A Hidden Markov Model of Wages and Employment Mobility with Worker and Firm Heterogeneity: Evidence from Danish and Italian Register Data

Long Hong* Rasmus Lentz† Jean-Marc Robin‡

February 15, 2025

Abstract

In this paper we formulate a model of joint wage and mobility outcomes with two-sided heterogeneity. The paper extends the finite mixture approaches of Bonhomme, Lamadon, and Manresa (2019) and Lentz, Piyapromdee, and Robin (2023) to identify worker type dynamics to follow a hidden Markov model. Furthermore, the analysis applies a variational expectation maximization (VEM) algorithm to the firm classification that ensures that both worker and firm latent type identifications are informed by the full likelihood of the wage and mobility data. The paper documents significant classification improvements over existing methods.

The hidden Markov model allows the analysis to flexibly identify models where the worker's current type is at least in part affected by the workers' employment history. In particular, this includes models with human capital dynamics, where these dynamics can depend on the current employment status and the type of the employer. The type dynamics also contribute to the understanding of sorting by allowing that the type dynamics of a match need not require worker reallocation, but rather that the worker's type can change.

The model is identified on Danish and Italian register data.

Keywords: Heterogeneity; Wage distributions; Employment and job mobility; Matched employer-employee data; Finite mixtures; EM algorithm; Classification algorithm; Sorting; Decomposition of wage inequality

JEL codes: E24; E32; J63; J64

*W. P. Carey School of Business, Arizona State University. E-mail: long.hong@asu.edu.

†University of Wisconsin-Madison. E-mail: rlentz@wisc.edu. Rasmus Lentz gratefully acknowledges financial support from NSF grant SES-1459897.

‡SciencesPo, Paris. Email `jeanmarc.robin@sciencespo.fr`. Jean-Marc Robin gratefully acknowledge the support from the European Research Council (grant reference ERC-2020-ADG-101018130).

1 Introduction

2 The model

We start by describing the matched employer-employee data used in this study. Then we construct a model for the data.

2.1 Data and setup

The panel starts in year $t = 1$ and ends in year $t = T$, a rather small, fixed value. If a worker quits her job within a year, the employment sequence is recorded. Say that the first 25 weeks are spent with the incumbent employer, the next 12 weeks the worker is unemployed until she is hired by another firm for the remaining 15 weeks. In this case, we say that the year contains three spells, two employed and one unemployed, and two different total earnings are recorded–one for each employment spell.

We thus construct for each worker $i \in \{1, ..., I\}$ in the panel a sequence of observations indexed by the spell index $s \in \{1, ..., S_i\}$, comprising the following information: the unique employer ID, total earnings $EARN$, the spell length $D \in \{1, ..., 52\}$ in weeks, together with starting calendar time $TIME$ (in year and week), potential experience EXP (cumulated number of weeks spent working since leaving school) and job tenure TEN (cumulated number of weeks with the current employer) at the beginning of the spell. We write $W = \ln(EARN/D)$ for log weekly earnings, that we call log wages (or simply wages).

Workers are drawn in the stock in the first survey year $t = 1$. They enter the panel with a non zero tenure and experience in this case. For $t > 1$, there is a flow of workers entering the labor market after school and an initial search period. For these workers we record trajectories from the first job, and tenure and experience are initially equal to zero.

2.2 Initial network

The initial network of employment matches is modeled as a Degree Corrected Stochastic Block Model (DC-SBM) (see Dasgupta, Hopcroft, and McSherry, 2004; Karrer and Newman, 2011; Larremore, Clauset, and Jacobs, 2014).

Firm nodes. We divide J employers indexed by $j \in \{1, ..., I\}$ into L groups $\ell \in \{1, ..., L\}$. Let $Z_j^f = \ell$ if firm j draws type ℓ and define the indicator variables $Z_{j\ell}^f \in \{0, 1\}$, with $\sum_{\ell=1}^{L} Z_{j\ell}^f = 1$ and $Z_{j\ell}^j = 1$ if and only if $Z_j^f = \ell$. Firms draw a type $Z_j^f = \ell$ with probability π_ℓ^f. Lastly, we add a special "firm" node $j = 0$ that denotes unemployment and a corresponding firm type $\ell = 0$.

Worker nodes. We divide I workers indexed by $i \in \{1, ..., I\}$ into K groups $k \in \{1, ..., K\}$. Workers initially draw a type $Z_i^w(1) = k$, independently, with probability π_k^w. We also define indicator variables $Z_{ik}^w(1)$, with $\sum_{k=1}^K Z_{ik}^w(1) = 1$ and $Z_{ik}^w(1) = 1$ if $Z_i^w(1) = k$.

