A Distributional Framework for Matched Employed Employee Data

S. Bonhomme, T. Lamadon, and E. Manresa, ECMA2019

Presenter: Zhongji Wei

September 24, 2023

Question Description & Concepts

- Some questions in Labor
  - What causes earning dispersion : individual or firm?
  - The nature of sorting pattern : why someone in some firm?
- Some key concepts for explanation
  - Matched Data : who in which firm
  - Heterogeneity and Interaction
  - Sorting and Complementarity, Becker (1973)
  - Two-sided Unobserved Heterogeneity
  - Distributional framework: explain from identified distribution

Contents

1. Introduction
2. Framework of Analysis
3. Identification
4. Two-Step Estimation

Introduction

Introduction: Previous Approach

- Approach: identify the contribution of worker and firm (2-sided unobserved) heterogeneity to earning dispersion
- Two angles: reduced and structural
- Reduced: Two Way Fixed Effect, AKM(1999)

$$y_{it} = \mu_y + (x_{it} - \mu_x) + \underbrace{\theta_i}_{\text{Pure Person Effect}} + \underbrace{\psi_{J_{(i,t)}}}_{\text{Pure Firm Effect}} + \varepsilon_{it}$$

  - Lack of firm-worker interaction, restrics complementarity
  - Static: lack of previous earning & firm dependence

Introduction: Previous Approach

- Structural: full-specified theoretical models
  - Example: wage posting, bargaining
  - Portray the interaction between worker and firms
  - Empirical challenge: worker $\times$ firms, curse of dimensionality
  - May be driven by functional form

Introduction: This Paper

- This paper: an empirical framework to reconcile both angles
  - Allowing complementarities, sorting, and dynamics
  - Dimension reduction: discrete firm class and worker type

$$Y_{it} = \underbrace{\rho_t Y_{i,t-1}}_{Dynamic} + a_{1t}(k_{it}) + \underbrace{a_{2t}(k_{i,t-1})}_{Dynamic} + \underbrace{b_t(k_{it})}_{Interact} \alpha_i + X_{it}' c_t + v_{it}$$

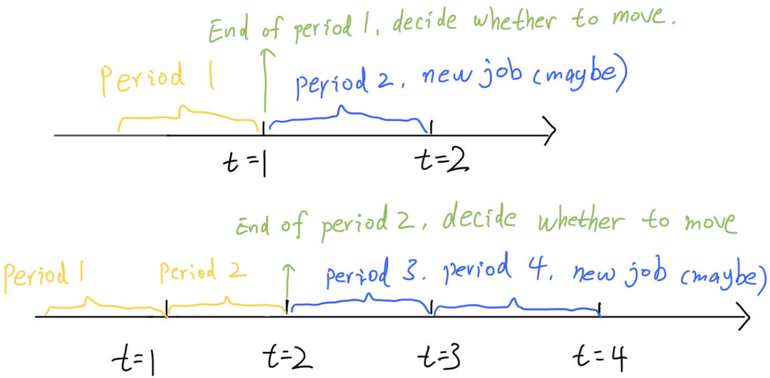
- Identification of income and worker distribution
- 2-step estimation
  - Classification: k-means for firm grouping, given worker type
  - Estimation: estimate key parameters using MLE

Framework of Analysis

Framework: basic settings

- Firms
  - J firms into K classes, $k_{it} = k(j_{it}) \in \{1, 2, ..., K\}$, $j_{it} \in \{1, 2, ..., J\}$
- Workers
  - N workers, at period t, worker i's "state": $(Y_{it}, j_{it}, m_{it}, X_{it})$
  - Earning, firm class, job moving choice at the end of period, other characteristics
  - Each worker belongs to a type α_i time-invariantly
- History of Z_{it}: $Z_i^t = (Z_{i1}, ..., Z_{it})$
- What to find (identify)
  - Given (k, α), the earning distribution
  - The proportion of type α worker in class k firm

Framework: Timing



Framework: Static Model

- "Model": what determines (the distribution of) a variable
- **Assumption 1.1 (Mobility determinant)**:

$$m_{it} \sim F_m(\cdot \mid \alpha_i, k_i^t, m_i^{t-1}, X_i^t), \perp Y_i^t$$

  Similar for $k_{i,t+1}$ and $X_{i,t+1}$

- **Assumption 1.2 (Serially Independence)**

$$Y_{it+1} \sim F_Y(\cdot \mid \alpha_i, k_{it+1}, X_{it+1}, m_{it} = 1), \perp Y_i^t, k_i^t, m_i^{t-1}, X_i^t$$

