Incorporating Unobserved Heterogeneity into Structural Models

Stéphane Bonhomme
University of Chicago

Econometric Society Summer School
in Dynamic Structural Econometrics

University of Wisconsin Madison, August 4–9 2024

Outline

1. Mixture representation

2. Strategies for identification

3. Random-effects estimation

4. Grouped fixed-effects estimation

5. Applications

1. Mixture representation

A dynamic choice model

- Finite horizon. Choices c_t. States s_t.
- Utility $u(c_t, s_t)$. Discount factor β.
- State transition $s_{t+1} \,|\, s_t, c_t$.
- Agents maximize expected utility.

$$\max_{c_1,\ldots,c_T} \mathbb{E}\left[\sum_{t=1}^{T} \beta^{t-1} u(c_t, s_t)\right]$$

- Bellman's principle implies the value functions

$$V_t(s_t) = \max_{c_t} u(c_t) + \beta \mathbb{E}\left[V_{t+1}(s_{t+1}) \,|\, s_t, c_t\right].$$

Unobserved heterogeneity

• Heterogeneity in preferences u (also: technology, ability, ...), and discounting β.

• Beliefs may also be heterogeneous.

• State transitions can be heterogeneous as well.

• We mostly focus on permanent unobserved heterogeneity (simpler!).

• Unobserved heterogeneity matters quantitatively. An example is Keane and Wolpin's (1997) analysis of career choices.

Discrete or continuous types

• A common approach in structural work is to assume latent types take a small number of values:

$$a \in \mathcal{A} = \{a_1, ..., a_K\}, \text{ where K is "small".}$$

• This is a tractable approach to allow for multi-dimensional heterogeneity.

• For example, utility $u(c_t, s_t, a)$ can be type-specific. Equivalently, each type of agent has her own utility function $u_a(c_t, s_t)$.

• Continuous-type modeling has the advantage of being more realistic, since it allows for within-group as well as between-group heterogeneity. However, it is also more challenging.

Mixture representation

• Let Y be a vector containing all endogenous variables in the model. This includes choices c_t, (endogenous) state variables s_t, as well additional payoff variables.

• Let X be a vector containing all exogenous variables, including initial conditions and time-invariant characteristics.

• Let A be a vector containing the latent types.

• Finally, let θ contain all structural parameters. We assume Y, X, A are all discrete.

• We can write, using the law of total probability,

$$\mathsf{Pr}(Y = y \,|\, X = x) = \sum_{a \in \mathcal{A}} \mathsf{Pr}(Y = y \,|\, X = x, A = a; \theta)\, \mathsf{Pr}(A = a \,|\, X = x).$$

Linear system

• In compact form, we can equivalently write (at each x):

$$D_x = M_x(\theta)\pi_x.$$

• Here D_x is the "data" vector. It has as many elements as Y has points of support. Dimension $n \times 1$.

• In turn, $M_x(\theta)$ is the "model" matrix, with y's in rows and a's in columns. Dimension $n \times K$.

• Lastly, π_x is the "heterogeneity" vector. Dimension $K \times 1$.

2. Strategies for identification

Differencing (I): recovering θ

• This slide and the next two follow the functional differencing approach (B., 2012).

• Starting from $D_x = M_x(\theta)\pi_x$, a "residual" (or "within") projection gives

$$\left(I_n - M_x(\theta)M_x(\theta)^\dagger\right) D_x = 0.$$

• This provides moment restrictions on θ. Any $n \times 1$ vector h_x gives a moment function

$$\varphi_x(y;\theta) = e_y' \left(I_n - M_x(\theta)M_x(\theta)^\dagger\right) h_x,$$

where e_y is the canonical vector with a one in position y, which satisfies

$$\mathbb{E}\left[\varphi_X(Y;\theta)\right] = 0.$$

• Identification typically requires larger support of Y compared to A ("overidentification").

Background: SVD and Moore-Penrose pseudo-inverse

• Let M be an $n \times K$ matrix. We can always write the singular value decomposition (SVD)

$$M = USV',$$

where U is $n \times n$ orthogonal, V is $K \times K$ orthogonal, and D has a diagonal upper left block and zeros everywhere else.

• The Moore-Penrose pseudo-inverse of M is

$$M^\dagger = V\widetilde{D}U',$$

where $\widetilde{D}$ has a diagonal upper left block whose elements are the inverses of those of D, and zeros everywhere else.

• $MM^\dagger$ and $I_n - MM^\dagger$ are orthogonal projectors onto the image of M and its ortho-complement, respectively.

Differencing (II): recovering π_x

- Suppose that $M_x(\theta)$ has rank K.
- Then $M_x(\theta)^\dagger M_x(\theta) = I_K$.
- This implies that π_x is identified (given θ), as

$$\pi_x = M_x(\theta)^\dagger D_x.$$

- For any $K \times 1$ vector g_x, we can construct the functions

$$g_x(a) = e_a' g_x,$$

and

$$\psi_x(a;\theta) = e_a' \left(M_x(\theta)^\dagger\right)' g_x,$$

which satisfy

$$\mathbb{E}[\psi_X(A;\theta)] = \mathbb{E}\left[g(A,X)\right].$$

Independence restrictions

• While the above arguments require parametric restrictions on $Y \mid X, A$. structural models often imply conditional independence restrictions with strong identifying power.

• A general approach for discrete data was proposed by Hu (2008) and subsequently generalized (see Hu, 2015, for a monograph).

• Suppose $Y = (Y_1, Y_2, Y_3)$ are independent given A, X. Then

$$\begin{aligned}
&\Pr(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3 \mid X = x) \\
&= \sum_{a \in \mathcal{A}} \Pr(Y_1 = y_1 \mid X = x, A = a) \Pr(Y_2 = y_2 \mid X = x, A = a) \\
&\qquad \times \Pr(Y_3 = y_3 \mid X = x, A = a) \Pr(A = a \mid X = x).
\end{aligned}$$

• In matrix form, for given x and y_2,

$$D_x^{13}(y_2) = M_x^1 \Delta_x^2(y_2) \left(M_x^3\right)'.$$

Eigenvalue decomposition

- The model implies the system

$$D_x^{13}(y_2) = M_x^1 \Delta_x^2(y_2) \left(M_x^3\right)',$$

where M_x^1 and M_x^3 have rank K, and $\Delta_x^2(y_2)$ is diagonal.

- Suppose that $y_2 \in \{0, 1\}$. Let $D_x^{13}(0) = USV'$ be the singular value decomposition of $D_x^{13}(0)$.

- Let $\Omega_x = S^{-\frac{1}{2}} U' M_x^1 \Delta_x^2(0)^{\frac{1}{2}}$. Since

$$S^{-\frac{1}{2}} U' D_x^{13}(0) V' S^{-\frac{1}{2}} = I_K,$$

we have

$$\Delta_x^2(0)^{\frac{1}{2}} \left(M_x^3\right)' V' S^{-\frac{1}{2}} = \Omega_x^{-1}.$$

Eigenvalue decomposition (cont.)

- Moreover, we have

$$S^{-\frac{1}{2}}U'D_x^{13}(1)V'S^{-\frac{1}{2}} = \Omega_x \left(\Delta_x^2(0)^{-1}\Delta_x^2(1)\right)\Omega_x^{-1}.$$

- This is an eigenvalue decomposition.

- Hence, if the diagonal elements of $\Delta_x^2(0)^{-1}\Delta_x^2(1)$ are all distinct, then the columns of Ω_x are uniquely determined, as the corresponding eigenvectors. It then follows that M_x^1 and M_x^3 are identified.

- Identification is up to labeling, i.e., permutation of the columns of M_x^1 and M_x^3. Note this labeling is x-specific. Recovering a common labeling across x's requires exploiting more of the model's structure (Aguirregabiria and Mira, 2019).

Identifying content of covariates

• In discrete choice outcomes have limited variation. However, state variables and other covariates often exhibit more variation, which can be exploited under additional assumptions.

• As an example, suppose X is independent of A. Then

$$\Pr(Y = y \,|\, X = x) = \sum_{a \in \mathcal{A}} \Pr(Y = y \,|\, X = x, A = a; \theta) \Pr(A = a).$$

• Stacking all restrictions across x values then gives

$$D = M(\theta)\pi,$$

where D is $(n \dim x) \times 1$ and π is $K \times 1$.

• The assumption that the law of motion of dynamic state variables does not depend on heterogeneity has identifying content, as shown by Kasahara and Shimotsu (2009).

Extensions (I): Time-varying heterogeneity

• Time-varying unobservables are important in structural models: human capital, firm productivity, workers' effort.

• Existing identification strategies rely on Markovian assumptions:

$$A_t \,|\, A_{t-1}, ..., A_1, Y_{t-1}, ..., Y_1, X_{t-1}, ..., X_1 \overset{d}{=} A_t \,|\, A_{t-1}, Y_{t-1}, X_{t-1}.$$

• Hu and Shum (2012) provide identification results, with applications to structural models. See also Arcidiacono and Miller (2011).

• The statistics literature on Hidden Markov Models (HMM) is vast, see for example the tutorial by Rabiner (1989).

Extensions (II): Continuous heterogeneity

- Allowing for continuous heterogeneity complicates identification.

- The identity

$$D_x = M_x(\theta)\pi_x$$

continues to hold, but now π_x is a function of a, and $M_x(\theta)$ is a linear operator; i.e., a linear mapping between functions.

- Identification is unlikely to hold when Y is discrete, as in structural dynamic discrete choice models (although there are some exceptions, as in Aguirregabiria, Gu and Luo, 2021).

- When outcomes are continuous, Hu and Schennach (2008) provide identification conditions generalizing the analysis of Hu (2008).

- The full-rank assumptions on M_x^1 and M_x^3 become nonparametric "completeness" conditions.

3. Random-effects estimation

Maximum likelihood

• Parameterize $\pi_x(a) = \pi(a, x; \gamma)$, where γ is a low-dimensional parameter.

• The likelihood function takes the form

$$L(\theta, \gamma) = \sum_{i=1}^{N} \ln \left(\sum_{a \in \mathcal{A}} \Pr(Y = Y_i \,|\, X = X_i, A = a; \theta) \Pr(A = a \,|\, X = X_i; \gamma) \right).$$

• A possibility is to maximize this likelihood using a nested fixed point approach (Rust, 1987). The MPEC approach is an alternative (Su and Judd, 2012).

• In mixture models, the presence of local optima is a pervasive issue, and it is important to sample multiple starting values for the parameters.

EM algorithm

• A common way to maximize the likelihood in the presence of unobserved heterogeneity is to use the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977).