Initial matching. The initial block-structure of the network is governed by the probabilities $m(k, \ell) = m_{k\ell} \geq 0$, with $\sum_{\ell=0}^L m_{k\ell} = 1$. There is also a parameter θ_j, that we treat like a firm fixed effect and that controls the expected degree (size) of firm $j \neq 0$. To simplify expressions, we define $\theta_{j\ell} = \sum_j Z_{j\ell}\theta_j / \sum_j Z_{j\ell}$ that we interpret as the probability that having drawn a firm group ℓ, we draw firm j from all the firms in this group. We also normalize $\theta_{00} = 1$.

Let $Y_i(1) = j$ and $Y_{ij}(1) = 1$ if worker i matches with firm j initially; $Y_{ij}(1) = 0$ otherwise. Moreover, $Y_i(1) = 0$ or $Y_{i0}(1) = 1$ indicates unemployment. We have $\sum_{j=0}^J Y_{ij}(1) = 1$, as employment states are mutually exclusive. We suppose that $Y_i(1) = j$ is drawn with probability $m_{k\ell}\theta_{j\ell}$ given $Z_i^w = k$ and $Z_j^f = \ell$.[1]

2.3 Dynamics

The network of worker-firm matches is exposed to two separate dynamics experienced by workers: employment transitions and worker type changes. We assume that further realizations of worker wages and employment transitions are subject to the following Markovian conditional independence assumption between wages, employment transitions and type changes: for $s = 1, ..., S_i - 1$,

$$W_i(s) \perp\!\!\!\perp (D_i(s), Y_i(s+1)) \perp\!\!\!\perp Z_i^w(s+1) \mid Z^f, Z_i^w(s), Y_i(s),$$

where past observations beyond $Z^f, Z_i^w(s), Y_i(s)$ are irrelevant.

Matching dynamics. Employment transitions occur between employment states (to and from unemployment, or employer change) with probability $M(\ell'|k, \ell) := M_{k\ell\ell'} \geq 0$ per week. Not that $M_{k\ell\ell}$ is the probability of changing employer within the same group $\ell \geq 1$, but $M_{k00} = 0$. Let also $M(\neg|k, \ell) := M_{k\ell\neg} = 1 - \sum_{\ell'=0}^L M_{k\ell\ell'}$ denote the probability of staying one additional week in the same employment state. Specific employers are drawn within firm groups as in the initial spell with probabilities $\theta_{j\ell}$. We denote as $D_i(s)$ the length of the sth spell in weeks. All durations $D_i(s)$ are censored by the number of weeks left until the end of the current year. Suppose that the period starts in week 30. Then $D_i(s) \leq 52 - 29 = 23$ and the minimum duration is one week. For all $1 \leq d < 23$, the joint

[1] In a standard SBM, the typical assumption is that (i, j) are drawn independently from a Bernoulli distribution (or Poisson) with parameter $m_{k\ell}$ given $k_i(1) = k$ and $\ell_j = \ell$. So there is no normalization $\sum_{\ell=0}^L m_{k\ell} = 1$. If there are many firms, the dependence in the first case is negligible.

probability of the recorded duration and destination (d, j', ℓ') given the origin (k, j, ℓ) is $(M_{k\ell\neg})^{d-1} M_{k\ell\ell'}\theta_{j'\ell'}$. The probability of censoring — $D_i(s) = 23$ — is $(M_{k\ell\neg})^{23}$. These probabilities do not depend on the current employer ID j beyond its type ℓ.

The renewal process of worker types. The type of worker i may change with the period index s. Let $Z_i^w(s)$ denote the sequence of worker i's types. The probability that $Z_i^w(s) = k'$ given $Z_i^w(s-1) = k$, $Y_i(s-1) = j$ and $Z_j^f = \ell$ is only function of k, ℓ, k' and is denoted $A(k'|k, \ell) := A_{k\ell k'}$. We allow types to change during unemployment ($j = \ell = 0$). In this case, we expect a loss of skill. Various constrained specifications are possible. It could thus be $A_{k\ell k'}$ only depends on whether $\ell = 0$ (unemployment) or $\ell \neq 0$ (employment). We could also assume that only incremental type changes are possible: $A_{k\ell k'} = 0$ if $|k' - k| > 1$.

