Supplementary Appendix to

“Discretizing Unobserved Heterogeneity”

Stéphane Bonhomme, Thibaut Lamadon and Elena Manresa

S1 Complements to the econometrics

S1.1 Conditions for Assumptions 2 and 3 in Example 2

Let us verify Assumptions 2 and 3 in Example 2 under the following conditions.

Assumption S1. *(regularity in Example 2)*

(i) Observations are i.i.d. across individuals conditional on the α_{i0}'s and μ_{i0}'s. The parameter spaces Θ for $\theta_0 = (\rho_0, \beta_0')'$ and $\mathcal{A}$ for $(\alpha_{i0}, \mu_{i0}')'$ are compact, and θ_0 belongs to the interior of Θ.

(ii) $|\rho_0| < 1$, and $(Y_{it}, X_{it}')'$ is stationary for every i. $\mathbb{E}(U_{it}) = 0$, $\mathbb{E}(U_{it}Y_{i,t-1}) = 0$, and $\mathbb{E}(U_{it}X_{it}) = 0$. In addition, letting $W_{it} = (Y_{i,t-1}, X_{it}')'$, the minimum eigenvalue of $\mathbb{E}\left((W_{it} - \mathbb{E}(W_{it}))\,(W_{it} - \mathbb{E}(W_{it}))'\right)$ is bounded away from zero.

(iii) Let $V_{it} = X_{it} - \mu_{i0}$. $\mathbb{E}(V_{it}) = 0$. Moreover, for every i, $Z_{it} = (U_{it}, V_{it}')'$ is a stationary mixing sequence such that, for some $0 < a < 1$ and $C > 0$:

$$\sup_i \left| \sup_t \sup_{B \in \mathcal{B}_t^i, D \in \mathcal{D}_{t+m}^i} |\Pr(B \cap D) - \Pr(B)\Pr(D)| \right| \leq C a^m,$$

where $\mathcal{B}_t^i$ and $\mathcal{D}_t^i$ denote the sigma-algebras generated by $(Z_{it}, Z_{it-1}, ...)$ and $(Z_{it}, Z_{it+1}, ...)$, respectively. Z_{it} has finite $(8+\eta)$ moments uniformly in i, t, for some $\eta > 0$. $N = O(T)$.

Consider the quasi-likelihood function: $\ell_i(\alpha_i, \theta) = -\frac{1}{2T}\sum_{t=1}^{T}(Y_{it} - \rho Y_{i,t-1} - X_{it}'\beta - \alpha_i)^2$. Third-order differentiability in Assumption 2 (i) is immediate. Furthermore we have, using stationarity:

$$\begin{aligned}&\mathbb{E}\left(\ell_i(\alpha_{i0},\theta_0)-\ell_i(\alpha_i,\theta)\right)\\&\quad=\frac{1}{2}\mathbb{E}\left(2U_{it}\left(W_{it}'(\theta_0-\theta)+\alpha_{i0}-\alpha_i\right)+\left(W_{it}'(\theta_0-\theta)+\alpha_{i0}-\alpha_i\right)^2\right)\\&\quad=\frac{1}{2}\mathbb{E}\left(\left(W_{it}'(\theta_0-\theta)+\alpha_{i0}-\alpha_i\right)^2\right).\end{aligned}$$

Using Assumption S1 (ii) thus implies the first condition in Assumption 2 (ii).

Next we have: $\mathbb{E}\left(v_i\left(\alpha_i,\theta\right)\right)=\frac{1-\rho}{1-\rho_0}\left(\alpha_{i0}+\beta_0'\mu_{i0}\right)-\beta'\mu_{i0}-\alpha_i$, so:

$$\overline{\alpha}_i(\theta)=\frac{1-\rho}{1-\rho_0}\alpha_{i0}+\left(\frac{1-\rho}{1-\rho_0}\beta_0-\beta\right)'\mu_{i0},$$

and $\overline{\alpha}_i(\theta)$ is unique. Moreover $v_i^\alpha=-1$, so $\inf_i\inf_\theta\mathbb{E}(-\frac{\partial^2\ell_i(\overline{\alpha}_i(\theta),\theta)}{\partial\alpha_i\partial\alpha_i'})=1$. Finally, the function $\frac{1}{N}\sum_{i=1}^N\mathbb{E}(\ell_i(\overline{\alpha}_i(\theta),\theta))$ is quadratic in $\theta=(\rho,\beta')'$, and its partial derivatives with respect to ρ and β are, respectively:

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

and:

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

It is easy to verify that those are zero at θ_0. Moreover, the second derivative $-H$ is negative definite by Assumption S1 (ii). This completes the verification of Assumption 2 (ii).

Next, since $(U_{it},V_{it}')'$ has finite second moments and $\mathcal{A}$ and Θ are compact it is easy to see that $\sup_i\sup_{(\alpha_i,\theta)}|\mathbb{E}(\ell_i(\alpha_i,\theta))|=O(1)$, and similarly for the first three derivatives of ℓ_i. From the assumptions on time-series mixing and moment existence it follows (as in Lemma 1 in Hahn and Kuersteiner, 2011) that, for all (α_i,θ), $\max_{i=1,\ldots,N}|\ell_i(\alpha_i,\theta)-\mathbb{E}\left(\ell_i(\alpha_i,\theta)\right)|=o_p\left(1\right)$. Combining the latter lemma with the compactness of the parameter space as in Lemma 4 in Hahn and Kuersteiner (2011) one can show that $\max_{i=1,\ldots,N}\sup_{(\alpha_i,\theta)}|\ell_i(\alpha_i,\theta)-\mathbb{E}\left(\ell_i(\alpha_i,\theta)\right)|=o_p\left(1\right)$. The same argument can be applied to all first three derivatives of ℓ_i. Moreover, the rate on $\frac{1}{N}\sum_{i=1}^N(\ell_i(\alpha_{i0},\theta_0)-\mathbb{E}(\ell_i(\alpha_{i0},\theta_0)))^2$, and the corresponding rates on the derivatives of ℓ_i, come from the fact that $(U_{it},V_{it}')'$ has finite second moments and satisfies suitable mixing conditions.

Next, we have:

$$\begin{aligned}&\mathbb{E}_{(\alpha,\mu)}\left(v_i(\overline{\alpha}_i(\theta),\theta)\right)=\mathbb{E}_{(\alpha,\mu)}\left(Y_{it}-\frac{\alpha_{i0}+\mu_{i0}'\beta_0}{1-\rho_0}-\rho\left(Y_{i,t-1}-\frac{\alpha_{i0}+\mu_{i0}'\beta_0}{1-\rho_0}\right)-(X_{it}-\mu_{i0})'\beta\right)\\&\quad=(1-\rho)\frac{\alpha-\alpha_{i0}+(\mu-\mu_{i0})'\beta_0}{1-\rho_0}-(\mu-\mu_{i0})'\beta,\end{aligned}$$

so:

$$\frac{\partial}{\partial \alpha}\bigg|_{(\alpha_{i0},\mu_{i0})} \mathbb{E}_{(\alpha,\mu)}\left(v_i(\overline{\alpha}_i(\theta),\theta)\right) = \frac{1-\rho}{1-\rho_0}, \quad \frac{\partial}{\partial \mu}\bigg|_{(\alpha_{i0},\mu_{i0})} \mathbb{E}_{(\alpha,\mu)}\left(v_i(\overline{\alpha}_i(\theta),\theta)\right) = \frac{1-\rho}{1-\rho_0}\beta_0 - \beta,$$

which are uniformly bounded. Likewise:

$$\mathbb{E}_{(\alpha,\mu)}\left(v_i^{\theta}(\alpha_{i0},\theta_0)\right) = \left(-\mathbb{E}_{(\alpha,\mu)}(Y_{i,t-1})\,,\,-\mathbb{E}_{(\alpha,\mu)}(X_{it})'\right)' = \left(-\frac{\alpha+\mu'\beta_0}{1-\rho_0}\,,\,-\mu'\right)',$$

and $\mathbb{E}_{(\alpha,\mu)}\left(v_i^{\alpha}(\alpha_{i0},\theta_0)\right) = -1$, both of which have uniformly bounded derivatives. This shows the last part of Assumption 2 (iii).

Turning to the part (iv) in Assumption 2 we have:

$$\widehat{\alpha}(\widehat{k}_i,\theta) = \overline{Y}(\widehat{k}_i) - \rho\overline{Y}_{-1}(\widehat{k}_i) - \overline{X}(\widehat{k}_i)'\beta,$$

where $\overline{Y}(\widehat{k}_i)$, $\overline{Y}_{-1}(\widehat{k}_i)$ and $\overline{X}(\widehat{k}_i)$ are group-specific means of Y_{it}, $Y_{i,t-1}$ and X_{it}, over individuals and time periods. This implies that $\widehat{\ell}_i(\theta) = \ell_i(\widehat{\alpha}(\widehat{k}_i,\theta),\theta)$ is quadratic in θ. Assumption 2 (iv) directly follows.

Finally, when using $h_i = (\overline{Y}_i, \overline{X}_i')'$ in the classification step, it follows from the expressions in the main text that φ is injective and both φ and ψ are Lipschitz, since $|\rho_0| < 1$. This shows that Assumption 3 holds.