• The EM algorithm iterates between:

1. E-step: compute the posterior probabilities

$$\begin{aligned} p_i(a;\theta,\gamma) &:= \Pr(A = a \,|\, X = X_i, Y = Y_i; \theta, \gamma) \\ &\propto \Pr(Y = Y_i \,|\, X = X_i, A = a; \theta) \Pr(A = a \,|\, X = X_i; \gamma). \end{aligned}$$

2. M-step: maximize with respect to θ', γ'

$$\sum_{i=1}^{N} \sum_{a \in \mathcal{A}} p_i(a;\theta,\gamma) \ln \Big(\Pr(Y = Y_i \,|\, X = X_i, A = a; \theta') \Pr(A = a \,|\, X = X_i; \gamma') \Big).$$

• Arcidiacono and Miller (2011) develop an EM approach for dynamic discrete choice models based on conditional choice probabilities.

Joint diagonalization

• Under independence restrictions, nonparametric estimation of distributions is possible.

• Consider the case where $Y = (Y_1, Y_2, Y_3)$ are independent given A, X. Then we have, for all y_2,

$$S^{-\frac{1}{2}} U' D_x^{13}(y_2) V' S^{-\frac{1}{2}} = \Omega_x \left(\Delta_x^2(0)^{-1} \Delta_x^2(y_2) \right) \Omega_x^{-1}.$$

• This suggests estimating Ω_x as the joint eigenvectors to a set of matrices.

• B., Jochmans and Robin (2015, 2016) propose and study such nonparametric estimators.

• Implementation relies on algorithms for approximate joint diagonalization of matrices based on elementary rotations (e.g., Givens).

Continuous heterogeneity

• In the continuous case, if one is willing to parameterize the distribution of heterogeneity, then one can maximize the log of the integrated likelihood,

$$L(\theta, \gamma) = \sum_{i=1}^{N} \ln \left(\int_{\mathcal{A}} \Pr(Y = Y_i \,|\, X = X_i, A = a; \theta) \pi(a \,|\, X = X_i; \gamma) da \right).$$

• Implementation of the EM algorithm requires approximating the integrals. An alternative is to draw $A_i^{(s)}$ values from the posterior distribution in every E step ("stochastic EM").

• Allowing for nonparametric π_x, and possibly nonparametric outcome distributions as well, is more challenging. For the latter, nonparametric maximum likelihood (Kiefer and Wolfowitz, 1957) is an option. Sieve-MLE provides a general-purpose approach.

4. Grouped fixed-effects estimation

The GFE model

• An alternative view of types is given by the Grouped Fixed Effects (GFE) approach (B. and Manresa, 2015).

• In this approach, all types A_i, for $i = 1, ..., n$, are viewed as discrete parameters to estimate.

• No probability model of the heterogeneity is needed.

• In addition, the dependence of the types A_i on covariates X_i is fully unrestricted.

• For example, one can analyze ex post how estimated types $\widehat{A}_i$ depend on covariates X_i.

The role of T

• To learn about the type A_i one needs the dimension of Y_i (typically, the time dimension T) to be relatively large.

• For an intuition, suppose we want to estimate types based on the model $\overline{Y}_i = A_i + \overline{U}_i$, where $\overline{U}_i \,|\, A_i \sim \mathcal{N}\left(0, \frac{\sigma^2}{T}\right)$.

• For two individuals i and j whose types are such that $A_i > A_j$, the probability of a classification mistake is

$$\Pr\left(\overline{Y}_i < \overline{Y}_j\right) = \Phi\left(-\frac{A_i - A_j}{\sqrt{\frac{2\sigma^2}{T}}}\right).$$

• This only tends to zero if T tends to infinity...

• ... but tends to zero exponentially fast as T grows.

Joint GFE estimation

• Likelihood model $f(Y_i \,|\, X_i, A_i, \theta)$ (structural model given the types).

• A joint GFE estimator maximizes

$$\sum_{i=1}^{n} \ln f(Y_i \,|\, X_i, A_{k_i}, \theta),$$

with respect to θ, $A_1, ..., A_K$ (the "centroids"), and $k_1, ..., k_N$ (group membership).

• In regression settings, there is an extensive literature in computer science on algorithms for k-means clustering and clusterwise regression. Recent grouping methods include Chetverikov and Manresa (2021), Mugnier (2023), and Gu *et al.* (2024).

• However, joint classification and estimation is computationally expensive in structural models.

Two-step grouped fixed-effects estimation

- B., Lamadon and Manresa (2022) propose two-step GFE:

1. First step: Estimate a *partition* of individual units, $\{\widehat{k}_i\}$, by applying *kmeans* to individual-specific moments $h_i = h(Y_i, X_i)$; that is:

$$\left(\widehat{h}, \widehat{k}_1, ..., \widehat{k}_n\right) = \underset{(\tilde{h}, k_1, ..., k_n)}{\text{argmin}} \sum_{i=1}^{n} \left\| h_i - \tilde{h}(k_i) \right\|^2,$$

where $(k_1, ..., k_n) \in \{1, ..., K\}^n$.

2. Second step: Maximize the log-likelihood function with respect to common parameters and *group-specific* individual effects, where the groups are given by the $\widehat{k}_i$ estimated in the first step; that is:

$$\left(\widehat{\theta}, \widehat{A}\right) = \underset{(\theta, A)}{\text{argmax}} \sum_{i=1}^{n} \ln f\left(Y_i \,|\, X_i, A_{\widehat{k}_i}, \theta\right).$$

Theoretical justifications for grouping in a grouped world

• In a grouped data generating process, the groups are consistent as n and T tend to infinity (B. and Manresa, 2015, Hahn and Moon, 2010, Lin and Ng, 2012, Su *et al.*, 2016).

• The asymptotic distribution of parameter estimates is not affected by group estimation.

• Conditions for group consistency allow n to grow polynomially faster than T. This does not require long panels.

• Group consistency hinges on groups being sufficiently well separated.

• The number of groups K can be consistently estimated using information criteria such as BIC, or using testing methods (e.g., Lu and Su, 2017).

Theoretical justifications for grouping in a continuous world

• When heterogeneity A_i is continuous, type estimates converge to some "pseudo-true values" as n tends to infinity for T fixed (Pollard, 1981, 1982).

• Moreover, parameter estimates remain consistent for the true parameter values provided K, n, T tend to infinity (B., Lamadon and Manresa, 2022).

• In that case the groups are a regularization device, with K being a tuning parameter.

• The convergence rate depends crucially on the dimensionality of heterogeneity.

5. Applications

A. Dynamic discrete model of location choice

• Choices j_{it}, payoffs W_{it}, states (X_{it}, A_i).

• The likelihood function is (under standard assumptions):

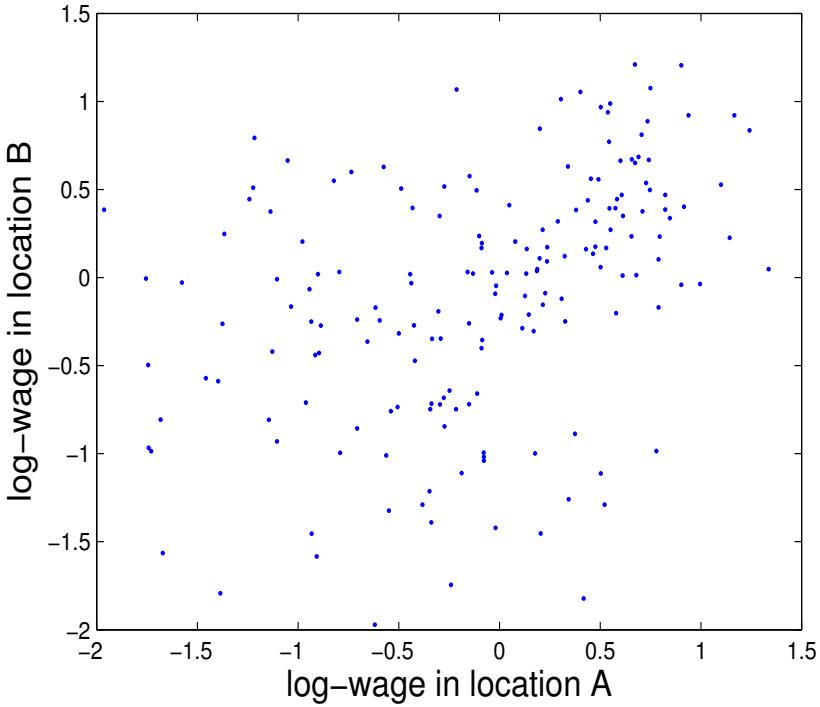
$$\prod_t \underbrace{f\left(j_{it} \mid X_{it}, A_i, \theta\right)}_{\text{choices}} \underbrace{f\left(X_{it} \mid j_{i,t-1}, X_{i,t-1}, A_i, \theta\right)}_{\text{states}} \underbrace{f\left(W_{it} \mid j_{it}, X_{it}, A_i, \theta\right)}_{\text{payoffs}}.$$

• The vector h_i may contain moments of payoff variables or observed state variables, or individual choice probabilities.

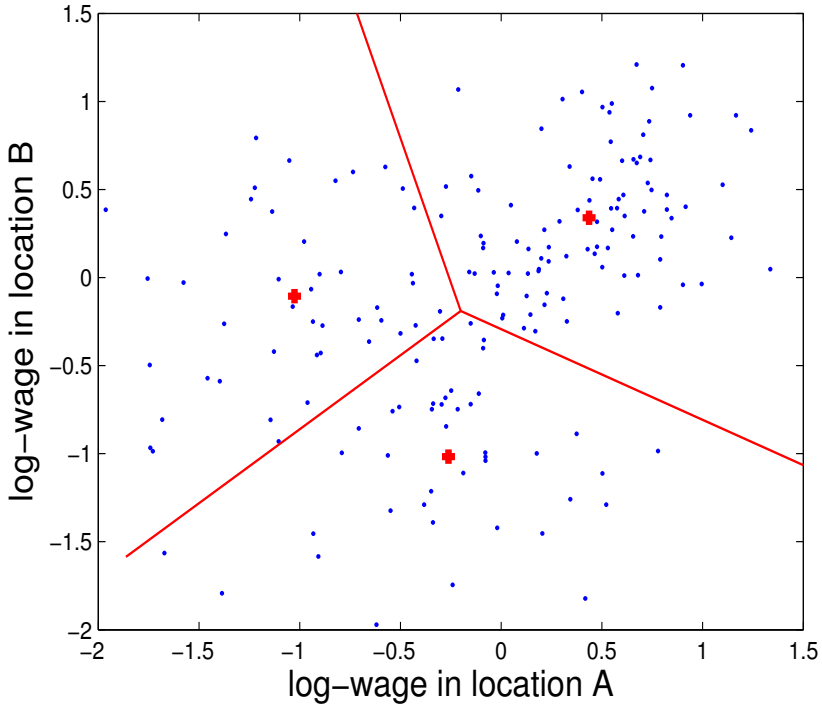
• B., Lamadon and Manresa (2022) estimate a structural model of location choice in the spirit of Kennan and Walker (2011).