Wages. As stated above, a unique wage is observed for each separate period of employment within a year. We keep the main job in case of several simultaneous jobs. We assume the wage in spell s is Gaussian:

$$W_i(s) \mid Z_i^w(s) = k, Z_j^f = \ell \sim \mathcal{N}\left(\mu_{k\ell}, \sigma_{k\ell}^2\right).$$

We could have a different equation for within-job wages and hiring wages, add autoregressive effects, etc. We prefer to focus on the main innovation of this paper: worker type dynamics. Also, errors need not be Gaussian. For example, Brault, Keribin, and Mariadassou (2020) consider the case of the regular univariate exponential family. The normal is bivariate exponential, which also greatly simplifies the maximum likelihood calculations. We can write the density of observation $W_i(s) = w$ conditional on $Z_i^w(s) = k, Z_j^f = \ell$ as $f(w|k, \ell) = \varphi(T(w), \eta_{k\ell})$ with

$$\varphi(T, \eta) = h(w) \exp\left(\eta^\top T - \psi(\eta)\right),$$

where $T(w) = (w, w^2)^\top$ is the sufficient statistic, $\eta = (\eta_1, \eta_2)^\top = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^\top$ is the natural parameter and $\psi(\eta) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\eta_2)$. Using ψ we can derive moments of the sufficient statistics by differentiation:

$$\mathbb{E}T = \frac{\partial\psi(\eta)}{\partial\eta}, \quad \mathrm{Var}(T) = \frac{\partial^2\psi(\eta)}{\partial\eta\partial\eta^\top}.$$

3 Identification of parameters given firm types

In this section we show identification using at least three workers per firm and at least three consecutive periods of observation per worker. We present a simple proof of identification for the discrete time case with no control variables and where the timings of wage

observations and employment transition coincide. We also suppose to simplify that the wage distributions are discrete.

3.1 Identification of firm types

Suppose that for each firm j we observe three wages. Then, assuming independent wage draws given firm type, we can identify the proportion of each firm type, and the distributions of observed firm characteristics and wages given firm types in each year, using standard arguments for the identification of random mixtures. This was the argument in BLM and it still applies in the present context. So let us now assume that we know the type of each firm in every year.

3.2 Identifying matrices

We start by constructing a collection of identifying matrices. Fix one firm type $\ell = 1, ..., L$.

Within spell wage dynamics. The probability $\mathbb{P}(\ell, w, \neg, w')$ of two wages w and w' in periods 1 and 2 in the same firm of type ℓ ($\neg$ sign means no firm change between the two periods) is, with obvious notations,

$$\begin{aligned}\mathbb{P}(\ell, w, \neg, w') &= \sum_{k,k'} m(k,\ell)\, f(w|k,\ell)\, M(\neg|k,\ell)\, A(k'|k,\ell)\, f(w'|k',\ell) \\ &= \sum_{k'} \left(\sum_{k} m(k,\ell)\, f(w|k,\ell)\, M(\neg|k,\ell)\, A(k'|k,\ell) \right) f(w'|k',\ell),\end{aligned}$$

where the sums are over $\{1, ..., K\}$. Let us stack the parameters into the following matrices,

$$\Pi_\ell = \text{diag}[m(k,\ell)]_k, \quad F_\ell = [f(w|k,\ell)]_{w\times k}, \quad A_\ell = [A(k'|k,\ell)]_{k\times k'}, \quad D_{\ell\neg} = \text{diag}[M(\neg|k,\ell)]_k.$$

We have $Q_{\ell\neg} \equiv [\mathbb{P}(\ell, w, \neg, w')]_{w\times w'} = F_\ell \Pi_\ell D_{\ell\neg} A_\ell F_\ell^\top$.

Between spell wage dynamics. The probability of wages w and w' in periods 1 and 2 with a job change from ℓ to ℓ' is

$$\mathbb{P}(\ell, w, \ell', w') = \sum_{k,k'} m(k,\ell)\, f(w|k,\ell)\, M(\ell'|k,\ell)\, A(k'|k,\ell)\, f(w'|k',\ell').$$

Denoting $D_{\ell\ell'} = \text{diag}[M(\ell'|k,\ell)]_k$, we have

$$Q_{\ell\ell'} \equiv [\mathbb{P}(\ell, w, \ell', w')]_{w\times w'} = F_\ell \Pi_\ell D_{\ell\ell'} A_\ell F_{\ell'}^\top.$$

Transition to unemployment. The probability of wage w in period 1 in a firm of type $\ell = 1, ..., L$ followed by a transition to non-employment in period 2 is

$$\mathbb{P}(\ell, w, 0) = \sum_k m(k, \ell)\, f(w|k, \ell)\, M(0|k, \ell).$$

In matrix notations we have

$$Q_{\ell 0} = [\mathbb{P}(\ell, w, 0)]_w = F_\ell \Pi_\ell D_{\ell 0} e_K,$$

for $D_{\ell 0} = \text{diag}[\delta_{k\ell}]_k$ and e_K is the K-vector of ones.