S1.2 Sequential estimation based on a partial likelihood

Consider the following grouped fixed-effects estimator. Instead of jointly maximizing the likelihood function in the second step, one sequentially estimates (α_1, θ_1) based on $\sum_{i=1}^N \ell_{i1}(\alpha_{i1},\theta_1)$, and (α_2,θ_2) given (α_1,θ_1) based on $\sum_{i=1}^N \ell_{i2}(\alpha_{i1},\alpha_{i2},\theta_1,\theta_2)$. Under similar assumptions as in Theorem 1, $\widehat{\theta}_1$ and $\widehat{\alpha}_1(\widehat{k}_i)$ follow the same expansions as in (6) and (7) in Theorem 1, up to adapting the notation. The grouped fixed-effects estimator of α_2's and θ_2 is then:

$$\left(\widehat{\theta}_2, \widehat{\alpha}_2\right) = \underset{(\theta_2,\alpha_2)}{\operatorname{argmax}} \sum_{i=1}^N \ell_{i2}\left(\widehat{\alpha}_1\left(\widehat{k}_i\right), \alpha_2\left(\widehat{k}_i\right), \widehat{\theta}_1, \theta_2\right).$$

Next, let $\widehat{\alpha}_{i2}\left(\alpha_1,\theta_1,\theta_2\right) = \operatorname{argmax}_{\alpha_2}\, \ell_{i2}\left(\alpha_1,\alpha_2,\theta_1,\theta_2\right)$. Replacing ℓ_i by:

$$\widehat{\ell}_{i2}\left(\alpha_{i2},\theta_2\right) = \ell_{i2}\left(\widehat{\alpha}_1\left(\widehat{k}_i\right), \alpha_{i2}, \widehat{\theta}_1, \theta_2\right)$$

in the proof of Theorem 1, we obtain the following counterpart to (A3):

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\widehat{\ell}_{i2}(\widehat{\alpha}_2(\widehat{k}_i,\theta_{20}),\theta_{20})}{\partial\theta_2} &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial\ell_{i2}\left(\widehat{\alpha}_1\left(\widehat{k}_i\right),\widehat{\alpha}_{i2}\left(\widehat{\alpha}_1\left(\widehat{k}_i\right),\widehat{\theta}_1,\theta_{20}\right),\widehat{\theta}_1,\theta_{20}\right)}{\partial\theta_2}+O_p\left(\delta\right)\\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial\ell_{i2}\left(\alpha_{i10},\alpha_{i20},\theta_{10},\theta_{20}\right)}{\partial\theta_2}+\frac{\partial^2\ell_{i2}\left(\alpha_{i10},\alpha_{i20},\theta_{10},\theta_{20}\right)}{\partial\theta_2\partial\alpha_{i1}'}\left(\widehat{\alpha}_{i1}-\alpha_{i10}\right)\\
&\quad+\frac{\partial^2\ell_{i2}\left(\alpha_{i10},\alpha_{i20},\theta_{10},\theta_{20}\right)}{\partial\theta_2\partial\theta_1'}\left(\widehat{\theta}_1-\theta_{10}\right)\\
&\quad+\frac{\partial^2\ell_{i2}\left(\alpha_{i10},\alpha_{i20},\theta_{10},\theta_{20}\right)}{\partial\theta_2\partial\alpha_{i2}'}\left(\frac{\partial\widehat{\alpha}_{i2}\left(\alpha_{i10},\theta_{10},\theta_{20}\right)}{\partial\alpha_{i1}'}\left(\widehat{\alpha}_{i1}-\alpha_{i10}\right)\right.\\
&\quad\left.+\frac{\partial\widehat{\alpha}_{i2}\left(\alpha_{i10},\theta_{10},\theta_{20}\right)}{\partial\theta_1'}\left(\widehat{\theta}_1-\theta_{10}\right)+\widehat{\alpha}_{i2}\left(\alpha_{i10},\theta_{10},\theta_{20}\right)-\alpha_{i20}\right)+O_p\left(\delta\right),
\end{aligned}$$

where the last identity follows as in the proof of Theorem 1 (see also (A32) in the proof of Corollary 3).

We also have the following counterpart to (A4):

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\left.\frac{\partial^2}{\partial\theta_2\partial\theta_2'}\right|_{\theta_{20}}\widehat{\ell}_{i2}\left(\widehat{\alpha}_2(\widehat{k}_i,\theta_2),\theta_2\right) &= \frac{1}{N}\sum_{i=1}^{N}\left.\frac{\partial^2}{\partial\theta_2\partial\theta_2'}\right|_{\theta_{20}}\widehat{\ell}_{i2}\left(\widehat{\alpha}_{i2}\left(\widehat{\alpha}_1\left(\widehat{k}_i\right),\widehat{\theta}_1,\theta_{20}\right),\theta_2\right)+o_p(1)\\
&= \frac{1}{N}\sum_{i=1}^{N}\left.\frac{\partial^2}{\partial\theta_2\partial\theta_2'}\right|_{\theta_{20}}\ell_{i2}\left(\alpha_{i10},\widehat{\alpha}_{i2}\left(\alpha_{i10},\theta_{10},\theta_2\right),\theta_{10},\theta_2\right)+o_p(1).
\end{aligned}$$

Let us define, omitting references to true values for conciseness:

$$\begin{aligned}
s_{i1} &= \frac{\partial\ell_{i1}}{\partial\theta_1}+\mathbb{E}\left(\frac{\partial^2\ell_{i1}}{\partial\theta_1\partial\alpha_{i1}'}\right)\left[\mathbb{E}\left(-\frac{\partial^2\ell_{i1}}{\partial\alpha_{i1}\partial\alpha_{i1}'}\right)\right]^{-1}\frac{\partial\ell_{i1}}{\partial\alpha_{i1}},\\
H_1 &= \lim_{N,T\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(-\frac{\partial^2\ell_{i1}}{\partial\theta_1\partial\theta_1'}\right)-\mathbb{E}\left(\frac{\partial^2\ell_{i1}}{\partial\theta_1\partial\alpha_{i1}'}\right)\left[\mathbb{E}\left(-\frac{\partial^2\ell_{i1}}{\partial\alpha_{i1}\partial\alpha_{i1}'}\right)\right]^{-1}\mathbb{E}\left(\frac{\partial^2\ell_{i1}}{\partial\alpha_{i1}\partial\theta_1'}\right),\\
s_{i2} &= \frac{\partial\ell_{i2}}{\partial\theta_2}+\mathbb{E}\left(\frac{\partial^2\ell_{i2}}{\partial\theta_2\partial\alpha_{i1}'}\right)\left[\mathbb{E}\left(-\frac{\partial^2\ell_{i1}}{\partial\alpha_{i1}\partial\alpha_{i1}'}\right)\right]^{-1}\frac{\partial\ell_{i1}}{\partial\alpha_{i1}}+\mathbb{E}\left(\frac{\partial^2\ell_{i2}}{\partial\theta_2\partial\theta_1'}\right)H_1^{-1}\frac{1}{N}\sum_{j=1}^{N}s_{j1}\\
&\quad+\mathbb{E}\left(\frac{\partial^2\ell_{i2}}{\partial\theta_2\partial\alpha_{i2}'}\right)\left[\mathbb{E}\left(-\frac{\partial^2\ell_{i2}}{\partial\alpha_{i2}\partial\alpha_{i2}'}\right)\right]^{-1}\left(\frac{\partial\ell_{i2}}{\partial\alpha_{i2}}+\mathbb{E}\left(\frac{\partial^2\ell_{i2}}{\partial\alpha_{i2}\partial\alpha_{i1}'}\right)\left[\mathbb{E}\left(-\frac{\partial^2\ell_{i1}}{\partial\alpha_{i1}\partial\alpha_{i1}'}\right)\right]^{-1}\frac{\partial\ell_{i1}}{\partial\alpha_{i1}}\right.\\
&\quad\left.+\mathbb{E}\left(\frac{\partial^2\ell_{i2}}{\partial\alpha_{i2}\partial\theta_1'}\right)H_1^{-1}\frac{1}{N}\sum_{j=1}^{N}s_{j1}\right),\\
H_2 &= \lim_{N,T\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(-\frac{\partial^2\ell_{i2}}{\partial\theta_2\partial\theta_2'}\right)-\mathbb{E}\left(\frac{\partial^2\ell_{i2}}{\partial\theta_2\partial\alpha_{i2}'}\right)\left[\mathbb{E}\left(-\frac{\partial^2\ell_{i2}}{\partial\alpha_{i2}\partial\alpha_{i2}'}\right)\right]^{-1}\mathbb{E}\left(\frac{\partial^2\ell_{i2}}{\partial\alpha_{i2}\partial\theta_2'}\right).
\end{aligned}$$

We thus have:

$$\begin{aligned}
\widehat{\theta}_1 &= \theta_{10} + H_1^{-1}\frac{1}{N}\sum_{i=1}^{N} s_{i1} + O_p\left(\frac{1}{T}\right) + O_p\left(B_\alpha(K)\right) + o_p\left(\frac{1}{\sqrt{NT}}\right), \\
\widehat{\theta}_2 &= \theta_{20} + H_2^{-1}\frac{1}{N}\sum_{i=1}^{N} s_{i2} + O_p\left(\frac{1}{T}\right) + O_p\left(B_\alpha(K)\right) + o_p\left(\frac{1}{\sqrt{NT}}\right).
\end{aligned}$$

S1.3 Properties of classification based on empirical distributions

Let $F_i(w) = \Pr\left(W_{it} \leq w \,|\, \alpha_{i0}\right) = G(w;\alpha_{i0})$ denote the population cdf of W_{it}.[1] Similarly as in Lemma 1, the following convergence rate is achieved:

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

provided (i) $\frac{1}{N}\sum_{i=1}^{N}\|\widehat{F}_i - F_i\|_\omega^2 = O_p(T^{-1})$, and (ii) $G(\cdot;\alpha_{i0})$ is Lipschitz with respect to its second argument. Both conditions are satisfied quite generally. For (i) 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 $\int \frac{\partial \ln f(y,x,\alpha_i)}{\partial \alpha_i}\frac{\partial \ln f(y,x,\alpha_i)}{\partial \alpha_i'} f(y,x,\alpha_i)dydx$ is uniformly bounded. Here $\alpha_i \mapsto G(\cdot;\alpha_i)$ maps individual-specific parameters to $L^2(\omega)$.

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, $\alpha_i \mapsto G(\cdot,\alpha_i)$ being injective is equivalent to $G(\cdot,\alpha_{i2}) = G(\cdot,\alpha_{i1}) \Rightarrow \alpha_{i2} = \alpha_{i1}$, which in turn is equivalent to α_{i0} being identified given knowledge of the function G (hence in particular given knowledge of θ_0).

S1.4 Iterated grouped fixed-effects estimator

In this subsection we consider a fully specified likelihood model, where $f_i(Y_i, X_i)$ is indexed by α_{i0}. We have the following result for $\widehat{\theta}^{(2)}$ in (10). Similar results hold for $\widehat{\alpha}^{(2)}(\widehat{k}_i^{(2)})$ and average effects, although we omit them for brevity.

Corollary S1. *Let the assumptions of Theorem 1 hold. Let $\widehat{\theta}$ be the two-step grouped fixed-effects estimator of θ_0. Then, as N, T, K tend to infinity:*

$$\widehat{\theta}^{(2)} = \widehat{\theta} + O_p\left(\frac{1}{T}\right) + O_p\left(B_\alpha(K)\right) + o_p\left(\frac{1}{\sqrt{NT}}\right).$$

[1] In conditional models where the data also depends on μ_{i0} we will write $F_i(w) = G(w;\alpha_{i0},\mu_{i0})$.

