehat{\alpha}_i))\left(\sum_{i'=1}^{N}\mathbf{1}\{\widehat{k}_{i'}=\widehat{k}_i\}(-v_{i'}^{\alpha}(\widehat{\alpha}_{i'}))\right)^{-1}\\
&\qquad\qquad\times\sum_{i'=1}^{N}\mathbf{1}\{\widehat{k}_{i'}=\widehat{k}_i\}v_{i'}^{\alpha\alpha}\left(a_{i'}(\widehat{k}_{i'})\right)\left(\widehat{\alpha}(\widehat{k}_{i'})-\widehat{\alpha}_{i'}\right)^{\otimes 2}+o_p\left(\frac{1}{T}\right)\\
&= \frac{1}{2N}\sum_{i=1}^{N}\widetilde{g}(\widehat{k}_i)v_i^{\alpha\alpha}\left(a_i(\widehat{k}_i)\right)\left(\widehat{\alpha}(\widehat{k}_i)-\widehat{\alpha}_i\right)^{\otimes 2}+o_p\left(\frac{1}{T}\right)\\
&= \frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}_{\eta_{i0}}\left(v_i^{\theta}(\alpha_{i0})\right)\left[\mathbb{E}_{\eta_{i0}}(-v_i^{\alpha}(\alpha_{i0}))\right]^{-1}\mathbb{E}_{\eta_{i0}}\left[v_i^{\alpha\alpha}(\alpha_{i0})\right]\nu_i^{\otimes 2}+o_p\left(\frac{1}{T}\right).
\end{aligned}$$

Combining results, we get:

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

This shows (E14), since $\frac{\partial\widehat{\alpha}_i(\theta_0)}{\partial\theta'}=\widehat{g}_i'$, and:

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

Equation (E14) readily delivers an expression for the first-order bias term of the GFE estimator.

Indeed, using that (e.g., Arellano and Hahn, 2007):

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

it follows that Corollary E2 holds, with:

$$C=H^{-1}\operatorname*{plim}_{N,T\to\infty}\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\theta}\Big|_{\theta_0}T\,b_i(\theta),\tag{E24}$$

for:

$$\begin{aligned}b_i(\theta)=&-\frac{1}{2}\left(\widehat{\alpha}_i(\theta)-\mathbb{E}_{h_i}\left(\widehat{\alpha}_i(\theta)\right)\right)'\mathbb{E}_{\eta_{i0}}\left[-v_i^{\alpha}\left(\overline{\alpha}_i(\theta),\theta\right)\right]\left(\widehat{\alpha}_i(\theta)-\mathbb{E}_{h_i}\left(\widehat{\alpha}_i(\theta)\right)\right)\\&+\frac{1}{2}v_i\left(\overline{\alpha}_i(\theta),\theta\right)'\mathbb{E}_{\eta_{i0}}\left[-v_i^{\alpha}\left(\overline{\alpha}_i(\theta),\theta\right)\right]^{-1}v_i\left(\overline{\alpha}_i(\theta),\theta\right).\end{aligned}\tag{E25}$$

Bias in a regression example. Consider the following model for a scalar outcome:

$$Y_{it}=\rho_0Y_{i,t-1}+X_{it}'\beta_0+\alpha_{i0}+U_{it},\tag{E26}$$

where $|\rho_0|<1$. In this case, $\widehat{\alpha}_i(\theta)=(1-\rho)\overline{Y}_i-\overline{X}_i'\beta+o_p\left(T^{-\frac{1}{2}}\right)$. Hence, when classifying individuals based on $h_i=\left(\overline{Y}_i,\overline{X}_i'\right)'$, $\widehat{\alpha}_i(\theta)$ belongs to the span of h_i, up to small-order terms. It follows that C/T is identical to the first-order bias C^{FE}/T of the fixed-effects estimator, and fixed-effects and two-step GFE are first-order equivalent.

However, this equivalence does not hold generally. As an example, let us suppose the unobservables $(\alpha_{i0},\mu_{i0}')'$ follow a one-factor structure with $\mu_{i0}=\lambda\alpha_{i0}$ for a vector λ, and let us classify individual units based on $h_i=\overline{Y}_i$ only. Injectivity is satisfied in this example, due to the low underlying dimensionality of $\eta_{i0}=(\alpha_{i0},\mu_{i0}')'$. In this case it can be verified that:

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

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

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

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

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

so $C\neq C^{\mathrm{FE}}$ in general.

F Verification of assumptions in linear models

In this section we verify the assumptions of Theorem 1 in the linear model:

$$\begin{cases} Y_{it} = \rho_0 Y_{i,t-1} + X_{it}'\beta_0 + \alpha_{it0} + U_{it}, \\ X_{it} = \mu_{it0} + V_{it}, \end{cases} \tag{F1}$$

where Y_{it} is a scalar outcome and X_{it} is a vector of strictly exogenous covariates. We consider two cases: time-invariant α_{i0}, μ_{i0}, and time-varying α_{it0}, μ_{it0}. Throughout, we use the notation $Z_i = (Z_{i1}', ..., Z_{iT}')'$ and $\overline{Z}_i = \frac{1}{T}\sum_{t=1}^T Z_{it}$.

F.1 Time-invariant heterogeneity

Assumption F1. *(regularity in linear model, time-invariant heterogeneity)*

(i) *The parameter space Θ for $\theta_0 = (\rho_0, \beta_0')'$, and the space for (α_{i0}, μ_{i0}'), are compact.*
θ_0 belongs to the interior of Θ, and $|\rho_0| < 1$.

(ii) *$U_{it} \sim \mathcal{N}(0, \sigma_0^2)$, i.i.d. across individuals and over time, independent of $X_i, \alpha_{i0}, \mu_{i0}$ and Y_{i0}.*
$V_{it} \sim \mathcal{N}(0, \Sigma_0^2)$, i.i.d. across individuals and over time, independent of $X_i, \alpha_{i0}, \mu_{i0}$, U_i, and Y_{i0}.
Y_{i0} is drawn from its stationary distribution conditional on α_{i0} and μ_{i0}.

(iii) *Let $W_{it} = (Y_{i,t-1}, X_{it}')'$. $\mathbb{E}(\left(W_{it} - \mathbb{E}_{\alpha_{i0},\mu_{i0}}(W_{it})\right)\left(W_{it} - \mathbb{E}_{\alpha_{i0},\mu_{i0}}(W_{it})\right)')$ is positive definite.*

(iv) *T tends to infinity, and GFE is based on the moment $h_i = (\overline{Y}_i, \overline{X}_i')'$.*

In Assumption F1 we suppose that U_{it} and V_{it} are normal homoskedastic. This can be relaxed, and we could work instead under stationary mixing conditions as in Hahn and Kuersteiner (2011). In that case the likelihood function would be interpreted as a pseudo-likelihood, and the formula for the asymptotic distribution should be adapted.

We are now going to verify the assumptions of Theorem 1, part (b). Assumption 1 holds letting $\xi_{i0} = (\alpha_{i0}, \mu_{i0}')'$, which has compact support. Note that in this model with time-invariant heterogeneity we can abstract from λ_{t0} without loss of generality. Next, $\varphi(\xi_{i0}) = (\frac{\alpha_{i0} + \mu_{i0}'\beta_0}{1-\rho_0}, \mu_{i0}')'$, and it is easy to check that φ is Lipschitz-continuous. Moreover, by part (ii) in Assumption F1 we have $\frac{1}{N}\sum_{i=1}^N \|h_i - \varphi(\xi_{i0})\|^2 = O_p(1/T)$. Lastly, letting $\psi(h_1, h_2) = ((1-\rho_0)h_1 - h_2'\beta_0, h_2')'$, we have $(\alpha_{i0}, \mu_{i0}')' = \psi(\varphi(\xi_{i0}))$. ψ is Lipschitz-continuous by part (i) in Assumption F1. We have thus verified Assumptions 1 and 2, and the conditions of Lemmas 1 and 2.

Let us now verify Assumption 3, with part (iib). In the present case, $p = 1$ and $R = T$. We consider the log-likelihood function:

$$\ell_i(\alpha, \theta) = -\frac{1}{2T}\sum_{t=1}^T \left(Y_{it} - W_{it}'\theta - \alpha\right)^2,$$

where we have set $\sigma_0^2 = 1$ without loss of generality. In addition, we have, for all $\xi = (\alpha, \mu')'$:

$$\overline{\alpha}(\theta, \xi) = \mathbb{E}_{\xi_{i0}=\xi}(\overline{Y}_i - \overline{W}_i'\theta) = \frac{1-\rho}{1-\rho_0}\alpha + \mu'\left(\frac{1-\rho}{1-\rho_0}\beta_0 - \beta\right).$$

Note that $\overline{\alpha}(\theta_0, \xi_{i0}) = \alpha_{i0}$.

It is easy to see that Assumption 3 part (i) is satisfied. Next, by part (i) in Assumption F1 the expected log-likelihood, and the expected log-likelihood derivatives, are bounded. Moreover, $\frac{\partial^2 \ell_i(\alpha,\theta)}{\partial \alpha^2} = -1$, and the third derivatives of ℓ_i are zero. From the assumptions on U_{it}:

$$\frac{1}{N}\sum_{i=1}^{N}[\ell_i(\alpha_{i0}, \theta_0) - \mathbb{E}_{\xi_{i0}}(\ell_i(\alpha_{i0}, \theta_0))]^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}U_{it}^2 - \mathbb{E}_{\xi_{i0}}(U_{it}^2)\right)^2 = O_p(1/T).$$

For all first three derivatives of ℓ_i, we obtain similar $O_p(1/T)$ rates by using the conditions on U_{it}, V_{it} and heterogeneity. This verifies Assumption 3 part (iib).

Next, the expected "target" log-likelihood $\mathbb{E}\left[\ell_i(\overline{\alpha}(\theta, \xi_{i0}), \theta)\right]$ is quadratic in θ, and its partial derivatives with respect to ρ and β are, respectively:

$$\mathbb{E}\left(\left(Y_{i,t-1} - \frac{\alpha_{i0} + \mu_{i0}'\beta_0}{1-\rho_0}\right)\left(Y_{it} - \frac{\alpha_{i0} + \mu_{i0}'\beta_0}{1-\rho_0} - \rho\left(Y_{i,t-1} - \frac{\alpha_{i0} + \mu_{i0}'\beta_0}{1-\rho_0}\right) - (X_{it} - \mu_{i0})'\beta\right)\right),$$

and:

$$\mathbb{E}\left((X_{it} - \mu_{i0})\left(Y_{it} - \frac{\alpha_{i0} + \mu_{i0}'\beta_0}{1-\rho_0} - \rho\left(Y_{i,t-1} - \frac{\alpha_{i0} + \mu_{i0}'\beta_0}{1-\rho_0}\right) - (X_{it} - \mu_{i0})'\beta\right)\right).$$

It is easy to verify that those are zero at $\theta_0 = (\rho_0, \beta_0')'$. Moreover, the second derivative $-H$ is negative definite by Assumption F1 (iii). Lastly, we have:

$$\sup_{\theta}\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\partial^2 \ell_i(\overline{\alpha}(\theta, \xi_{i0}), \theta)}{\partial\theta\partial\alpha}\right\|^2 = \frac{1}{N}\sum_{i=1}^{N}\|\overline{W}_i\|^2 = O_p(1),$$

where we have used the assumptions on U_{it}, V_{it}, and the heterogeneity. This verifies Assumption 3 part (iii).

Turning to Assumption 3 part (iv), we have:

$$\left.\frac{\partial}{\partial\xi'}\right|_{\xi=\widetilde{\xi}}\mathbb{E}_{\xi_{i0}=\xi}\left(\frac{\partial^2 \ell_i(\alpha, \theta_0)}{\partial\theta\partial\alpha}\right) = -\left.\frac{\partial}{\partial\xi'}\right|_{\xi=\widetilde{\xi}}\mathbb{E}_{\xi_{i0}=\xi}\left(\overline{W}_i\right) = -\begin{pmatrix} \frac{1}{1-\rho_0} & \frac{\beta_0}{1-\rho_0} \\ 0 & I \end{pmatrix},$$

where I is the identity matrix of size $\dim \mu_{i0}$. Similarly, we have:

$$\left.\frac{\partial}{\partial\xi'}\right|_{\xi=\widetilde{\xi}}\mathbb{E}_{\xi_{i0}=\xi}\left(\frac{\partial^2 \ell_i(\alpha, \theta_0)}{\partial\alpha^2}\right) = 0,$$

and:

$$\left.\frac{\partial}{\partial\xi'}\right|_{\xi=\widetilde{\xi}}\mathbb{E}_{\xi_{i0}=\xi}\left(\frac{\partial\ell_i(\overline{\alpha}(\theta, \widetilde{\xi}), \theta)}{\partial\alpha}\right) = \left.\frac{\partial}{\partial\xi'}\right|_{\xi=\widetilde{\xi}}\left(\overline{\alpha}(\theta, \xi) - \overline{\alpha}(\theta, \widetilde{\xi})\right) = \left(\frac{1-\rho}{1-\rho_0}, \frac{1-\rho}{1-\rho_0}\beta_0' - \beta'\right),$$

which is bounded by Assumption F1 (i). This verifies Assumption 3 part (iv).

Lastly, turning to Assumption 3 part (v), we have:

$$\mathbb{E}_{h_i=h,\xi_{i0}=\xi}\left(\frac{\partial \ell_i(\overline{\alpha}(\theta,\xi),\theta)}{\partial \alpha}\right) = \mathbb{E}_{\varepsilon_i=h-\varphi(\xi),\xi_{i0}=\xi}\left(\overline{Y}_i - \overline{W}_i'\theta - \overline{\alpha}(\theta,\xi)\right).$$

It is easy to verify that, by Assumption F1 (ii), this is a linear function of $h-\varphi(\xi)$ whose coefficients are uniformly bounded.

Likewise:

$$\mathbb{E}_{h_i=h,\xi_{i0}=\xi}\left(\text{vec}\,\frac{\partial}{\partial \theta}\bigg|_{\theta_0}\frac{\partial \ell_i(\overline{\alpha}(\theta,\xi),\theta)}{\partial \alpha}\right)$$

is also linear in $h-\varphi(\xi)$, with coefficients that are uniformly bounded.

Moreover:

$$\text{Var}_{h_i=h,\xi_{i0}=\xi}\left(\frac{\partial \ell_i(\overline{\alpha}(\theta,\xi),\theta)}{\partial \alpha}\right) = \text{Var}_{h_i=h,\xi_{i0}=\xi}\left(\overline{Y}_i - \overline{W}_i'\theta\right)$$

is non-stochastic by Assumption F1 (ii), and it can be shown that is satisfies:

$$\text{Var}_{h_i=h,\xi_{i0}=\xi}\left(\overline{Y}_i - \overline{W}_i'\theta\right) = O_p\left(\frac{1}{T}\right),$$

uniformly in h, ξ, and θ.

Finally, we have:

$$\text{Var}_{h_i=h,\xi_{i0}=\xi}\left(\text{vec}\,\frac{\partial}{\partial \theta}\bigg|_{\theta_0}\frac{\partial \ell_i(\overline{\alpha}(\theta,\xi),\theta)}{\partial \alpha}\right) = \text{Var}_{h_i=h,\xi_{i0}=\xi}\left(\overline{W}_i\right).$$

Now, $\text{Var}_{h_i=h,\xi_{i0}=\xi}\left(\overline{X}_i\right)=0$. Moreover:

$$\text{Var}_{h_i=h,\xi_{i0}=\xi}\left(\overline{Y}_{i,-1}\right) = O_p\left(\frac{1}{T}\right),$$

uniformly in h and ξ, where $\overline{Y}_{i,-1}=\frac{1}{T}\sum_{t=1}^{T}Y_{i,t-1}$.

This verifies Assumption 3 part (v).

Applying Theorem 1, we obtain:

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

where $H=\mathbb{E}\left(\left(W_{it}-\mathbb{E}_{\xi_{i0}}(W_{it})\right)\left(W_{it}-\mathbb{E}_{\xi_{i0}}(W_{it})\right)'\right)$, and $s_i=\frac{1}{T}\sum_{t=1}^{T}\left(W_{it}-\mathbb{E}_{\xi_{i0}}(W_{it})\right)U_{it}$.

F.2 Time-varying heterogeneity

We will verify the conditions in the time-varying case under the following assumptions, where for simplicity we focus on a scalar covariate X_{it}. The two unobservables have a factor structure, with a single factor in each equation.

Assumption F2. *(regularity in linear model, time-varying heterogeneity)*

(i) $\alpha_{it0} = \alpha_{i0}\lambda^{\alpha}_{t0}$ *and* $\mu_{it0} = \mu_{i0}\lambda^{\mu}_{t0}$*, where* α_{i0}*,* μ_{i0}*,* λ^{α}_{t0}*, and* λ^{μ}_{t0} *are scalar.*

(ii) λ^{α}_{t0} *and* λ^{μ}_{t0} *are stationary with bounded supports and non-zero means, and are independent of* U_i *and* V_i*. Moreover,* $\|\frac{1}{T}\sum_{t=1}^{T}\lambda^{\alpha}_{t0} - \mathbb{E}(\lambda^{\alpha}_{t0})\|^2 = O_p(1/T)$*, and* $\|\frac{1}{T}\sum_{t=1}^{T}\lambda^{\mu}_{t0} - \mathbb{E}(\lambda^{\mu}_{t0})\|^2 = O_p(1/T)$*.*

(iii) Assumption F1 holds, with part (iii) replaced by $\mathbb{E}(\left(W_{it} - \mathbb{E}_{\alpha_{i0},\mu_{i0},\lambda_0}(W_{it})\right)\left(W_{it} - \mathbb{E}_{\alpha_{i0},\mu_{i0},\lambda_0}(W_{it})\right)')$ *being positive definite.*

Assumption 1 holds by part (i) in Assumption F2. In addition, we have:

$$\varphi(\xi_{i0}) = \operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix} \alpha_{i0}\lambda^{\alpha}_{t0} + \mu_{i0}\lambda^{\mu}_{t0}\beta_0 + \sum_{s=1}^{+\infty}\rho_0^s\left[\alpha_{i0}\lambda^{\alpha}_{t-s,0} + \mu_{i0}\lambda^{\mu}_{t-s,0}\beta_0\right] \\ \mu_{i0}\lambda^{\mu}_{t0}\end{pmatrix},$$

where, as in the time-invariant case, we define $\xi_{i0} = (\alpha_{i0}, \mu_{i0})'$. Hence, by stationarity of the factors:

$$\varphi(\xi_{i0}) = \begin{pmatrix} \frac{\alpha_{i0}\mathbb{E}[\lambda^{\alpha}_{t0}] + \mu_{i0}\mathbb{E}[\lambda^{\mu}_{t0}]\beta_0}{1-\rho_0} \\ \mu_{i0}\mathbb{E}[\lambda^{\mu}_{t0}]\end{pmatrix}.$$

It follows that φ is Lipschitz-continuous. Injectivity follows from $\mathbb{E}[\lambda^{\alpha}_{t0}] \neq 0$ and $\mathbb{E}[\lambda^{\mu}_{t0}] \neq 0$.