Wage dynamics with an unemployment interruption. Lastly, consider the probability of wages w and w' and employer types ℓ and ℓ' in periods 1 and 3 with an intermediate non-employment spell in period 2:

$$\mathbb{P}(\ell, w, 0, \ell', w') = \sum_{k,k',k''} m(k, \ell)\, f(w|k, \ell)\, M(0|k, \ell)\, A(k'|k, \ell)\, M(\ell'|k', 0)\, A(k''|k', 0)\, f(w'|k'', \ell').$$

In matrix terms, we have

$$Q_{\ell 0 \ell'} = [\mathbb{P}(\ell, w, 0, \ell', w')]_{w \times w'} = F_\ell \Pi_\ell D_{\ell 0} A_\ell D_{0\ell'} A_0 F_{\ell'}^\top,$$

with $A_0 = [A(k'|k, 0)]_{k \times k'}$, $D_{\ell 0} = \text{diag}[M(0|k, \ell)]_k$ and $D_{0\ell'} = \text{diag}[M(\ell'|k, 0)]_k$.

3.3 Assumptions and main identification result

We make the following identifying restrictions.

Assumption 1. *The following conditions hold.*

1. *For all $(k, \ell) \in \{1, ..., K\} \times \{1, ..., L\}$, all matches are initially possible, i.e. $m(k, \ell) \neq 0$, making the matrices $\Pi_\ell = \text{diag}[m(k, \ell)]_k$ non singular for all $\ell \in \{1, ..., L\}$.*
2. *The matrices $F_\ell = [f(w|k, \ell)]_{w \times k}$ have rank K for all $\ell \in \{1, ..., L\}$.*
3. *The matrices A_ℓ are square-invertible for all $\ell = 1, ..., L$.*
4. *For all $(k, \ell) \in \{1, ..., K\} \times \{1, ..., L$, no mobility is always possible, i.e. $M(\neg|k, \ell) \neq 0$, making the matrices $D_{\ell\neg} = \text{diag}[M(\neg|k, \ell)]_k$ non singular for all $\ell \in \{1, ..., L\}$.*
5. *For all (k, k') there exists $\ell \in \{1, ..., L\}$ such that $\frac{M(\ell|k,\ell)}{M(\neg|k,\ell)} \neq \frac{M(\ell|k',\ell)}{M(\neg|k',\ell)}$, which means that the diagonal matrices $D_{\ell\neg}^{-1} D_{\ell\ell} = \text{diag}[M(\ell|k, \ell)/M(\neg|k, \ell)]$ have distinct diagonal entries.*
6. *In unemployment, there is a non zero probability of not changing type: $A(k|k, 0) > 0$.*

Conditions 1-4 are rank restrictions implying that the matrices $Q_{\ell\neg}$, $\ell \in \{1, ..., L\}$, have rank K. Hence, the number of worker types is identified as the rank of an observable matrix. This is because we are making more assumptions than Kasahara and Shimotsu (2014) who only identify a lower bound of the number of groups. In particular, we restrict the conditional wage distributions $(f(w|k,\ell))_k$ to be linearly independent across worker types for all firm types and the wage supports is large enough for the row rank to be greater that K. Condition 5 is not a rank restriction. It ensures that certain matrices have all their eigenvalues simple, making eigenspaces unidimensional. Condition 6 allows to recover the parameters corresponding to unemployment spell once all other parameters have been recovered. Conditions 1-6 are far from being minimal, but they facilitate the proof of identification. And drawing all parameters from a continuous distribution unrestrictedly, they are satisfied with probability one.

The following proposition shows that it suffices to observe worker wages and employer types over three consecutive periods to identify all parameters.

Theorem 2. *The model is identified under Assumption 1 up to a group label permutation from all matrices $Q_{\ell\neg}, Q_{\ell\ell'}$, $Q_{\ell 0}, Q_{\ell 0\ell'}$, for $\ell, \ell' = 1, ..., L$.*

The proof is in Appendix 1. It is constructive. It requires two consecutive employments spells with and without intermediate period of unemployment. The latter case is the only reason why we need three consecutive spells. The main idea of the proof, which is specific to our setup and greatly simplifies the argument, is that we can learn a lot on the parameters by contrasting wage changes within firms and between firms of the same type.