Proof. Let $\delta = 1/T + B_\alpha(K)$. We start by noting that, by definition of $\{\widehat{k}_i^{(2)}\}$:

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

By (A22) we have:

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

from which it follows that:

$$0 \leq \frac{1}{N}\sum_{i=1}^N \ell_i\left(\widehat{\alpha}_i, \widehat{\theta}\right) - \frac{1}{N}\sum_{i=1}^N \ell_i\left(\widehat{\alpha}(\widehat{k}_i^{(2)}), \widehat{\theta}\right) = O_p(\delta).$$

Then, following the first part of the proof of Theorem 1 (using that $\widehat{\theta}$ is consistent for θ_0) we then obtain, similarly as in (A12):

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

Hence: $\frac{1}{N}\sum_{i=1}^N \|\widehat{\alpha}(\widehat{k}_i^{(2)}) - \alpha_{i0}\|^2 = O_p(\delta)$. This establishes that there exists a function of $\{\widehat{k}_i^{(2)}\}$ which approximates the true α_{i0} on average at the desired rate.

Let us then define: $a(k,\theta) = \overline{\alpha}(\theta, \widehat{\alpha}(k))$. Note that:

$$\frac{1}{N}\sum_{i=1}^N \ell_i\left(a(\widehat{k}_i^{(2)},\theta),\theta\right) \leq \frac{1}{N}\sum_{i=1}^N \ell_i\left(\widehat{\alpha}^{(2)}(\widehat{k}_i^{(2)},\theta),\theta\right) \leq \frac{1}{N}\sum_{i=1}^N \ell_i\left(\widehat{\alpha}_i(\theta),\theta\right) = \frac{1}{N}\sum_{i=1}^N \ell_i\left(\overline{\alpha}_i(\theta),\theta\right) + O_p\left(\frac{1}{T}\right).$$

The rest of the proof is identical as in the proof of Theorem 1, up to a change in notation consisting in adding (2) superscripts.

∎

S1.5 Bias of the one-step estimator in Example 2

Write (4) in compact form as $Y_{it} = W_{it}'\theta_0 + \alpha_{i0} + U_{it}$, where $W_{it} = (Y_{i,t-1}, X_{it}')'$ and $\theta_0 = (\rho_0, \beta_0')'$. Pollard (1981, 1982a) provides conditions under which, for fixed K, T and as N tends to infinity, the one-step grouped fixed-effects estimator is root-N consistent and asymptotically normal for the minimizer θ^* of the following population objective function:[2]

$$Q(\theta) = \operatorname*{plim}_{N\to\infty} \min_{(\alpha,\{k_i\})} \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \left(Y_{it} - W_{it}'\theta - \alpha(k_i)\right)^2.$$

[2] Pollard focuses on the standard kmeans estimator, without covariates. See the supplementary appendix in Bonhomme and Manresa (2015) for an analysis with covariates.

Now:

$$Q(\theta) = \operatorname*{plim}_{N\to\infty} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(Y_{it}-\overline{Y}_i-\left(W_{it}-\overline{W}_i\right)'\theta\right)^2 + Q_B(\theta),$$

where:

$$Q_B(\theta) = \operatorname*{plim}_{N\to\infty} \min_{(\alpha,\{k_i\})} \frac{1}{N}\sum_{i=1}^{N}\left(\overline{Y}_i-\overline{W}_i'\theta-\alpha(k_i)\right)^2.$$

From Theorem 6.2 in Graf and Luschgy (2000) we have, as K tends to infinity for fixed T, and provided the density f_θ of $\overline{Y}_i-\overline{W}_i'\theta$ is non-singular with respect to the Lebesgue measure:

$$Q_B(\theta) = \frac{1}{12K^2}\left(\int [f_\theta(y)]^{\frac{1}{3}}\,dy\right)^3 + o\left(\frac{1}{K^2}\right).$$

As an example, consider the case where the $\overline{Y}_i-\overline{W}_i'\theta$ are i.i.d. normal with mean $\mu(\theta)$ and variance $\sigma^2(\theta)$. Then direct calculations show that:

$$\frac{1}{12K^2}\left(\int [f_\theta(y)]^{\frac{1}{3}}\,dy\right)^3 = \frac{\pi\sqrt{3}}{2K^2}\sigma^2(\theta).$$

Moreover:

$$\frac{\partial\sigma^2(\theta)}{\partial\theta} = 2\operatorname{Var}(\overline{W}_i)\theta - 2\operatorname{Cov}\left(\overline{W}_i,\overline{Y}_i\right).$$

This suggests that, up to an $o(K^{-2})$ term, the pseudo-true value θ^* solves:

$$\mathbb{E}\left[-\frac{1}{T}\sum_{t=1}^{T}\left(W_{it}-\overline{W}_i\right)\left(Y_{it}-\overline{Y}_i-\left(W_{it}-\overline{W}_i\right)'\theta\right)\right] + \frac{\pi\sqrt{3}}{K^2}\left(\operatorname{Var}(\overline{W}_i)\theta - \operatorname{Cov}\left(\overline{W}_i,\overline{Y}_i\right)\right) = 0.$$

This gives:

$$\begin{aligned}\theta^* &= \left(\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(W_{it}-\overline{W}_i\right)\left(W_{it}-\overline{W}_i\right)'\right] + \frac{\pi\sqrt{3}}{K^2}\operatorname{Var}(\overline{W}_i)\right)^{-1} \\ &\quad \times\left(\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(W_{it}-\overline{W}_i\right)\left(Y_{it}-\overline{Y}_i\right)\right] + \frac{\pi\sqrt{3}}{K^2}\operatorname{Cov}\left(\overline{W}_i,\overline{Y}_i\right)\right) + o\left(\frac{1}{K^2}\right).\end{aligned}$$

Hence:

$$\begin{aligned}\theta^*-\theta_0 &= \left(\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(W_{it}-\overline{W}_i\right)\left(W_{it}-\overline{W}_i\right)'\right] + \frac{\pi\sqrt{3}}{K^2}\operatorname{Var}(\overline{W}_i)\right)^{-1} \\ &\quad \times\left(\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(W_{it}-\overline{W}_i\right)U_{it}\right] + \frac{\pi\sqrt{3}}{K^2}\operatorname{Cov}\left(\overline{W}_i,\alpha_i+\overline{U}_i\right)\right) + o\left(\frac{1}{K^2}\right).\end{aligned}$$

As K tends to infinity θ^* converges to the probability limit of the within estimator. The convergence rate is $1/K^2$. Moreover, the approximation bias depends on the "between" moments $\operatorname{Var}(\overline{W}_i)$ and $\operatorname{Cov}\left(\overline{W}_i,\alpha_i+\overline{U}_i\right)$.

S1.6 Conditions for Assumptions 7 and 8 in Example 3

Let us verify Assumptions 7 and 8 in Example 3 under the following conditions.

Assumption S2. *(regularity in Example 3)*

(i) Observations are i.i.d. across individuals conditional on the $\alpha_{i0}(t)$'s and $\mu_{i0}(t)$'s. The parameter spaces Θ for β_0 and $\mathcal{A}$ for $(\alpha_{i0}(t), \mu_{i0}(t)')'$ are compact, and β_0 belongs to the interior of Θ.

(ii) $(Y_{it}, X_{it}')'$ is stationary conditional on the $\alpha_{i0}(t)$'s. Let $V_{it} = X_{it} - \mu_{i0}(t)$. $\mathbb{E}(U_{it}) = 0$, $\mathbb{E}(V_{it}) = 0$, and $\mathbb{E}(U_{it}V_{it}) = 0$. The minimum eigenvalue of $\mathbb{E}\left((X_{it} - \mathbb{E}(X_{it}))\,(X_{it} - \mathbb{E}(X_{it}))'\right)$ is bounded away from zero. X_{it} have bounded support.

(iii) Let $Z_{it} = (U_{it}, V_{it}')'$. $(Z_{it})_{i,t}$ satisfies Definition 1.

Take $h_i = (Y_i, X_i')'$. Then $\varphi(\alpha_{i0}(t)) = (\alpha_{i0}(t) + \mu_{i0}(t)'\beta_0, \mu_{i0}(t)')'$ is Lipschitz since β_0 belongs to a compact set. As in Example 2 it is easy to see that φ is injective, and that ψ is Lipschitz. Moreover, $\varepsilon_{it} = (U_{it} + V_{it}'\beta_0, V_{it}')'$ satisfies Definition 1 since $(U_{it}, V_{it}')'$ is sub-Gaussian and $(1, \beta_0')'$ belongs to a compact set. This verifies Assumptions 7.

Consider next the quasi-likelihood: $\ell_{it}(\alpha_i(t), \beta) = -\frac{1}{2}(Y_{it} - X_{it}'\beta - \alpha_i(t))^2$. $\overline{\alpha}_i(\beta, t)$ is uniquely defined, equal to:

$$\overline{\alpha}_i(\beta, t) = \alpha_{i0}(t) + \mu_{i0}(t)'(\beta_0 - \beta).$$

$\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \mathbb{E}(\ell_{it}(\overline{\alpha}_i(\beta), \beta))$ is a quadratic function of β. Moreover, it derivative is:

$$\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \mathbb{E}\Bigg(V_{it}\left(U_{it} + V_{it}'(\beta_0 - \beta)\right)\Bigg).$$

This derivative is zero at β_0, and the second derivative $-H$ is negative definite by Assumption S2 (ii).

Since $v_{it}^{\alpha} = -1$ the last part of Assumption 8 (i) follows.

Next, it is easy to check that $\sup_{i,t} \sup_{(\alpha_i(t), \beta)} |\mathbb{E}(\ell_{it}(\alpha_i(t), \beta))| = O(1)$, and similarly for the first three derivatives of ℓ_{it}, since $(U_{it}, V_{it}')'$ being sub-Gaussian implies it has finite moments at any order.

Third derivatives of ℓ_{it} are zero. As for second derivatives we have $v_{it}^{\alpha} = -1$, $\frac{\partial^2 \ell_{it}}{\partial\beta\partial\alpha_i(t)} = -X_{it}$, and $\frac{\partial^2 \ell_{it}}{\partial\beta\partial\beta'} = -X_{it}X_{it}'$. Those are uniformly bounded since X_{it} have bounded support.