Let $\nu^{\alpha}_{t0} = \lambda^{\alpha}_{t0} - \mathbb{E}[\lambda^{\alpha}_{t0}]$, and $\nu^{\mu}_{t0} = \lambda^{\mu}_{t0} - \mathbb{E}[\lambda^{\mu}_{t0}]$. We have:

$$\begin{aligned} h_i - \varphi(\xi_{i0}) =& \frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix} U_{it} + V_{it}\beta_0 + \sum_{s=1}^{+\infty}\rho_0^s\left[U_{i,t-s} + V_{i,t-s}\beta_0\right] \\ V_{it}\end{pmatrix} \\ &+ \frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix} \alpha_{i0}\nu^{\alpha}_{t0} + \mu_{i0}\nu^{\mu}_{t0}\beta_0 + \sum_{s=1}^{+\infty}\rho_0^s\left[\alpha_{i0}\nu^{\alpha}_{t-s,0} + \mu_{i0}\nu^{\mu}_{t-s,0}\beta_0\right] \\ \mu_{i0}\nu^{\mu}_{t0}\end{pmatrix}.\end{aligned}$$

Using Assumption F1 and part (ii) in Assumption F2, we verify that $\frac{1}{N}\sum_{i=1}^{N}\|h_i - \varphi(\xi_{i0})\|^2 = O_p(1/T)$.

Let us now verify Assumption 3, with part (iib). In the present case, $p = T$ and $R = 1$. We consider the log-likelihood function:

$$\ell_{it}(\alpha, \theta) = -\frac{1}{2}\left(Y_{it} - W_{it}'\theta - \alpha\right)^2,$$

where we have again set $\sigma_0^2 = 1$. We have, for all $\xi = (\alpha, \mu)'$ and t:

$$\begin{aligned}&\overline{\alpha}^t(\theta, \xi) = \mathbb{E}_{\xi_{i0}=\xi,\lambda^{\alpha}_0,\lambda^{\mu}_0}(Y_{it} - W_{it}'\theta) \\ &= \alpha\lambda^{\alpha}_{t0} + (\beta_0 - \beta)\mu\lambda^{\mu}_{t0} + (\rho_0 - \rho)\alpha\sum_{s=1}^{\infty}\rho_0^{s-1}\lambda^{\alpha}_{t-s,0} + (\rho_0 - \rho)\beta_0\mu\sum_{s=1}^{\infty}\rho_0^{s-1}\lambda^{\mu}_{t-s,0}.\end{aligned}$$

Note that $\overline{\alpha}^t(\theta_0, \xi_{i0}) = \alpha_{i0}\lambda^{\alpha}_{t0} = \alpha_{it0}$.

It is easy to see that Assumption 3 part (i) is satisfied. To check Assumption 3 part (iib), note that the expected log-likelihood is uniformly bounded by Assumption F1 and part (ii) in Assumption

F2, and similarly for its first three derivatives. As in the time-invariant case, $\frac{\partial^2 \ell_{it}(\alpha,\theta)}{\partial \alpha^2} = -1$, and the third derivatives of ℓ_{it} are zero. Moreover:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}[\ell_{it}(\alpha_{it0},\theta_0) - \mathbb{E}_{\xi_{i0},\lambda_0}(\ell_{it}(\alpha_{it0},\theta_0))]^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(U_{it}^2 - \mathbb{E}_{\alpha_{i0},\mu_{i0},\lambda_0^\alpha,\lambda_0^\mu}(U_{it}^2)\right)^2 = O_p(1),$$

and all the first three derivatives of ℓ_{it} are similarly $O_p(1)$. This verifies Assumption 3 part (iib).

Turning to part (iii), the function:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\ell_{it}(\overline{\alpha}^t(\theta,\xi_{i0}),\theta)\right]$$

is quadratic in θ, and its partial derivatives with respect to ρ and β are, respectively:

$$\mathbb{E}\left(\left(Y_{i,t-1} - \alpha_{i0}\sum_{s=1}^{+\infty}\rho_0^{s-1}\lambda_{t-s,0}^{\alpha} - \beta_0\mu_{i0}\sum_{s=1}^{+\infty}\rho_0^{s-1}\lambda_{t-s,0}^{\mu}\right)\left(Y_{it} - \rho Y_{i,t-1} - \beta X_{it} - \overline{\alpha}^t(\theta,\xi_{i0})\right)\right),$$

and:

$$\mathbb{E}\left(\left(X_{it} - \mu_{i0}\lambda_{t0}^{\mu}\right)\left(Y_{it} - \rho Y_{i,t-1} - \beta X_{it} - \overline{\alpha}^t(\theta,\xi_{i0})\right)\right).$$

These are zero at θ_0. Moreover, the second derivative $-H$ is negative definite by Assumption F2 (iii). Lastly, we have:

$$\sup_{\theta}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\frac{\partial^2 \ell_{it}(\overline{\alpha}^t(\theta,\xi_{i0}),\theta)}{\partial\theta\partial\alpha}\right\|^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\|W_{it}\|^2 = O_p(1).$$

This verifies Assumption 3 part (iii).

Let us now check Assumption 3 part (iv). We have:

$$\begin{aligned}
&\left.\frac{\partial}{\partial\xi'}\right|_{\xi=\widetilde{\xi}}\mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}\left(\frac{\partial^2\ell_{it}(\alpha,\theta_0)}{\partial\theta\partial\alpha}\right) = -\left.\frac{\partial}{\partial\xi'}\right|_{\xi=\widetilde{\xi}}\mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}\left(W_{it}\right)\\
&= -\begin{pmatrix} \sum_{s=1}^{\infty}\rho_0^{s-1}\lambda_{t-s}^{\alpha} & \beta_0\sum_{s=1}^{\infty}\rho_0^{s-1}\lambda_{t-s}^{\mu} \\ 0 & \lambda_t^{\mu}\end{pmatrix},
\end{aligned}$$

which is uniformly bounded using Assumption F1 and part (ii) in Assumption F2. The second condition in Assumption 3 part (iv) is immediate to verify. Lastly, we have:

$$\begin{aligned}
&\left.\frac{\partial}{\partial\xi'}\right|_{\xi=\widetilde{\xi}}\mathbb{E}_{\xi_{i0}=\xi,\lambda_0=\lambda}\left(\frac{\partial\ell_{it}(\overline{\alpha}^t(\theta,\widetilde{\xi}),\theta)}{\partial\alpha}\right) = \left.\frac{\partial}{\partial\xi'}\right|_{\xi=\widetilde{\xi}}\left(\overline{\alpha}^t(\theta,\xi) - \overline{\alpha}^t(\theta,\widetilde{\xi})\right)\\
&= \left(\lambda_t^{\alpha} + (\rho_0-\rho)\sum_{s=1}^{+\infty}\rho_0^{s-1}\lambda_{t-s}^{\alpha},\ (\beta_0-\beta)\lambda_t^{\mu} + (\rho_0-\rho)\beta_0\sum_{s=1}^{+\infty}\rho_0^{s-1}\lambda_{t-s}^{\mu}\right)
\end{aligned}$$

which is also uniformly bounded. This verifies Assumption 3 part (iv).

Finally, let us verify Assumption 3 part (v). We have:

$$\mathbb{E}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}\left(\frac{\partial \ell_{it}(\overline{\alpha}^t(\theta,\xi_{i0}),\theta)}{\partial \alpha}\right) = \mathbb{E}_{\varepsilon_i=h-\varphi(\xi),\xi_{i0}=\xi,\lambda_0=\lambda}\left(Y_{it} - W_{it}\theta - \overline{\alpha}^t(\theta,\xi_{i0})\right).$$

This is a linear function of $h-\varphi(\xi)$ whose coefficients are uniformly bounded.

Likewise:

$$\mathbb{E}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}\left(\left.\frac{\partial}{\partial \theta}\right|_{\theta_0}\frac{\partial \ell_{it}(\overline{\alpha}^t(\theta,\xi_{i0}),\theta)}{\partial \alpha}\right)$$

is also linear in $h-\varphi(\xi)$, with coefficients that are uniformly bounded.

Moreover:

$$\mathrm{Var}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}\left(\frac{\partial \ell_{it}(\overline{\alpha}^t(\theta,\xi_{i0}),\theta)}{\partial \alpha}\right) = \mathrm{Var}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}\left(Y_{it} - W_{it}'\theta\right)$$

is non-stochastic by Assumption F1 (ii), and it is easy to see that it is bounded uniformly in h, ξ, λ, t, and θ.

Similarly:

$$\mathrm{Var}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}\left(\left.\frac{\partial}{\partial \theta}\right|_{\theta_0}\frac{\partial \ell_{it}(\overline{\alpha}^t(\theta,\xi),\theta)}{\partial \alpha}\right) = \mathrm{Var}_{h_i=h,\xi_{i0}=\xi,\lambda_0=\lambda}\left(W_{it}\right)$$

is uniformly bounded.

This verifies Assumption 3 part (v).

Applying Theorem 1, we obtain:

$$\widehat{\theta} \;=\; \theta_0 + \widetilde{H}^{-1}\frac{1}{N}\sum_{i=1}^{N}\widetilde{s}_i + O_p\left(\frac{1}{T}\right) + O_p\left(\frac{K}{N}\right) + O_p\left(K^{-\frac{2}{\dim X_{it}+1}}\right) + o_p\left(\frac{1}{\sqrt{NT}}\right),$$

where $\widetilde{H} = \mathbb{E}\left(\left(W_{it} - \mathbb{E}_{\xi_{i0},\lambda_0}(W_{it})\right)\left(W_{it} - \mathbb{E}_{\xi_{i0},\lambda_0}(W_{it})\right)'\right)$, and $\widetilde{s}_i = \frac{1}{T}\sum_{t=1}^{T}\left(W_{it} - \mathbb{E}_{\xi_{i0},\lambda_0}(W_{it})\right)U_{it}$.

G Additional simulation results for the probit models

In this section we present additional results on the simulations of probit models. We start by describing the DGP in detail. We then show results about inference and estimation of K. Finally, we present two exercises related to computational aspects of GFE.

G.1 Monte Carlo designs and estimators

We consider the following static probit model:

$$\left\{\begin{array}{rcl} Y_{it} &=& \mathbf{1}\left\{X_{it}'\theta_{it0} + \alpha_{it0} + U_{it} > 0\right\}, \\ X_{it} &=& \mu_{it0} + V_{it}, \end{array}\right. \tag{G1}$$

where U_{it} are i.i.d. standard normal, independent of $X_i, \alpha_{i0}, \mu_{i0}$, and V_{it} are i.i.d. normal with zero mean and identity covariance matrix, independent of all U_{it}'s, X_i, α_{i0}, and μ_{i0}.

We consider three different specifications for θ_{it0}, α_{it0}, and μ_{it0}. In the *fixed-effects probit* specification, all components of θ_{it0} are constant, equal to 1, and $\mu_{it0} = \mu_{i0}$ are normal with zero mean and identity covariance matrix. In DGP 2 to 4, we set $\alpha_{i0} = \frac{1}{\sqrt{r}} \sum_{r'=1}^{r} \mu_{i0}^{(r')} + \zeta_i$, where ζ_i are standard normal independent of μ_{i0}, and $\mu_{i0}^{(r')}$, $r' = 1, ..., r$, denote the components of μ_{i0}. In DGP 1 we set $\alpha_{i0} = \mu_{i0}$.

In the *time-varying* specification, all components of θ_{it0} are constant, equal to 1, $\alpha_{it0} = \alpha_{i0} f_t$ and $\mu_{it0} = \mu_{i0} f_t$, where α_{i0} and μ_{i0} are as in the fixed-effects specification, and $f_t = -t$ is a time trend. Note that, given the one-dimensional factor structure, the underlying dimension of α_{it0} is still one in this case.

Finally, in the *random coefficients* specification, α_{it0} and μ_{it0} are modeled as for fixed-effects probit. In addition, the first component of θ_{it0} is constant, equal to 1, and all other components are identical, equal to $\theta_{it0}^{(r')} = \theta_{i0}^{(r')} = 1 + \frac{\alpha_{i0}}{2}$. While the number of heterogeneous parameters is increasing relative to the fixed-effects specification, the underlying dimension of $(\alpha_{it0}, \theta_{it0}')$ is equal to one.

We estimate the fixed-effects probit specification using GFE methods based on the moment vectors $h_i = (\overline{Y}_i, \overline{X}_i')'$, running a probit on group indicators in the second step. When estimating the time-varying specification we take $h_i = (\overline{Y}_i, \overline{X}_i')'$ as moments, and interact the group indicators with time dummies in the second step. Lastly, we estimate the random coefficients specification using GFE based on $h_i = (\overline{Y}_i, \overline{X}_i', \overline{YX}_i')'$, where in the second step we interact the group indicators with all covariates except the first one. We will compare GFE with fixed-effects estimators in two of the three specifications: fixed-effects probit and random coefficients. However, fixed-effects is not feasible in the time-varying specification. For every specification, we will vary the dimension of heterogeneity by varying the number of covariates.

In all simulations we take $N = 1000$ and $T = 20$, and report estimates of the first component of θ_{it0}, which is always constant and equal to 1 irrespective of the design. We perform 500 simulations, and estimate kmeans using the matlab implementation, which augments Lloyd's updates with a local search rountine, using 100 randomly generated starting values. For iterated GFE, we start the iteration at the two-step estimate and iterate 10 times. To implement conditional kmeans we initialize Algorithm 2 at two-step estimates. In the cubic specification we include all linear, quadratic, and cubic powers of the elements of X_{it}, as well as a set of interactions, for $(\dim X_{it} + 1)^2$ coefficients in total. In the neural network specification we use a single hidden layer feedforward model with three nodes, with a sigmoid input link and a linear output link. We train the neural network using matlab's "trainscg". Note that, in this case the algorithm is not guaranteed to decrease the objective function in each step. We stop the algorithm after 10 iterations.

G.2 Inference

In Table S1 we report the analytical coverage for two-step GFE, iterated GFE, and fixed-effects, both uncorrected and using half-panel jackknife bias correction. To compute confidence intervals we use two methods: an analytical method based on $\widehat{H}$ in (17), and a method based on the bootstrap clustered at the individual level.

Consider first the fixed-effects probit model in Panel A. The results show that coverage accuracy for two-step GFE strongly depends on the dimension of heterogeneity. In DGP 1, where the dimension is one, both analytical and bootstrap methods give close to nominal coverage when using bias correction. This is true both when $K = 10$ and when we estimate K using our rule (DGP 1*). This finding is in line with Corollary 4. However, in the other DGP the dimension of heterogeneity is larger and two-step GFE is substantially biased, as shown by Table 1. As a result, coverage probabilities are heavily distorted. The intermediate case of DGP 2 is interesting, since under-coverage is somewhat less severe when using our rule to estimate K, especially when using the bootstrap to construct confidence intervals.

In contrast, the iterated GFE estimator, while showing some coverage distortions as the dimension of X grows (even after bias correction), outperforms two-step GFE, and in fact it tends to outperform bias-corrected fixed-effects in terms of coverage when using the bootstrap. According to our simulations, bootstrap inference is more reliable than inference based on the analytical method. It is interesting to see that bootstrap inference based on iterated GFE with bias correction has relatively small finite sample distortions, even in DGP 3 and 4 where the dimension of heterogeneity is such that Corollary 4 does not apply.

Consider next the model with time-varying unobservables in Panel B of Table S1. In DGP 1, bias-corrected two-step GFE tends to undercover. Note that in this case the conditions of Corollary 4 are not met, since there are $p = T$ time-varying parameters, and the DGP features a non-stationary time trend. In DGP 2 to 4 the dimension of heterogeneity is larger, and biases are higher, so the two-step method undercovers heavily. An exception is DGP 2 when we estimate K (DGP 2*), in which case bias-reduced two-step GFE is not substantially biased, and the bootstrap coverage probability is good. In the various DGP, iterated GFE gives less distorted coverage probabilities. Indeed, while the analytical inference method undercovers, bias-reduced iterated GFE tends to be conservative when based on the bootstrap. Lastly, in this time-varying specification, fixed-effects is no longer consistent, hence we do not report inference in this case.

Finally, consider the random coefficients probit model in Panel C. In both DGP 3 and 4, where the dimension of heterogeneity is relatively large, coverage based on two-step GFE is heavily distorted. In such DGP, Corollary 4 is not applicable, since (16) is not met. In contrast, although iterated GFE undercovers substantially when using analytical confidence intervals, the bootstrap provides more accurate coverage. Note that for DGP 4 the bias-corrected fixed-effects estimator undercovers

less than in DGP 3, but in both cases the estimator is substantially biased and the higher coverage probability in DGP 4 only comes from the very large standard deviation of the estimator (see Table 2 in the main text).

G.3 Additional results with estimated K

We next present results on the performance of two-step estimators when we estimate K. When the dimension of heterogeneity is small, such as in DGP 1 and 2, we use (4) with $\gamma = 1$ to set K. We have used this method in DGP 1^* and 2^* in Tables 1, 2, and S1. In DGP with a larger dimension of heterogeneity, such as DGP 3 and 4, conditional methods perform substantially better than two-step GFE, and we use a different rule based on a conditional first step to set K, as described in Appendix B.2. Both rules are based on the first step alone.

In Table S2 we show the results for DGP 1 and 2, using the unconditional rule (4). The results illustrate how the choice of K adapts to the dimension of heterogeneity. A larger dimension d increases approximation error, hence leads to a larger value of K. We also report the estimated K in half-panels, since we compute them as part of the bias correction strategy. In half-panels the estimates of K decrease, since the noise level increases. Notice that, in the time-varying specification, K estimates differ for the two half-panels. This is probably due to the lack of stationarity of the DGP in this case. We commented on mean estimates when discussing the results of Tables 1 and 2 in the main text. In addition, standard deviations do not seem to increase when estimating K.

We turn now to the performance of GFE methods when the choice of K is based on a conditional first step, as described in Appendix B.2. We first compute the kmeans estimator with $K_{\max} = 30$. We use this unconditional classification to initialize a linear conditional kmeans algorithm that uses first-order polynomials as a basis, also with $K_{\max} = 30$. In the case of the time-varying probit model, we add a linear trend to the set of covariates. This step provides the value of the noise level. We then estimate the conditional kmeans estimator, and we compare the noise level to the value of the objective, for different values of K. The selected K is then the smallest value such that the noise level is larger than the conditional kmeans objective. This choice of K balances the time-series noise versus the approximation error solely due to α_{i0}.

We show the results in Tables S3 and S4. Notice that the estimated K is now lower than when using our unconditional rule (see Table S2). As an example, in the fixed-effects probit model in DGP 2, K is around 6 in the conditional case, whereas K is around 20 in the unconditional case. In addition, the estimated K values remain stable when the number of covariates increases. In fact, we see a mild decrease in K, which might be due to the nonparametric approximation to the conditional expectation of Y given X: since we keep the number of terms in the series fixed, the nonparametric error increases with the number of covariates, hence driving the noise level up and giving a lower K value.

Tables S3 and S4 show that the performance of GFE methods is quite similar to the case with

$K = 10$. The case of DGP 2 in Table S3 is interesting, since when using the conditional rule to set K two-step GFE is severely biased in this case. This is due to the fact that the estimated number of groups ($K \approx 5 - 6$) is small relative to the one given by our unconditional rule in Table S2 ($K \approx 20$). Indeed, the conditional rule targets a uni-dimensional approximation error in this case, but the bias of two-step GFE is affected by a bi-dimensional approximation error. In contrast, we see that GFE methods with a conditional first step – for which the conditional rule is designed – tend to perform very well.