- Example, reduced to AKM when $b_t(k) = 1$, $K = J$

$$Y_{it} = a_t(k_{it}) + b_t(k_{it})\alpha_t + X_{it}'c_t + \varepsilon_{it} \tag{1}$$

  where $E(\varepsilon_{it} \mid \alpha_i, k_i^T, m_i^T, X_i^T) = 0$

Framework: Dynamic Model

- Introduce dynamic using first-order Markov Property
- **Assumption 2.1 (Mobility determinant)**

$$m_{it} \sim F_m(\cdot \mid Y_{it}, \alpha_i, k_i^t, m_i^{t-1}, X_i^t), \perp Y_i^{t-1}, k_i^{t-1}, m_i^{t-1}, X_i^{t-1}$$

  Similar for k_{it+1} and X_{it+1}

- **Assumption 2.2 (Serial Dependence)**

$$Y_{it+1} \sim F_Y^{t+1}(\cdot \mid Y_{it}, \alpha_i, k_{it+1}, k_{it}, X_{it+1}, m_{it}), \perp Y_i^{t-1}, k_i^{t-1}, m_i^{t-1}, X_i^t$$

- Example

$$Y_{it} = \rho_t Y_{it-1} + a_{1t}(k_{it}) + a_{2t}(k_{it-1}) + b_t(k_{it})\alpha_i + X_{it}' c_t + v_{it} \tag{2}$$

  where $E(v_{it} | \alpha_i, k_i^t, m_i^{t-1}, Y_i^{t-1}, X_i^t) = 0$, $m_{it-1} = 1$.

Framework: Theoretical base

It can include varieties of models

- Non-linear wage function: $w(\alpha_t, k_t, \varepsilon_t)$
  - Just static setting
- Time effect: boom or bust
  - Let Markovian to be non-homogeneous
- Match-specific heterogeneity and observable potential wage $\tilde{Y}_{t+1}$
  - Jointly Markov: $(Y^*_{t+1}, k^*_{t+1}, \tilde{Y}_{t+1}) \sim F_{Y,k,\tilde{Y}}(\cdot, \cdot, \cdot | \alpha, Y_t, k_t)$
- Outside this framework
  - Non-markov: permanent-transitory earning;
  - Comment: distribution may be useful, but story of power absent
  - Unemployment state not considered

Identification

Identification: what to identify?

Generally speaking, four targets to identify:

- Move from here to there or not: $p_{kk'}(\alpha)$
- The proportion of each type of worker in a firm class: $q_k(\alpha)$
  - Above two: sorting pattern
- Earning if leave: $F^m_{k'\alpha}(y)$
- If not leave: $F_{k\alpha}(y)$
  - Above two: complementarities

Note: Class k to be estimated, type α can be arbitrary labeled

Identification: intuition of conditions

The key to identification this is a rank condition

- Consider workers moving from class k' to k and vice versa, then:

$$Y_{i1} = a(k') + b(k')\alpha_i + \varepsilon_{i1} \quad Y_{i2} = a(k) + b(k)\alpha_i + \varepsilon_{i2}$$

and

$$Y_{i1} = a(k) + b(k)\alpha_i + \varepsilon_{i1} \quad Y_{i2} = a(k') + b(k')\alpha_i + \varepsilon_{i2} \tag{3}$$

then

$$\frac{b(k')}{b(k)} = \frac{E_{kk'}(Y_{i2}) - E_{k'k}(Y_{i1})}{E_{kk'}(Y_{i1}) - E_{k'k}(Y_{i2})} \tag{4}$$

where $E_{kk'} = E(\cdot | k_{i1} = k, k_{i2} = k', m_{i1} = 1)$

Identification: intuition of conditions

- To identify the interaction effect $\frac{b(k')}{b(k)}$(explain), we need:

$$E_{kk'}(\alpha_i) \neq E_{k'k}(\alpha_i) \tag{5}$$

i.e.