Next, we have:

$$\begin{aligned}\mathbb{E}_{(\alpha(t),\mu(t))}\left(v_{it}(\overline{\alpha}_i(\beta, t), \beta)\right) &= \mathbb{E}_{(\alpha(t),\mu(t))}\left(Y_{it} - X_{it}'\beta - \alpha_{i0}(t) - \mu_{i0}(t)'(\beta_0 - \beta)\right)\\ &= \alpha(t) - \alpha_{i0}(t) + (\mu(t) - \mu_{i0}(t))'(\beta_0 - \beta),\end{aligned}$$

so $\frac{\partial}{\partial\alpha(t)}\Big|_{(\alpha_{i0}(t),\mu_{i0}(t))}\mathbb{E}_{(\alpha(t),\mu(t))}\left(v_{it}(\overline{\alpha}_i(\beta,t),\beta)\right)$ is uniformly bounded. Similar arguments end the verification of Assumption 8 (ii).

Lastly:

$$v_{it}(\overline{\alpha}_i(\beta,t),\beta) = Y_{it} - X_{it}'\beta - \alpha_{i0}(t) - \mu_{i0}(t)'(\beta_0 - \beta) = U_{it} + V_{it}'(\beta_0 - \beta).$$

Since $(1,(\beta_0-\beta)')'$ is bounded and the $(U_{it}, V_{it}')'$ satisfy Definition 1, the vector stacking all $v_{it}(\overline{\alpha}_i(\beta,t),\beta)'$ satisfies the sub-Gaussian requirement of Definition 1 uniformly in β. Likewise:

$$\frac{\partial}{\partial\beta}\bigg|_{\beta_0} v_{it}\left(\overline{\alpha}_i(\beta,t),\beta\right) = -V_{it},$$

which also satisfies Definition 1.

This ends the verification of Assumption 8.

S2 Complements to the applications

S2.1 Dynamic model of location choice

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

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

By Bellman's principle the alternative-specific value functions are:

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

where $S_{it} = \left(j, \mathcal{J}^j_{i,t-1}, \alpha_i\left(\mathcal{J}^j_{i,t-1}\right)\right)$ 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):

$$\overline{V}_t(S_{i,t-1}) = \ln\left(\sum_{j=1}^{J} \exp V_t(j, S_{i,t-1})\right) + \gamma, \tag{S1}$$

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

$$\begin{aligned} &V_t(j, S_{i,t-1}) = \\ &\quad \mathbb{E}\left[\rho \exp\left(\alpha_i(j) + \frac{\sigma^2}{2}\right) - c(j_{i,t-1}, j) + \beta \overline{V}_t\left(j, \mathcal{J}^j_{i,t-1}, \alpha_i\left(\mathcal{J}^j_{i,t-1}\right)\right) \,\middle|\, j_{it} = j, S_{i,t-1}\right], \end{aligned} \tag{S2}$$

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

Computation. Computation of the solution proceeds in a recursive manner. In the case where all locations have been visited, $\mathcal{J}_{it} = \{1, ..., J\}$ so $S_{it} = (j_{it}, \{1, ..., J\}, \{\alpha_i(1), ..., \alpha_i(J)\})$. Denote the corresponding integrated value function given most recent location j as $\overline{V}^J(i, j)$. From (S1) and (S2) we have:

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

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:

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

where $\mathbb{E}_{\mathcal{J},j,j'}$ is taken with respect to the distribution of $\alpha_i(j')$ given $\alpha_i(\mathcal{J})$, 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 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 (18), where the estimated value functions $\widehat{V}_t\left(j, j_{i,t-1}, \mathcal{J}_{i,t-1}, \widehat{\alpha}(\widehat{k}_i, \mathcal{J}_{i,t-1}), \theta\right)$ solve the system (S1)-(S2). We estimate the conditional expectation in (S2) as a conditional mean given $\widehat{\alpha}(\widehat{k}_i, \mathcal{J}_{i,t-1})$, 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 experimented with both a Nadaraya Watson kernel estimator and a polynomial series estimator. We use an exponential regression estimator in the illustration.

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

$$\widehat{\sigma}^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\ln W_{it} - \widehat{\alpha}(\widehat{k}_i, j_{it})\right)^2. \tag{S3}$$

Then, individual groups are assigned as:

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

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

Given group assignments, parameters can be updated 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 (S3), and:

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

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, 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. We use this approach for computational reasons; see Rust (1994) and Arcidiacono and Jones (2003) for related approaches. In Section S1 we study properties of two-step grouped fixed-effects in a partial likelihood setting.

S2.2 Dynamic model of location choice: additional results

In this subsection we show additional estimation results for the illustration in Section 6.

Fixed-K grouped fixed-effects: results. We start by reporting results based on fixed values of K, from $K = 2$ to $K = 8$, in Figure S1. 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 3, where the average value for $\widehat{K}$ is 7).

Fixed-effects estimation: results. In this DGP, 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 7 groups (the average value of $\widehat{K}$), although it becomes 9 times slower in a sample with 10 times as many individuals. The results for the fixed-effects estimator, and the bias-reduced fixed-effects estimator based on the half-panel jackknife method of Dhaene and Jochmans (2015), are shown in Figure S2. We see that the results do not differ markedly from the grouped fixed-effects results in Figure 3, consistently with Theorem 1.

EM algorithm: results. As a comparison, we next report the results of random-effects estimation 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, 67, and 294 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. The results in Figure S3 show that the estimates with $K = 2$ types are severely biased, and have large variances. The quality of estimation improves substantially when taking $K = 8$ groups. In the latter case, performance seems roughly comparable to the bias-corrected two-step results shown in Figure 3.

S2.3 Firm and worker heterogeneity: estimation

Following Bonhomme *et al.* (2015) 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:

$$\mathbb{E}\left(Y_{i2} - Y_{i1} \,|\, m_i = 1, j(i,1), j(i,2)\right) = \psi_{j(i,2)} - \psi_{j(i,1)}, \tag{S4}$$

$$\mathrm{Var}\left(Y_{i2} - Y_{i1} \,|\, m_i = 1, j(i,1), j(i,2)\right) = \mathrm{Var}\left(\varepsilon_{i2}\right) + \mathrm{Var}\left(\varepsilon_{i1}\right) = 2\,s^2. \tag{S5}$$

Then, in the first cross-section we have:

$$\mathbb{E}\left(Y_{i1} \,|\, j(i,1)\right) = \psi_{j(i,1)} + \mathbb{E}\left(\eta_i \,|\, j(i,1)\right) = \psi_{j(i,1)} + \mu_{j(i,1)}, \tag{S6}$$

$$\mathrm{Var}\left(Y_{i1} \,|\, j(i,1)\right) = \mathrm{Var}\left(\eta_i \,|\, j(i,1)\right) + \mathrm{Var}\left(\varepsilon_{i1}\right) = \sigma^2_{j(i,1)} + s^2. \tag{S7}$$

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 (S4).
2. Compute $\widehat{s}^2$ based on (S5).
3. Given $\widehat{\psi}(\widehat{k}_j)$, compute $\widehat{\mu}(\widehat{k}_j)$ based on (S6).
4. Given $\widehat{s}^2$, compute $\widehat{\sigma}^2(\widehat{k}_j)$ based on (S7). 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 (22) by aggregation across types.

The fixed-effects estimator in Table 1 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 *et al.* (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.[3]

S2.4 Firm and worker heterogeneity: additional results

In this part of the supplementary appendix we report the results for additional DGPs.

[3] We compute the connected set in an initial step, and use sparse matrix coding for efficient computation.

Monte Carlo designs. We consider four additional DGPs, in addition to DGP1 reported in Table 1. In Table S1 we show the sample sizes that we use in all designs, including the average number of job movers per firm. DGP2 has one-dimensional underlying heterogeneity, with different parameter values: the variance of firm effects is larger than in DGP1, while the correlation between firm effects and worker effects is smaller, the relative magnitudes being close to the AKM estimates of Card *et al.* (2013). DGP3 and DGP4 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 $\sigma_j^2 = (a + bV_j)^2$ 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 in Sorkin, 2016). As displayed in Table S2, the two-dimensional DGPs differ in terms of parameter values.[4] Lastly, DGP5 has discrete heterogeneity. Specifically, there are $K^* = 10$ "true" groups in the population. The groups are chosen by approximating the firm heterogeneity of DGP1.

Alternative DGP with one-dimensional heterogeneity: results. In Table S3 we report the results of two-step grouped fixed-effects and its bias-corrected version, as well as fixed-effects, in DGP2 with one-dimensional heterogeneity and a larger variance of firm effects than in Table 1. The performance of the estimators is comparable to Table 1.

Bias-corrected fixed-effects. In Table S4 we report the results of bias-corrected fixed-effects estimation in DGP1 (top panel, see Table 1) and DGP2 (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.[5]

Inferring the underlying dimension of firm heterogeneity. As a motivation for considering DGPs with an underlying dimension higher than one, but still relatively low, we first attempt

[4] The last row of the table shows the correlation between the wage firm effect ψ_j and the present value V_j in all DGPs.

[5] 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.

to learn the underlying dimension of firm heterogeneity on the Swedish matched employer-employee data set used in Section 7. 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 the method for selecting the number of groups outlined in Subsection 4.2, the method we use here consists in comparing the length of the panel T with the number of groups $\widehat{K}$ estimated from (12). If the underlying dimension of $\varphi(\alpha_{i0})$ is $d > 0$, then we expect $\widehat{Q}(K)$ to decrease at a rate $O_p(K^{-\frac{2}{d}}) + o_p(T^{-1})$. This suggests that $\widehat{K}^{\frac{2}{d}}$ and T 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 $\ln T$, the slope of which is $2/\widehat{d}$.

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 $\ln(\widehat{V}_h/T)$.[6] On the right graph we then plot $\ln \widehat{K}$ against the logarithm of the average firm size in each sample.[7] 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 S5 we report the simulation results for DGP3 with continuous two-dimensional firm heterogeneity. The results for DGP4, with a smaller variance of firm effects, are reported in Table S7. 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 $\xi = 1$ in (12)), we see that the two-step estimators show more substantial 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 bias is still substantial. At the same time, as shown by the two bottom panels of the tables, taking $\xi = .5$ and $\xi = .25$ improves the performance of the two-step estimator.[8] Performance is further improved when using the bias-reduced

[6] 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.