K-means clustering with $K = 3$ groups

(a) Data



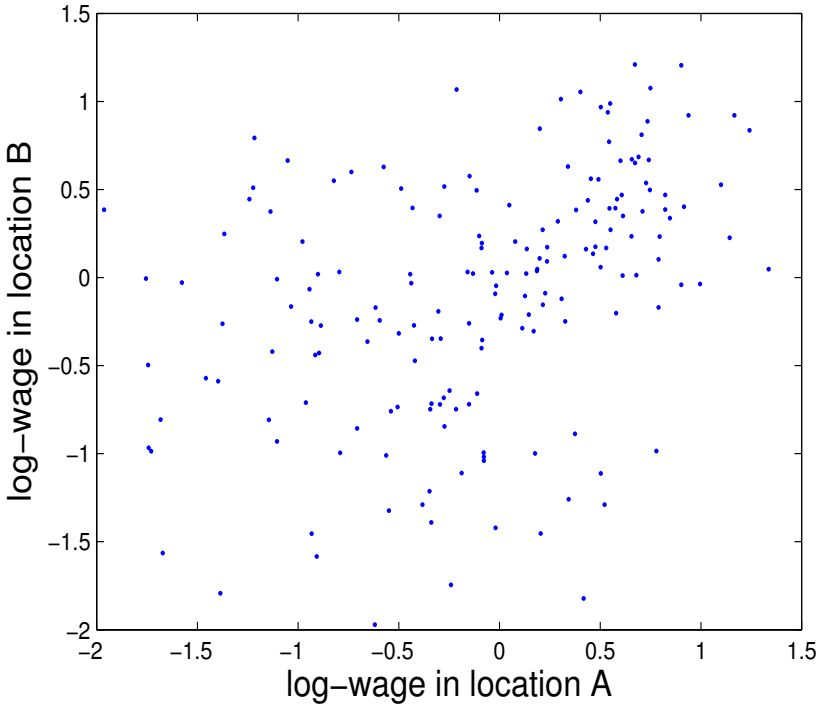
(b) Partition



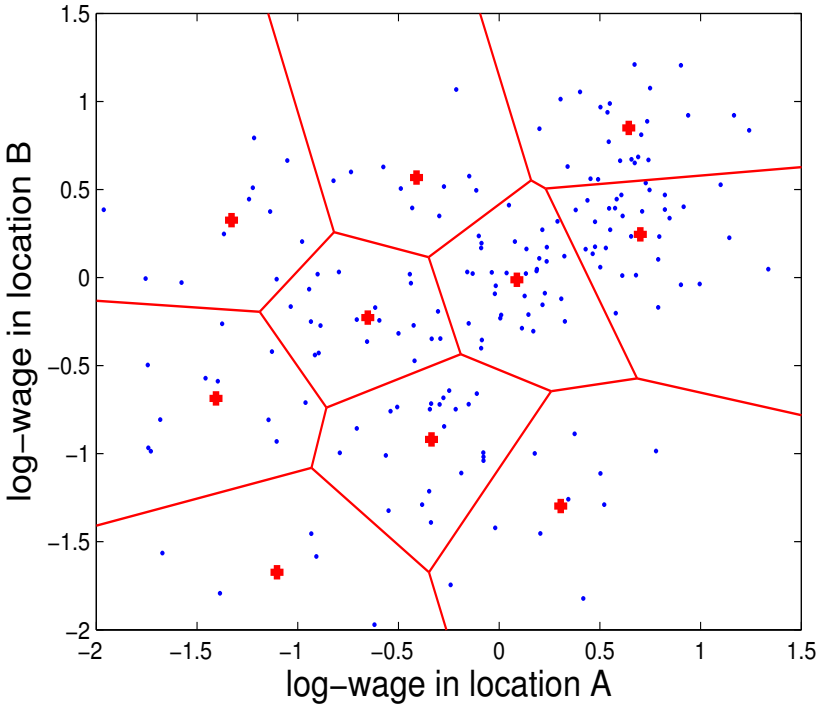
Notes: Average log-wages in locations A (East) and B (West), for male movers in NLSY79.

K-means clustering with $K = 10$ groups

(a) Data



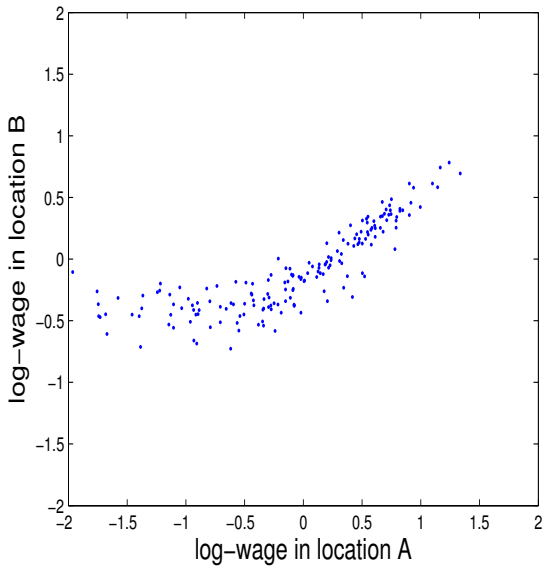
(b) Partition



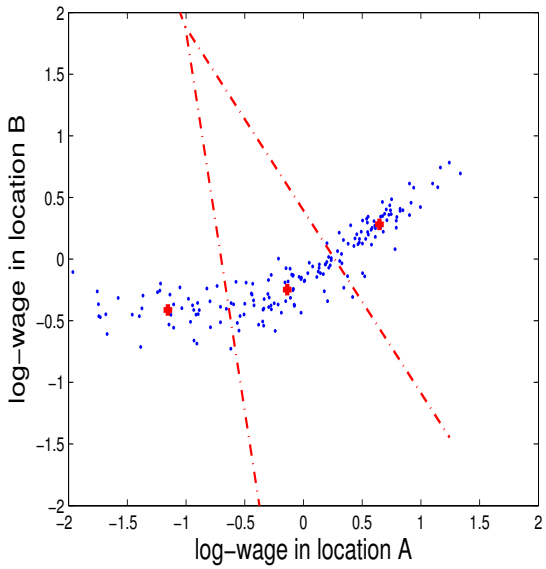
Notes: Average log-wages in locations A (East) and B (West), for male movers in NLSY79.

K-means in the presence of a low underlying dimension

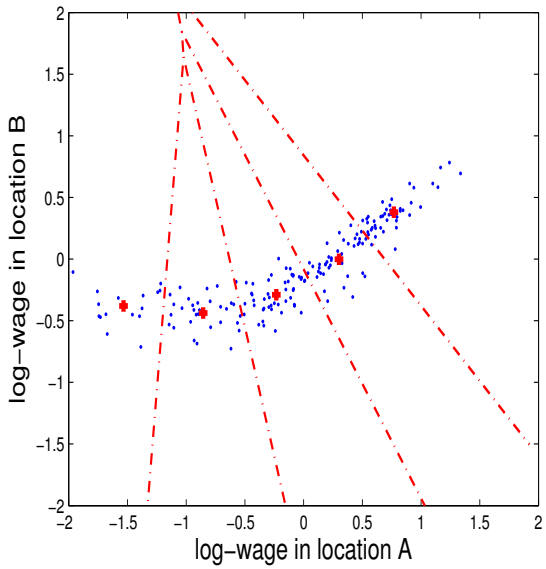
(a) Data



(b) $K = 3$ groups



(c) $K = 5$ groups



Notes: Sample with the same conditional mean as in the NLSY79, and one third of the conditional standard deviation.

Model (in the spirit of Kennan and Walker, 2011)

- Locations: $j_{it} \in \{1, ..., J\}$. Log-wages in j: $\ln W_{it} = A_i(j) + \varepsilon_{it}(j)$.

- Utility: $U_{it}(j) = \rho W_{it}(j) + \xi_{it}(j)$. Mobility costs: $c_{j,j'}$.

- Agents face uncertainty about their types $A_i(\cdot)$ in future locations. They have rational expectations, and solve an infinite-horizon optimization problem with continuous heterogeneity.

- Discrete estimation: in the first step, estimate $\widehat{A}(j_{it}, \widehat{k}_i)$ by discretizing location-specific log-wage means. In the second step, estimate utility and cost parameters by solving the dynamic problem.

- There are K state variables given any history of locations.

Calibration and simulation

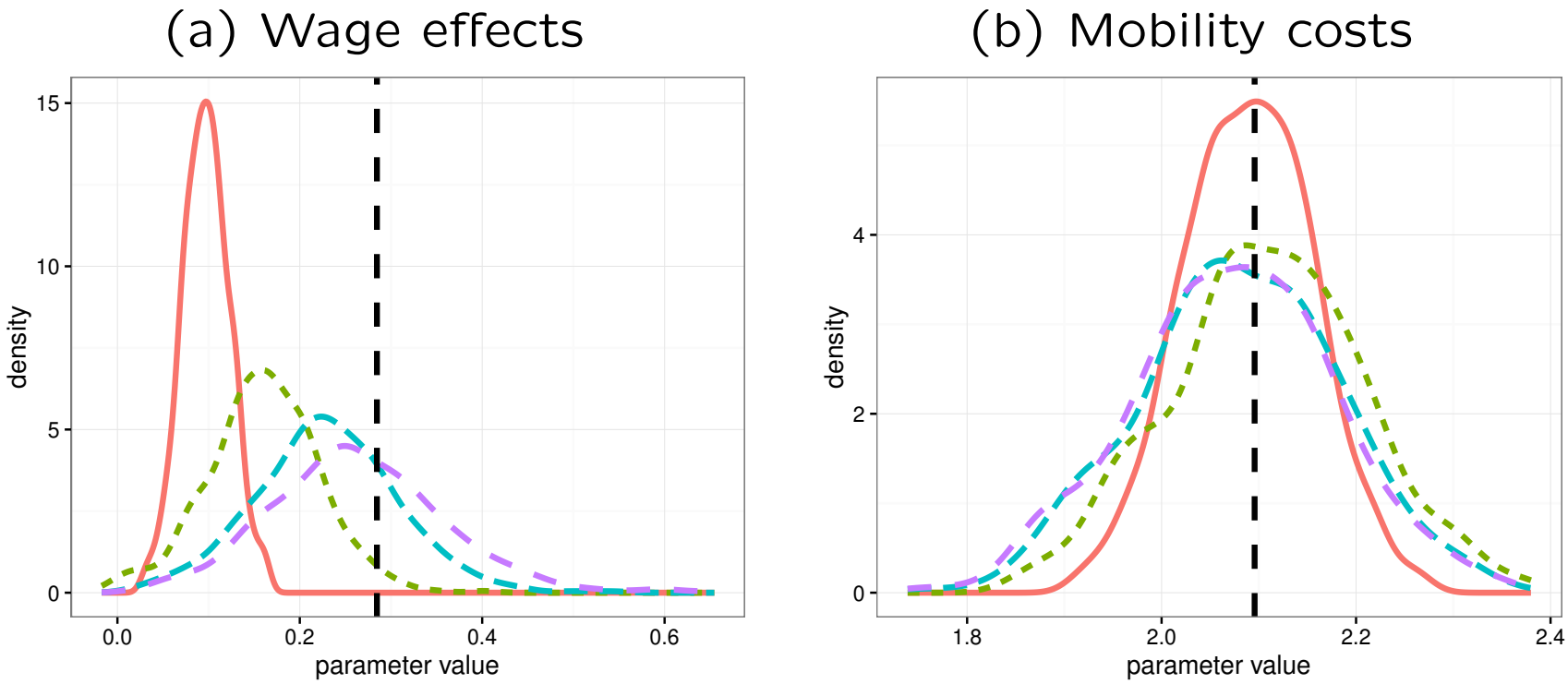
• NLSY79 data used to calibrate parameter values. Males older than 22, followed until 1994. Log-wage residuals net of age, race and education dummies. Two regions: A (North-East and South) and B (North-Central and West). Wages are 9% higher in A. Mobility rates are low (1.5% yearly).