$$E_{kk'}(Y_{i1} + Y_{i2}) \neq E_{k'k}(Y_{i1} + Y_{i2}) \tag{6}$$

which can be empirically tested

- 6 in fact is a rank condition

Identification: static

Let type α to be discrete, $F_z^m(\cdot) = F(\cdot|z, m = 1)$

- For a job mover from k to k', we have:

$$Pr[Y_{i1} \leq y_1, Y_{i2} \leq y_2 | k_{i1} = k, k_{i2} = k', m_{i1} = 1]$$
$$= \overbrace{\sum_{\alpha=1}^{L} \underbrace{F_{k\alpha}(y_1) F_{k'\alpha}^m(y_2)}_{independence} p_{kk'}(\alpha)}^{\text{Take expectation}} \tag{7}$$

- $F_{k'\alpha}^m(y_2)$: log-earnings' cdf in period 2, for α worker, k' firm
- $p_{kk'}(\alpha)$: proportion of α workers among those from k to k'
- $F_{k\alpha}(y_1)$: log-earnings' cdf in period 1, for α worker in k firm

Identification: static

- And log-earnings' cdf in period 1 in k firm:

$$Pr[Y_{i1} \leq y_1 | k_{i1} = k] = \sum_{\alpha=1}^{L} F_{k\alpha}(y_1) q_k(\alpha) \tag{8}$$

- $q_k(\alpha)$: proportion of α workers in k firm
- Question: In what conditions can they be well-identified?

Identification: static

Definition (Connecting cycle of length R)

A pair of sequences of classes $(k_1, ..., k_R)$ in period 1, $(\tilde{k}_1, ..., \tilde{k}_R)$ in period 2,$k_{R+1} = k_1$, s.t.

- $p_{k_r, \tilde{k}_r}(\alpha) \neq 0$
- $p_{k_{r+1}, \tilde{k}_r}(\alpha) \neq 0$

, $\forall (r, \alpha) \in \{1, ..., R\} \times \{1, ..., L\}$

- How to understand this?
- Both "stay" and "leave" are possible
- Communicate and accessible
- $\tilde{k}$ is like a re-ordering

Identification: static

Assumption 3:Mixture Model, Static

- **Assumption 3.1**:(Accessibility and communicativeness) $\forall k \neq k' \in \{1, ..., K\}$, $\exists$ connecting cycle $(k_1, ..., k_R)$ and $(\tilde{k}_1, ..., \tilde{k}_R)$,s.t. $\exists r, k_1 = k,\ k_r = k'$, and scalar $a(1), ..., a(L)$ are distinct, where:

$$a(\alpha) = \frac{p_{k_1,\tilde{k}_1}(\alpha)...p_{k_R,\tilde{k}_R}(\alpha)}{p_{k_2,\tilde{k}_1}(\alpha)...p_{k_1,\tilde{k}_R}(\alpha)}$$

Identification: static

- **Assumption 3.2**:(Rank condition) $\exists$ finite sets including M y_1,y_2, s.t. $\forall r \in \{1, ..., R\}$, matrix $A(k_r, \tilde{k}_r)$ and $A(k_{r+1}, \tilde{k}_r)$ have rank L, where $A_{R\times R}(k, k')$ has (y_1, y_2) element:

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

Identification: static

Theorem (Well-identification)

Let $T = 2$ and Assumptions 1,3 hold. Suppose firm classes are observed. Then, up to labeling of types α, $F_{k\alpha}$ and $F^m_{k'\alpha}$ are identified for $\forall(\alpha, k, k')$.
$\forall(k, k'), k, p_{kk'}(\alpha), q_k(\alpha)$ is identified for all α, for the same labeling

- Label of α can be arbitrary
- Identification is up to (α, k, k') with high degree of freedom

Identification: dynamic

- Backward and forward cdf $G^b_{y_3,k',\alpha}(y_4)$ and $G^f_{y_2,k,\alpha}(y_1)$
  - Impact of previous and future job and earning
  - Can be recovered from data; Interpret "forward"
- Proportion $p_{y_2,y_3,k,k'}(\alpha)$, then we have:

$$Pr[Y_{i1} \leq y_1, Y_{i4} \leq y_4 | Y_{i2} = y_2, Y_{i3} = y_3,$$
$$k_{i1} = k_{i2} = k, k_{i3} = k_{i4} = k', m_{i1} = 0, m_{i2} = 1, m_{i3} = 0]$$
$$= \sum_{\alpha=1}^{L} G^b_{y_3,k',\alpha}(y_4) G^f_{y_2,k,\alpha}(y_1) p_{y_2,y_3,k,k'}(\alpha) \quad (9)$$