4 Complete information

Let $X = (Y, W, D)$ denote the observation sample. The latent worker and firm types are gathered in $Z = (Z^f, Z^w)$. The log-likelihood of the complete data $\left(X, Z^f, Z^w\right)$ is

$$\begin{aligned}\ln \mathcal{L}\left(X, Z^f, Z^w\right) &= \ln \mathcal{L}\left(Z^f\right) + \ln \mathcal{L}\left(X, Z^w \mid Z^f\right) \\ &= \ln \mathcal{L}\left(Z^f\right) + \sum_{i=1}^{I} \ln \mathcal{L}_i\left(X_i, Z_i^w \mid Z^f\right),\end{aligned}$$

with

$$\ln \mathcal{L}\left(Z^f\right) = \sum_{j=1}^{J}\sum_{\ell=1}^{L} Z_{j\ell}^f \ln \pi_\ell^f.$$

To calculate $\mathcal{L}_i\left(X_i, Z_i^w \mid Z^f\right)$, let us split the whole sequence of individual observations (emissions) as

$$\begin{aligned} X_i(1) &= (Y_i(1), W_i(1), D_i(1), Y_i(2)), \\ X_i(s) &= (W_i(s), D_i(s), Y_i(s+1)), \ s = 2, ..., S_i - 1, \\ X_i(S_i) &= (W_i(S_i), D_i(S_i)). \end{aligned}$$

Given firm types Z^f, the probability that the worker is initially of type k is

$$\alpha_k(1) = \mathbb{P}\left(Z_i^w(1) = k\right) = \pi_k^w.$$

The probability of the first emission $X_i(1)$ given that $Z_i^w(1) = k$ is

$$\begin{aligned} \beta_k(1|Z^f) = \exp\Bigg[\sum_{j=0}^{J} Y_{ij}(1) \sum_{\ell=0}^{L} Z_{j\ell}^f \Bigg(&\ln m_{k\ell} + \ln \theta_{j\ell} \\ &+ \mathbf{1}\{j \neq 0\} \ln\left[\frac{1}{\sigma}\varphi\left(\frac{W_i(1) - \mu_{k\ell}}{\sigma}\right)\right] \\ &+ D_i(1) \ln M_{k\ell\neg} + \sum_{j'=0}^{J} Y_{ij'}(2) \sum_{\ell'=0}^{L} Z_{j'\ell'}^f \mathbf{1}\{j' \neq j\} \left[\ln M_{k\ell\ell'} + \ln \theta_{j'\ell'}\right] \Bigg)\Bigg]. \end{aligned}$$

Then, for all following spells $s = 2, ..., S_i$, the probability that $Z_i^w(s) = k'$ given Z^f and $Z_i^w(s-1) = k$, irrespective of the whole past, is

$$\alpha_{kk'}(s|Z^f) = \exp\left[\sum_{j=1}^{J} Y_{ij}(s-1) \sum_{\ell=0}^{L} Z_{j\ell}^f \ln A_{k\ell k'}\right].$$

The likelihood of emissions $X_i(s)$ is, for $s = 2, ..., S_i - 1$,

$$\begin{aligned} \beta_k(s|Z^f) = \exp\Bigg[\sum_{j=0}^{J} Y_{ij}(s) \sum_{\ell=1}^{L} Z_{j\ell}^f \Bigg(&\mathbf{1}\{j \neq 0\} \ln\left[\frac{1}{\sigma}\varphi\left(\frac{W_i(s) - \mu_{k\ell}}{\sigma}\right)\right] \\ &+ D_i(s) \ln M_{k\ell\neg} + \sum_{j'=0}^{J} Y_{ij'}(s+1) \sum_{\ell'=0}^{L} Z_{j'\ell'}^f \mathbf{1}\{j' \neq j\} \left[\ln M_{k\ell\ell'} + \ln \theta_{j'\ell'}\right] \Bigg)\Bigg], \end{aligned}$$

and for the last spell,

$$\beta_k(S_i|Z^f) = \exp\left[\sum_{j=0}^{J} Y_{ij}(S_i) \sum_{\ell=1}^{L} Z_{j\ell}^f \left(\mathbf{1}\{j \neq 0\} \ln\left[\frac{1}{\sigma}\varphi\left(\frac{W_i(S_i) - \mu_{k\ell}}{\sigma}\right)\right] + D_i(s) \ln M_{k\ell\neg} \right)\right].$$