G.4 Computational aspects

Finding the global optimum in kmeans is challenging, and most available algorithms only provide approximate solutions. In Table S5 we study the impact of using different numbers of random starting values in the kmeans heuristic on the statistical performance of GFE estimators. We focus on a fixed-effects probit specification with $N = 1000$ and $T = 20$, and use $K = 10$ groups. Let n be a number of starting values. In each of the 500 simulated samples, we compute 100 kmeans partitions starting at n randomly generated values, and the resulting two-step GFE estimators. We then report the total variance across samples and kmeans runs, as well as the between-sample and within-sample variances. The share of within to total variance measures the contribution of numerical error to overall variability of the estimator.

Unsurprisingly, as the number of starting values increases, the share of variance due to numerical error decreases. Interestingly, it is quite small as soon as we use more than 100 starting values. With 100 values, the numerical variability of kmeans increases the variance of two-step GFE by 1.9% in DGP 2, and by 12.3% in DGP 4. With 1000 starting values, the increase is 0.3% and 7.5%., respectively. Although increasing the number of starting values adds to the overall computational burden, off-the-shelf parallelization is available in various kmeans routines.

Finally, in Table S6 we compare the computational cost of fixed-effects, two-step GFE, and iterated GFE estimators, for different cross-sectional sizes, from $N = 100$ to $N = 10000$, in a random coefficients probit model with three covariates (DGP 3). We separately record the time used by the kmeans algorithm in the total computation of GFE. We use 100 starting values to compute the kmeans partition in the first step. When the sample size is small, fixed-effects is faster than GFE, due to the overhead cost of starting the computation in kmeans. However, as N increases to moderate sizes, such as $N = 1000$, GFE becomes faster than fixed-effects. The reason is that kmeans computation only increases moderately with the sample size, and second-step computation involves much fewer parameters to estimate than fixed-effects. Iterating adds some computational cost, and fixed-effects is faster than iterated GFE (where we iterate 10 times) up to $N = 1000$. However, for $N = 5000$ and $N = 10000$, iterated GFE is faster than fixed-effects in this DGP.

H Complements on the dynamic model of location choice

In this section of the appendix we provide details on our illustration to the dynamic structural model of migration, and we report additional results.

H.1 Details on computation and estimation

Value functions. Let us denote the integrated value function as:

$$\overline{V}_t(S_{i,t-1}) = \mathbb{E}\left[\max_{j\in\{1,...,J\}} V_t(j, S_{i,t-1}) + \xi_{it}(j) \,\middle|\, S_{i,t-1}\right].$$

The alternative-specific value functions are:

$$V_t(j, S_{i,t-1}) = \mathbb{E}\left[\rho W_{it}(j) - c_i(j_{i,t-1})\mathbf{1}\{j \neq j_{i,t-1}\} + \beta\overline{V}_t(S_{it}) \,\middle|\, j_{it} = j, S_{i,t-1}\right],$$

where $S_{it} = \left(j, \mathcal{J}^j_{i,t-1}, \alpha_i\left(\mathcal{J}^j_{i,t-1}\right), c_i\left(\mathcal{J}^j_{i,t-1}\right)\right)$ when $j_{it} = j$, for $\mathcal{J}^j_{i,t-1} = \mathcal{J}_{i,t-1} \cup \{j\}$. From the functional forms we obtain (as in Rust, 1994):

$$\overline{V}_t(S_{i,t-1}) = \ln\left(\sum_{j=1}^{J} \exp V_t(j, S_{i,t-1})\right) + \gamma, \tag{H1}$$

where $\gamma \approx .57$ is Euler's constant. Moreover:

$$V_t(j, S_{i,t-1}) = \\ \mathbb{E}\left[\rho \exp\left(\alpha_i(j) + \frac{\sigma^2}{2}\right) - c_i(j_{i,t-1})\mathbf{1}\{j \neq j_{i,t-1}\} + \beta\overline{V}_t\left(j, \mathcal{J}^j_{i,t-1}, \alpha_i\left(\mathcal{J}^j_{i,t-1}\right), c_i\left(\mathcal{J}^j_{i,t-1}\right)\right) \,\middle|\, j_{it} = j, S_{i,t-1}\right], \tag{H2}$$

where the expectation is taken with respect to the belief distribution, which is the conditional distribution of $\alpha_i(j)$ given $\alpha_i\left(\mathcal{J}_{i,t-1}\right)$ and $c_i\left(\mathcal{J}_{i,t-1}\right)$, conditional on $j_{i,t-1}$ and $j_{it} = j$.

Computation. To compute the solution, we first calculate the full-information value functions and then proceed using backward induction. In the case where all locations have been visited, $\mathcal{J}_{it} = \{1, ..., J\}$ so $S_{it} = (j_{it}, \{1, ..., J\}, \{\alpha_i(1), ..., \alpha_i(J)\}, \{c_i(1), ..., c_i(J)\})$. Denote the corresponding integrated value function given the most recent location j as $\overline{V}^J(i, j)$. From (H1) and (H2) we have:

$$\overline{V}^J(i, j) = \ln\left(\sum_{j'=1}^{J} \exp\left[\rho \exp\left(\alpha_i(j') + \frac{\sigma^2}{2}\right) - c_i(j)\mathbf{1}\{j' \neq j\} + \beta\overline{V}^J(i, j')\right]\right) + \gamma, \quad j = 1, ..., J.$$

We solve this fixed-point system by successive iterations.

Consider now a case where the agent has visited s states in set $\mathcal{J} \subsetneq \{1, ..., J\}$, and is currently at location j. Let $\overline{V}^s(i, j, \mathcal{J})$ denote her integrated value function. The latter solves:

$$\begin{aligned}\overline{V}^s(i,j,\mathcal{J}) &= \ln\Bigg(\sum_{j'\notin\mathcal{J}}\exp\left[\mathbb{E}_{\mathcal{J},j,j'}\left(\rho\exp\left(\alpha_i(j')+\frac{\sigma^2}{2}\right)-c_i(j)+\beta\overline{V}^{s+1}(i,j',\mathcal{J}^{j'})\right)\right] \\ &\quad+\sum_{j'\in\mathcal{J}}\exp\left[\rho\exp\left(\alpha_i(j')+\frac{\sigma^2}{2}\right)-c_i(j)\mathbf{1}\{j'\neq j\}+\beta\overline{V}^{s}(i,j',\mathcal{J})\right]\Bigg)+\gamma,\end{aligned}$$

where $\mathbb{E}_{\mathcal{J},j,j'}$ is taken with respect to the distribution of $\alpha_i(j')$ given $\alpha_i(\mathcal{J})$ and $c_i(\mathcal{J})$, conditional on moving from j to j'. In practice we discretize the values of each $\alpha_i(j)$ on a 50-point grid. In the model with heterogeneous costs, this also implies a discretization of mobility costs given our modeling of costs and returns (see below). In the computation of the fixed points we set a 10^{-11} numerical tolerance.

Estimation. The choice probabilities entering the likelihood are given by an estimated counterpart to (26), where the estimated value functions $\widehat{V}_t\left(j, j_{i,t-1}, \mathcal{J}_{i,t-1}, \widehat{\alpha}(\widehat{k}_i, \mathcal{J}_{i,t-1}), c(\widehat{k}_i, \mathcal{J}_{i,t-1}), \theta\right)$ solve the system (H1)-(H2). Notice that, in the model with homogeneous costs, c does not depend on the group or the history of locations. We estimate the conditional expectation in (H2) as a conditional mean given $\widehat{\alpha}(\widehat{k}_i, \mathcal{J}_{i,t-1})$ and $c(\widehat{k}_i, \mathcal{J}_{i,t-1})$, based on all job movers from $\mathcal{J}_{i,t-1}$ to $j_{it} = j$. Nonparametric or semi-parametric methods could be used for this purpose. We use an exponential regression estimator in the illustration, with degree one to compute expected returns, and degree two for expected value functions. We checked these specifications provided a good fit to the conditional expectations.

Iteration. To perform the iteration, we first estimate the idiosyncratic variance of log-wages σ^2 as:

$$\widehat{\sigma}^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\ln W_{it} - \widehat{\alpha}(\widehat{k}_i, j_{it})\right)^2. \tag{H3}$$

Then, individual groups are assigned as:

$$\begin{aligned}\widehat{k}_i^{(2)} &= \underset{k\in\{1,...,K\}}{\operatorname{argmax}}\sum_{t=1}^{T}\sum_{j=1}^{J}\mathbf{1}\{j_{it}=j\}\Bigg(\ln\Pr\left(j_{it}=j\,|\,j_{i,t-1},\mathcal{J}_{i,t-1},\widehat{\alpha}(k,\mathcal{J}_{i,t-1}),\widehat{c}(k,\mathcal{J}_{i,t-1}),\widehat{\theta}\right) \\ &\quad+\ln\phi(\ln W_{it};\widehat{\alpha}(k,j),\widehat{\sigma}^2)\Bigg),\end{aligned}$$

where ϕ denotes the normal density. Note that information on both wages and choices is used to reclassify individuals.

Given group assignments, we update parameters as:

$$\widehat{\alpha}^{(2)}(k,j) = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{1}\{\widehat{k}_i^{(2)}=k\}\mathbf{1}\{j_{it}=j\}\ln W_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{1}\{\widehat{k}_i^{(2)}=k\}\mathbf{1}\{j_{it}=j\}},$$

with an update for σ^2 analogous to (H3), and:

$$(\widehat{\theta}^{(2)}, \widehat{c}^{(2)}) = \underset{(\theta,c)}{\text{argmax}} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{j=1}^{J} \mathbf{1}\{j_{it} = j\} \ln \Pr\left(j_{it} = j \,|\, j_{i,t-1}, \mathcal{J}_{i,t-1}, \widehat{\alpha}^{(2)}(\widehat{k}_i^{(2)}, \mathcal{J}_{i,t-1}), c(\widehat{k}_i^{(2)}, \mathcal{J}_{i,t-1}), \theta\right).$$

This procedure may be iterated further. Note that in the update step we do not maximize the full likelihood as a function of parameters α, c, σ^2, θ. Rather, we use a partial likelihood estimator by which we first estimate wage parameters α and σ^2, and then estimate utility and cost parameters θ and c. We use this approach for computational reasons; see Rust (1994) and Arcidiacono and Jones (2003) for related approaches.

Specification of cost heterogeneity. In the DGP with heterogeneous costs, we specify $c_i(j) = \exp(c_0 + c_1 \alpha_i(j))$, where we estimate c_0 and c_1 using an exponential regression of the estimates $\widehat{c}(\widehat{k}_i, j)$ on $\widehat{\alpha}(\widehat{k}_i, j)$ and a constant, across both groups and regions. We find $\widehat{c}_0 = 0.85$ and $\widehat{c}_1 = 0.32$. A positive relationship between returns and mobility costs could reflect a link between costs and investments. In the Monte Carlo, GFE recovers c_0 quite well, although c_1 estimates are somewhat biased downwards.

H.2 Additional results

In this subsection we show additional estimation results for the illustration in Section 5.

Parameters a and b. In Figure S1 we report analogous Monte Carlo results as in Figures 4 and 5, for the parameters a and b that govern the probability of being a "mover type". In particular, we see that two-step GFE and its bias-reduced and iterated versions perform quite well.

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

Fixed-K GFE. Until the end of this subsection we focus on the model with homogeneous costs. We start by reporting results based on fixed values of K, from $K = 2$ to $K = 8$, in Figure S2. We see that taking $K = 2$ yields imprecise estimates, in particular for ρ. In comparison, taking $K = 4$, $K = 6$ or $K = 8$ results in better performance. The most accurate results are obtained taking $K = 6$ or $K = 8$ and using bias reduction and one or three iterations. Those results are close to the ones using our method to select K (see Figure 4, where the average value for $\widehat{K}$ is 6.4).

Complements: fixed-effects estimates. In the DGP with homogeneous costs, fixed-effects estimation is computationally tractable. This is due to the fact that the α's and the structural parameters can be estimated sequentially. One fixed-effects estimation of the structural parameters is about 2.5 times slower than one estimation of the model with $6-7$ groups (an average value of $\widehat{K}$), although it becomes 5 times slower in a sample with 5 times as many individuals.

Complements: random-effects estimates. To compute random-effects estimators based a finite mixture with $K=2$, $K=4$, and $K=8$ types, respectively, we use the EM algorithm of Arcidiacono and Jones (2003), where wage-specific parameters and structural parameters are estimated sequentially in each M-step of the algorithm. Setting a tolerance of 10^{-6} on the change in the likelihood, the algorithm stops after 27, 80, and 339 iterations with $K=2$, $K=4$, and $K=8$ types, respectively. Estimation is substantially more time-consuming than when using two-step grouped fixed-effects.

I Empirical illustration: firm and worker heterogeneity

In this section of the appendix we present an empirical illustration, where we consider the question of assessing the sources of wage dispersion across workers and firms.

I.1 Setup and main results

We consider an additive model in worker and firm heterogeneity:

$$Y_{it} = \eta_i + \psi_{j(i,t)} + \varepsilon_{it}, \tag{I1}$$

where Y_{it} denote log-wages, worker i works in firm $j(i,t)$ at time t, and η_i and ψ_j denote unobserved attributes of worker i and firm j, respectively. Equation (I1) corresponds to the model of Abowd, Kramarz and Margolis (1999) for matched employer-employee data, where we abstract from observed covariates for simplicity. Our interest centers on the decomposition of the variance of log-wages into a worker component, a firm component, a component reflecting the sorting of workers into heterogeneous firms, and an idiosyncratic match component:

$$\mathrm{Var}\left(y_{i1}\right) = \mathrm{Var}\left(\eta_i\right) + \mathrm{Var}\left(\psi_{j(i,1)}\right) + 2\,\mathrm{Cov}\left(\eta_i, \psi_{j(i,1)}\right) + \mathrm{Var}\left(\varepsilon_{i1}\right). \tag{I2}$$

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

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

which is uninformative for workers who remain in the same firm in the two periods. When the number of job movers into and out of firm j is low, ψ_j may be poorly estimated; see Andrews, Gill, Schank

and Upward (2008) and Jochmans and Weidner (2016). This source of incidental parameter bias may be particularly severe in short panels.

To alleviate this "low-mobility bias", Bonhomme, Lamadon and Manresa (2019, BLM) propose to reduce the number of firm-specific parameters by grouping firms based on firm-level observables in a first step. Then, in a second step, the ψ_j's are recovered at the group level, thus pooling information across job movers within firm groups. Specifically, given a kmeans-based classification $\{\widehat{k}_j\}$ of firms, the $\psi(k)$'s are estimated based on the following criterion:

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

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

This estimator is a two-step GFE estimator based on external moments. Here N is the number of firms, T is the number of available observations to estimate the firm-specific parameters ψ_j (that is, T is the number of job movers per firm), and S is the number of measurements on firm heterogeneity. In BLM, firms are classified based on their empirical wage distribution functions. Note that a difference with the setting of Theorem 1 is that here the likelihood function is not separable across firms, due to the fact that two firms are linked by the workers who move between them. We conjecture that Theorem 1 could be extended to such network settings, although formally developing this extension exceeds the scope of this paper.

Classifying firms based on wage distribution functions. Here we outline the distribution-based approach we use to classify firms. Let Y_{is} denote the wage of worker s in firm i. We denote $\widehat{F}_i(y) = \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\{Y_{is} \leq y\}$ the empirical cumulative distribution function of Y_{is}. We classify firms based on $h_i = \widehat{F}_i$, using the norm $\|g\|_\omega^2 = \int g(w)^2 \omega(w) dw$, where ω is an integrable function. The classification step then is: $\min_{(\widetilde{h},\{k_i\})} \sum_{i=1}^{N} \|h_i - \widetilde{h}(k_i)\|_\omega^2$, where the $\widetilde{h}(k)$'s are functions. In practice we discretize the integral, leading to a weighted kmeans objective function.

Simulation exercise. We now present simulation results based on model (I1). We focus on a two-period model, where ε_{it} are independent of $j(i,1)$, $j(i,2)$, η's, and ψ's, have zero means, and are i.i.d. across workers and time. Following BLM we adopt a correlated random-effects approach to model worker heterogeneity within firms. The parameters of the model are the firm fixed-effects ψ_j, the means and variances of worker effects in each firm $\mu_j = \mathbb{E}\left(\eta_i \,|\, j(i,1) = j\right)$ and $\sigma_j^2 = \mathrm{Var}\left(\eta_i \,|\, j(i,1) = j\right)$, and the variance of idiosyncratic errors $s^2 = \mathrm{Var}\left(\varepsilon_{i1}\right)$. We will be estimating the components of the variance decomposition in (I2). In addition we will report estimates of the correlation between worker and firm effects, $\mathrm{Corr}\left(\eta_i, \psi_{j(i,1)}\right)$, which is commonly interpreted as a measure of sorting.

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

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

so all firm-specific parameters are (nonlinear) functions of the scalar firm effects ψ_j. This specification is consistent with theoretical models of worker-firm sorting where firms are characterized by their scalar productivity level.[2] Below we report simulations using several alternative designs. We study cases where the underlying dimension of firm heterogeneity is equal to two, which allows for a second dimension of latent firm heterogeneity in addition to the wage effects ψ_j (below we provide evidence suggesting that the underlying dimension of firm heterogeneity is low in the data). We also consider a DGP where firm heterogeneity is discrete in the population.

We start by estimating model (I1) on Swedish register data, following BLM. We select male workers full-year employed in 2002 and 2004, and define as job movers the workers whose firm IDs change between the two years. We focus on firms that are present throughout the period. There are about 20,000 job movers in the sample. We use two-step GFE with a classification based on the firms' empirical distributions of log-wages in 2002, evaluated at 20 percentiles of the overall log-wage distribution, with $K = 10$ groups. In the second step, we estimate the model's parameters $\widehat{\psi}(\widehat{k}_j)$, $\widehat{\mu}(\widehat{k}_j)$, $\widehat{\sigma}^2(\widehat{k}_j)$, and $\widehat{s}^2$.

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

In Table S9 and Figure S3 we report the mean and 95% confidence intervals of GFE and fixed-effects estimators of the components of the variance decomposition (I2), across 500 simulations. We estimate the number of groups in every simulation. We see that biases of two-step GFE estimators decrease quite rapidly when firm size grows, although biases are not negligible when firms are small. As an example, the variance of firm effects is two thirds of the true value on average when firm size equals 10, and 75% of the true value for a firm size of 20. Moreover, although we do not have a theoretical justification for it in this setting, bias correction using the half-panel jackknife method of Dhaene and Jochmans (2015) tends to provide performance improvements: for example, biases for the variance of firm effects become 25% and 5% for firm sizes of 10 and 20, respectively. Note that here bias correction is not associated with large increases in dispersion. To implement the bias-correction

[2]Here Assumption 2 requires the mapping $\psi_j \mapsto \left(\psi_j + \mathbb{E}\left(\eta_i \,|\, \psi_j \right), \text{Var}\left(\eta_i \,|\, \psi_j \right) \right)$ to be injective. Firm-specific means of log-wages being monotone in ψ_j is sufficient, though not necessary, for this to hold.

method we select two halves within each firm at random, and we re-estimate the number of groups in each half-sample. In addition, the last column in the table shows that the estimated number of groups is rather small, and close to proportional to the square root of firm size (which is to be expected in this DGP with one-dimensional underlying heterogeneity).

Lastly, in the bottom panel of Table S9 we report the results for a fixed-effects estimator, which is computationally feasible in this linear setting. We see that fixed-effects is substantially biased. This shows that incidental parameter bias due to low mobility is particularly acute in this DGP. The contrast between fixed-effects and GFE is in line with Theorem 1, since here the number T of job movers per firm is small relative to the total number S of workers in the firm which we use to group firms. Hence, GFE, possibly combined with bias reduction, provides an effective regularization in this context.