[7] Here we report results based on empirical cdfs of log-wages evaluated at 20 points. We checked that using 40 points instead did not affect the results.

[8] Notice that while the selected number of groups $\widehat{K}$ is monotone in firm size for $\xi = 1$ and $\xi = .5$, it is not monotone for $\xi = .25$. This is a finite sample issue: when taking $\xi = .25$ and focusing on large firms the number

estimator.

As pointed out in Section 4, features of the model may be exploited to improve the classification. In the two-dimensional designs DGP3 and DGP4, 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 T tends to infinity. We then apply two-step grouped fixed-effects to the moments h_{1j} and h_{2j}. In Tables S7 and S8 we report the results for the iterated estimator (only iterated once) and its bias-corrected version, for DGP3 and DGP4, respectively. We see that the iteration improves performance substantially for DGP3, although it has small effects on performance in DGP4.

Low mobility bias and regularization. As shown by Theorem 2, 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 grouped fixed-effects 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 grouped fixed-effects with bias correction, and iterated two-step grouped fixed-effects 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 DGPs: DGP3 and DGP4, respectively. We see a striking difference between fixed-effects and grouped fixed-effects: 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 grouped fixed-effects remain moderate. This is in line with Theorem 2. 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

of groups is no longer negligible with respect to the number of firms in the sample.

discrete regularization achieved in grouped fixed-effects may result in practical improvements relative to fixed-effects in data sets of realistic dimensions.

Discrete heterogeneity: results. Finally, in Table S9 we report results for a discrete DGP (DGP5) where all firm population parameters are constant within groups $\widehat{k}_j$, with $K^* = 10$. In this case the results of two-step grouped fixed-effects with $K = K^*$ turn out to be quite similar to those obtained in the continuous DGP in Table 1. 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.[9] 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.

[9] We computed misclassification frequencies by solving a linear assignment problem using the simplex algorithm in every simulation.

Table S1: Firm and worker effects, 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 DGPs.

Table S2: Different DGPs

	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%
$Cov(\psi, \eta)$	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 DGP1, DGP4, DGP2, and DGP3, respectively.

Table S3: Estimates of firm and worker heterogeneity across simulations, one-dimensional firm heterogeneity, large variance of firm effects

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

Notes: See notes to Table 1. Results for DGP2.

Table S4: Bias-corrected fixed-effects estimators, one-dimensional firm heterogeneity

Firm size	$\text{Var}(\eta_i)$	$\text{Var}(\psi_j)$	$\text{Cov}(\eta_i, \psi_j)$	$\text{Corr}(\eta_i, \psi_j)$	$\text{Var}(\varepsilon_{i1})$
	one-dimensional, small firm effect				
-	0.0758	0.0017	0.0057	0.4963	0.0341
	fixed-effects, bias-corrected				
10	0.0065	-0.0717	0.0791	-0.0976	0.0300
	[-0.004,0.016]	[-0.082,-0.064]	[0.071,0.089]	[-0.125,-0.072]	[0.029,0.031]
20	0.0645	-0.0098	0.0172	0.0973	0.0339
	[0.062,0.067]	[-0.011,-0.008]	[0.016,0.019]	[0.073,0.125]	[0.033,0.035]
50	0.0733	-0.0007	0.0082	0.3069	0.0341
	[0.072,0.075]	[-0.001,-0.000]	[0.008,0.009]	[0.279,0.335]	[0.033,0.035]
100	0.0748	0.0007	0.0067	0.4173	0.0341
	[0.073,0.076]	[0.000,0.001]	[0.006,0.007]	[0.388,0.447]	[0.033,0.035]
200	0.0753	0.0012	0.0062	0.4822	0.0341
	[0.074,0.077]	[0.001,0.002]	[0.006,0.007]	[0.451,0.512]	[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	-0.0533	0.0788	-0.0077	0.0301
	[-0.013,0.006]	[-0.062,-0.045]	[0.070,0.088]	[-0.034,0.019]	[0.029,0.031]
20	0.0547	0.0089	0.0166	0.1163	0.0339
	[0.053,0.057]	[0.007,0.011]	[0.015,0.018]	[0.096,0.137]	[0.033,0.035]
50	0.0636	0.0180	0.0075	0.1561	0.0341
	[0.062,0.065]	[0.017,0.019]	[0.007,0.008]	[0.139,0.173]	[0.033,0.035]
100	0.0650	0.0194	0.0061	0.1554	0.0341
	[0.064,0.066]	[0.019,0.020]	[0.006,0.007]	[0.139,0.171]	[0.033,0.035]
200	0.0656	0.0199	0.0055	0.1487	0.0341
	[0.064,0.067]	[0.019,0.021]	[0.005,0.006]	[0.133,0.164]	[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 DGP1, with a small variance of firm effects, while the bottom panel shows the results for DGP2, with a larger variance of firm effects. 500 *simulations.*

Table S5: Firm and worker effects, two-dimensional firm heterogeneity, large $Var(\psi)$, different choices of ξ

	two-step estimator						two-step 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	
	$\xi = 1.0$						$\xi = 1.0$					
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]
	$\xi = 0.5$						$\xi = 0.5$					
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]
	$\xi = 0.25$						$\xi = 0.25$					
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 ξ in (12).* 500 *simulations. Results for DGP3.*

Table S6: Firm and worker effects, two-dimensional firm heterogeneity, small $Var(\psi)$, different choices of ξ

	two-step estimator						two-step 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	
	$\xi = 1.0$						$\xi = 1.0$					
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]
	$\xi = 0.5$						$\xi = 0.5$					
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]
	$\xi = 0.25$						$\xi = 0.25$					
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 ξ in (12).* 500 *simulations. Results for DGP4.*

Table S7: Firm and worker effects, two-dimensional firm heterogeneity, large $Var(\psi)$, different choices of ξ, 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	
	$\xi = 1.0$						$\xi = 1.0$					
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]
	$\xi = 0.5$						$\xi = 0.5$					
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]
	$\xi = 0.25$						$\xi = 0.25$					
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 ξ in (12).* 500 *simulations. Results for DGP3.*

Table S8: Firm and worker effects, two-dimensional firm heterogeneity, small $Var(\psi)$, different choices of ξ, 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	
	$\xi = 1.0$						$\xi = 1.0$					
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]
	$\xi = 0.5$						$\xi = 0.5$					
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]
	$\xi = 0.25$						$\xi = 0.25$					
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]

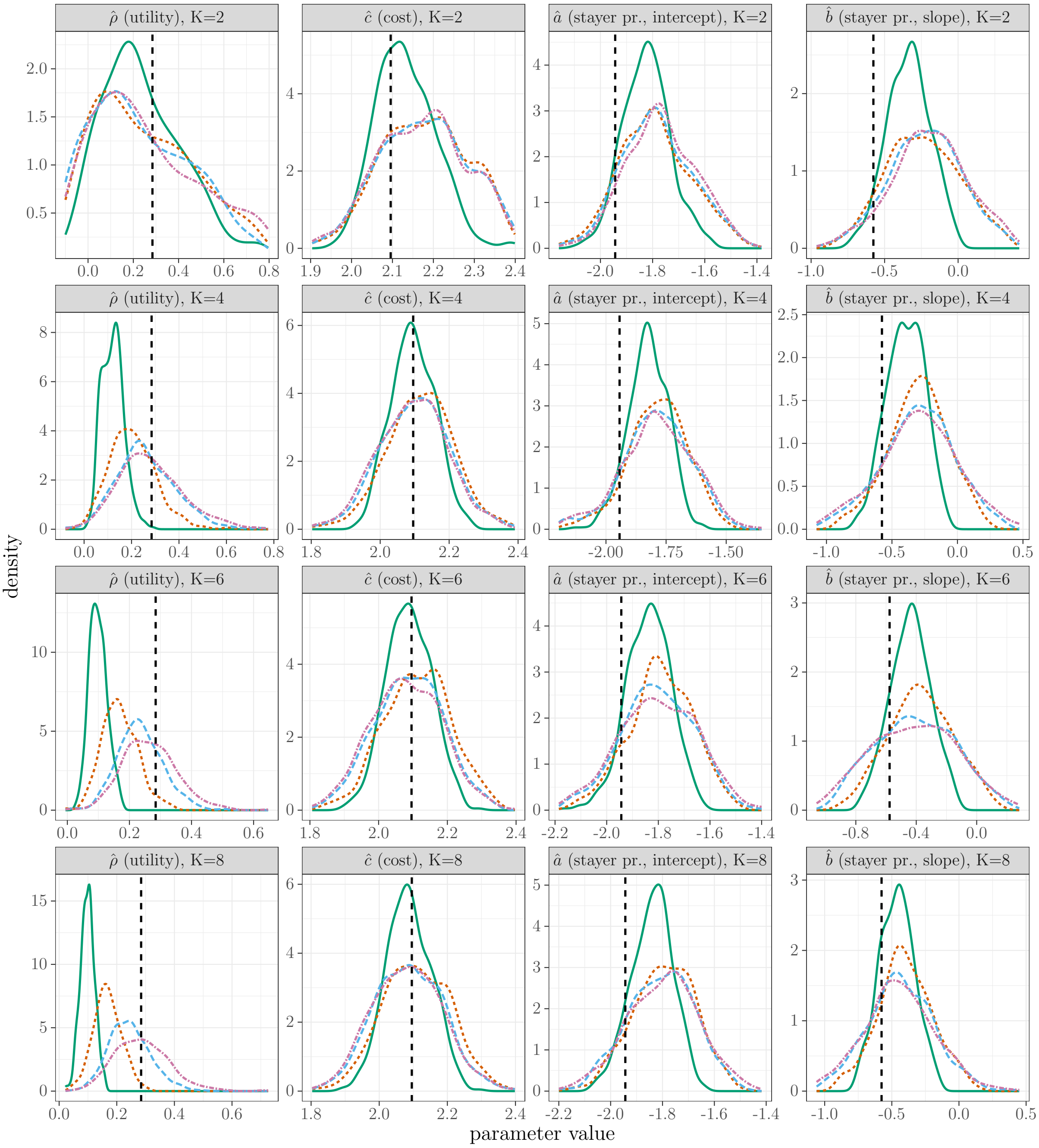
Notes: Means and 95% confidence intervals. Unobserved heterogeneity is continuously distributed in the DGP, with underlying dimension equal to 2*. The number of groups K is estimated in every replication, with different choices for ξ in (12).* 500 *simulations. Results for DGP4.*