And finally

$$Pr[Y_{i1} \leq y_1, Y_{i2} \leq y_2 | k_{i1} = k_{i2} = k, m_{i1} = 0]$$
$$= \sum_{\alpha=1}^{L} G^f_{y_2,k\alpha}(y_1) F_{k\alpha}(y_2) q_k(\alpha) \quad (10)$$

Identification: dynamic

- This identify the pattern of
  - Earning distribution for job movers who move at the end of period 2
  - Income distribution of all workers in period 1
- With similar conditions to static case, it well-identifies the distribution

Two-step Estimation

What to estimate

- Estimate the classification of firms
- Using the estimated class, recover the distribution of earnings and workers by parameter estimation
- EM algorithms plays an important role for both steps

Introduction to EM algorithm

Steps (**E**xpectation **M**aximization algorithm)

Input: observed data $x = (x^{(1)}, ..., x^{(m)})$*, joint distribution* $p(x, z|\theta)$*, conditional distribution* $p(z|x, \theta)$*, maximum iteration* J

- *Step 1: randomly initialize parameter* θ *by* θ^0
- *Step 2: for* j *in* $1 : J$*:*
  - **E-step***: calculate*

$$Q_i(z^{(i)}) \equiv P(z^{(i)}|x^{(i)}, \theta)$$

  - **M-step***: get maximal* θ*:*

$$\theta \equiv arg \max_{\theta} L(\theta) = \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) log P(x^{(i)}, z^{(i)}|\theta)$$

Repeat E and M until converge, output θ

Discussion about EM algorithm

- EM is a kind of heuristic algorithm, all estimation below use it
- May be sensitive to initial parameter: initialize many times
- Must converge to a stationary point
- If $L(\theta, \theta^j) \equiv \sum_{i=1}^{m} \sum_{z^{(i)}} P(z^{(i)}|x^{(i)}, \theta^j) log P(z^{(i)}|x^{(i)}, \theta)$ convex, then global maximal

Firm Classification: k-Means clustering

- Assume that firms' heterogeneity is only in class level, we have:

$$Pr[Y_{i1} \leq y_1 | j_{i1} = j] = \sum_{\alpha=1}^{L} F_{k\alpha}(y_1) q_k(\alpha) \tag{11}$$

- Partition J firms into K classes, K exogenous by solving the weighted k-means problem

$$\min_{k(1),\ldots,k(J),H_1,\ldots,H_K} \sum_{j=1}^{J} n_j \int (\hat{F}_j(y) - H_{k(j)}(y))^2 d\mu(y) \tag{12}$$

- $\hat{F}$: empirical distribution of j, In practice: get empirical distribution by griding the support of j by percentiles
- $H_{k(j)}$: targeted distribution we want to find
- y can be log-earning, or other variables

Recover distribution: static

- Already have estimated $\hat{k}_{it}$
  - $\hat{k}(j) \overset{J\to\infty}{\to}$ population one, $\forall$ labeling
- Maximize the log-likelihood below:

$$\sum_{i=1}^{N_m}\sum_{k=1}^{K}\sum_{k'=1}^{K}\mathbf{1}\{\hat{k}_{i1}=k\}\mathbf{1}\{\hat{k}_{i2}=k'\}$$

$$ln(\sum_{\alpha=1}^{L}\underbrace{p_{kk'}(\alpha;\theta_p)}_{\theta_p:\text{ prop param}}\quad\underbrace{f_{k\alpha}(Y_{i1};\theta_f)f^m_{k'\alpha}(Y_{i2};\theta_{f^m})}_{\text{log-norm,}\theta_f,\ \theta^m_f:\ (k,\alpha)\text{ specific mean-var}}\quad) \tag{13}$$

- Interpret: Likelihood of worker in α moves from k to k' and gets income Y_{i1}, Y_{i2} before and after. N_m: number of job-movers.