I.2 Complements

Asymptotic properties of the distributional first step. Let $F_i(y) = \Pr(Y_{is} \leq y \,|\, \alpha_{i0}) = G(y; \alpha_{i0})$ denote the population cumulative distribution function of Y_{is}. Similarly to Lemma 1, the following convergence rate holds:

$$\frac{1}{N}\sum_{i=1}^{N} \left\| \widehat{h}(\widehat{k}_i) - G(\cdot; \alpha_{i0}) \right\|_{\omega}^2 = O_p\left(\frac{1}{S}\right) + O_p\left(B_\alpha(K)\right),$$

provided (i) $\frac{1}{N}\sum_{i=1}^{N} \|\widehat{F}_i - F_i\|_{\omega}^2 = O_p(S^{-1})$, and (ii) $G(\cdot; \alpha_{i0})$ is Lipschitz-continuous in its second argument. Here $\alpha \mapsto G(\cdot; \alpha)$ maps individual-specific parameters to functions in $L^2(\omega)$. In order for (i) to hold, a functional central limit theorem on $\widehat{F}_i$, together with ω being integrable, will suffice. The Lipschitz condition in (ii) will be satisfied provided $\int \frac{\partial \ln f(y \,|\, \alpha_i)}{\partial \alpha} \frac{\partial \ln f(y \,|\, \alpha_i)}{\partial \alpha'} f(y \,|\, \alpha_i) dy dx$ is uniformly bounded.

For the second step to deliver estimators with similar properties as in Theorem 1, an injectivity condition is needed. When classifying individuals based on empirical distributions, this condition does not impose further restrictions other than α_{i0} being identified. Indeed, $\alpha \mapsto G(\cdot, \alpha)$ being injective is equivalent to $[G(\cdot, \alpha_2) = G(\cdot, \alpha_1) \Rightarrow \alpha_2 = \alpha_1]$, which in turn is equivalent to α_{i0} being identified given knowledge of the function G.

Details on estimation. Following BLM we exploit the following restrictions, where we denote as $m_i = \mathbf{1}\{j(i,1) \neq j(i,2)\}$ the job mobility indicator. For job movers, using the fact that mobility does not depend on ε's, and that ε_{i1} is independent of ε_{i2}, we have:

$$\mathbb{E}\left(Y_{i2} - Y_{i1} \,|\, m_i = 1, j(i,1), j(i,2)\right) = \psi_{j(i,2)} - \psi_{j(i,1)}, \tag{I3}$$

$$\mathrm{Var}\left(Y_{i2} - Y_{i1} \,|\, m_i = 1, j(i,1), j(i,2)\right) = \mathrm{Var}\left(\varepsilon_{i2}\right) + \mathrm{Var}\left(\varepsilon_{i1}\right) = 2\, s^2. \tag{I4}$$

Then, in the first cross-section we have:

$$\mathbb{E}\left(Y_{i1} \,|\, j(i,1)\right) = \psi_{j(i,1)} + \mathbb{E}\left(\eta_i \,|\, j(i,1)\right) = \psi_{j(i,1)} + \mu_{j(i,1)}, \tag{I5}$$

$$\mathrm{Var}\left(Y_{i1} \,|\, j(i,1)\right) = \mathrm{Var}\left(\eta_i \,|\, j(i,1)\right) + \mathrm{Var}\left(\varepsilon_{i1}\right) = \sigma^2_{j(i,1)} + s^2. \tag{I6}$$

In estimation, we first compute a firm partition $\{\widehat{k}_j\}$ into K groups based on firm-specific empirical distributions of log-wages (evaluated at 20 points). In the second step, we use the following algorithm:

1. Compute $\widehat{\psi}(\widehat{k}_j)$ based on sample counterparts to (I3).
2. Compute $\widehat{s}^2$ based on (I4).
3. Given $\widehat{\psi}(\widehat{k}_j)$, compute $\widehat{\mu}(\widehat{k}_j)$ based on (I5).
4. Given $\widehat{s}^2$, compute $\widehat{\sigma}^2(\widehat{k}_j)$ based on (I6). In practice we impose non-negativity of the variances using a quadratic programming routine.

Given parameter estimates, we then estimates the variances and covariance in (I2) by aggregation across types.

The fixed-effects estimator in Table S9 is computed following the same algorithm, except that K is taken equal to N. Hence, the estimates of the firm effects ψ_j correspond to the estimator of Abowd, Kramarz and Margolis (1999). However, instead of relying on a fixed-effects approach on the worker side, in this two-period setting we rely on a correlated random-effects approach to deal with worker heterogeneity. In that specification, the mean and variance of worker effects η_i are firm-specific. We compute the connected set in an initial step, and use sparse matrix coding for efficient computation.

Monte Carlo designs. We consider four additional DGP, in addition to DGP 1 reported in Table S9. In Table S7 we show the sample sizes that we use in all designs, including the average number of job movers per firm. DGP 2 has one-dimensional underlying heterogeneity, with different parameter values: the variance of firm effects is larger than in DGP 1, while the correlation between firm effects and worker effects is smaller, the relative magnitudes being close to the estimates of Card, Heining and Kline (2013). DGP 3 and DGP 4 have two-dimensional underlying heterogeneity (ψ_j, V_j), where ψ_j is the wage firm effect and V_j drives workers' firm choice. (ψ_j, V_j) are drawn from a bivariate normal distribution, and the mean and variance of worker effects in the firm are set to $\mu_j = V_j$ and $\sigma^2_j = (a + bV_j)^2$ for some constants a, b which are calibrated to the Swedish sample. We interpret V_j as a present value driving workers' mobility decisions across firms, which may be only imperfectly correlated with ψ_j in the presence of non-pecuniary attributes valued by workers. As displayed in Table S8, the two-dimensional DGP differ in terms of parameter values. The last row of the table shows the correlation between the wage firm effect ψ_j and the present value V_j in all DGP. Lastly, DGP 5 has discrete heterogeneity. Specifically, there are $K^* = 10$ "true" groups in the population. The groups are chosen by approximating the firm heterogeneity of DGP 1.

Alternative DGP with one-dimensional heterogeneity: results. In Table S10 we report the results of two-step GFE and its bias-corrected version, as well as fixed-effects, in DGP 2 with one-dimensional heterogeneity and a larger variance of firm effects than in Table S9. The performance of the estimators is comparable to Table S9.

Bias-corrected fixed-effects. In Table S11 we report the results of bias-corrected fixed-effects estimation in DGP 1 (top panel, see Table S9) and DGP 2 (bottom panel). In order to implement the bias correction we use the half-panel jackknife of Dhaene and Jochmans (2015), splitting all workers in every firm into two random halves, including job movers. We see that, although bias correction improves relative to fixed-effects, the bias-corrected estimator is still substantially biased, even for moderately large firms. Notice that some of the variance estimates are in fact negative. This is due to the fact that the additive bias correction method does not enforce non-negativity.

Inferring the underlying dimension of firm heterogeneity. As a motivation for considering DGP with an underlying dimension higher than one, but still relatively low, we now attempt to learn the underlying dimension of firm heterogeneity on the Swedish matched employer-employee data set. In statistics, the literature on manifold learning aims at inferring the low intrinsic dimension of large dimensional data; see for example Levina and Bickel (2004) and Raginsky and Lazebnik (2005). Motivated by our method for selecting the number of groups, the method we use here consists in comparing the length of the panel S with the number of groups $\widehat{K}$ estimated from (4). If the underlying dimension of $\varphi(\xi_{i0})$ is $d > 0$, then we expect $\widehat{Q}(K)$ to decrease at a rate $O_p(K^{-\frac{2}{d}}) + o_p(S^{-1})$. This suggests that $\widehat{K}$ and $S^{\frac{d}{2}}$ will have a similar order of magnitude. In such a case the underlying dimension may be inferred by plotting the relationship, for panels of different lengths, between $\ln \widehat{K}$ and $\ln S$, the slope of which is $\widehat{d}/2$.

In Figure S4 we report the results of this exercise, taking firms with more than 50 employees, and then randomly selecting $x\%$ in each firm, where x varies between 5 and 100. The left graph shows the shape of the objective function $\widehat{Q}(K)$ as a function of K, in logs. In each sample the estimated number of groups $\widehat{K}$ lies at the intersection of that curve and the horizontal line $\ln(\widehat{V}_h)$. We use a slight modification of the $\widehat{V}_h$ formula to deal with the fact that here the "panel" is unbalanced, since different firms may have different sizes. On the right graph we then plot $\ln \widehat{K}$ against the logarithm of the average firm size in each sample (here we report results based on empirical cdfs of log-wages evaluated at 20 points, but we checked that using 40 points instead did not affect the results). We see that the relationship is approximately linear and the slope is close to one, suggesting that the underlying dimension is around $\widehat{d} = 2$.

Two-dimensional heterogeneity: results. In Table S12 we report the simulation results for DGP 3 with continuous two-dimensional firm heterogeneity. The results for DGP 4, with a smaller

variance of firm effects, are reported in Table S14. The results are shown graphically in Figures S5 and S6. Focusing on the first panel, which corresponds to our recommended choice for the selection rule of the number of groups (that is, taking $\gamma = 1$ in (4)), we see that the two-step estimators show larger biases than in the one-dimensional case, especially for the variance of firm effects and the correlation parameter. Moreover, bias correction does not succeed at reducing the bias substantially. This suggests that, for the selected number of groups, the approximation error is still substantial. At the same time, as shown by the two bottom panels of the tables, taking $\gamma = .5$ and $\gamma = .25$ improves the performance of the two-step estimator. Notice that while the selected number of groups $\widehat{K}$ is monotone in firm size for $\gamma = 1$ and $\gamma = .5$, it is not monotone for $\gamma = .25$. This is a finite sample issue: when taking $\gamma = .25$ and focusing on large firms the number of groups is no longer negligible with respect to the number of firms in the sample. Lastly, performance is further improved when using the bias-reduced estimator.

As pointed out in Section 4, features of the model may be exploited to improve the classification. In the two-dimensional designs DGP 3 and DGP 4, we perform the following moment-based iteration. The two-step method delivers estimates of the mean and variance of worker effects η_i in firm group $\widehat{k}_j$: $\widehat{\mu}(\widehat{k}_j)$ and $\widehat{\sigma}^2(\widehat{k}_j)$, respectively. Regressing $\sqrt{\widehat{\sigma}^2(\widehat{k}_j)}$ on $\widehat{\mu}(\widehat{k}_j)$ and a constant then gives estimates $\widehat{b}$ and $\widehat{a}$. Given those, we construct the (iterated) moments:

$$h_{1j} = \widehat{\mathbb{E}}(Y_{i1} \,|\, j) - \frac{\sqrt{\widehat{\mathrm{Var}}(Y_{i1} \,|\, j) - \widehat{s}^2} - \widehat{a}}{\widehat{b}}, \quad h_{2j} = \frac{\sqrt{\widehat{\mathrm{Var}}(Y_{i1} \,|\, j) - \widehat{s}^2} - \widehat{a}}{\widehat{b}},$$

where $\widehat{\mathbb{E}}$ and $\widehat{\mathrm{Var}}$ denote firm-specific means and variances. Those moments will be consistent for ψ_j and V_j, respectively, as S tends to infinity. We then apply two-step GFE to the moments h_{1j} and h_{2j}. In Tables S14 and S15 we report the results for the iterated estimator (only iterated once) and its bias-corrected version, for DGP 3 and DGP 4, respectively. We see that the iteration improves performance substantially for DGP 3, although it has small effects on performance in DGP 4.

Low mobility bias and regularization. As shown by Theorem 1, a benefit of discretizing unobserved heterogeneity is that it can reduce the incidental parameter bias of fixed-effects estimators. In the illustration on matched employer-employee data, fixed-effects estimators may be biased due to low rates of worker mobility between firms. In order to assess the impact of mobility rates on the performance of fixed-effects and GFE estimators, in Figures S7 and S8 we report the results of the estimated variance decomposition on 500 simulations, comparing fixed-effects, bias-corrected fixed-effects, two-step GFE with bias correction, and iterated two-step GFE with bias correction. We perform simulations for different number of job movers per firm, from 2 to 10 (shown on the x-axis), and a fixed firm size of 50. The two figures show the results for the two-dimensional DGP: DGP 3 and DGP 4, respectively. We see a striking difference between fixed-effects and GFE: while the former is very sensitive to the number of job movers, the latter is virtually insensitive. In particular, for low

numbers of job movers fixed-effects and its bias-corrected counterpart are severely biased, while the biases of GFE remain moderate. This is in line with Theorem 1. It is worth noting that the average number of job movers per firm is around 0.5 in the original Swedish sample. This suggests that, at least in short panels, the discrete regularization achieved in GFE may result in practical improvements relative to fixed-effects in data sets of realistic dimensions.

Discrete heterogeneity: results. Finally, in Table S16 we report results for a discrete DGP (DGP 5) where all firm population parameters are constant within groups $\widehat{k}_j$, with $K^* = 10$. In this case the results of two-step GFE with $K = K^*$ turn out to be quite similar to those obtained in the continuous DGP in Table S9. However, as the last column in the table shows, in this discrete DGP misclassification frequencies are sizable: 69% misclassification when firm size equals 10, and still 23% when size is 100. We computed misclassification frequencies by solving a linear assignment problem using the simplex algorithm in every simulation. This suggests that, for this DGP, an "oracle" asymptotic theory based on the premise that group misclassification is absent in the limit may not provide reliable guidance for finite sample inference, even when the true number of groups is known. Lastly, the table shows some evidence that bias correction (where the number of groups is estimated in every simulation) improves the performance of the estimator in this setting too.

References

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

[2] Andrews, M. J., L. Gill, T. Schank, and R. Upward (2008): "High Wage Workers and Low Wage Firms: Negative Assortative Matching or Limited Mobility Bias?" *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(3), 673–697.

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

[4] 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.

[5] 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.

[6] Banna, M., F. Merlevède, and P. Youssef (2016): "Bernstein-Type Inequality for a Class of Dependent Random Matrices," *Random Matrices: Theory and Applications*, 5(02), 1650006.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[23] Tropp, J. A. (2012): "User-Friendly Tail Bounds for Sums of Random Matrices," *Foundations of Computational Mathematics*, 12(4), 389–434.

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

Table S1: Coverage probabilities

A. Probit with fixed effects

	Two-step				Iterated				Fixed Effects			
	analytical		bootstrap		analytical		bootstrap		analytical		bootstrap	
DGP	UC	BC	UC	BC	UC	BC	UC	BC	UC	BC	UC	BC
1	0.874	0.914	0.870	0.944	0.010	0.898	0.014	0.934	0.010	0.906	0.012	0.902
1*	0.864	0.948	0.878	0.944	0.006	0.932	0.008	0.952	0.010	0.906	0.012	0.902
2	0.000	0.000	0.000	0.000	0.022	0.858	0.030	0.952	0.010	0.920	0.014	0.922
2*	0.092	0.544	0.142	0.852	0.006	0.866	0.020	0.930	0.010	0.920	0.014	0.922
3	0.000	0.000	0.000	0.000	0.002	0.700	0.004	0.884	0.000	0.848	0.002	0.858
4	0.000	0.000	0.000	0.000	0.000	0.634	0.008	0.816	0.002	0.758	0.002	0.796

B. Probit with time-varying unobservables

	Two-step				Iterated				Fixed Effects			
	analytical		bootstrap		analytical		bootstrap		analytical		bootstrap	
DGP	UC	BC	UC	BC	UC	BC	UC	BC	UC	BC	UC	BC
1	0.896	0.846	0.732	0.876	0.000	0.518	0.000	1.000	–	–	–	–
1*	0.876	0.730	0.680	0.670	0.000	0.226	0.000	0.952	–	–	–	–
2	0.000	0.000	0.000	0.000	0.018	0.268	0.074	0.994	–	–	–	–
2*	0.002	0.234	0.082	0.970	0.000	0.030	0.000	0.520	–	–	–	–
3	0.000	0.000	0.000	0.000	0.000	0.326	0.000	1.000	–	–	–	–
4	0.000	0.000	0.000	0.000	0.000	0.388	0.000	1.000	–	–	–	–

C. Probit with random coefficients

	Two-step				Iterated				Fixed Effects			
	analytical		bootstrap		analytical		bootstrap		analytical		bootstrap	
DGP	UC	BC	UC	BC	UC	BC	UC	BC	UC	BC	UC	BC
3	0.000	0.000	0.000	0.000	0.002	0.398	0.002	0.876	0.000	0.160	0.000	0.464
4	0.000	0.000	0.000	0.000	0.020	0.380	0.042	0.928	0.000	0.172	0.000	0.812

Notes: See notes to Tables 1 and 2. The analytical coverage is based on an estimate of the asymptotic variance matrix. The bootstrap is clustered at the individual level, computed using 100 bootstrap simulations. Coverage probabilities for a nominal level of 95%*. Results of* 500 *simulations.*

Table S2: Estimated K, unconditional rule

A. Probit with fixed effects						
	$\widehat{K}$		Two-step		Iterated	
DGP	full	half	UC	BC	UC	BC
1	11.8 (0.41)	8.45 (0.50)	0.985 (0.019)	1.002 (0.019)	1.087 (0.021)	0.999 (0.021)
2	20.31 (0.69)	12.48 (0.50)	0.940 (0.018)	0.970 (0.024)	1.088 (0.022)	0.998 (0.022)
	Conditional Methods					
	Linear		Cubic		Neural network	
DGP	UC	BC	UC	BC	UC	BC
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	1.001 (0.020)	0.971 (0.024)	1.017 (0.021)	0.991 (0.026)	1.035 (0.023)	1.016 (0.029)
B. Probit with time-varying heterogeneity						
	$\widehat{K}$		Two-step		Iterated	
DGP	full	half	UC	BC	UC	BC
1	11.92 (0.38)	12.0 (0.4), 8.0 (0.2)	1.015 (0.021)	1.028 (0.025)	1.319 (0.039)	1.093 (0.070)
2	21.69 (0.78)	20.5 (0.7), 11.1 (0.4)	0.889 (0.023)	0.935 (0.037)	1.278 (0.044)	1.196 (0.080)
	Conditional Methods					
	Linear		Cubic		Neural network	
DGP	UC	BC	UC	BC	UC	BC
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	1.041 (0.028)	1.056 (0.046)	1.096 (0.035)	1.098 (0.060)	1.085 (0.035)	1.144 (0.062)

Notes: See notes to Tables 1 and 2. We estimate K using (4), with $\gamma = 1$.

Table S3: Estimated K, conditional rule

A. Probit with fixed effects						
	$\widehat{K}$		Two-step		Iterated	
DGP	full	half	UC	BC	UC	BC
1	4.98 (0.13)	4.00 (0.00)	0.999 (0.018)	1.013 (0.019)	1.094 (0.021)	1.008 (0.022)
2	5.90 (0.30)	5.00 (0.06)	0.744 (0.034)	0.774 (0.066)	1.070 (0.022)	1.001 (0.025)
3	5.00 (0.00)	4.83 (0.38)	0.757 (0.021)	0.763 (0.028)	1.078 (0.026)	0.969 (0.033)
4	4.02 (0.15)	4.00 (0.00)	0.749 (0.022)	0.754 (0.028)	1.069 (0.029)	0.949 (0.038)
	Conditional Methods					
	Linear		Cubic		Neural network	
DGP	UC	BC	UC	BC	UC	BC
1	1.041 (0.019)	0.999 (0.020)	1.066 (0.020)	1.011 (0.022)	1.080 (0.021)	1.030 (0.024)
2	0.944 (0.019)	0.910 (0.024)	1.005 (0.020)	0.980 (0.025)	1.021 (0.022)	1.003 (0.028)
3	0.923 (0.023)	0.903 (0.035)	0.924 (0.026)	0.923 (0.040)	1.004 (0.027)	0.992 (0.044)
4	0.894 (0.024)	0.875 (0.038)	0.893 (0.026)	0.897 (0.041)	0.996 (0.034)	0.982 (0.059)

Notes: See notes to Table 1. We implement the choice of K described in Appendix B.2, based on a linear conditional kmeans specification, and $K_{\max} = 30$.