Table S9: Firm and worker effects, discrete firm heterogeneity ($K^* = 10$)

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

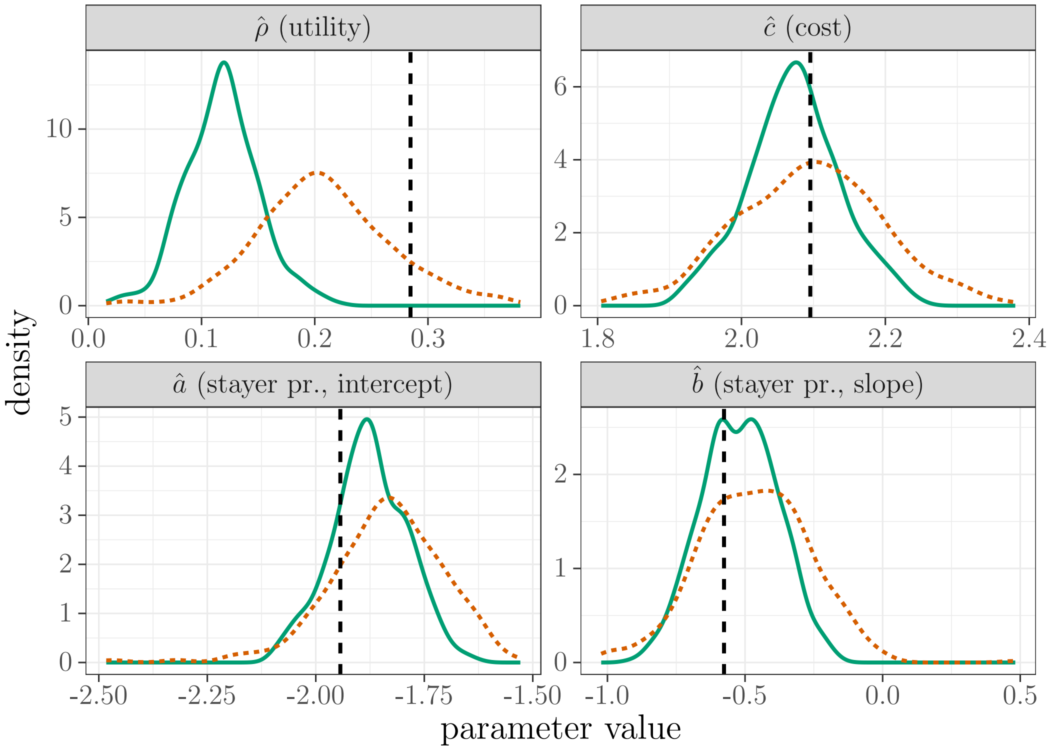
Notes: Means and 95% confidence intervals. Unobserved heterogeneity is discretely distributed in the DGP, with $K^* = 10$ groups. In the top panel the true number of groups is used. The last column shows frequencies of misclassification. In the bottom panel the number of groups is estimated in every replication. 500 *simulations. Results for DGP5.*

Figure S1: Parameter estimates across simulations, fixed K



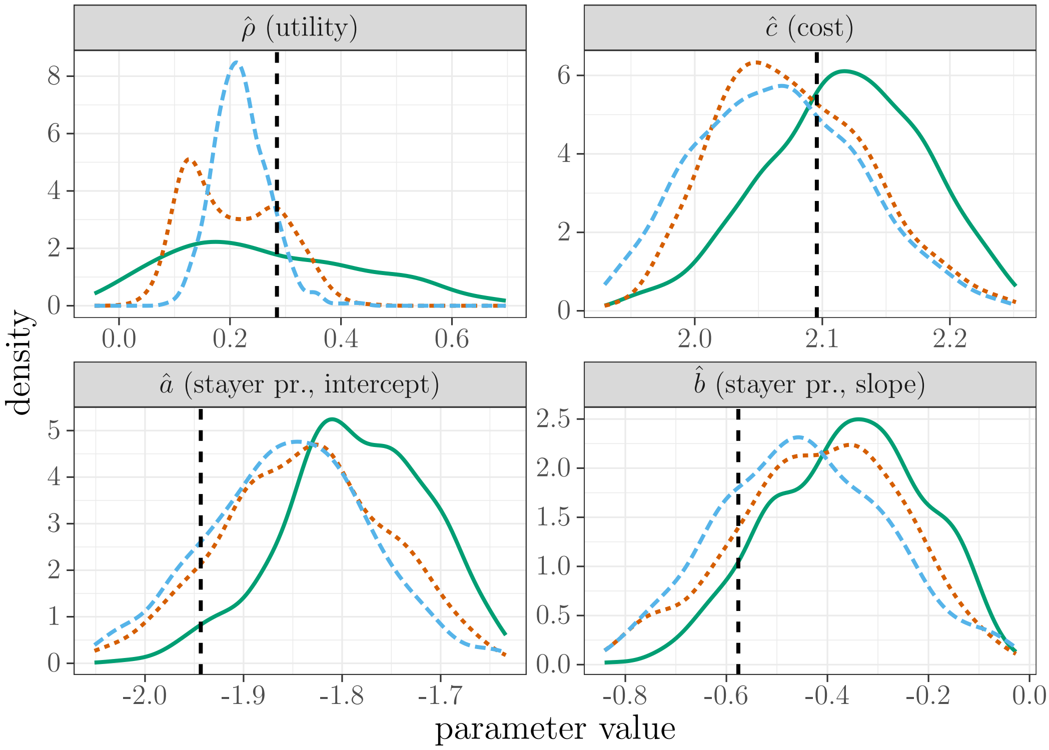
Notes: See note to Figure 3. K is kept fixed. 500 *replications.*

Figure S2: Parameter estimates across simulations, fixed-effects and bias-corrected fixed-effects estimators



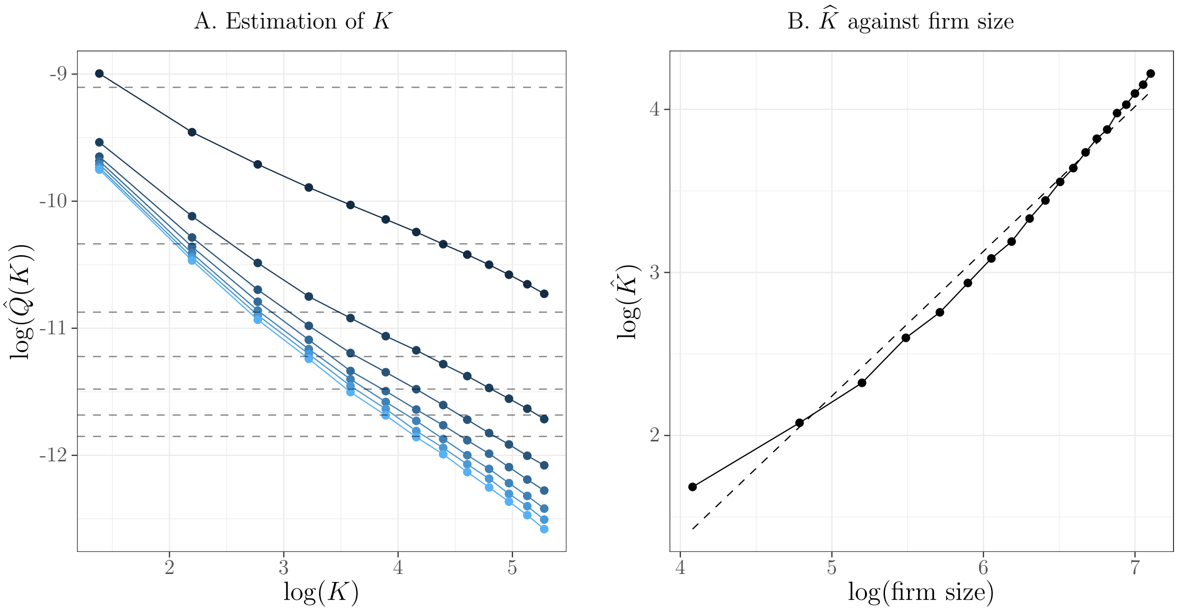
Notes: Solid is fixed-effects, dotted is bias-corrected fixed-effects. The vertical line indicates the true parameter value. $N = 1889$, $T = 16$. *Unobserved heterogeneity is continuously distributed in the DGP.* 500 *replications.*