We finally have,

$$\ln \mathcal{L}_i\left(X_i, Z_i^w \mid Z^f\right) = \sum_{k=1}^{K} Z_{ik}^w(1) \ln\left[\alpha_k(1)\beta_k(1|Z^f)\right] + \sum_{s=2}^{S_i}\sum_{k=1}^{K} Z_{ik}^w(s-1) \sum_{k'=1}^{K} Z_{ik'}^w(s) \ln \alpha_{kk'}(s|Z^f) + \sum_{s=2}^{S_i}\sum_{k=1}^{K} Z_{ik}^w(s) \ln \beta_k(s|Z^f)$$
$$= \ln\left[\alpha_{k_1}(1)\beta_{k_1}(1|Z^f)\alpha_{k_1k_2}(2|Z^f)\beta_{k_2}(2|Z^f)\alpha_{k_2k_3}(3|Z^f)\beta_{k_3}(3|Z^f)...\right].$$

for $Z_i^w = (k_1, k_2, k_3, ...)$.

Let b denote the parameter of a parametric specification of $\mathcal{L}(X, Z)$. It comprises the parameters of the likelihood of the latent variables: group shares π_k^w, π_ℓ^f and the worker type transition probabilities $A_{k\ell k'}$; and the parameters of the emission probabilities: the initial match probabilities $m_{k\ell}$, the match transition probabilities $M_{k\ell\ell'}$, the firm sampling probabilities $\theta_{j\ell}$, the wage parameters $\eta_{k\ell}$. We derive the Maximum Likelihood Estimator of the complete information model, maximizing $\mathcal{L}(X, Z; b)$, in Appendix B.

5 Estimation with the observed data

The complete information MLE is a useful benchmark, but is unfortunately not feasible because the types $Z = (Z^w, Z^f)$ are unobserved. Let us gather observed emissions in the array $X_i = (W_i, Y_i, D_i)$ and let $X = (X_i)$. Averaging $\mathcal{L}(X, Z)$ over Z to calculate the likelihood of the observed data $\mathcal{L}(X)$ is infeasible because of the large dimension of the vector of firm types Z^f. The EM algorithm comes to mind. Unfortunately, EM requires the computation of the conditional probability $\mathcal{L}(Z \mid X)$ which is also intractable.[2] We can calculate $\mathcal{L}(Z^w \mid X, Z^f)$ because worker trajectories are independent conditional on firm types, but not $\mathcal{L}(Z^f \mid X, Z^w)$ because the network of firms connected by common employees is essentially one big connected component.

Bonhomme, Lamadon, and Manresa (2019) solve this difficulty by pre-clustering firms using a k-means algorithm. Then, they estimate the latent worker model using a standard EM algorithm. Lentz, Piyapromdee, and Robin (2023) follow BLM and produce a hard classification of firms and a random clustering of workers. However, they imbed the firm classification inside the M-stage of the EM algorithm for workers, using the Classification EM algorithm of Celeux and Govaert (1992).

The problem with a hard classification of firms is that the estimation of the model can be used for simulation with only the firms used for estimation. The model remains incomplete as it does not say how firm types are drawn. In addition, some firms will be clustered accurately because they have many employees, but there are also many small

[2] To simplify notations, we use the same notations for the likelihood (function of parameters) and the density function or probability mass of observed and latent variables.

firms with little information to group them accurately. Abowd, McKinney, and Schmutte (2019) develop a model of worker-firm matching very similar to the one above, using a Bayesian approach to estimation. Drawing firm types from the conditional distributions with the Gibbs sampler is a formidable computing task that Abowd et al. speed up a bit by making use of the possibility that firms may belong to disconnected sub-networks, which allows for some degree of parallelization. But this parallelization is not massive enough — the firm network is too connected — to make a big difference. In the end, they reduce the estimation sample to a random subsample of about 400,000 person-year observations with 80,000 workers, 60,000 firms and 130,000 matches. Although a very active research exists and has been successful at increasing the efficiency of the moves in MCMC algorithms (see Lee and Wilkinson, 2019), an alternative to Monte Carlo methods is the class of variational expectation-maximization (VEM) methods, that aim at maximizing a lower bound of the observed log-likelihood $\ln \mathcal{L}(X)$ (Govaert and Nadif, 2008, Daudin, Picard, and Robin, 2008) which we now describe.