• $K = 10$, and, following Kennan and Walker (2011), there is a probability to be a "stayer type". Estimates show a positive wage effect $\widehat{\rho} = .28$ and large costs $\widehat{c} = 2.1$ (assumed symmetric).

• Given parameter values, the model with continuous heterogeneity is simulated. $(A_i(1), A_i(2))$ is assumed to be bivariate normal.

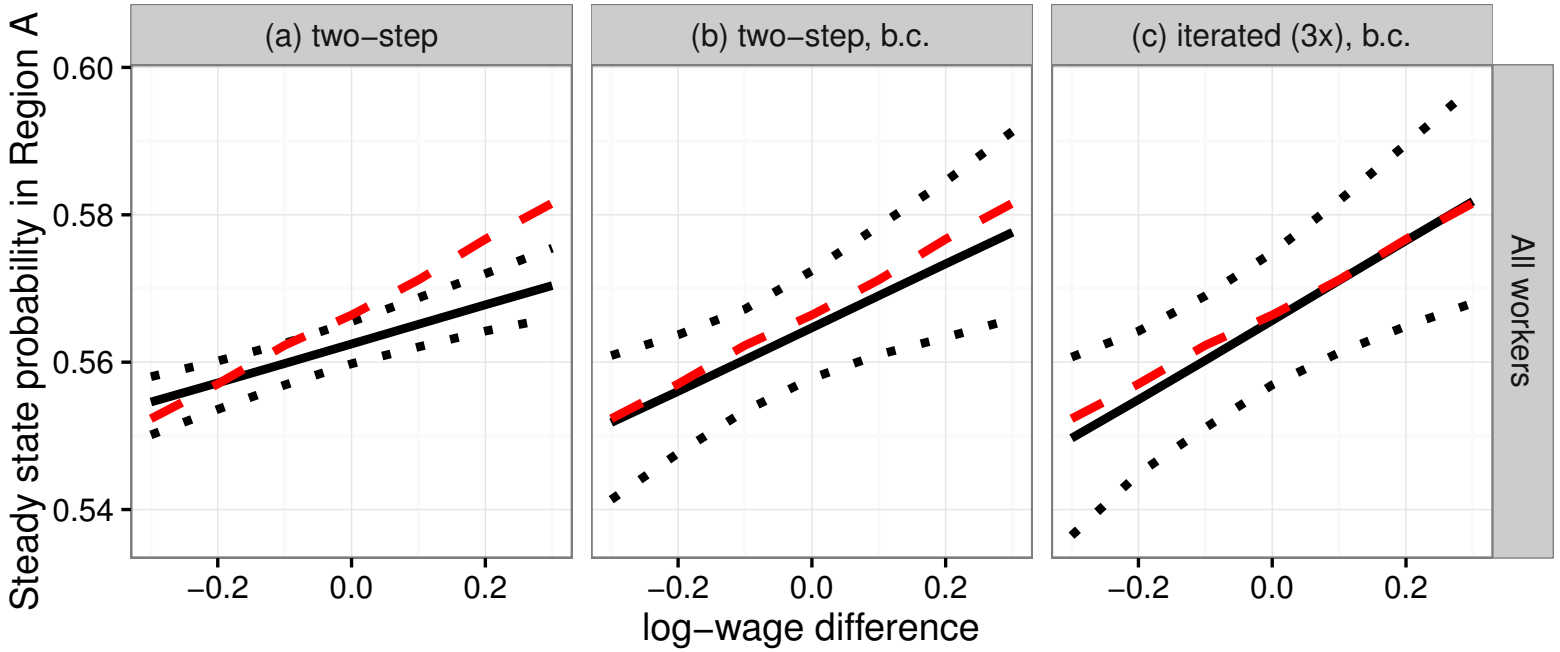
• The model is estimated using two-step GFE.

Parameter estimates (average $\widehat{K}$ is 7)



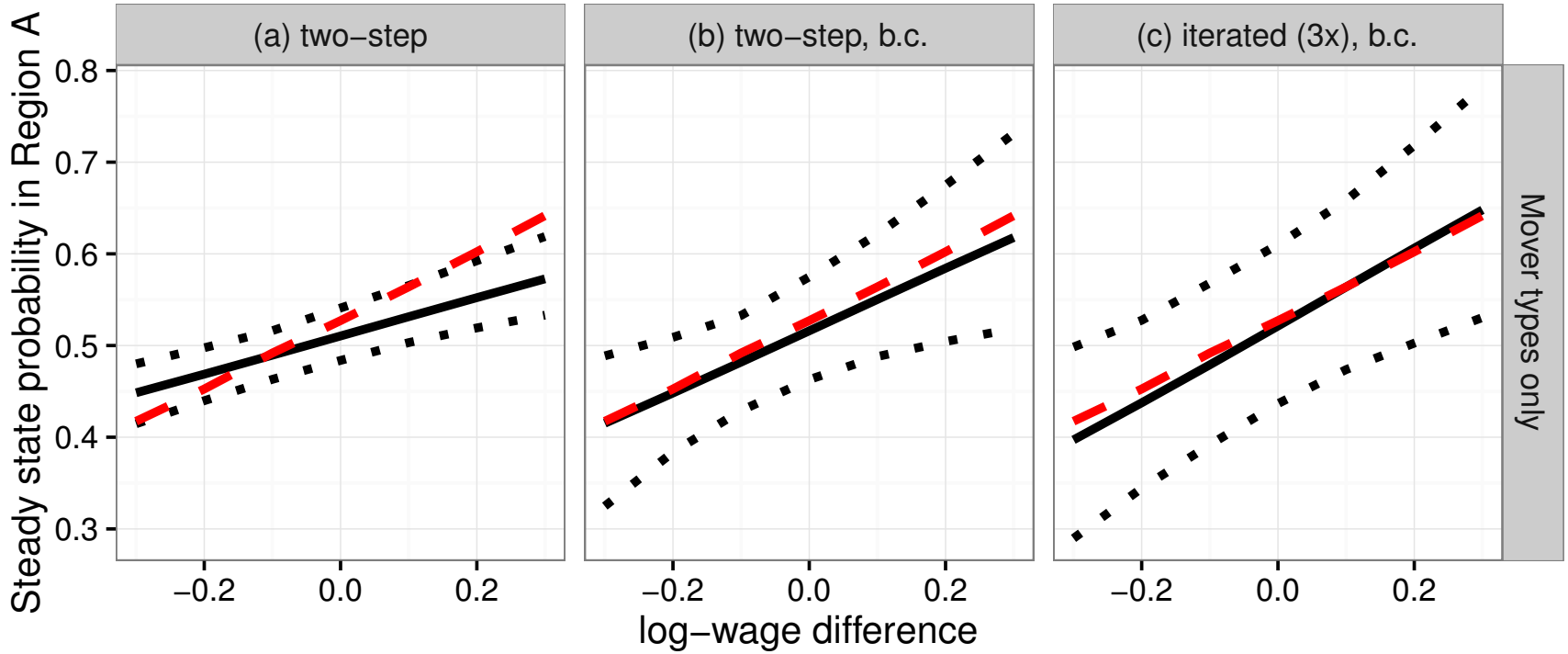
Notes: Solid is two-step grouped fixed-effects, dotted is bias-corrected, dashed is iterated once and bias-corrected, dashed-dotted is iterated three times and bias-corrected. The vertical line indicates the true parameter value. $N = 1889$, $T = 16$, 500 *replications.*

Long-run effects of wages on location



Notes: (a) *is two-step grouped fixed-effects,* (b) *is bias-corrected,* (c) *is iterated three times and bias-corrected. The dashed curve indicates the true value. Solid curves are means, and dotted curves are 97.5% and 2.5% percentiles, across simulations.* 500 *replications.*

Long-run effects of wages on location (mover types only)



Notes: (a) *is two-step grouped fixed-effects,* (b) *is bias-corrected,* (c) *is iterated three times and bias-corrected. The dashed curve indicates the true value. Solid curves are means, and dotted curves are 97.5% and 2.5% percentiles, across simulations.* 500 *replications.*

B. Worker and firm heterogeneity

• Important questions: firm/worker sorting and optimal allocations, sources of earnings inequality...

• Two literatures have approached these questions from different angles using matched employer employee data:

-Two-way fixed-effects regressions, since Abowd, Kramarz and Margolis (1999, AKM).

-Structural models of worker and firm sorting, inspired by Becker (1974).

• B., Lamadon and Manresa (2019) develop a framework with the aim of building a bridge between these two approaches.

Fixed-effects regressions

- AKM regression:

$$log\, earnings = worker\, FE + firm\, FE + covariates + error\, term.$$

- This allows documenting the association $cov(worker\, FE, firm\, FE)$, and more generally the contributions of workers and firms to earnings dispersion.

- Provides a tractable way of allowing for two-sided heterogeneity using matched data.

- Widely applied method, in labor economics and outside (schools, hospitals, cities...).

Two features of the fixed-effects model

1. Additive model (in logs). Hence a very specific form of complementarity between worker and firm unobservables.

-Feature emphasized in the theoretical and structural literature on sorting models (Becker 1974, Shimer and Smith 2000, Eeckhout and Kircher 2011, Hagedorn Law and Manovskii 2017, among others).