Table S4: Estimated K, conditional rule (cont.)

B. Probit with time-varying heterogeneity

	$\widehat{K}$		Two-step		Iterated	
DGP	full	half	UC	BC	UC	BC
1	5.04 (0.19)	4.5 (0.5), 4.0 (0.1)	1.029 (0.019)	1.020 (0.025)	1.191 (0.031)	1.119 (0.055)
2	7.25 (0.51)	6.0 (0.4), 5.0 (0.1)	0.631 (0.044)	0.692 (0.086)	1.082 (0.038)	1.159 (0.070)
3	5.05 (0.21)	5.0 (0.1), 4.0 (0.2)	0.562 (0.027)	0.547 (0.039)	1.082 (0.040)	1.060 (0.062)
4	4.17 (0.38)	4.3 (0.5), 3.1 (0.3)	0.559 (0.025)	0.542 (0.035)	1.046 (0.042)	1.026 (0.070)

	Conditional Methods					
	Linear		Cubic		Neural network	
DGP	UC	BC	UC	BC	UC	BC
1	1.065 (0.020)	1.012 (0.0269)	1.125 (0.022)	1.079 (0.034)	1.136 (0.027)	1.092 (0.046)
2	0.906 (0.028)	0.922 (0.046)	0.955 (0.031)	0.956 (0.057)	0.979 (0.035)	1.020 (0.063)
3	0.845 (0.035)	0.851 (0.064)	0.760 (0.045)	0.771 (0.080)	0.948 (0.041)	0.994 (0.083)
4	0.802 (0.032)	0.834 (0.056)	0.753 (0.037)	0.782 (0.066)	0.923 (0.054)	0.981 (0.104)

C. Probit with random coefficients

	$\widehat{K}$		Two-step		Iterated	
DGP	full	half	UC	BC	UC	BC
3	5.08 (0.28)	4.86 (0.40)	0.690 (0.023)	0.699 (0.028)	1.035 (0.025)	0.949 (0.037)
4	4.31 (0.46)	4.11 (0.32)	0.594 (0.029)	0.602 (0.034)	0.976 (0.034)	0.917 (0.058)

	Conditional Methods					
	Linear		Cubic		Neural network	
DGP	UC	BC	UC	BC	UC	BC
3	0.879 (0.028)	0.864 (0.038)	0.909 (0.026	0.889 (0.039)	0.982 (0.025)	0.947 (0.040)
4	0.765 (0.032)	0.756 (0.049)	0.798 (0.045)	0.805 (0.073)	0.923 (0.044)	0.893 (0.084)

Notes: See notes to Table 2. We implement the choice of K described in Appendix B.2, based on a linear conditional kmeans specification, and $K_{\max} = 30$.

Table S5: Starting values in kmeans

	DGP 1						
♯ starting values	1	5	10	50	100	500	1000
within	0.027	0.007	0.004	0.002	0.001	0.000	0.000
between	3.110	3.118	3.121	3.127	3.130	3.132	3.133
total	3.137	3.125	3.125	3.128	3.131	3.133	3.134
within (% total)	0.9%	0.2%	0.1%	0.0%	0.0%	0.0%	0.0%
	DGP 2						
♯ starting values	1	5	10	50	100	500	1000
within	1.730	0.634	0.384	0.121	0.068	0.018	0.010
between	2.815	3.192	3.387	3.568	3.589	3.603	3.606
total	4.522	3.814	3.761	3.681	3.650	3.614	3.609
within (% total)	38.3%	16.6%	10.2%	3.3%	1.9%	0.5%	0.3%
	DGP 3						
♯ starting values	1	5	10	50	100	500	1000
within	0.894	0.675	0.581	0.354	0.281	0.148	0.112
between	2.705	2.866	2.959	3.185	3.271	3.450	3.486
total	3.585	3.529	3.529	3.529	3.543	3.589	3.590
within (% total)	24.9%	19.1%	16.5%	10.0%	7.9%	4.1%	3.1%
	DGP 4						
♯ starting values	1	5	10	50	100	500	1000
within	1.023	0.847	0.781	0.621	0.549	0.395	0.338
between	3.553	3.618	3.689	3.853	3.909	4.090	4.158
total	4.560	4.449	4.455	4.460	4.445	4.473	4.485
within (% total)	22.4%	19.0%	17.5%	13.9%	12.3%	8.8%	7.5%

Notes: See notes to Table 1. Let n be a number of starting values. In each of the 500 simulated samples, we compute 100 kmeans partitions starting at n randomly generated values and the resulting two-step GFE estimators. We then report the total variance across samples and kmeans runs, as well as the between-sample and within-sample variances. All variances are multiplied by 10000*.*

Table S6: Computation Time (in seconds)

N	100	500	1000	5000	10000
Kmeans	0.11	0.12	0.16	0.66	1.51
Two-step GFE	0.12	0.15	0.21	1.11	2.71
Iterated GFE	0.28	0.58	0.94	7.07	20.1
Fixed-effects	0.02	0.16	0.50	11.7	93.5

Notes: Computation time for various estimators in the random coefficients probit model of DGP 3. See notes to Table 2. We use sparse matrix computation to compute the fixed-effects estimator, and we use parallel computing for kmeans. Computations are performed on a 20 CPU core machine.

Table S7: Firms and workers, sample sizes

Firm size	Number firms	Number job movers per firm
10	10000	2
20	5000	4
50	2000	10
100	1000	20
200	500	40

Notes: Sample sizes for different firm sizes, all DGP.

Table S8: Firms and workers, different DGP

	small $Var(\psi)$		large $Var(\psi)$	
	1D	2D	1D	2D
$Var(\psi)$	0.0017	0.0017	0.0204	0.0204
	2.0%	2.0%	21.2%	21.2%
$Var(\eta)$	0.0758	0.0758	0.0660	0.0660
	85.2%	85.2%	68.4%	68.4%
$2Cov(\psi, \eta)$	0.0057	0.0057	0.0050	0.0050
	12.8%	12.8%	10.4%	10.4%
$Corr(\psi, \eta)$	0.4963	0.4963	0.1373	0.1373
$Var(\varepsilon)$	0.0341	0.0341	0.0341	0.0341
$Corr(V, \psi)$	1.0000	0.7130	1.0000	0.2540

Notes: The four columns show the parameter values and overall shares of variance in DGP 1, DGP 4, DGP 2, and DGP 3, respectively.

Table S9: Estimates of firm and worker heterogeneity across simulations

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

Notes: Means and 95% confidence intervals. Unobserved heterogeneity is continuously distributed in the DGP. The number of groups $\widehat{K}$ is estimated in every replication, using (4) with $\gamma = 1$, and it is reported in the last column of the first panel. We use the kmeans routine from R, with 100 *starting values.* 500 *simulations.*

Table S10: Estimates of firm and worker heterogeneity across simulations, one-dimensional firm heterogeneity, large variance of firm effects

Firm size	$\text{Var}(\eta_i)$	$\text{Var}(\psi_j)$	$\text{Cov}(\eta_i, \psi_j)$	$\text{Corr}(\eta_i, \psi_j)$	$\text{Var}(\varepsilon_{i1})$	$\hat{K}$
	true values					
-	0.0660	0.0204	0.0050	0.1373	0.0341	
	two-step estimator					
10	0.0605 [0.059,0.062]	0.0124 [0.012,0.013]	0.0078 [0.008,0.008]	0.2868 [0.275,0.300]	0.0422 [0.041,0.043]	3.0 [3,3]
20	0.0626 [0.061,0.064]	0.0155 [0.015,0.016]	0.0068 [0.006,0.007]	0.2178 [0.205,0.230]	0.0392 [0.038,0.040]	4.0 [4,4]
50	0.0645 [0.063,0.066]	0.0180 [0.017,0.019]	0.0058 [0.005,0.006]	0.1714 [0.158,0.183]	0.0365 [0.036,0.037]	6.0 [6,6]
100	0.0653 [0.064,0.066]	0.0191 [0.018,0.020]	0.0054 [0.005,0.006]	0.1542 [0.141,0.166]	0.0354 [0.035,0.036]	8.0 [8,8]
200	0.0657 [0.065,0.067]	0.0198 [0.019,0.021]	0.0052 [0.005,0.006]	0.1448 [0.132,0.157]	0.0348 [0.034,0.035]	10.9 [10,12]
	two-step estimator, bias-corrected					
10	0.0650 [0.064,0.066]	0.0149 [0.014,0.016]	0.0056 [0.005,0.006]	0.1445 [0.127,0.163]	0.0397 [0.039,0.041]	
20	0.0647 [0.063,0.066]	0.0185 [0.018,0.019]	0.0057 [0.005,0.006]	0.1499 [0.133,0.167]	0.0361 [0.035,0.037]	
50	0.0656 [0.064,0.067]	0.0202 [0.019,0.021]	0.0053 [0.005,0.006]	0.1416 [0.126,0.155]	0.0344 [0.034,0.035]	
100	0.0661 [0.065,0.067]	0.0202 [0.019,0.021]	0.0050 [0.005,0.005]	0.1371 [0.122,0.150]	0.0344 [0.034,0.035]	
200	0.0661 [0.065,0.067]	0.0204 [0.020,0.021]	0.0050 [0.005,0.005]	0.1361 [0.123,0.149]	0.0342 [0.033,0.035]	
	fixed-effects estimator					
10	0.1252 [0.123,0.127]	0.0528 [0.051,0.055]	-0.0273 [-0.029,-0.026]	-0.3357 [-0.346,-0.324]	0.0173 [0.017,0.018]	
20	0.0908 [0.090,0.092]	0.0318 [0.031,0.033]	-0.0063 [-0.007,-0.006]	-0.1165 [-0.127,-0.105]	0.0256 [0.025,0.026]	
50	0.0752 [0.074,0.076]	0.0242 [0.023,0.025]	0.0013 [0.001,0.002]	0.0301 [0.019,0.041]	0.0307 [0.030,0.031]	
100	0.0705 [0.069,0.072]	0.0222 [0.021,0.023]	0.0033 [0.003,0.004]	0.0827 [0.071,0.095]	0.0324 [0.032,0.033]	
200	0.0683 [0.067,0.069]	0.0213 [0.021,0.022]	0.0041 [0.004,0.005]	0.1085 [0.096,0.120]	0.0333 [0.033,0.034]	

Notes: See notes to Table S9. Results for DGP 2.

Table S11: Bias-corrected fixed-effects estimators, one-dimensional firm heterogeneity

Firm size	Var (η_i)	Var (ψ_j)	Cov (η_i, ψ_j)	Corr (η_i, ψ_j)	Var (ε_{i1})
	one-dimensional, small firm effect				
-	0.0758	0.0017	0.0057	0.4963	0.0341
	fixed-effects, bias-corrected				
10	0.0065 [-0.004,0.016]	-0.0717 [-0.082,-0.064]	0.0791 [0.071,0.089]	-0.0976 [-0.125,-0.072]	0.0300 [0.029,0.031]
20	0.0645 [0.062,0.067]	-0.0098 [-0.011,-0.008]	0.0172 [0.016,0.019]	0.0973 [0.073,0.125]	0.0339 [0.033,0.035]
50	0.0733 [0.072,0.075]	-0.0007 [-0.001,-0.000]	0.0082 [0.008,0.009]	0.3069 [0.279,0.335]	0.0341 [0.033,0.035]
100	0.0748 [0.073,0.076]	0.0007 [0.000,0.001]	0.0067 [0.006,0.007]	0.4173 [0.388,0.447]	0.0341 [0.033,0.035]
200	0.0753 [0.074,0.077]	0.0012 [0.001,0.002]	0.0062 [0.006,0.007]	0.4822 [0.451,0.512]	0.0341 [0.033,0.035]
	one-dimensional, large firm effect				
-	0.0660	0.0204	0.0050	0.1373	0.0341
	fixed-effects, bias-corrected				
10	-0.0036 [-0.013,0.006]	-0.0533 [-0.062,-0.045]	0.0788 [0.070,0.088]	-0.0077 [-0.034,0.019]	0.0301 [0.029,0.031]
20	0.0547 [0.053,0.057]	0.0089 [0.007,0.011]	0.0166 [0.015,0.018]	0.1163 [0.096,0.137]	0.0339 [0.033,0.035]
50	0.0636 [0.062,0.065]	0.0180 [0.017,0.019]	0.0075 [0.007,0.008]	0.1561 [0.139,0.173]	0.0341 [0.033,0.035]
100	0.0650 [0.064,0.066]	0.0194 [0.019,0.020]	0.0061 [0.006,0.007]	0.1554 [0.139,0.171]	0.0341 [0.033,0.035]
200	0.0656 [0.064,0.067]	0.0199 [0.019,0.021]	0.0055 [0.005,0.006]	0.1487 [0.133,0.164]	0.0341 [0.033,0.035]

Notes: Means and 95% confidence intervals. Unobserved heterogeneity is continuously distributed in the DGP. Bias correction is based on splitting both job movers and job stayers into two sub-samples. The top panel shows the results on DGP 1, with a small variance of firm effects, while the bottom panel shows the results for DGP 2, with a larger variance of firm effects. 500 *simulations.*

Table S12: Firm and worker effects, two-dimensional firm heterogeneity, large $Var(\psi)$, different choices of γ

	two-step estimator						two-step estimator, bias corrected					
Firm size	$\text{Var}(\eta_i)$	$\text{Var}(\psi_j)$	$\text{Cov}(\eta_i, \psi_j)$	$\text{Corr}(\eta_i, \psi_j)$	$\text{Var}(\varepsilon_{i1})$	$\hat{K}$	$\text{Var}(\eta_i)$	$\text{Var}(\psi_j)$	$\text{Cov}(\eta_i, \psi_j)$	$\text{Corr}(\eta_i, \psi_j)$	$\text{Var}(\varepsilon_{i1})$	$\hat{K}$
	true values						true values					
-	0.0660	0.0204	0.0050	0.1373	0.0341		0.0660	0.0204	0.0050	0.1373	0.0341	
	$\gamma = 1.0$						$\gamma = 1.0$					
10	0.0513 [0.050,0.053]	0.0098 [0.009,0.010]	0.0124 [0.012,0.013]	0.5500 [0.539,0.563]	0.0448 [0.044,0.045]	4.0 [4,4]	0.0529 [0.050,0.055]	0.0112 [0.010,0.012]	0.0115 [0.011,0.012]	0.4574 [0.437,0.486]	0.0434 [0.042,0.044]	4.2 [4,5]
20	0.0515 [0.049,0.053]	0.0112 [0.010,0.012]	0.0124 [0.012,0.013]	0.5180 [0.498,0.536]	0.0433 [0.042,0.044]	5.7 [5,6]	0.0514 [0.049,0.053]	0.0126 [0.011,0.014]	0.0125 [0.012,0.013]	0.4856 [0.454,0.509]	0.0420 [0.041,0.043]	7.4 [6,8]
50	0.0514 [0.049,0.054]	0.0123 [0.012,0.013]	0.0124 [0.012,0.013]	0.4939 [0.471,0.512]	0.0423 [0.041,0.043]	8.9 [8,9]	0.0513 [0.049,0.054]	0.0131 [0.012,0.014]	0.0125 [0.012,0.013]	0.4797 [0.451,0.505]	0.0415 [0.041,0.043]	11.8 [10,12]
100	0.0519 [0.049,0.054]	0.0128 [0.012,0.014]	0.0124 [0.012,0.013]	0.4796 [0.453,0.503]	0.0416 [0.041,0.043]	13.3 [13,14]	0.0520 [0.049,0.055]	0.0133 [0.013,0.014]	0.0123 [0.011,0.013]	0.4664 [0.436,0.496]	0.0411 [0.040,0.042]	17.9 [17,20]
200	0.0548 [0.051,0.058]	0.0147 [0.014,0.016]	0.0104 [0.009,0.012]	0.3683 [0.303,0.426]	0.0399 [0.039,0.041]	21.4 [20,23]	0.0579 [0.053,0.062]	0.0165 [0.015,0.018]	0.0089 [0.006,0.011]	0.2713 [0.168,0.374]	0.0381 [0.037,0.040]	30.1 [28,33]
	$\gamma = 0.5$						$\gamma = 0.5$					
10	0.0498 [0.048,0.053]	0.0110 [0.010,0.012]	0.0134 [0.013,0.014]	0.5730 [0.555,0.589]	0.0435 [0.043,0.044]	12.8 [12,13]	0.0386 [0.031,0.046]	0.0126 [0.012,0.013]	0.0123 [0.012,0.013]	0.5385 [0.494,0.578]	0.0419 [0.041,0.043]	14.0 [12,15]
20	0.0510 [0.049,0.052]	0.0123 [0.012,0.013]	0.0125 [0.012,0.013]	0.4997 [0.482,0.519]	0.0423 [0.041,0.043]	16.1 [16,17]	0.0520 [0.049,0.054]	0.0136 [0.013,0.014]	0.0116 [0.011,0.012]	0.4297 [0.402,0.460]	0.0410 [0.040,0.042]	19.7 [19,22]
50	0.0536 [0.052,0.056]	0.0140 [0.013,0.015]	0.0113 [0.011,0.012]	0.4134 [0.389,0.437]	0.0407 [0.040,0.042]	26.0 [24,28]	0.0556 [0.053,0.058]	0.0152 [0.015,0.016]	0.0104 [0.010,0.011]	0.3484 [0.312,0.378]	0.0394 [0.039,0.040]	34.4 [30,38]
100	0.0563 [0.053,0.059]	0.0153 [0.014,0.016]	0.0099 [0.009,0.011]	0.3371 [0.300,0.376]	0.0392 [0.038,0.040]	38.8 [36,41]	0.0589 [0.056,0.062]	0.0168 [0.016,0.018]	0.0086 [0.007,0.010]	0.2635 [0.205,0.314]	0.0377 [0.036,0.039]	52.9 [48,57]
200	0.0596 [0.056,0.063]	0.0171 [0.016,0.018]	0.0085 [0.007,0.010]	0.2662 [0.214,0.314]	0.0375 [0.037,0.039]	53.2 [49,57]	0.0626 [0.058,0.067]	0.0187 [0.018,0.020]	0.0070 [0.005,0.009]	0.1948 [0.129,0.247]	0.0359 [0.035,0.037]	71.8 [65,78]
	$\gamma = 0.25$						$\gamma = 0.25$					
10	0.0595 [0.058,0.062]	0.0119 [0.011,0.013]	0.0132 [0.013,0.014]	0.4988 [0.477,0.514]	0.0428 [0.042,0.043]	125.6 [121,130]	0.0504 [0.047,0.054]	0.0136 [0.013,0.014]	0.0117 [0.011,0.012]	0.4379 [0.400,0.465]	0.0411 [0.040,0.042]	152.6 [145,160]
20	0.0567 [0.055,0.058]	0.0134 [0.013,0.014]	0.0119 [0.011,0.012]	0.4318 [0.409,0.447]	0.0412 [0.040,0.042]	138.3 [132,143]	0.0536 [0.051,0.057]	0.0149 [0.014,0.016]	0.0106 [0.010,0.011]	0.3677 [0.335,0.394]	0.0397 [0.039,0.040]	163.7 [153,172]
50	0.0574 [0.055,0.059]	0.0155 [0.015,0.016]	0.0099 [0.009,0.011]	0.3316 [0.300,0.357]	0.0391 [0.038,0.040]	154.0 [146,163]	0.0582 [0.055,0.061]	0.0170 [0.016,0.018]	0.0085 [0.007,0.009]	0.2622 [0.217,0.301]	0.0376 [0.037,0.039]	190.1 [177,204]
100	0.0601 [0.057,0.063]	0.0171 [0.016,0.018]	0.0083 [0.007,0.010]	0.2598 [0.222,0.303]	0.0374 [0.037,0.038]	151.1 [142,163]	0.0624 [0.059,0.066]	0.0186 [0.018,0.019]	0.0069 [0.006,0.008]	0.1932 [0.151,0.240]	0.0359 [0.035,0.037]	186.1 [172,202]
200	0.0626 [0.058,0.066]	0.0186 [0.018,0.020]	0.0069 [0.005,0.009]	0.2027 [0.156,0.251]	0.0361 [0.035,0.037]	133.7 [127,141]	0.0649 [0.060,0.069]	0.0199 [0.019,0.021]	0.0056 [0.004,0.007]	0.1512 [0.095,0.205]	0.0348 [0.034,0.036]	162.3 [151,176]