Recover distribution: static

- After getting $\hat{\theta}_f$, maximize:

$$\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbf{1}\{k_{i1}\hat{=}k\}ln(\sum_{\alpha=1}^{L}\underbrace{q_k(\alpha;\theta_q)}_{\theta_q:\text{ prop parameter}} f_{k\alpha}(Y_{i1};\hat{\theta}_f)) \tag{14}$$

- Interpret: Likelihood of all workers' earning pattern in period 1
- Parameter vector $(\hat{\theta}_f, \hat{\theta}_{f^m}, \hat{\theta}_p, \hat{\theta}_q)$ characterizes the sorting and complementarity pattern

Recover distribution: dynamic

- A specific parametric form for the G^f, G^b defined before:

$$E[Y_{i1}|Y_{i2},k,\alpha] = \mu_{1k\alpha} + \rho_{1|2}Y_{i2}$$
$$E[Y_{i4}|Y_{i3},k',\alpha] = \mu_{4k'\alpha} + \rho_{4|3}Y_{i3}$$

  - μ:(k,α) specific heterogeneity, ρ: lasting effect of earning

$$E[Y_{i2}|\alpha,k,k'] = \mu_{2k\alpha} + \xi_2(k')$$
$$E[Y_{i3}|\alpha,k,k'] = \mu_{3k'\alpha} + \xi_3(k)$$

  - ξ: effect of future and previous job on α worker
- All similar to static, but this conditional expectation change

Recover distribution: dynamic

Similarly, we maximize this log-likelihood:

$$\sum_{i=1}^{N_m} \sum_{k=1}^{K} \sum_{k'=1}^{K} \mathbf{1}\{\hat{k}_{i2} = k\}\mathbf{1}\{\hat{k}_{i3} = k'\} \times ... \times$$

$$ln(\sum_{\alpha=1}^{L} p_{kk'}(\alpha; \theta_p) \underbrace{f^f_{Y_{i2},k\alpha}(Y_{i1}; \hat{\rho}_{1|2}, \theta_{f^f})}_{\text{dist. w. forward effect}} \underbrace{f^m_{kk'\alpha}(Y_{i2}, Y_{i3}; \theta_{f^m})}_{\text{job-mover}}$$

$$\underbrace{f^b_{Y_{i3},k',\alpha}(Y_{i4}; \hat{\rho}_{4|3}, \theta_{f^b})}_{\text{dist. w. backward effect}}) \tag{15}$$

- The likelihood of job-mover's pattern throughout 4 periods

Recover distribution: dynamic

And after getting $(\hat{\rho}, \hat{\theta}_{f^b}, \hat{\theta}_{f^s})$, we have non-mover's likelihood:

$$\sum_{i=1}^{N} \sum_{k=1}^{K} \mathbf{1}\{\hat{k}_{i2} = k\} \times ln(\sum_{\alpha=1}^{L} q_k(\alpha; \theta_q)$$
$$f^f_{Y_{i2k\alpha}}(Y_{i1}; \hat{\rho}_{1|2}, \hat{\theta}_{f^f}) f^s_{k\alpha}(Y_{i2}, Y_{i3}; \theta_{f^s}) f^b_{Y_{i3},k',\alpha}(Y_{i4}; \hat{\rho}_{4|3}, \hat{\theta}_{f^b}) \tag{16}$$

- Parameter vector $(\hat{\theta}_p, \hat{\theta}_{f^f}, \hat{\theta}_{f^m}, \hat{\theta}_{f^b}, \hat{\theta}_q, \hat{\theta}_{f^s}, \hat{\rho})$ characterizes the sorting, complementarity, dynamic pattern
- $\hat{\xi}$ is also about dynamic, can be estimated similarly in $f^m_{kk'\alpha}(Y_{i2}, Y_{i3}; \xi, \theta_{f^m})$

Recover distribution: dynamic

- An effect distribution based on the estimation
$$\underbrace{Var(E(Y_{i3}|k_{i2}))}_{\text{total}} = Var(E[E(Y_{i3}|k_{i3}, k_{i2})|k_{i2}])$$
$$= \underbrace{Var(E[E(Y_{i3}|k_{i3})|k_{i2}])}_{\text{network effect}}$$
$$+ \underbrace{Var(E(Y_{i3}|k_{i2})) - Var(E[E(Y_{i3}|k_{i3})|k_{i2}])}_{\text{state dependence effect}}$$
where conditional expectations are from estimation
  - Network effect: from the link between current and previous job
  - State dependent effect: from the current job

Comments are welcome