2. Static model. In particular, earnings after a job move do not depend on the previous employer given the current one.

-This is potentially inconsistent with wage posting models, or models with bargaining for example.

Heterogeneity

- N workers, J firms, T periods.

- Firms are characterized by the *class* they belong to, $k_{it} \in \{1, ..., K\}$ denoting the class of firm j_{it}.

- $k_{it} = k(j_{it})$ could be the firm itself (when $K = J$), or a firm characteristic such as industry. B., Lamadon and Manresa (2019) assume discrete firm heterogeneity and use two-step GFE.

- Unobserved worker types A_i may be discrete or continuous.

Static model, two periods

• Period 1:

-A type-a worker in a class-k firm draws log-earnings Y_{i1} from $F_{ka}(y_1)$.

• Period 2:

-The worker moves to a class-k' firm with a probability that depends on a and k (and k'), not on Y_{i1}.

-If she moves, the worker draws log-earnings Y_{i2} from a distribution $F^m_{k'a}(y_2)$ that depends on a and k', not on (k, Y_{i1}).

• Two assumptions: 1) mobility is driven by types/classes, 2) serial independence upon job change.

Dynamic model, four periods

• Periods 1 and 2: A type-a worker in a class-k firm draws log-earnings (Y_{i1}, Y_{i2}) from a bivariate distribution that depends on (a, k).

• Period 3:

-The worker moves to a class-k' firm with a probability that depends on a, k and Y_{i2} (and k'), not on Y_{i1}.

-If she moves, the worker draws log-earnings Y_{i3} from a distribution that depends on a, k', k, Y_{i2}, not on Y_{i1}.

• Period 4: The worker then draws log-earnings Y_{i4} from a distribution that depends on a, k', Y_{i3}, not on (k, Y_{i2}, Y_{i1}).

Link to theoretical models

• Static model:

-Example: Shimer and Smith (2000), without or with on-the-job search (workers' threat points being the value of unemployment).
-No role for match-specific draws, unless independent over time or measurement error. No sequential auctions.

• Dynamic model:

-Models where state variables (a, k_t, Y_t) are first-order Markov.
-Examples: wage posting, sequential auctions (Lamadon, Lise, Meghir and Robin 2015), with aggregate shocks (Lise and Robin 2014).
-No latent human capital accumulation (a_t), no permanent+transitory within-job earnings dynamics (example: random walk+i.i.d. shock).

Main equations (static model)

• Let m_{it} denote mobility between t and $t+1$. In the static model on two periods we have:

$$\Pr\left[Y_{i1} \leq y_1\,,\, Y_{i2} \leq y_2 \,|\, m_{i1} = 1, k_{i1} = k, k_{i2} = k'\right] \tag{1}$$

$$= \int p_{kk'}(a)\, F_{ka}(y_1) F^m_{k'a}(y_2) da, \tag{MOV}$$

$$\Pr\left[Y_{i1} \leq y_1 \,|\, k_{i1} = k\right] = \int F_{ka}(y_1) q_k(a)\, da. \tag{CROSS}$$

• Firm classes $k_{it} = k(j_{it})$ can be recovered using (MOV) and/or (CROSS).

• BLM show how to recover F_{ka}, $F^m_{k'a}$, and $p_{kk'}(a)$ from (MOV) using job movers, and $q_k(a)$ from (CROSS) using the first cross-section.

Identification under discrete worker types (sketch)

• Let $k \neq k'$ be firm classes, and write (MOV) for (k, k') transitions in matrix form:

$$A(k, k') = F(k)D(k, k')F(k')',$$

where for simplicity assume that $F_{ka} = F_{ka}^m$, and that the matrices $F(k) = [F_{ka}(y_1)]_{(y_1,a)}$ are squared.

• If there are movers of all types making $k \mapsto k'$ and $k' \mapsto k$ transitions, and under suitable rank conditions:

$$A(k, k')A(k', k)^{-1} = F(k)D(k, k')D(k', k)^{-1}F(k)^{-1}$$

identifies F_{ka} up to scale (pinned down as it is a cdf) and labeling of the types.

• With continuous types these become operator restrictions, and identification requires completeness (as in Hu and Schennach, 2008).

Recovering firm classes

• Identification results hold at the class level k_{it}, which could in principle coincide with the firm j_{it}.

• However, in typical matched datasets allowing for firm-specific parameters increases computational cost and raises statistical challenges (e.g., incidental parameter "low mobility" bias).

• BLM propose a dimension reduction method, which may be performed in an initial estimation step. We provide conditions for consistency of the firm classification under discrete firm heterogeneity.

• The classification is based on earnings, although BLM use a variety of other approaches, using other data (worker flows, value-added).

Classification using k-means

• Unobservable firm heterogeneity operates at the level of firm classes in the model, not at the level of individual firms.

• For example, the log-earnings cdfs F_j should be the same for all firms j in class $k = k(j)$.

• This motivates the following k-means estimator:

$$\min_{k(1),...,k(J),H_1,...,H_K} \sum_{j=1}^{J} n_j \sum_{d=1}^{D} \left(\widehat{F}_j(y_d) - H_{k(j)}(y_d) \right)^2,$$

where $\widehat{F}_j$ denotes the empirical cdf of log-earnings in firm j, n_j is the number of workers in firm j, and $y_1, ..., y_D$ are wage quantiles.

Two-step GFE estimation

1. Estimate the classes $\widehat{k}_{it} = \widehat{k}(j_{it})$ using clustering.

2. Estimate the model's parameters given the $\widehat{k}_{it}$.

• Wage cdfs F_{ka} and F^m_{ka} are Gaussian with (k, a)-specific means and variances. BLM use the EM algorithm for computation.

• A key feature is that, conditional on the firm classes, the likelihood factors across workers.

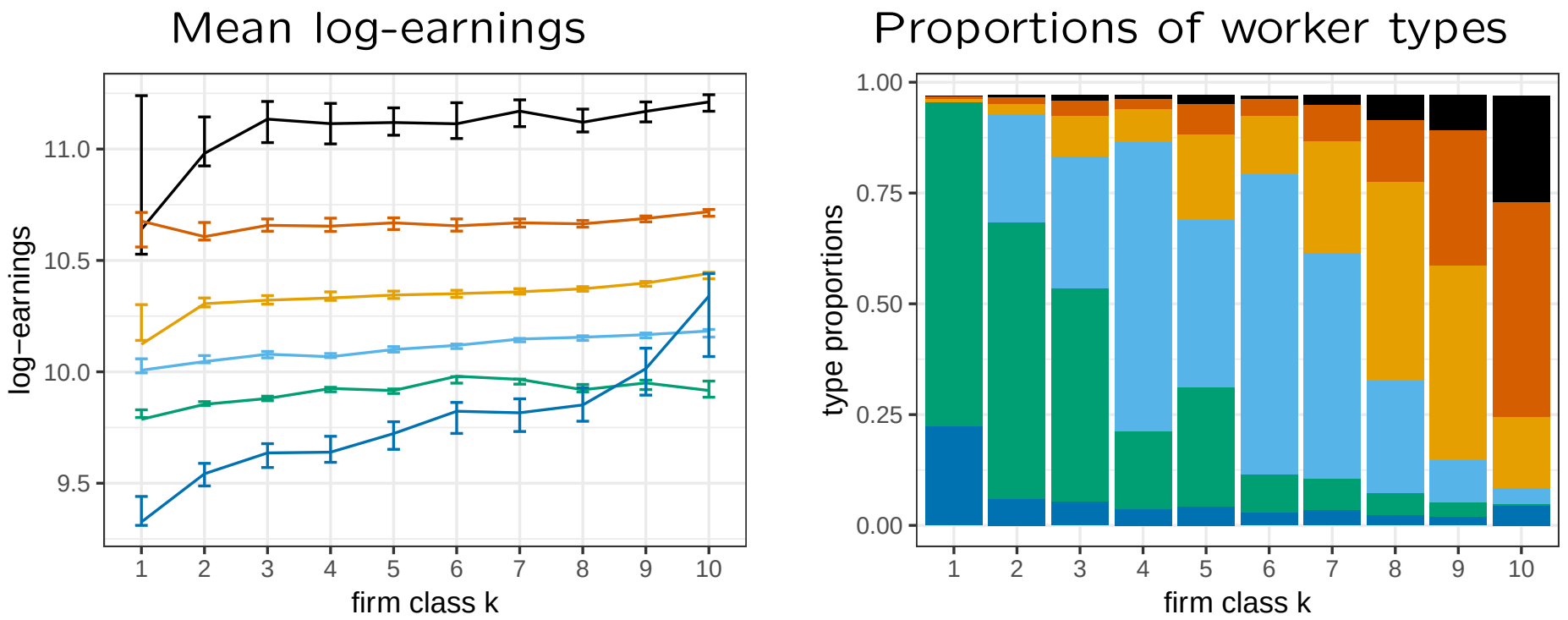
• Hence, estimating firm classes in a first step gets us back to a single-agent problem!

Descriptive statistics on estimated firm classes

class:	1	2	3	4	5	6	7	8	9	10	all
number of workers	16,868	50,906	74,073	76,616	80,562	66,120	105,485	61,272	47,164	20,709	599,775
number of firms	5,808	6,832	4,983	5,835	3,507	4,149	3,672	3,467	2,886	2,687	43,826
mean firm reported size	12.43	20.92	42.68	28.47	65.06	32.30	60.08	51.24	54.16	50.86	37.59
number of firms ≥ 10 (actual size)	160	1,034	1,519	1,357	1,192	930	999	855	632	415	9,093
number of firms ≥ 50 (actual size)	7	87	260	225	270	162	245	183	147	52	1,638
% high school drop out	28.5%	27.8%	25.9%	26.8%	22.2%	23.8%	18.9%	12.9%	6.1%	3.2%	20.6%
% high school graduates	61.3%	63.4%	62.3%	63.3%	59.1%	62.7%	58.4%	49.3%	34.9%	25.6%	56.7%
% some college	10.2%	8.8%	11.8%	9.9%	18.7%	13.5%	22.8%	37.8%	59.0%	71.2%	22.7%
% workers younger than 30	24.3%	19.5%	19.8%	17.5%	18.6%	15.4%	13.8%	14.3%	15.0%	14.3%	16.8%
% workers between 31 and 50	54.1%	54.6%	55.0%	56.2%	56.0%	57.6%	58.5%	58.9%	60.0%	64.2%	57.2%
% workers older than 51	21.7%	25.9%	25.1%	26.3%	25.5%	27.0%	27.6%	26.8%	25.0%	21.5%	26.0%
% workers in manufacturing	24.3%	39.3%	46.8%	53.0%	51.5%	52.0%	53.0%	40.3%	31.5%	7.6%	45.4%
% workers in services	39.3%	32.1%	23.3%	19.7%	14.4%	15.0%	16.0%	29.7%	52.1%	72.6%	25.3%
% workers in retail and trade	26.4%	19.0%	24.9%	10.6%	29.3%	7.9%	8.4%	17.7%	14.8%	18.7%	16.7%
% workers in construction	9.9%	9.6%	5.1%	16.8%	4.9%	25.1%	22.5%	12.3%	1.5%	1.1%	12.6%
mean log-earnings	9.69	9.92	10.01	10.06	10.15	10.16	10.24	10.36	10.50	10.77	10.18
variance of log-earnings	0.101	0.054	0.085	0.051	0.102	0.051	0.077	0.096	0.109	0.173	0.124
skewness of log-earnings	-1.392	-0.709	0.345	0.019	0.576	0.433	0.474	0.703	0.385	1.001	0.582
kurtosis of log-earnings	7.780	14.093	9.017	15.565	7.788	14.763	10.033	8.141	6.651	6.984	7.400
between-firm variance of log-earnings	0.0462	0.0044	0.0036	0.0018	0.0032	0.0016	0.0016	0.0045	0.0057	0.0435	0.0475
mean log-value-added per worker	12.40	12.58	12.69	12.69	12.84	12.75	12.87	12.94	13.03	13.18	12.74

Notes: Males, fully employed in the same firm 2002 and 2004, continuously existing firms. Figures for 2002.