5.1 Variational EM

For any probability distribution $R(Z)$ (summing to one), define

$$\mathcal{J}(R, X) = \ln \mathcal{L}(X) - D_{KL}\left(R \parallel \mathcal{L}(Z \mid X)\right)$$

where $D_{KL}\left(R \parallel \mathcal{L}(Z \mid X)\right)$ is the Kullback-Leiber divergence

$$D_{KL}\left(R \parallel \mathcal{L}(Z \mid X)\right) = \sum_{Z} R\left(Z\right) \ln\left(\frac{R(Z)}{\mathcal{L}(Z \mid X)}\right) \geq 0.$$

Then,

$$\begin{aligned}
\mathcal{J}(R, X) &= \ln \mathcal{L}(X) - D_{KL}\left(R \parallel \mathcal{L}(Z \mid X)\right) \\
&= \sum_{Z} R(Z) \ln \mathcal{L}(X) - \sum_{Z} R\left(Z\right) \ln\left(\frac{R(Z)}{\mathcal{L}(Z \mid X)}\right) \\
&= \sum_{Z} R(Z) \ln \mathcal{L}(X, Z) - \mathcal{H}(R),
\end{aligned}$$

where $\mathcal{H}(R) = -\sum_{Z} R(Z) \ln R(Z) = -\mathbb{E}_R \ln R(Z)$ is the entropy of distribution R. Hence, a good distribution $R(Z)$ is one that maximizes the expected complete log-likelihood while keeping the entropy $\mathcal{H}(R)$ as large as possible (no small groups).

Variational EM is similar to standard EM. Given $\mathcal{L}(Z \mid X)$ and data X, a pseudo

E-step optimizes on $R(Z)$:

$$\begin{aligned}\widehat{R} &= \arg\max_{R\in\mathcal{R}} \mathcal{J}(R, X) \\ &= \arg\max_{R\in\mathcal{R}} \sum_Z R(Z) \ln \mathcal{L}(Z \mid X) - \sum_Z R(Z) \ln R(Z)\end{aligned}$$

where $\mathcal{R}$ is a class of distributions making this optimization problem tractable.[3] If $\mathcal{L}(Z \mid X)$ is computable, then we do not need to restrict the feasibility set $\mathcal{R}$, and we get $\widehat{R}(Z) = \mathcal{L}(Z \mid X)$ as in standard EM. Otherwise, the optimal solution can only be approximated within a class of manageable distributions.

Notice that, as long as $\mathcal{R}$ contains the Dirac mass δ_Z,

$$\ln \mathcal{L}(X, Z) = \sum_z \delta_Z(z) \ln \mathcal{L}(X, z) = \mathcal{J}(\delta_Z, X) \leq \max_{R\in\mathcal{R}} \mathcal{J}(R, X) \leq \ln \mathcal{L}(X).$$

The parameter b is implicit everywhere. The Variational Estimator is

$$\widehat{b} = \arg\max_b \max_{R\in\mathcal{R}} \mathcal{J}(R, X) = \arg\max_b \sum_Z \widehat{R}(Z) \ln \mathcal{L}(X, Z; b).$$

Consider for example the case of a finite mixture model:

$$\mathcal{L}(X, Z; b) = \prod_{i=1}^{I} \sum_{k=1}^{K} Z_{ik} \pi_k f(X_i; \beta_k),$$

where $Z_{ik} = 1$ if i is in group k. We could restrict $R(Z)$ to be a hard classification of observations: $R(Z) = \prod_{i=1}^{I} \prod_{k=1}^{K} \tau_{ik}^{Z_{ik}}$, where $\tau_i = (\tau_{ik})$ are multi-indexes of length 1 (e.g. $(0, 1)$ and $(1, 0)$). In other words, $R(Z) = 1$ if and only if $Z_{ik} = \tau_{ik}$ for all i, k. Then, using the convention $0 \ln(0) = 0$ and since $1 \ln(1) = 0$,

$$\sum_Z R(Z) \ln \mathcal{L}(X, Z) - \sum_Z R(Z) \ln R(Z) = \sum_{i=1}^{I} \sum_{k=1}^{K} \tau_{ik} \pi_k f(X_i; \beta_k).$$

Thus, given $b = (\pi_k, \beta_k)$, the pseudo E-step selects for each i the k that maximizes $\pi_k f(X_i; \beta_k)$. This is the Classification EM algorithm of Celeux and Govaert (1992) that was used in Lentz, Piyapromdee, and Robin (2023).