Notes: Means and 95% confidence intervals. Unobserved heterogeneity is continuously distributed in the DGP, with underlying dimension equal to 2*. The number of groups K is estimated in every replication, with different choices for γ in (4).* 500 *simulations. Results for DGP 3.*

Table S13: Firm and worker effects, two-dimensional firm heterogeneity, small $Var(\psi)$, different choices of γ

Firm size	two-step estimator: $\text{Var}(\eta_i)$	$\text{Var}(\psi_j)$	$\text{Cov}(\eta_i, \psi_j)$	$\text{Corr}(\eta_i, \psi_j)$	$\text{Var}(\varepsilon_{i1})$	$\hat{K}$	two-step estimator, bias corrected: $\text{Var}(\eta_i)$	$\text{Var}(\psi_j)$	$\text{Cov}(\eta_i, \psi_j)$	$\text{Corr}(\eta_i, \psi_j)$	$\text{Var}(\varepsilon_{i1})$	$\hat{K}$
	true values						true values					
-	0.0758	0.0017	0.0057	0.4963	0.0341		0.0758	0.0017	0.0057	0.4963	0.0341	
	$\gamma = 1.0$						$\gamma = 1.0$					
10	0.0759 [0.074,0.078]	0.0008 [0.001,0.001]	0.0056 [0.005,0.006]	0.7010 [0.691,0.709]	0.0350 [0.034,0.036]	4.0 [4,4]	0.0760 [0.074,0.078]	0.0009 [0.001,0.001]	0.0056 [0.005,0.006]	0.6487 [0.636,0.660]	0.0349 [0.034,0.036]	4.0 [4,4]
20	0.0754 [0.073,0.077]	0.0009 [0.001,0.001]	0.0059 [0.005,0.006]	0.6927 [0.680,0.707]	0.0349 [0.034,0.036]	5.5 [5,6]	0.0749 [0.073,0.077]	0.0011 [0.001,0.001]	0.0061 [0.006,0.007]	0.6846 [0.666,0.705]	0.0348 [0.034,0.035]	6.9 [6,8]
50	0.0750 [0.073,0.078]	0.0011 [0.001,0.001]	0.0061 [0.006,0.007]	0.6877 [0.674,0.701]	0.0348 [0.034,0.035]	8.0 [8,8]	0.0747 [0.072,0.078]	0.0011 [0.001,0.001]	0.0062 [0.006,0.007]	0.6841 [0.669,0.701]	0.0347 [0.034,0.035]	10.0 [10,10]
100	0.0752 [0.072,0.079]	0.0011 [0.001,0.001]	0.0062 [0.006,0.007]	0.6848 [0.668,0.701]	0.0347 [0.034,0.035]	11.1 [11,12]	0.0750 [0.072,0.078]	0.0011 [0.001,0.001]	0.0063 [0.006,0.007]	0.6816 [0.660,0.702]	0.0347 [0.034,0.035]	14.2 [14,16]
200	0.0746 [0.069,0.079]	0.0011 [0.001,0.001]	0.0062 [0.006,0.007]	0.6765 [0.654,0.697]	0.0347 [0.034,0.035]	15.2 [14,16]	0.0745 [0.069,0.079]	0.0012 [0.001,0.001]	0.0062 [0.006,0.007]	0.6720 [0.647,0.694]	0.0347 [0.034,0.035]	19.3 [17,21]
	$\gamma = 0.5$						$\gamma = 0.5$					
10	0.0748 [0.073,0.076]	0.0010 [0.001,0.001]	0.0062 [0.006,0.007]	0.7333 [0.720,0.746]	0.0349 [0.034,0.036]	12.2 [12,13]	0.0731 [0.070,0.076]	0.0011 [0.001,0.001]	0.0062 [0.006,0.007]	0.6905 [0.664,0.715]	0.0348 [0.034,0.035]	12.7 [12,14]
20	0.0747 [0.073,0.076]	0.0010 [0.001,0.001]	0.0062 [0.006,0.007]	0.7076 [0.696,0.717]	0.0348 [0.034,0.035]	15.1 [15,16]	0.0746 [0.073,0.076]	0.0011 [0.001,0.001]	0.0063 [0.006,0.007]	0.6814 [0.664,0.695]	0.0347 [0.034,0.035]	18.0 [18,20]
50	0.0744 [0.072,0.077]	0.0011 [0.001,0.001]	0.0062 [0.006,0.007]	0.6858 [0.671,0.700]	0.0347 [0.034,0.035]	21.6 [20,23]	0.0744 [0.072,0.077]	0.0012 [0.001,0.001]	0.0062 [0.006,0.007]	0.6717 [0.643,0.691]	0.0347 [0.034,0.035]	27.0 [24,30]
100	0.0743 [0.071,0.078]	0.0011 [0.001,0.001]	0.0062 [0.006,0.007]	0.6709 [0.649,0.690]	0.0347 [0.034,0.035]	28.2 [26,31]	0.0743 [0.071,0.078]	0.0012 [0.001,0.001]	0.0062 [0.006,0.007]	0.6584 [0.626,0.684]	0.0347 [0.034,0.035]	35.7 [32,40]
200	0.0751 [0.071,0.079]	0.0012 [0.001,0.001]	0.0062 [0.006,0.007]	0.6542 [0.631,0.683]	0.0347 [0.034,0.035]	35.0 [31,39]	0.0751 [0.071,0.080]	0.0012 [0.001,0.001]	0.0062 [0.006,0.007]	0.6409 [0.603,0.681]	0.0346 [0.034,0.035]	44.0 [38,50]
	$\gamma = 0.25$						$\gamma = 0.25$					
10	0.0796 [0.078,0.082]	0.0012 [0.001,0.001]	0.0062 [0.006,0.007]	0.6355 [0.605,0.657]	0.0346 [0.034,0.035]	124.3 [121,127]	0.0699 [0.066,0.073]	0.0013 [0.001,0.002]	0.0061 [0.006,0.007]	0.6190 [0.566,0.658]	0.0345 [0.034,0.035]	148.7 [142,154]
20	0.0758 [0.074,0.078]	0.0013 [0.001,0.001]	0.0061 [0.006,0.007]	0.6242 [0.602,0.643]	0.0346 [0.034,0.035]	131.0 [125,137]	0.0721 [0.070,0.074]	0.0014 [0.001,0.002]	0.0060 [0.006,0.006]	0.6086 [0.567,0.647]	0.0345 [0.034,0.035]	150.3 [140,160]
50	0.0752 [0.072,0.077]	0.0014 [0.001,0.002]	0.0061 [0.006,0.007]	0.6020 [0.572,0.624]	0.0345 [0.034,0.035]	134.7 [127,142]	0.0749 [0.072,0.077]	0.0014 [0.001,0.002]	0.0060 [0.006,0.006]	0.5800 [0.529,0.622]	0.0344 [0.034,0.035]	159.0 [146,171]
100	0.0752 [0.072,0.078]	0.0014 [0.001,0.002]	0.0061 [0.006,0.006]	0.5879 [0.562,0.614]	0.0345 [0.034,0.035]	125.1 [116,132]	0.0752 [0.072,0.078]	0.0015 [0.001,0.002]	0.0060 [0.006,0.006]	0.5651 [0.516,0.611]	0.0344 [0.034,0.035]	146.7 [132,158]
200	0.0754 [0.071,0.079]	0.0014 [0.001,0.002]	0.0060 [0.005,0.007]	0.5781 [0.545,0.606]	0.0344 [0.034,0.035]	105.4 [99,113]	0.0755 [0.071,0.079]	0.0015 [0.001,0.002]	0.0060 [0.005,0.006]	0.5569 [0.498,0.604]	0.0344 [0.034,0.035]	121.7 [107,134]

Notes: Means and 95% confidence intervals. Unobserved heterogeneity is continuously distributed in the DGP, with underlying dimension equal to 2*. The number of groups K is estimated in every replication, with different choices for γ in (4).* 500 *simulations. Results for DGP 4.*

Table S14: Firm and worker effects, two-dimensional firm heterogeneity, large $Var(\psi)$, different choices of γ, iterated estimators

	iterated estimator						iterated estimator, bias corrected					
Firm size	Var (η_i)	Var (ψ_j)	Cov (η_i, ψ_j)	Corr (η_i, ψ_j)	Var (ε_{i1})	$\hat{K}$	Var (η_i)	Var (ψ_j)	Cov (η_i, ψ_j)	Corr (η_i, ψ_j)	Var (ε_{i1})	$\hat{K}$
	true values						true values					
-	0.0660	0.0204	0.0050	0.1373	0.0341		0.0660	0.0204	0.0050	0.1373	0.0341	
	$\gamma = 1.0$						$\gamma = 1.0$					
10	0.0661 [0.063,0.073]	0.0045 [0.002,0.006]	0.0050 [0.001,0.006]	0.2847 [0.115,0.329]	0.0501 [0.049,0.053]	4.0 [4,4]	0.0592 [0.055,0.073]	0.0079 [0.003,0.010]	0.0084 [0.001,0.010]	0.4058 [0.085,0.482]	0.0468 [0.045,0.052]	4.2 [4,5]
20	0.0632 [0.061,0.065]	0.0080 [0.007,0.009]	0.0066 [0.006,0.007]	0.2928 [0.265,0.319]	0.0466 [0.046,0.048]	5.7 [5,6]	0.0592 [0.055,0.063]	0.0117 [0.010,0.014]	0.0085 [0.007,0.011]	0.3192 [0.257,0.421]	0.0429 [0.040,0.045]	7.4 [6,8]
50	0.0608 [0.058,0.064]	0.0127 [0.012,0.013]	0.0077 [0.007,0.009]	0.2785 [0.251,0.312]	0.0420 [0.041,0.043]	8.9 [8,9]	0.0589 [0.056,0.062]	0.0163 [0.015,0.017]	0.0086 [0.007,0.010]	0.2702 [0.222,0.316]	0.0383 [0.037,0.040]	11.8 [10,12]
100	0.0617 [0.059,0.065]	0.0152 [0.014,0.016]	0.0074 [0.006,0.009]	0.2424 [0.207,0.286]	0.0392 [0.038,0.040]	13.3 [13,14]	0.0624 [0.059,0.065]	0.0179 [0.017,0.019]	0.0071 [0.006,0.009]	0.2044 [0.158,0.256]	0.0366 [0.036,0.038]	17.9 [17,20]
200	0.0628 [0.059,0.066]	0.0174 [0.017,0.018]	0.0064 [0.004,0.008]	0.1932 [0.139,0.242]	0.0371 [0.036,0.038]	21.4 [20,23]	0.0642 [0.061,0.068]	0.0196 [0.018,0.021]	0.0057 [0.004,0.008]	0.1552 [0.098,0.210]	0.0350 [0.034,0.036]	30.1 [28,33]
	$\gamma = 0.5$						$\gamma = 0.5$					
10	0.0549 [0.053,0.057]	0.0093 [0.009,0.010]	0.0106 [0.010,0.011]	0.4708 [0.451,0.489]	0.0452 [0.044,0.046]	12.8 [12,13]	0.0473 [0.038,0.053]	0.0141 [0.012,0.017]	0.0144 [0.012,0.018]	0.5393 [0.433,0.684]	0.0404 [0.037,0.042]	14.0 [12,15]
20	0.0555 [0.054,0.057]	0.0117 [0.011,0.012]	0.0102 [0.010,0.011]	0.4027 [0.385,0.424]	0.0429 [0.042,0.044]	16.1 [16,17]	0.0560 [0.054,0.058]	0.0141 [0.013,0.015]	0.0100 [0.009,0.011]	0.3396 [0.311,0.375]	0.0405 [0.039,0.041]	19.7 [19,22]
50	0.0579 [0.055,0.060]	0.0148 [0.014,0.015]	0.0092 [0.008,0.010]	0.3135 [0.288,0.349]	0.0399 [0.039,0.041]	26.0 [24,28]	0.0599 [0.057,0.062]	0.0170 [0.016,0.018]	0.0082 [0.007,0.009]	0.2436 [0.210,0.287]	0.0376 [0.037,0.039]	34.4 [30,38]
100	0.0605 [0.057,0.063]	0.0167 [0.016,0.018]	0.0077 [0.006,0.009]	0.2437 [0.206,0.281]	0.0378 [0.037,0.039]	38.8 [36,41]	0.0630 [0.060,0.066]	0.0188 [0.018,0.020]	0.0065 [0.005,0.008]	0.1780 [0.135,0.221]	0.0358 [0.035,0.037]	52.9 [48,57]
200	0.0631 [0.059,0.067]	0.0185 [0.018,0.019]	0.0067 [0.005,0.008]	0.1976 [0.147,0.248]	0.0362 [0.036,0.037]	53.2 [49,57]	0.0653 [0.061,0.069]	0.0200 [0.019,0.021]	0.0056 [0.004,0.008]	0.1505 [0.097,0.209]	0.0346 [0.034,0.035]	71.8 [65,78]
	$\gamma = 0.25$						$\gamma = 0.25$					
10	0.0653 [0.063,0.068]	0.0118 [0.011,0.012]	0.0129 [0.012,0.013]	0.4641 [0.443,0.481]	0.0428 [0.042,0.044]	125.6 [121,130]	0.0391 [0.034,0.045]	0.0141 [0.013,0.015]	0.0121 [0.011,0.013]	0.4642 [0.432,0.494]	0.0405 [0.040,0.041]	152.6 [145,160]
20	0.0574 [0.056,0.060]	0.0137 [0.013,0.014]	0.0114 [0.011,0.012]	0.4078 [0.391,0.425]	0.0409 [0.040,0.042]	138.3 [132,143]	0.0496 [0.047,0.053]	0.0155 [0.015,0.016]	0.0100 [0.009,0.011]	0.3537 [0.326,0.380]	0.0391 [0.038,0.040]	163.7 [153,172]
50	0.0583 [0.056,0.060]	0.0162 [0.016,0.017]	0.0090 [0.008,0.010]	0.2939 [0.262,0.321]	0.0384 [0.037,0.039]	154.0 [146,163]	0.0599 [0.057,0.063]	0.0180 [0.017,0.019]	0.0073 [0.006,0.008]	0.2117 [0.174,0.245]	0.0365 [0.036,0.037]	190.1 [177,204]
100	0.0618 [0.059,0.065]	0.0179 [0.017,0.019]	0.0074 [0.006,0.009]	0.2222 [0.186,0.270]	0.0366 [0.036,0.037]	151.1 [142,163]	0.0646 [0.062,0.068]	0.0195 [0.019,0.020]	0.0058 [0.004,0.008]	0.1549 [0.115,0.205]	0.0350 [0.034,0.036]	186.1 [172,202]
200	0.0642 [0.060,0.068]	0.0193 [0.018,0.020]	0.0061 [0.004,0.008]	0.1734 [0.127,0.225]	0.0354 [0.035,0.036]	133.7 [127,141]	0.0663 [0.062,0.070]	0.0205 [0.020,0.022]	0.0050 [0.003,0.007]	0.1306 [0.082,0.186]	0.0342 [0.033,0.035]	162.3 [151,176]

Notes: Means and 95% confidence intervals. Unobserved heterogeneity is continuously distributed in the DGP, with underlying dimension equal to 2*. The number of groups K is estimated in every replication, with different choices for γ in (4).* 500 *simulations. Results for DGP 3.*