Estimated mean log-earnings and proportions of worker types



Notes: Static model, 2002-2004. The left graph plots estimates of the means of log-earnings distributions, by worker type and firm class. The right graph shows estimates of the proportions of worker types in each firm class. Left: brackets indicate parametric bootstrap 2.5% and 97.5% quantiles (200 replications).

C. Quantifying human inputs in teams

• How much do individuals contribute to team output? This is a central question in research and innovation, labor economics, sports...

• B. (forthcoming) proposes an econometric framework to quantify individual contributions when only the output of their teams is observed.

• The identification strategy relies on following individuals who work in different teams over time.

• Two applications: to estimate the impact of economists on research quality, and the contributions of inventors to the quality of their patents.

Framework: workers and teams

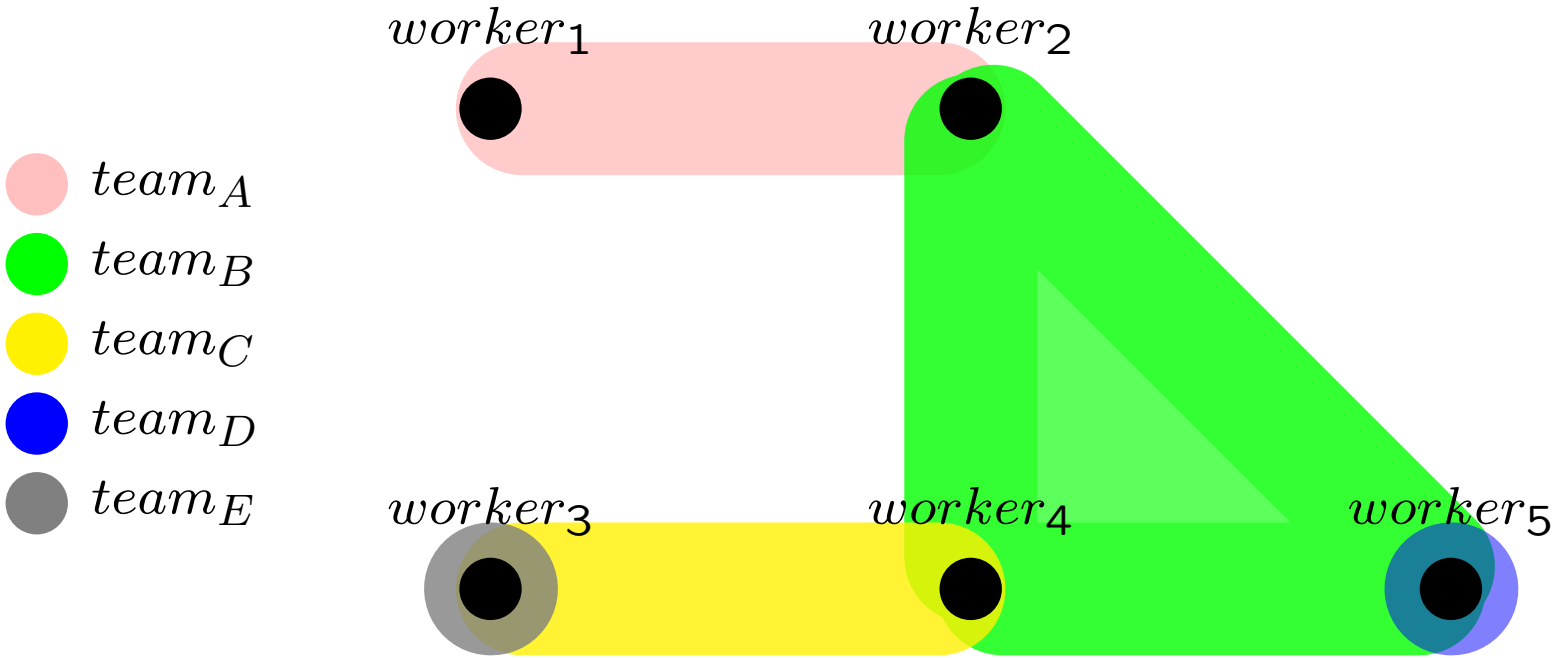
• Workers $i = 1, ..., N$ are nodes in the hypergraph. Workers collaborate and produce output in teams.

• Workers are linked to one another by hyperedges.

• Hyperedges are referred to as teams, and the set of workers in a team j of size n us denoted as $\{i_1(j), ..., i_n(j)\}$.

• Worker i contributes to the team a quantity A_i. The type A_i is constant across collaborations.

Collaboration hypergraph: an example



Framework: production

• The output Y_{nj} of a team j with n workers is given by

$$Y_{nj} = \phi_n(A_{i_1(j)}, ..., A_{i_n(j)}, \varepsilon_{nj}),$$

where ϕ_n is the production function of an n-worker team, symmetric with respect to its first n arguments.

• ε_{nj} represent other factors, or shocks, that affect team output beyond workers' inputs A_i.

• The model has no "team effects". Here a team is simply a collection of workers, plus an ε shock.

Framework: main assumptions

• Assumption 1 (Network exogeneity):
ε_{nj}'s are independent of $(i_1(j), ..., i_n(j))$'s conditional on A_i's.

- Assumption 1 restricts team formation. It will fail if, before joining a team, workers have advanced information about ε_{nj}.

• Assumption 2 (Independent shocks):
ε_{nj}'s are independent, and independent of $(i_1(j), ..., i_n(j))$'s and A_i's.

- Assumption 2 implies in particular that shocks to workers who collaborate repeatedly over time are serially independent.

Nonlinear production

• A_i takes at most K values.

• Random-effects approach: $A_1, ..., A_N$ are drawn from the joint distribution $\prod_i \pi(A_i)$, where $\pi(A_i)$ are type probabilities.

• An important question is how types A_i correlate with collaborations $(i_1(j), ..., i_n(j))$. B. (forthcoming) reports results based on three approaches:

- Independent random-effects (RE).

- Correlated RE (using degrees as worker characteristics).

- Joint RE, modeling both production and team formation (using a stochastic blockmodel of team formation).

Nonlinear production: identification

• Suppose A_i's are independent of $(i_1(j), ..., i_n(j))$'s, and team size is $n \in \{1, 2\}$.

• Focusing on workers who produce at least three times on their own gives:

$$\Pr\left[Y_{1j_1} \leq y_1, Y_{1j_2} \leq y_2, Y_{1j_3} \leq y_3 \,|\, i_1(j_1) = i_1(j_2) = i_1(j_3)\right]$$
$$= \sum_a \underbrace{\pi(a)}_{\text{Type prop.}} \prod_{r=1}^{3} \underbrace{\Pr\left[Y_{1j_r} \leq y_r \,|\, A_{i_1(j_r)} = a\right]}_{\text{Solo production}}.$$

• This identifies $\Pr\left[Y_{1j} \leq y \,|\, A_{i_1(j)} = a\right]$ and $\pi(a)$ under suitable rank conditions (Allman *et al.*, 2009, Hu, 2008).

Nonlinear production: identification (cont.)

• Pairwise productions are then identified by focusing on pairs of workers who produce once on their own and once together:

$$\begin{aligned}
&\Pr\left[Y_{1j_1} \leq y_1, Y_{1j_2} \leq y_2, Y_{2j_3} \leq y_3 \,|\, \{i_1(j_1), i_1(j_2)\} = \{i_1(j_3), i_2(j_3)\}\right] \\
&= \sum_{a,a'} \pi(a)\pi(a') \Pr\left[Y_{1j_1} \leq y_1 \,|\, A_{i_1(j_1)} = a\right] \Pr\left[Y_{1j_2} \leq y_2 \,|\, A_{i_1(j_2)} = a'\right] \\
&\qquad \times \underbrace{\Pr\left[Y_{2j_3} \leq y_3 \,|\, A_{i_1(j_3)} = a, A_{i_2(j_3)} = a'\right]}_{\text{Pairwise production}}.
\end{aligned}$$

• This argument relies on workers who produce on their own (though see Allman *et al.*, 2011).

Nonlinear production: random-effects (RE) likelihood

• Suppose that output density depends on a finite-dimensional θ.

• In the applications: log-normal specification in the sample of economists, and negative binomial specification in the sample of inventors.

• However, the RE likelihood

$$\mathcal{L}(\theta, \pi) = \sum_{a_1} ... \sum_{a_N} \prod_i \pi(a_i) \prod_n \prod_j f^n_{a_{i_1(j)},...,a_{i_n(j)}} \left(Y_{nj}; \theta \right)$$

involves an intractable N-dimensional sum over all possible worker type realizations.

• The likelihood does not factor in simple ways, except in the special case where all teams consist of one worker.

Nonlinear production: variational approximation

• To reduce computational complexity, B. (forthcoming) follows a mean-field variational approach (Bishop, 2006).

• The idea is to introduce an auxiliary distribution $\prod_{i=1}^{N} q_i(a_i)$, and set it to be as close as possible to the posterior density $p(a_1, ..., a_N; \theta, \pi)$ of $a_1, ..., a_N$.

• Unlike the posterior density p, its variational approximation q factors across i. This makes estimation feasible even in large data sets.