Figure S3: Parameter estimates across simulations, random-effects estimators



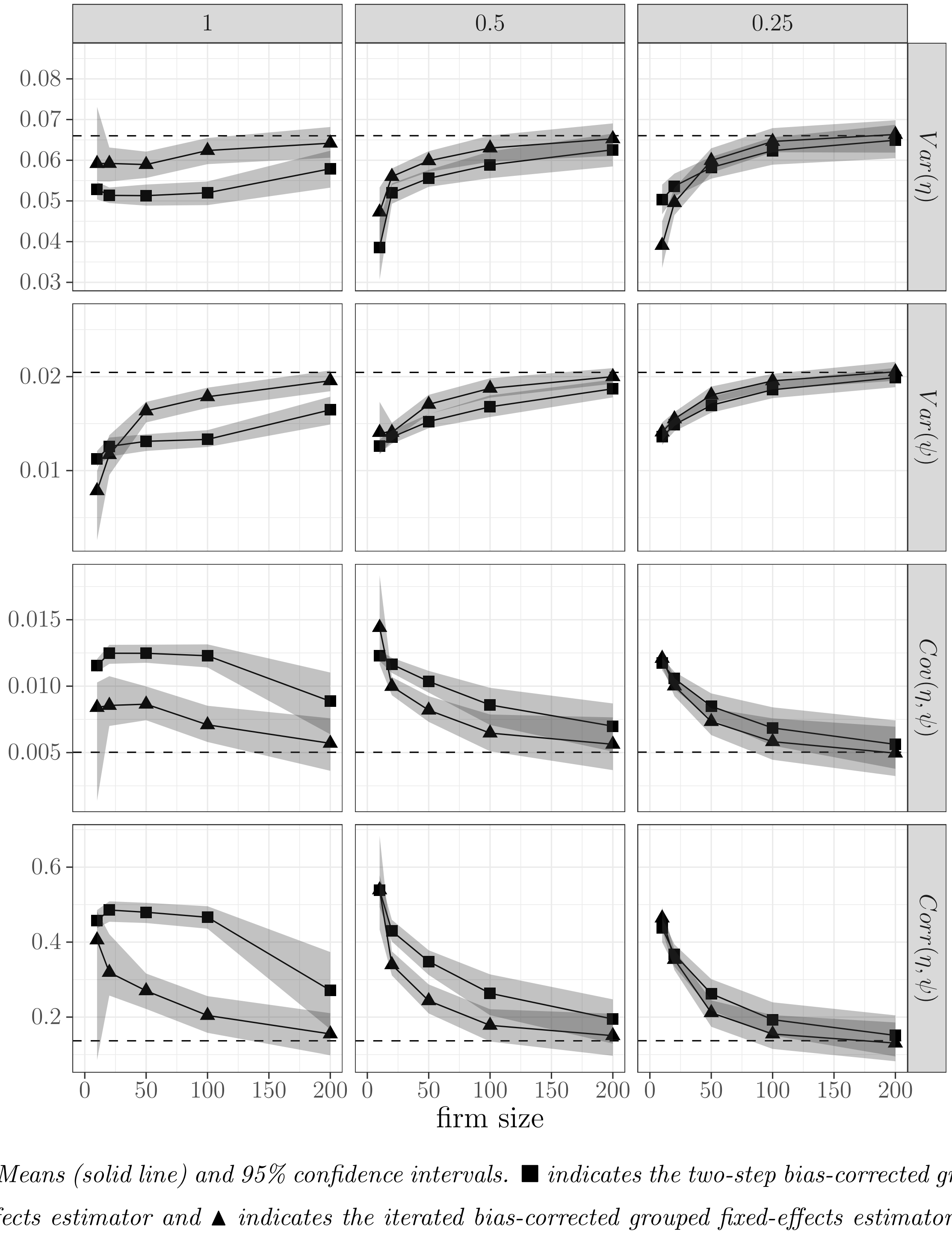
Notes: Solid is $K = 2$*, dotted is* $K = 4$*, dashed is* $K = 8$*. The vertical line indicates the true parameter value.* $N = 1889$*,* $T = 16$*. Unobserved heterogeneity is continuously distributed in the DGP.* 500 *replications.*

Figure S4: Dimension of firm heterogeneity



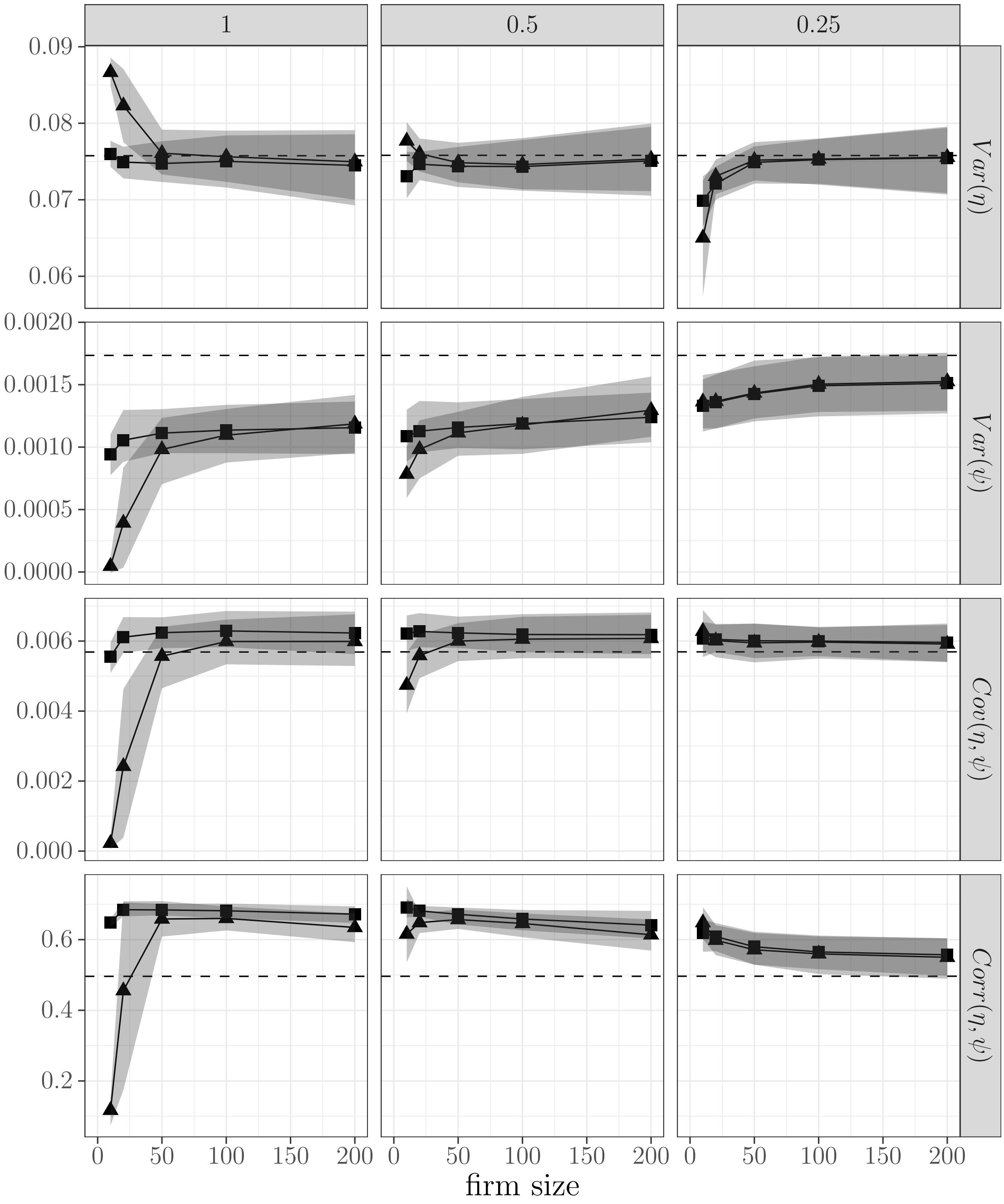
Notes: Source Swedish administrative data. Left graph shows the logarithm of $\widehat{Q}(K)$ as a function of K, for different average firm sizes T. Horizontal lines show the corresponding value of $\ln(\widehat{V}_h/T)$. The right graph shows the relationship between the log of $\widehat{K}$ and the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



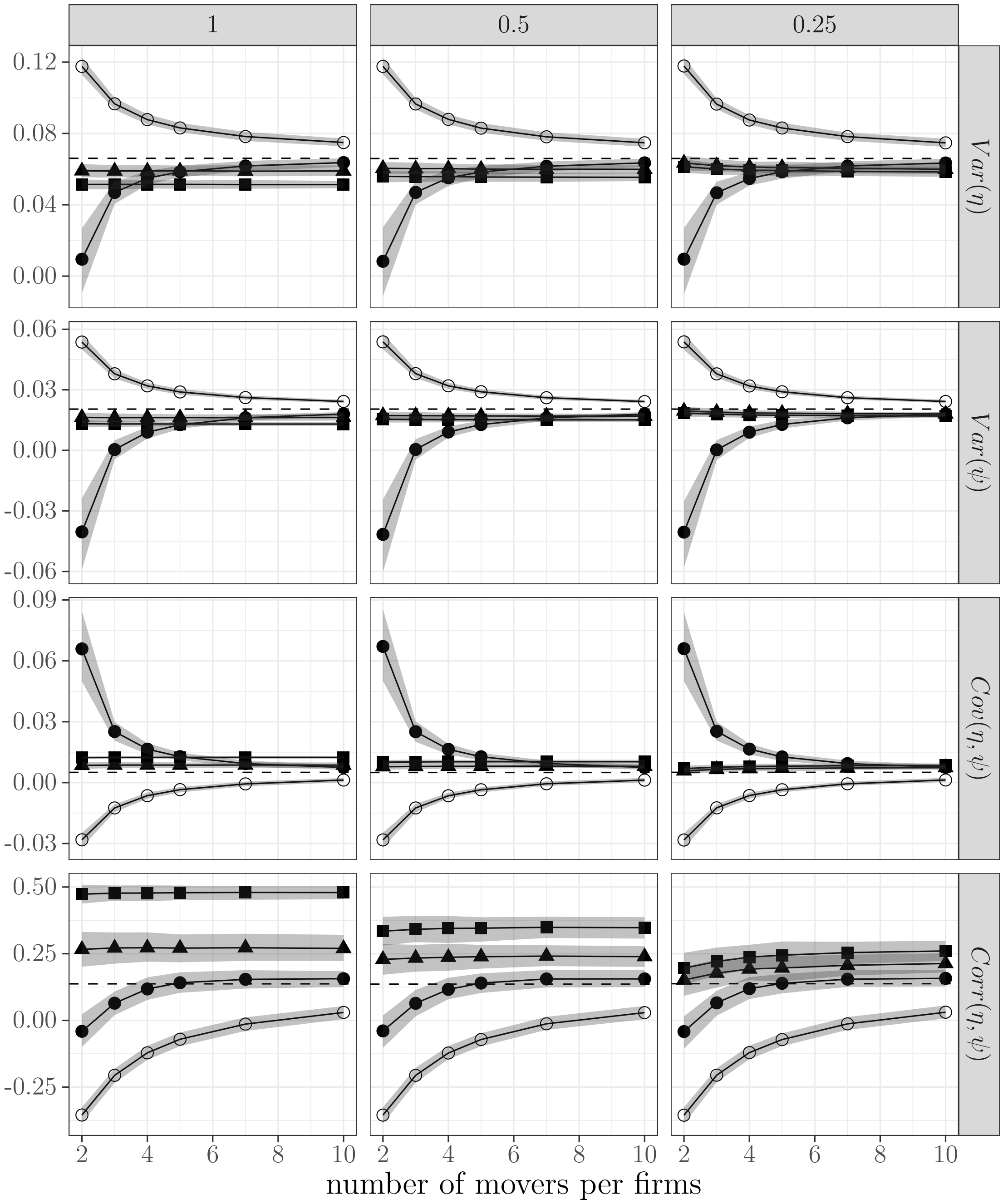
Notes: Means (solid line) and 95% confidence intervals. ■ *indicates the two-step bias-corrected grouped fixed-effects estimator and* ▲ *indicates the iterated bias-corrected grouped fixed-effects estimator. The different columns represent different values of* ξ *(that is, different selection rules for 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 DGP3.*