Table S15: Firm and worker effects, two-dimensional firm heterogeneity, small $Var(\psi)$, different choices of γ, iterated estimators

	iterated estimator						iterated estimator, bias corrected					
Firm size	Var (η_i)	Var (ψ_j)	Cov (η_i, ψ_j)	Corr (η_i, ψ_j)	Var (ε_{i1})	$\hat{K}$	Var (η_i)	Var (ψ_j)	Cov (η_i, ψ_j)	Corr (η_i, ψ_j)	Var (ε_{i1})	$\hat{K}$
	true values						true values					
-	0.0758	0.0017	0.0057	0.4963	0.0341		0.0758	0.0017	0.0057	0.4963	0.0341	
	$\gamma = 1.0$						$\gamma = 1.0$					
10	0.0866 [0.085,0.088]	0.0000 [0.000,0.000]	0.0003 [0.000,0.000]	0.1290 [0.108,0.150]	0.0358 [0.035,0.036]	4.0 [4,4]	0.0867 [0.085,0.089]	0.0000 [-0.000,0.000]	0.0002 [0.000,0.000]	0.1169 [0.075,0.153]	0.0358 [0.035,0.036]	4.0 [4,4]
20	0.0845 [0.081,0.087]	0.0002 [0.000,0.000]	0.0013 [0.000,0.002]	0.2921 [0.149,0.420]	0.0356 [0.035,0.036]	5.5 [5,6]	0.0823 [0.078,0.087]	0.0004 [0.000,0.001]	0.0024 [0.000,0.005]	0.4556 [0.174,0.709]	0.0354 [0.035,0.036]	6.9 [6,8]
50	0.0791 [0.077,0.082]	0.0007 [0.001,0.001]	0.0041 [0.004,0.005]	0.5444 [0.516,0.573]	0.0352 [0.035,0.036]	8.0 [8,8]	0.0761 [0.073,0.079]	0.0010 [0.001,0.001]	0.0056 [0.005,0.006]	0.6584 [0.609,0.709]	0.0349 [0.034,0.036]	10.0 [10,10]
100	0.0775 [0.074,0.081]	0.0009 [0.001,0.001]	0.0050 [0.004,0.006]	0.6035 [0.583,0.627]	0.0349 [0.034,0.036]	11.1 [11,12]	0.0756 [0.072,0.079]	0.0011 [0.001,0.001]	0.0060 [0.005,0.007]	0.6600 [0.626,0.694]	0.0347 [0.034,0.035]	14.2 [14,16]
200	0.0759 [0.071,0.080]	0.0011 [0.001,0.001]	0.0055 [0.005,0.006]	0.6174 [0.589,0.648]	0.0348 [0.034,0.035]	15.2 [14,16]	0.0750 [0.070,0.079]	0.0012 [0.001,0.001]	0.0060 [0.005,0.007]	0.6345 [0.593,0.674]	0.0347 [0.034,0.035]	19.3 [17,21]
	$\gamma = 0.5$						$\gamma = 0.5$					
10	0.0799 [0.078,0.082]	0.0006 [0.000,0.001]	0.0037 [0.003,0.004]	0.5408 [0.520,0.559]	0.0353 [0.035,0.036]	12.2 [12,13]	0.0777 [0.075,0.080]	0.0008 [0.001,0.001]	0.0047 [0.004,0.006]	0.6157 [0.536,0.751]	0.0351 [0.034,0.036]	12.7 [12,14]
20	0.0780 [0.076,0.080]	0.0008 [0.001,0.001]	0.0046 [0.004,0.005]	0.5940 [0.578,0.613]	0.0351 [0.034,0.036]	15.1 [15,16]	0.0760 [0.074,0.078]	0.0010 [0.001,0.001]	0.0056 [0.005,0.006]	0.6482 [0.619,0.676]	0.0349 [0.034,0.035]	18.0 [18,20]
50	0.0760 [0.073,0.079]	0.0010 [0.001,0.001]	0.0054 [0.005,0.006]	0.6299 [0.613,0.650]	0.0349 [0.034,0.036]	21.6 [20,23]	0.0748 [0.072,0.077]	0.0011 [0.001,0.001]	0.0060 [0.005,0.007]	0.6568 [0.630,0.685]	0.0347 [0.034,0.035]	27.0 [24,30]
100	0.0753 [0.072,0.079]	0.0011 [0.001,0.001]	0.0057 [0.005,0.006]	0.6364 [0.612,0.660]	0.0348 [0.034,0.035]	28.2 [26,31]	0.0746 [0.071,0.078]	0.0012 [0.001,0.001]	0.0061 [0.006,0.007]	0.6457 [0.607,0.674]	0.0347 [0.034,0.035]	35.7 [32,40]
200	0.0756 [0.071,0.080]	0.0012 [0.001,0.001]	0.0059 [0.005,0.007]	0.6231 [0.594,0.656]	0.0347 [0.034,0.035]	35.0 [31,39]	0.0753 [0.071,0.080]	0.0013 [0.001,0.002]	0.0061 [0.006,0.007]	0.6137 [0.569,0.657]	0.0346 [0.034,0.035]	44.0 [38,50]
	$\gamma = 0.25$						$\gamma = 0.25$					
10	0.0781 [0.076,0.080]	0.0011 [0.001,0.001]	0.0059 [0.005,0.006]	0.6217 [0.602,0.642]	0.0347 [0.034,0.035]	124.3 [121,127]	0.0650 [0.057,0.072]	0.0014 [0.001,0.002]	0.0063 [0.006,0.007]	0.6481 [0.602,0.692]	0.0345 [0.034,0.035]	148.7 [142,154]
20	0.0755 [0.074,0.077]	0.0012 [0.001,0.001]	0.0059 [0.006,0.006]	0.6122 [0.586,0.635]	0.0346 [0.034,0.035]	131.0 [125,137]	0.0731 [0.071,0.075]	0.0014 [0.001,0.002]	0.0060 [0.006,0.006]	0.5983 [0.556,0.641]	0.0345 [0.034,0.035]	150.3 [140,160]
50	0.0754 [0.073,0.078]	0.0013 [0.001,0.002]	0.0060 [0.005,0.006]	0.5928 [0.568,0.619]	0.0345 [0.034,0.035]	134.7 [127,142]	0.0752 [0.073,0.078]	0.0014 [0.001,0.002]	0.0059 [0.005,0.007]	0.5713 [0.529,0.619]	0.0344 [0.034,0.035]	159.0 [146,171]
100	0.0754 [0.072,0.078]	0.0014 [0.001,0.002]	0.0060 [0.006,0.006]	0.5806 [0.547,0.609]	0.0345 [0.034,0.035]	125.1 [116,132]	0.0753 [0.072,0.078]	0.0015 [0.001,0.002]	0.0060 [0.006,0.006]	0.5597 [0.503,0.609]	0.0344 [0.034,0.035]	146.7 [132,158]
200	0.0755 [0.071,0.080]	0.0014 [0.001,0.002]	0.0060 [0.005,0.006]	0.5704 [0.532,0.602]	0.0344 [0.034,0.035]	105.4 [99,113]	0.0756 [0.071,0.080]	0.0015 [0.001,0.002]	0.0059 [0.005,0.007]	0.5499 [0.490,0.604]	0.0343 [0.034,0.035]	121.7 [107,134]

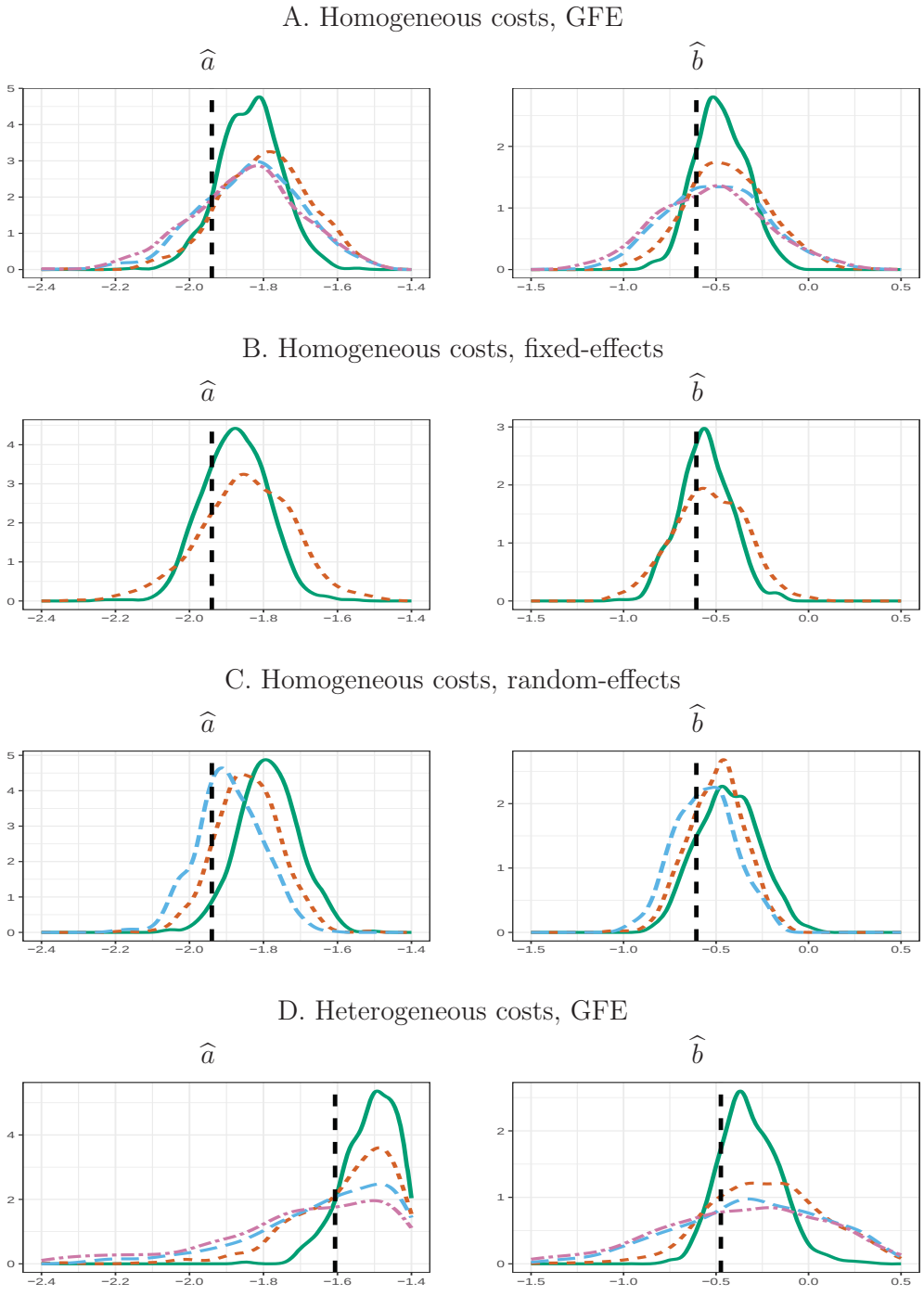
Notes: Means and 95% confidence intervals. Unobserved heterogeneity is continuously distributed in the DGP, with underlying dimension equal to 2*. The number of groups K is estimated in every replication, with different choices for γ in (4).* 500 *simulations. Results for DGP 4.*

Table S16: Firm and worker effects, discrete firm heterogeneity ($K^* = 10$)

Firm size	$\text{Var}(\eta_i)$	$\text{Var}(\psi_j)$	$\text{Cov}(\eta_i, \psi_j)$	$\text{Corr}(\eta_i, \psi_j)$	$\text{Var}(\varepsilon_{i1})$	% misclass.
	true values					
-	0.0758	0.0017	0.0057	0.4963	0.0341	
	two-step with $K = K^* = 10$					
10	0.0758 [0.074,0.077]	0.0013 [0.001,0.001]	0.0057 [0.005,0.006]	0.5770 [0.566,0.586]	0.0346 [0.034,0.035]	69.0% [0.678,0.705]
20	0.0758 [0.074,0.077]	0.0015 [0.001,0.002]	0.0057 [0.005,0.006]	0.5355 [0.525,0.546]	0.0344 [0.034,0.035]	58.5% [0.560,0.614]
50	0.0759 [0.075,0.077]	0.0016 [0.001,0.002]	0.0056 [0.005,0.006]	0.5083 [0.499,0.517]	0.0342 [0.034,0.035]	39.3% [0.338,0.476]
100	0.0759 [0.075,0.077]	0.0017 [0.001,0.002]	0.0056 [0.005,0.006]	0.4981 [0.489,0.507]	0.0342 [0.033,0.035]	22.6% [0.171,0.359]
200	0.0759 [0.074,0.077]	0.0017 [0.002,0.002]	0.0056 [0.005,0.006]	0.4945 [0.484,0.504]	0.0341 [0.033,0.035]	7.5% [0.050,0.115]
	bias corrected with estimated K					
10	0.0778 [0.076,0.079]	0.0013 [0.001,0.002]	0.0047 [0.004,0.005]	0.4527 [0.441,0.465]	0.0346 [0.034,0.035]	
20	0.0762 [0.075,0.078]	0.0016 [0.001,0.002]	0.0055 [0.005,0.006]	0.4917 [0.478,0.502]	0.0342 [0.034,0.035]	
50	0.0760 [0.075,0.077]	0.0017 [0.001,0.002]	0.0056 [0.005,0.006]	0.4906 [0.478,0.503]	0.0342 [0.033,0.035]	
100	0.0759 [0.074,0.077]	0.0017 [0.002,0.002]	0.0057 [0.005,0.006]	0.4909 [0.480,0.501]	0.0341 [0.033,0.035]	
200	0.0757 [0.074,0.077]	0.0018 [0.002,0.002]	0.0057 [0.005,0.006]	0.4930 [0.483,0.503]	0.0341 [0.033,0.035]	

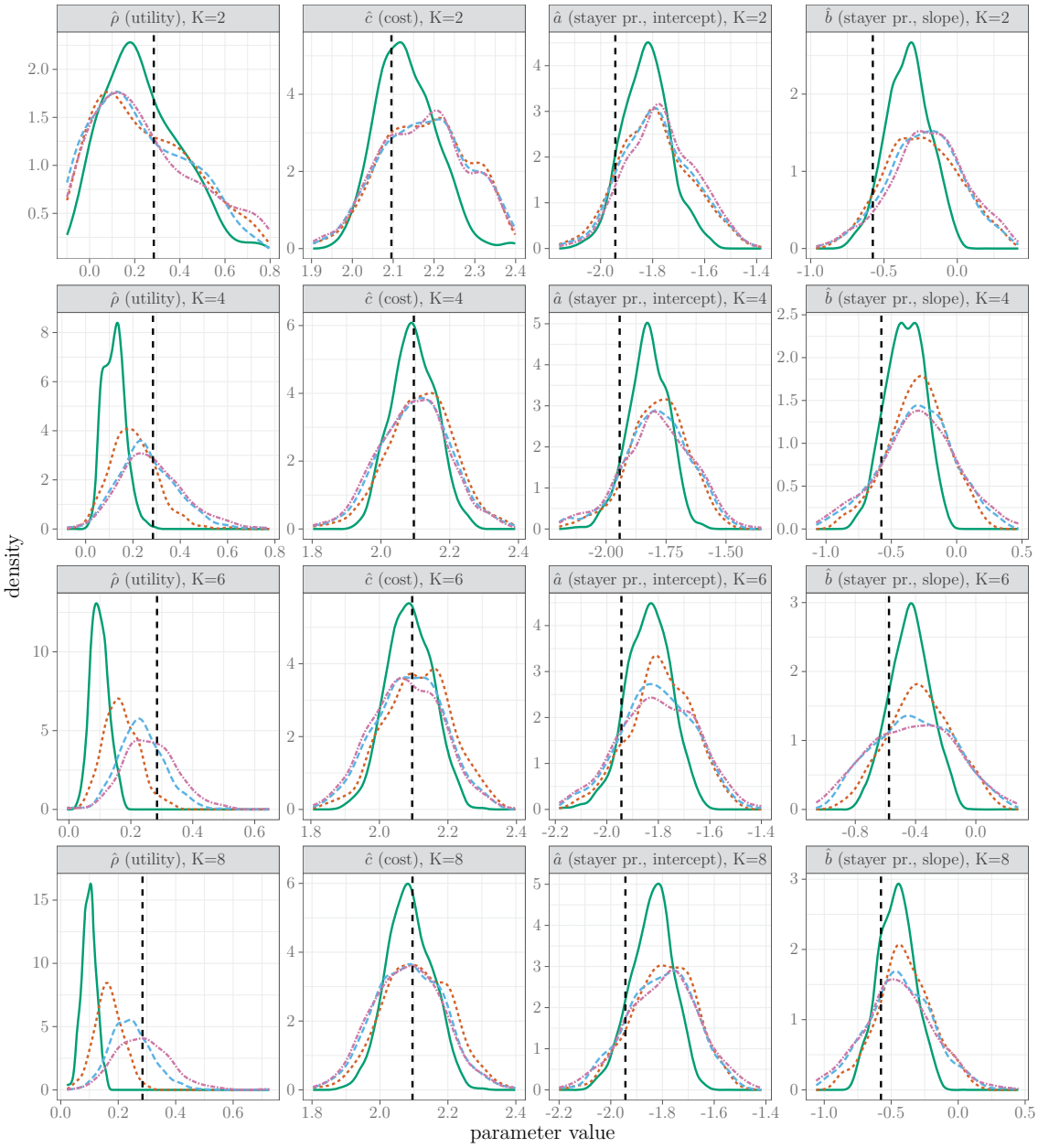
Notes: Means and 95% confidence intervals. Unobserved heterogeneity is discretely distributed in the DGP, with $K^* = 10$ groups. In the top panel the true number of groups is used. The last column shows frequencies of misclassification. In the bottom panel the number of groups is estimated in every replication. 500 *simulations. Results for DGP 5.*

Figure S1: Estimates of the parameters governing the probabilities of being a mover type across simulations



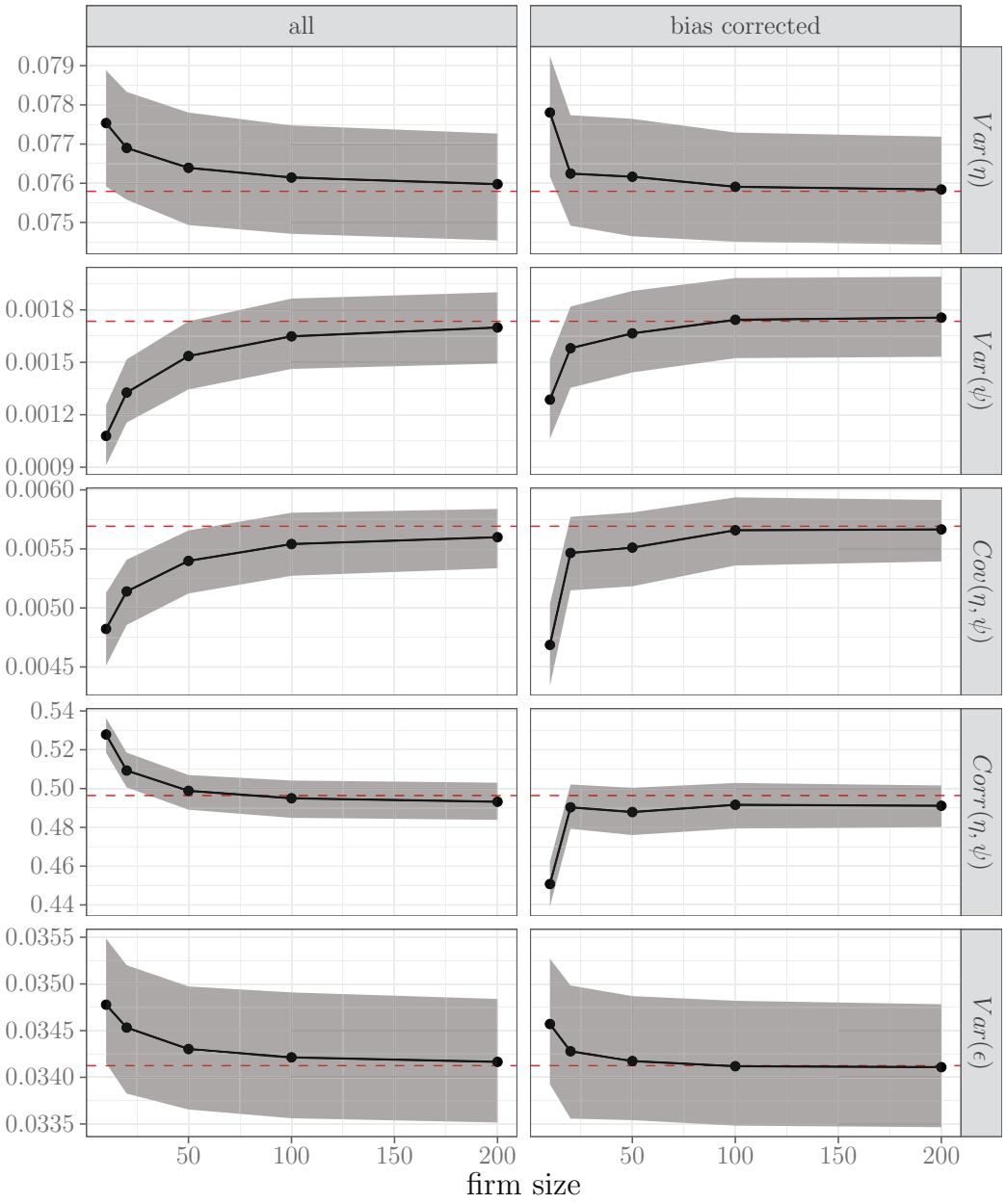
Notes: See notes to Figures 4 and 5. a and b are the intercept and slope in the probability of being a mover type.