• This idea is widely used in network models such as stochastic block-models (Daudin *et al.*, 2008), and more generally in settings with complex data (e.g., Blei *et al.*, 2003).

Variational approximation (cont.)

- The variational objective function is

$$\mathrm{ELBO}(\theta,\pi) = \max_{q_1,\ldots,q_N} \ln\left(\mathcal{L}(\theta,\pi)\right) - \underbrace{\mathbb{E}_{q_1\ldots q_N} \ln \frac{\prod_i q_i(a_i)}{p(a_1,\ldots,a_N;\theta,\pi)}}_{\text{KL divergence}}.$$

- Equivalently, we have

$$\mathrm{ELBO}(\theta,\pi) = \max_{q_1,\ldots,q_N} \sum_n \sum_j \sum_{a_1} \ldots \sum_{a_n} q_{i_1(j)}(a_1)\ldots q_{i_n(j)}(a_n) \ln f_{a_1,\ldots,a_n}(Y_{nj};\theta) + \sum_i \sum_a q_i(a) \ln \pi(a) - \sum_i \sum_a q_i(a) \ln q_i(a).$$

- This expression is feasible to compute whenever team size n is sufficiently small. One maximizes the "evidence lower bound" ELBO using the EM algorithm (Bishop, 2006, Mariadassou *et al.*, 2010).

Variational approximation: statistical properties

• In stochastic blockmodels, Bickel *et al.* (2013) provide conditions under which the mean-field variational estimator is consistent and asymptotically normal.

• Their conditions rely on the network growing in a particular way.

• Under their asymptotic, the variational estimator and the (intractable) maximum likelihood estimator are first-order equivalent.

• B. (forthcoming) conducts Monte Carlo simulation exercises to probe the accuracy of the variational estimator.

Monte Carlo simulation

	True	Smaller sample			Larger sample		
		Mean	p2.5%	p97.5%	Mean	p2.5%	p97.5%
Mean type 1	0.00	0.00	-0.05	0.06	0.00	-0.03	0.02
Mean type 2	2.00	2.00	1.92	2.07	2.00	1.97	2.03
Var. type 1	0.50	0.50	0.43	0.56	0.50	0.47	0.53
Var. type 2	0.50	0.50	0.42	0.57	0.50	0.47	0.53
Mean type (1,1)	0.00	0.05	-0.49	0.74	0.01	-0.10	0.12
Mean type (1,2)	1.00	1.01	0.59	1.42	1.01	0.92	1.10
Mean type (2,2)	4.00	3.71	0.78	4.72	4.00	3.85	4.16
Var. type (1,1)	0.50	0.46	0.11	0.93	0.50	0.40	0.62
Var. type (1,2)	0.50	0.46	0.16	0.84	0.50	0.42	0.59
Var. type (2,2)	0.50	0.38	0.00	1.24	0.49	0.35	0.69
Prop. type 1	0.60	0.60	0.53	0.67	0.60	0.57	0.63

Notes: Estimates of the random-effects model with $K = 2$ groups, in data generated according to that model with $K_0 = 2$ groups. 500 simulations. The smaller sample has 156 workers and 896 teams, the larger sample has 921 workers and 5447 teams.

Application to economists and research quality

• Data from Ductor *et al.* (2014), drawn from the EconLit database.

- Articles published between 1995 and 1999, with at most three co-authors.

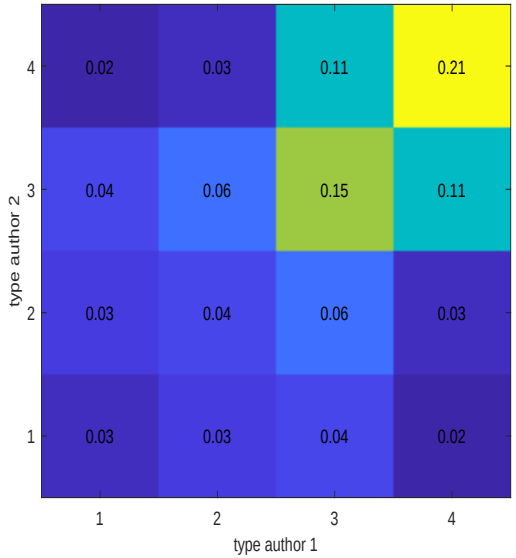
- Only authors who produced at least 5 articles during the period.

- There are 41150 articles and 6509 authors. 50% of authors produce at most 8 articles, while 1% of authors produces at least 39.

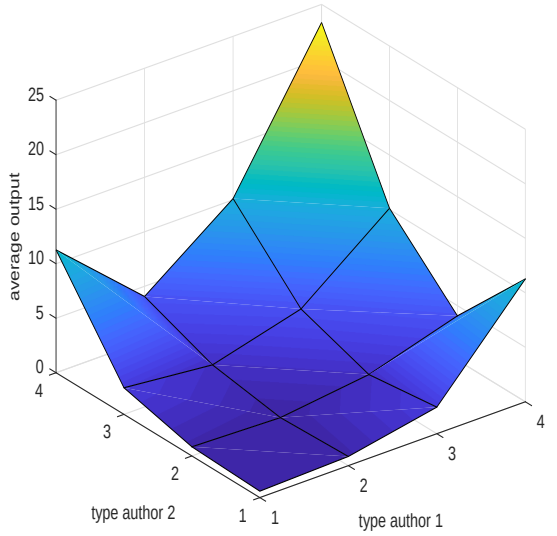
• Output is a measure of journal quality (Kodrzycki and Yu, 2006), which is a ranking between 0 and 100, net of multiplicative time effects.

Nonlinear model estimates, economic researchers ($K = 4$)

(a) Sorting



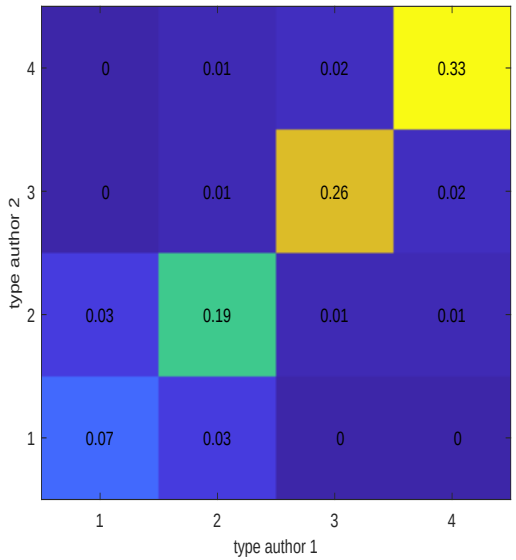
(b) Heterogeneity and complementarity



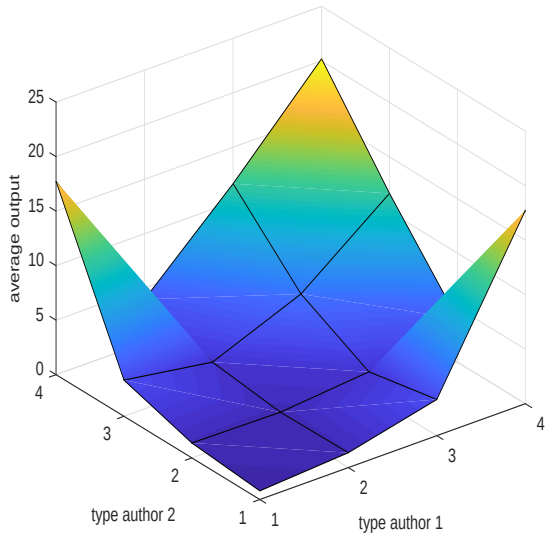
Notes: Random-effects estimates of a finite mixture model with $K = 4$ types. Panel (a) shows the type proportions for authors producing together in 2-author teams. Panel (b) shows average output for different type combinations. Overall type proportions: 15% (type I), 20% (II), 36% (III), 29% (IV).

Nonlinear model estimates, economic researchers: Joint RE

(a) Sorting



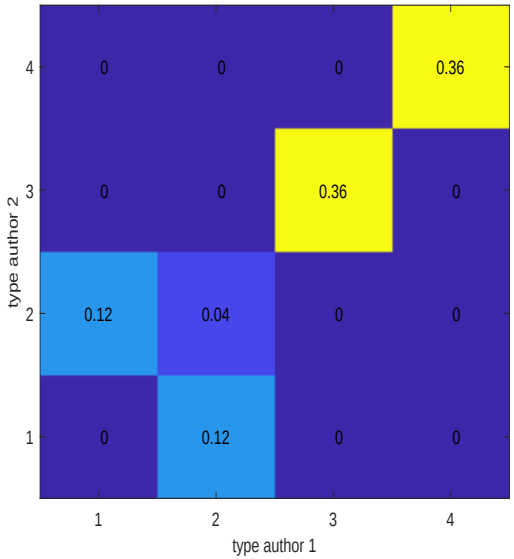
(b) Heterogeneity and complementarity



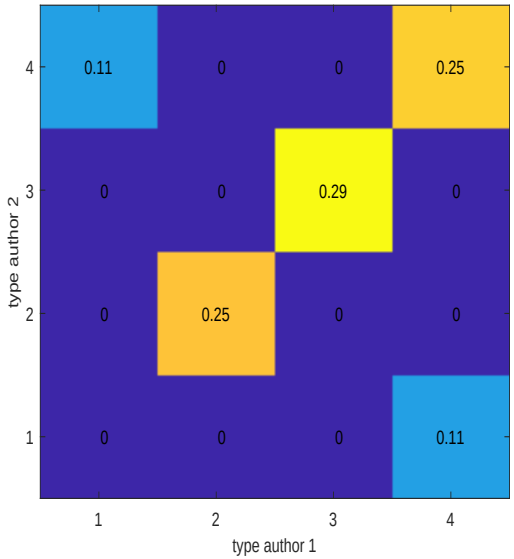
Notes: Random-effects estimates of a finite mixture model with $K = 4$ types, joint RE specification (using a Poisson blockmodel of team formation). Panel (a) shows the type proportions for authors producing together in 2-author teams. Panel (b) shows average output for different type combinations.

Optimal allocation, economic researchers ($K = 4$)

(a) Independent RE



(b) Joint RE



Notes: Proportions of types for authors producing together in a 2-author team, in the allocation that maximizes total output. In panel (a) and (b) the estimates are obtained using independent and joint random-effects, respectively.

Application to inventors and patent quality

• Data from Akcigit *et al.* (2016). Their main source is the disambiguated inventor data of Li *et al.* (2014), which identifies unique inventors in the USPTO data.

- Patents from the US and granted between 1995 and 1999, and in the class "Computers and Communications".