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



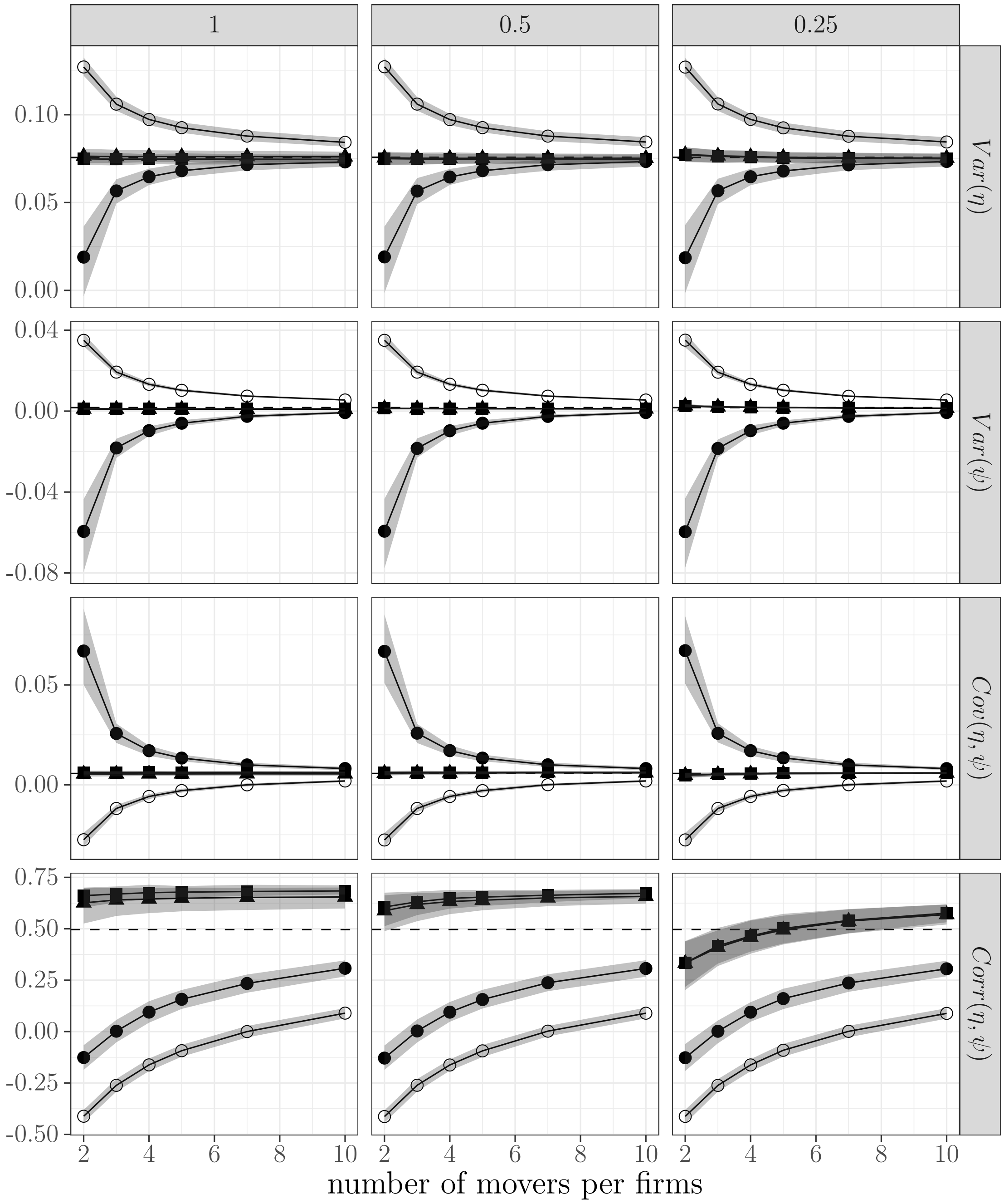
Notes: See the notes to Figure S5. Results for DGP4.

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 grouped fixed-effects estimator,* ▲ *the iterated bias-corrected grouped fixed-effects estimator,* ○ *the fixed-effects estimator, and* ● *the bias-corrected fixed-effects estimator. The different columns represent different values of ξ (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 DGP3.*

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 DGP4.

S3 Additional simulation exercises

S3.1 Time-varying unobserved heterogeneity

Let $Y_{it} = \alpha_{i0}(t) + U_{it}$. We focus on the mean squared error (MSE):

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

We use the following specification: U_{it} are i.i.d standard normal, and:

$$\alpha_{i0}(t) = \xi_{i1} + \xi_{i2}\frac{\Phi^{-1}\left(\frac{t}{T+1}\right)}{\Phi^{-1}\left(\frac{T}{T+1}\right)},$$

where Φ is the standard normal cdf, ξ_{i1} is standard normal, and $\ln \xi_{i2} \sim \mathcal{N}(.2, .04)$ independent of ξ_{i1}. We vary the sample size from $T = 5$ to $T = 40$, take $N = T^2$, and set $K = T$ in every sample.

Figure S9 shows the results for the grouped fixed-effects estimator where the kmeans algorithm is applied to the vectors of Y_{it}'s in the first step. The graph shows means and pointwise 95% confidence bands across 100 replications. The results align closely with Theorem 3. Indeed, according to (16) the rate of convergence consists of three terms: a term $O_p(K/N) = O_p(1/T)$ reflecting the estimation of the KT group-specific parameters, a term $O_p(B_\alpha(K)/T)$ reflecting the approximation bias, which in this two-dimensional case will be $O_p(1/K) = O_p(1/T)$, and a term $O_p((\ln K)/T) = O_p((\ln T)/T)$ reflecting the noise in estimating group membership for every individual. In this DGP the latter term is thus the dominant one. In Figure S9 the dashed line shows $(\widehat{c} \ln T)/T$ as a function of T, where $\widehat{c}$ is fitted to the solid line. We see that the MSE of grouped fixed-effects and the theoretical fit align closely. This suggests that the upper bound on the rate in (16) is very informative for this DGP.

S3.2 "Double grouped fixed-effects" in a probit model

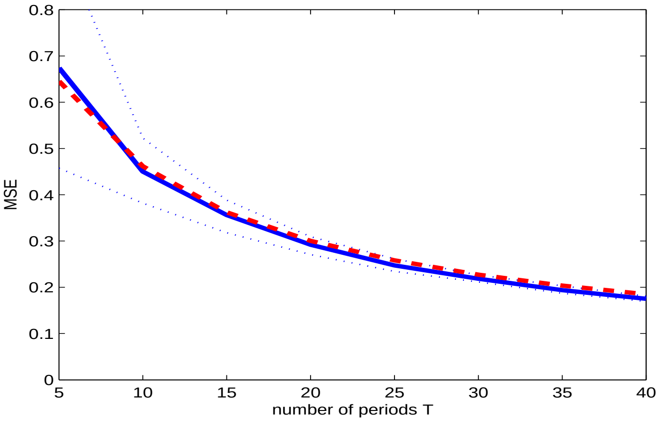
An alternative estimator, in linear or index models, is "double" grouped fixed-effects. As an example, consider the static probit model:

$$Y_{it} = \mathbf{1}\{X_{it}'\theta_0 + \alpha_{i0} + U_{it} \geq 0\},$$

where U_{it} are i.i.d standard normal, and $X_{it} = \mu_{i0} + V_{it}$, V_{it} i.i.d, independent of $\mu_{i0}, \alpha_{i0}, U_{is}$.

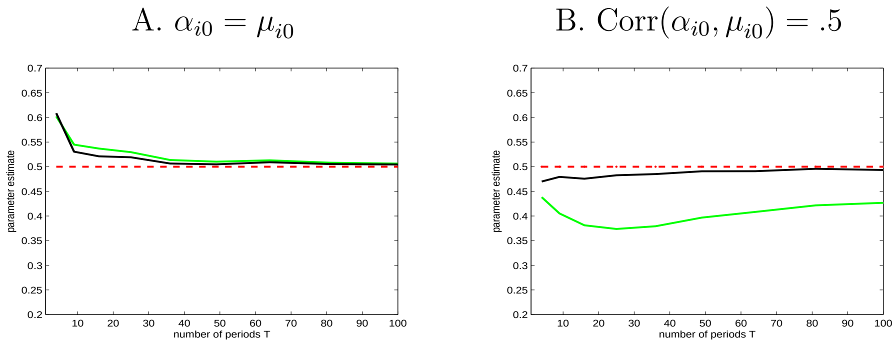
Consider the moments $h_i = (\overline{Y}_i, \overline{X}_i')'$. In the first step, we discretize each component of h_i separately, by applying kmeans to $\overline{Y}_i$ and all components of $\overline{X}_i$ in turn, with K groups. In the second step, we estimate the probit model by including all group indicators as additive controls. Figure S10 shows the results in two cases: one-dimensional (panel A) and two-dimensional heterogeneity (panel B). The results compare two-step grouped fixed-effects with "double" grouped fixed-effects estimators. For the sake of illustration, we set the number of groups $K = \lfloor \sqrt{T} \rfloor$ in every sample. This leads to a large approximation bias in the two-dimensional case, as shown by panel B. We see that double grouped fixed-effects performs significantly better than grouped fixed-effects in this environment.

Figure S9: Time-varying unobserved heterogeneity



Notes: Mean squared errors. Solid and dashed lines are the means and 95% confidence bands of grouped fixed-effects across 100 *replications. The dashed line is the fit based on a* $(\ln T)/T$ *rate.* $N = T^2, K = T$*.*

Figure S10: Two-step grouped fixed-effects and double grouped fixed-effects in a static probit model



A. $\alpha_{i0} = \mu_{i0}$

B. $\mathrm{Corr}(\alpha_{i0}, \mu_{i0}) = .5$

Notes: Averages over simulations. The dashed horizontal line is the true value. The curve further away from it is two-step grouped fixed-effects, the curve closer to it is double grouped fixed-effects. $N = 100$, 100 *replications.* $K = \lfloor\sqrt{T}\rfloor$.