Figure S2: Parameter estimates across simulations, fixed K



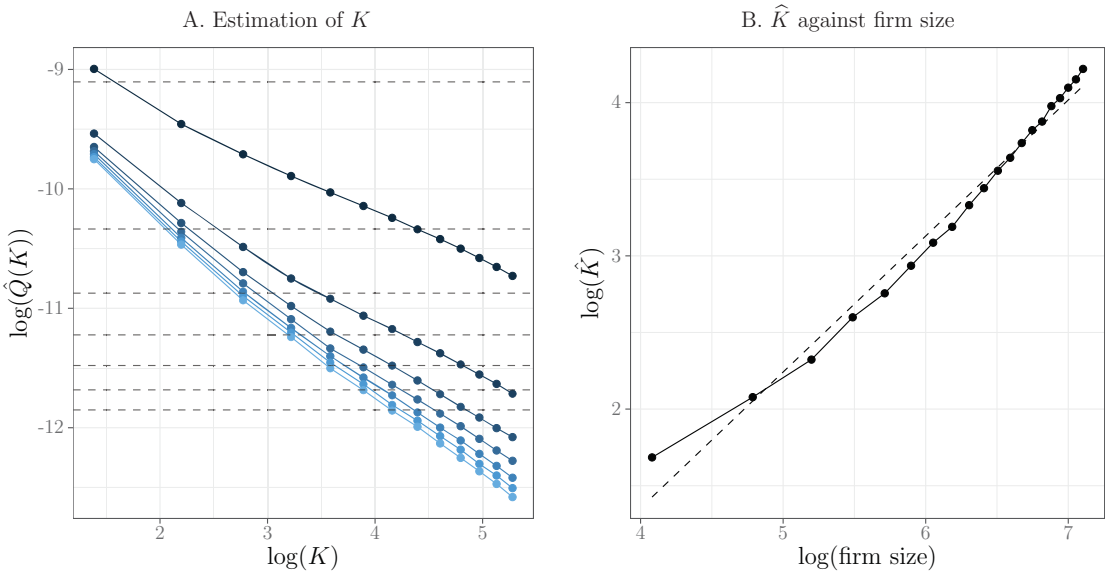
Notes: See note to Figure 4. K is kept fixed, and we focus on the model with homogeneous mobility costs. 500 *replications.*

Figure S3: Estimates of firm and worker heterogeneity across simulations



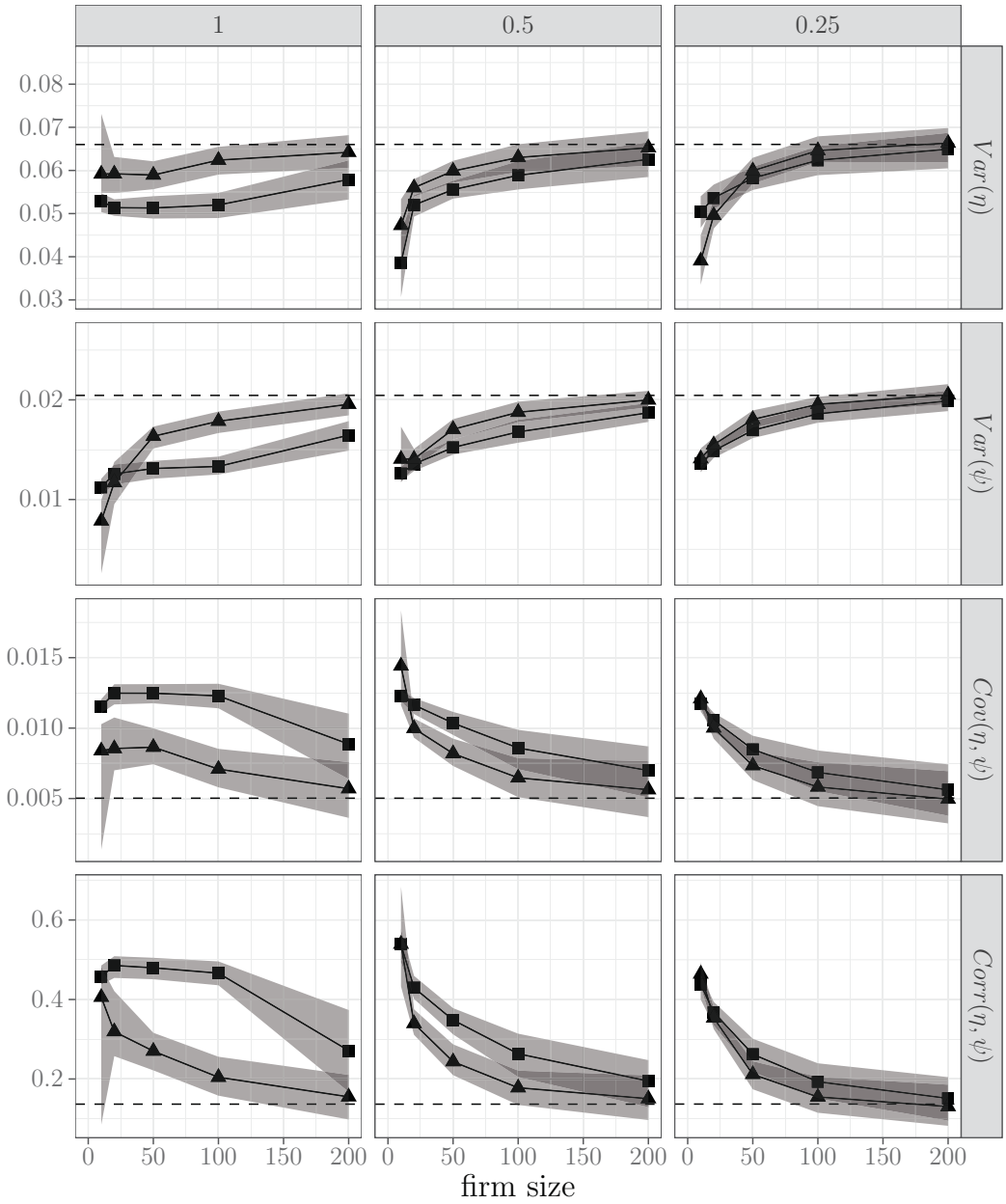
Notes: Means (solid line) and 95% confidence intervals. The dashed line indicates the true parameter value. Unobserved heterogeneity is continuously distributed in the DGP. The number of groups K is estimated in every replication. 500 *replications.*

Figure S4: Dimension of firm heterogeneity



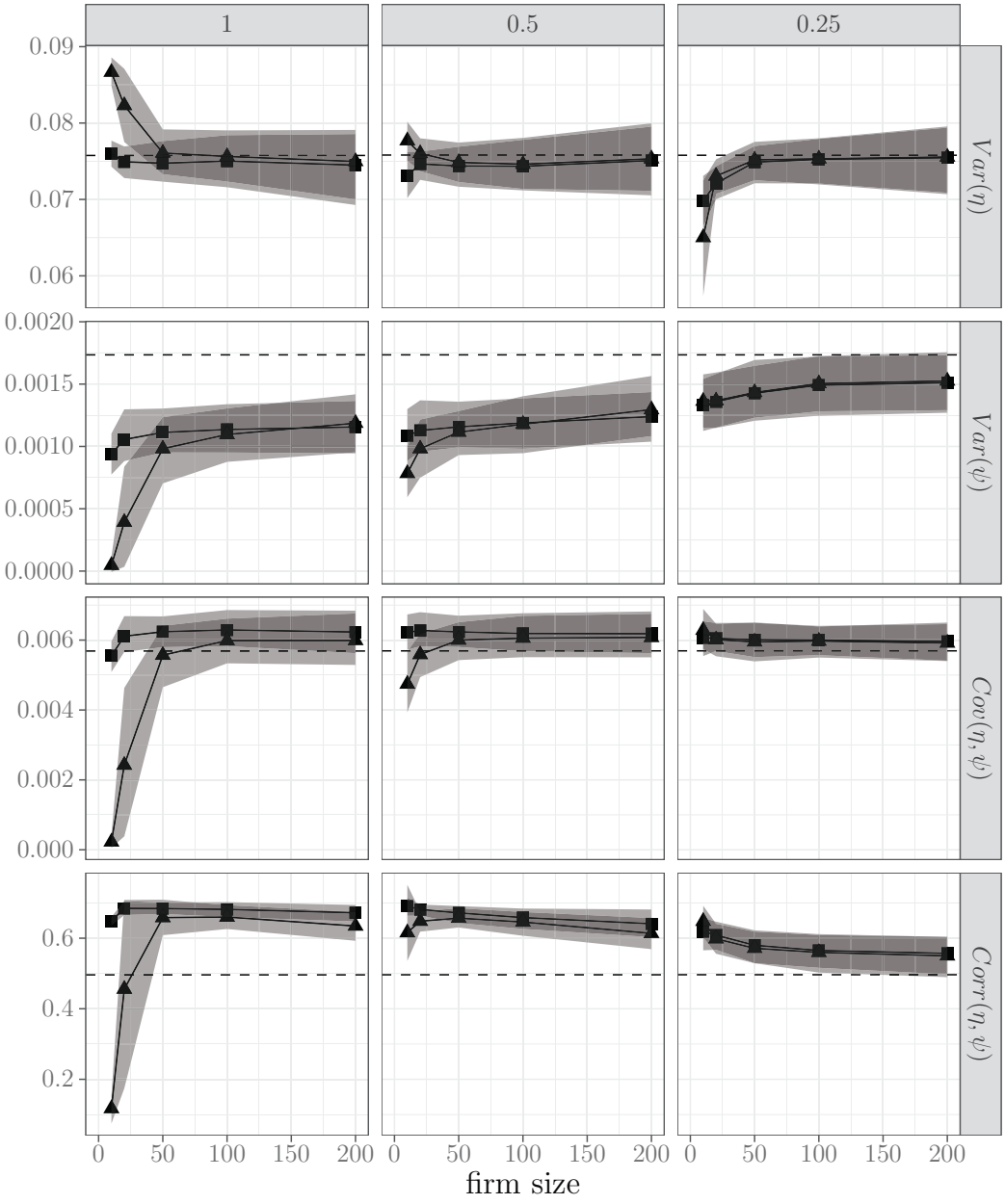
Notes: Source Swedish administrative data. Left graph shows the logarithm of $\widehat{Q}(K)$ *as a function of* K*, for different average firm sizes* S*. Horizontal lines show the corresponding value of* $\ln(\widehat{V}_h)$*. The right graph shows the relationship between the log of* $\widehat{K}$ *and the log of the average firm size in the sample, across samples.*

Figure S5: Estimates of firm and worker heterogeneity across simulations, two-dimensional firm heterogeneity, large variance of firm effects



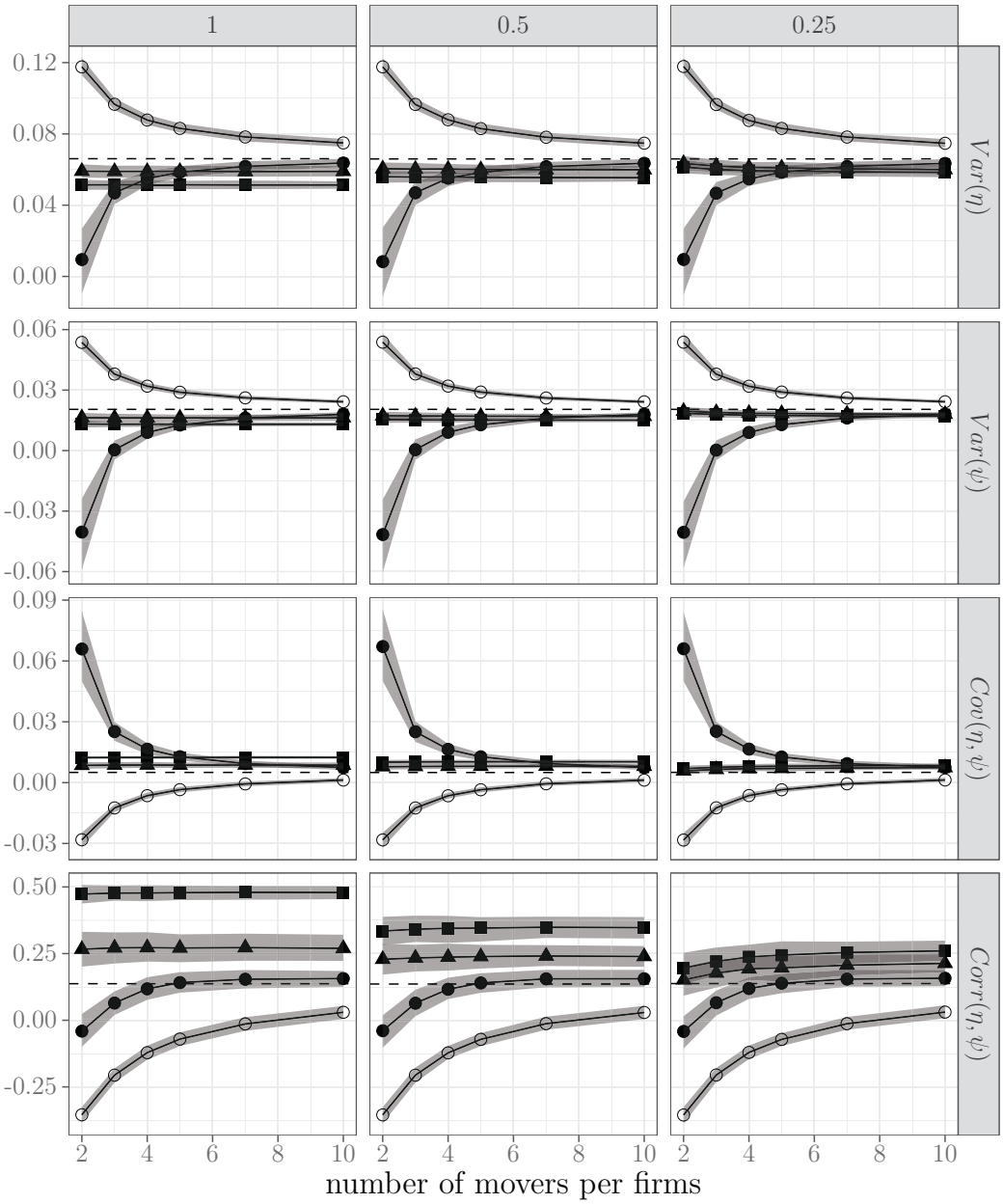
Notes: Means (solid line) and 95% confidence intervals. ■ *indicates the two-step bias-corrected GFE estimator and* ▲ *indicates the iterated bias-corrected GFE estimator. The different columns represent different values of* γ *(that is, different selection rules for the number of groups). Unobserved heterogeneity is continuously distributed in the DGP. The number of groups* K *is estimated in every replication.* 500 *replications. Results for DGP 3.*

Figure S6: Two-dimensional firm heterogeneity, small variance of firm effects



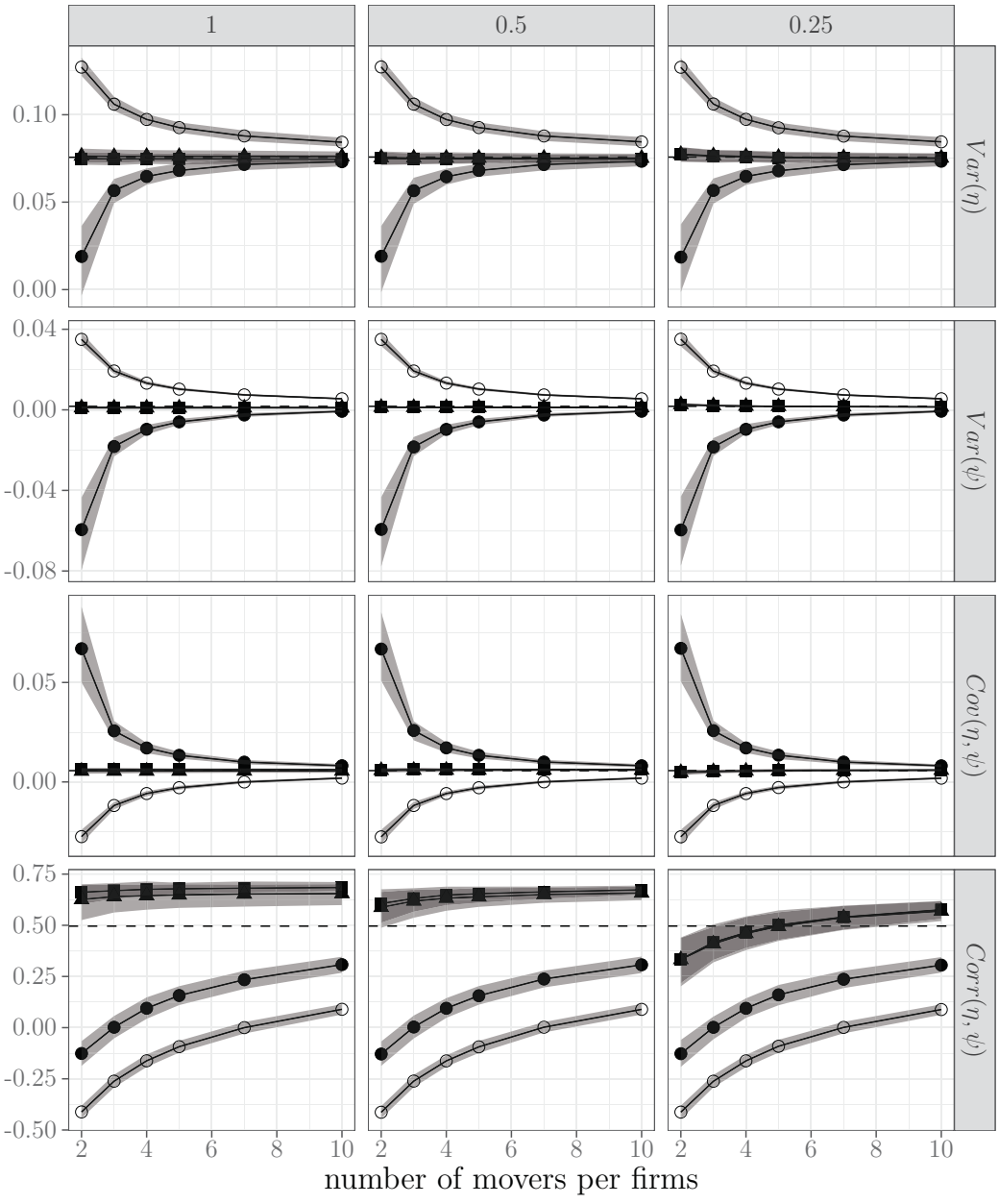
Notes: See the notes to Figure S5. Results for DGP 4.

Figure S7: Estimates of firm and worker heterogeneity across simulations, two-dimensional firm heterogeneity, large variance of firm effects, different number of job movers per firm



Notes: Means (solid line) and 95% confidence intervals. ■ *indicates the two-step bias-corrected GFE estimator,* ▲ *the iterated bias-corrected GFE estimator,* ○ *the fixed-effects estimator, and* ● *the bias-corrected fixed-effects estimator. The different columns represent different values of* γ *(that is, different selection rules for the number of groups). Unobserved heterogeneity is continuously distributed in the DGP. The number of groups* K *is estimated in every replication.* 500 *replications. Results for DGP 3.*

Figure S8: Estimates of firm and worker heterogeneity across simulations, two-dimensional firm heterogeneity, small variance of firm effects, different number of job movers per firm



Notes: See the notes to Figure S7. Results for DGP 4.