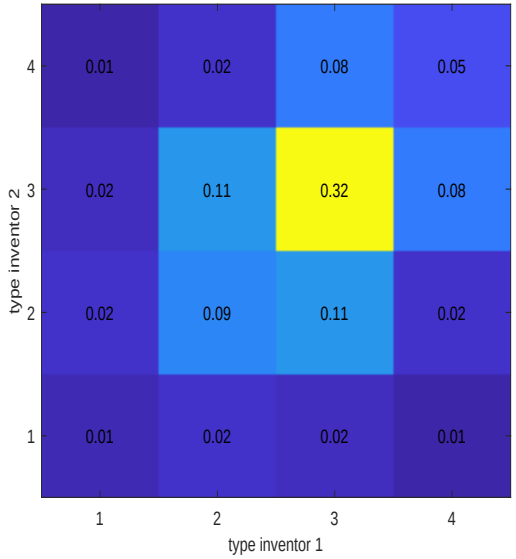
- Only inventors who produced at least 5 patents.

- There are 30068 patents and 5896 inventors.

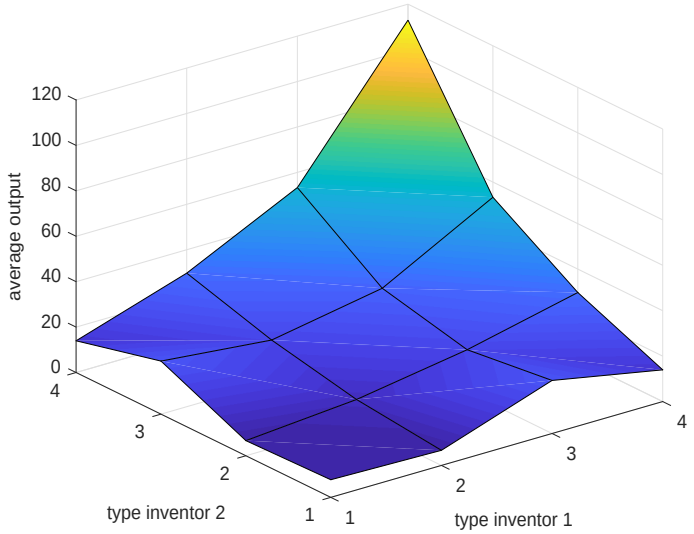
• Output is a Hall *et al.*'s (2001) measure of truncation-adjusted forward citations, net of multiplicative time effects.

Nonlinear model estimates, patents and inventors ($K = 4$)

(a) Sorting



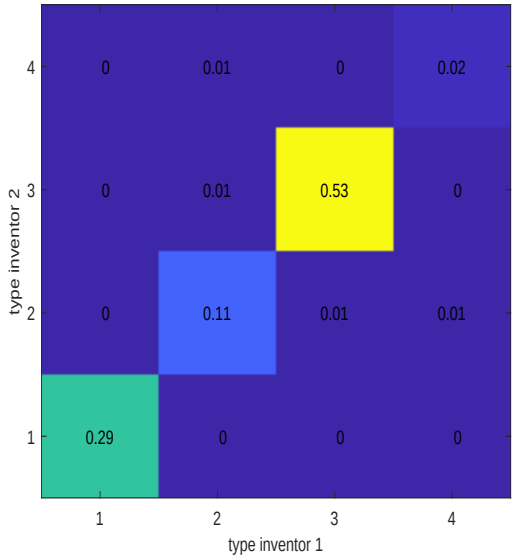
(b) Heterogeneity and complementarity



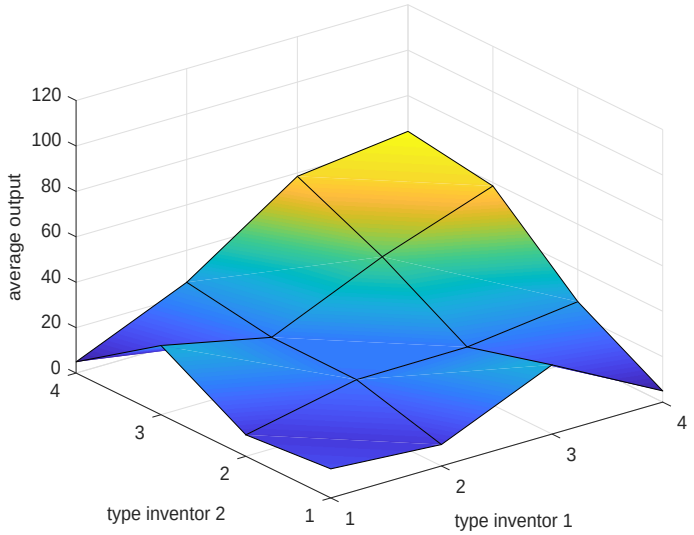
Notes: Panel (a) shows the type proportions for inventors in 2-worker teams. Panel (b) shows average output for different combinations of the types. Overall type proportions: 6% (type I), 25% (II), 52% (III), 17% (IV).

Nonlinear model estimates, patents and inventors: joint RE

(a) Sorting



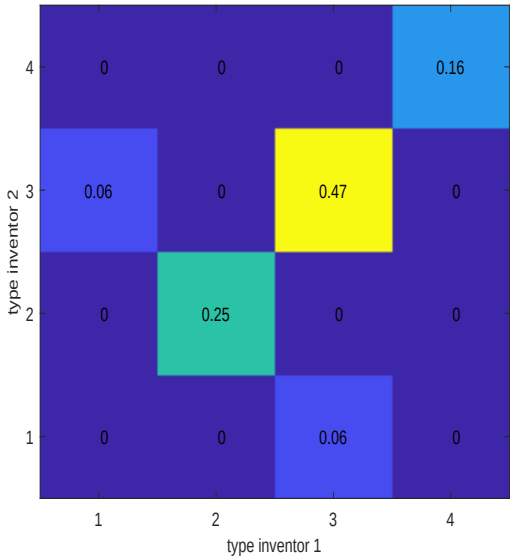
(b) Heterogeneity and complementarity



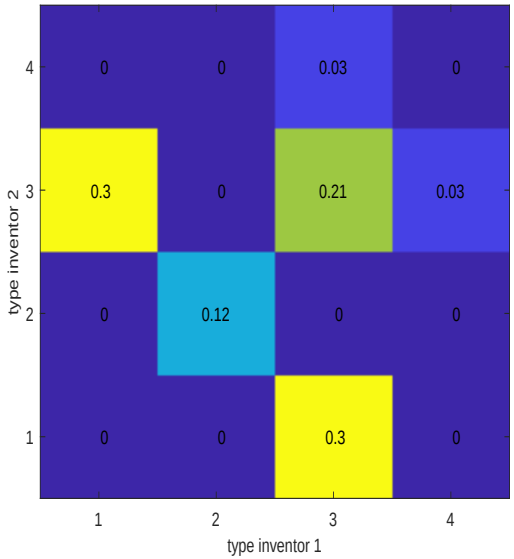
Notes: Joint RE specification (using a Poisson model of team formation). Panel (a) shows the type proportions for inventors in 2-worker teams. Panel (b) shows average output for different combinations of the types.

Optimal allocation, patents and inventors ($K = 4$)

(a) Independent RE



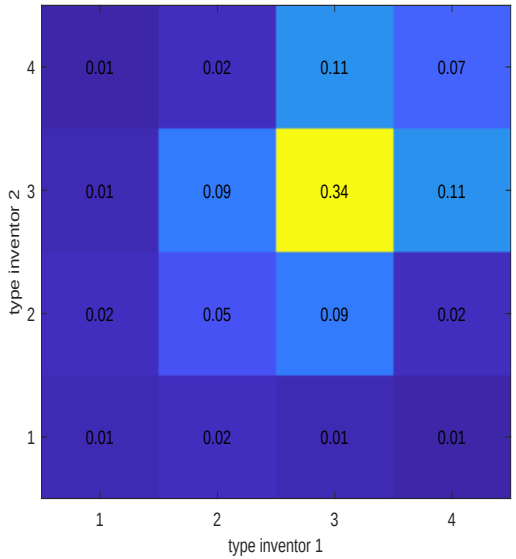
(b) Joint RE



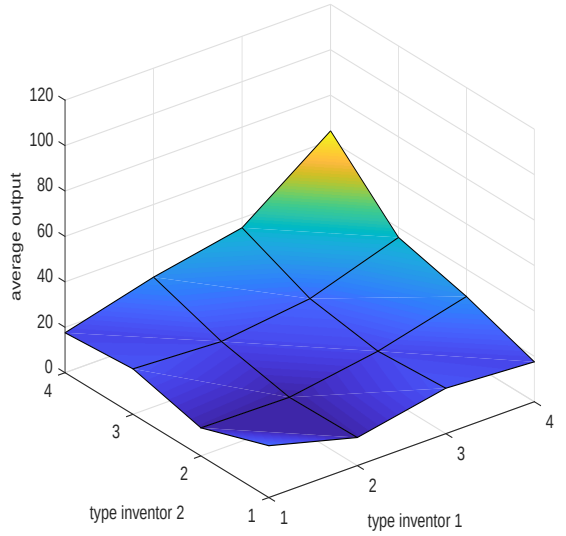
Notes: Proportions of types for inventors producing together in a 2-inventor team, in the allocation that maximizes total output. In panel (a) and (b) the estimates are obtained using independent and joint random-effects, respectively.

Application to inventors and patent quality: Heterogeneity, sorting and complementarity out of sample

(a) Sorting



(b) Heterogeneity and complementarity



Notes: Sample from Akcigit et al. (2016). The model is estimated on the 1995-1999 period, with $K = 4$ types. Inventors producing together in 2-worker teams between 2000 and 2005.

Some references from my own work

Bonhomme (forthcoming): "Teams: Heterogeneity, Sorting, and Complementarity", Proceedings of the 2020 World Congress of the Econometric Society.

Bonhomme (2012): "Functional Differencing", Econometrica.

Bonhomme and Manresa (2015): "Grouped Patterns of Heterogeneity in Panel Data", Econometrica.

Bonhomme, Lamadon and Manresa (2022): "Discretizing Unobserved Heterogeneity", Econometrica. + Working paper version

Bonhomme, Lamadon and Manresa (2019): "A Distributional Framework for Matched Employer-Employee Data", Econometrica.

More at: https://sites.google.com/site/stephanebonhommeresearch/

Some other references

Aguiregabiria and Mira (2010): "Dynamic discrete choice structural models: A survey", Journal of Econometrics.

Hu (2008): "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution", Journal of Econometrics.

Hu (2015): "Microeconomic models with latent variables: applications of measurement error models in empirical industrial organization and labor economics".

Kasahara and Shimotsu (2009): "Nonparametric identification of finite mixture models of dynamic discrete choices", Econometrica.