[3] For the last equality, write $\ln \mathcal{L}(X, Z) = \ln \mathcal{L}(Z \mid X) + \ln \mathcal{L}(X)$ and note that $\sum_Z R(Z) \ln \mathcal{L}(X) = \ln \mathcal{L}(X)$ is independent of R.

5.2 VEM for our model

We first approximate $\mathcal{L}(Z \mid X)$ as $R(Z) = R^w(Z^w)R^f(Z^f)$ with forced independence between worker and firm types. The integrated or expected log-likelihood is

$$\mathbb{E}_R \ln \mathcal{L}(X, Z^f, Z^w) = \sum_{Z^f} R^f(Z^f) \sum_{Z^w} R^w(Z^w) \ln \mathcal{L}(X, Z^f, Z^w).$$

The M-step maximizes $\mathbb{E}_R \ln \mathcal{L}(X, Z^f, Z^w; b)$ with respect to b given current R^f, R^w. The E-step is decomposed into a sequence of two steps. First, we maximize $\mathbb{E}_R \ln \mathcal{L}(X, Z^f, Z^w; b)$ with respect to R^w given the updated b and current R^f, with no restriction on R^w. Then we maximize the expected log-likelihood with respect to R^f given updated b and R^w subject to the following restriction:

$$R^f(Z^f) = \Pi_{j=1}^J R_j^f(Z_j^f) = \Pi_{j=1}^J \tau_{j\ell}^{Z_{j\ell}^f}.$$

In Appendix C, we show that the optimal value of R^w is of the form $R^w(Z^w) = \Pi_{i=1}^I R_i^w(Z_i^w)$, with independence between worker types. We also calculate the marginal distributions:

$$\mathbb{P}_{R^w}\left\{Z_i^w(s) = k\right\} = \zeta_{ik}(s), \quad \mathbb{P}_{R^w}\left\{Z_i^w(s) = k, Z_i^w(s+1) = k'\right\} = \xi_{ikk'}(s+1),$$

which are sufficient to calculate the expected log-likelihood.

6 Consistency of the ML and Variational estimators

Recent work derives conditions proving the identification and the consistency of VEM for stochastic block models depending on the network density (see Snijders and Nowicki (1997), Bickel and Chen (2009), Bickel, Choi, Chang, and Zhang (2013), Celisse, Daudin, and Pierre (2012), Mariadassou and Matias (2015), Longepierre and Matias (2019), Brault, Keribin, and Mariadassou (2020) and Zhao, Hao, and Zhu (2024)). The last three, most recent papers show consistency of Maximum Likelihood and Variational Estimators in different contexts. Longepierre and Matias (2019) consider the case of a Dynamic Stochastic Block Model (SBM) where the latent types of the network's nodes change like the worker types in our paper. Zhao, Hao, and Zhu (2024) study the case of a static Bipartite SBM. Brault, Keribin, and Mariadassou (2020) deal with a Latent Block Model, or a finite two-way mixture of continuous distributions. These three papers cover sufficiently many aspects of our framework that we can easily understand under which conditions our Variational EM estimator is also going to be consistent. We follow the method of proof of Brault et al. which is very clear and pedagogical.

The first result that is easy to prove is the consistency and the asymptotic normality of the complete information MLE when I and J tend to infinity, as well as the local

asymptotic normality of the MLE. The estimator exists in a closed form. The proof is straightforward.

The next step compares the observed likelihood ratio $\mathcal{L}(X;b)/\mathcal{L}(X;b^*)$, where b^* denotes the true value of the parameter, to the complete likelihood ratio $\mathcal{L}(X,Z;b)/\mathcal{L}(X,Z;b^*)$.

References

Abowd, J. M., K. L. McKinney, and I. M. Schmutte (2019): "Modeling Endogenous Mobility in Earnings Determination," *Journal of Business & Economic Statistics*, 37(3), 405–418.

Bickel, P., D. Choi, X. Chang, and H. Zhang (2013): "Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels," *The Annals of Statistics*, 41(4), 1922 – 1943.

Bickel, P. J., and A. Chen (2009): "A nonparametric view of network models and Newman–Girvan and other modularities," *Proceedings of the National Academy of Sciences*, 106, 21068 – 21073.

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

Brault, V., C. Keribin, and M. Mariadassou (2020): "Consistency and asymptotic normality of Latent Block Model estimators," *Electronic Journal of Statistics*, 14(1), 1234 – 1268.

Celeux, G., and G. Govaert (1992): "A classification {EM} algorithm for clustering and two stochastic versions," *Computational Statistics & Data Analysis*, 14(3), 315 – 332.

Celisse, A., J.-J. Daudin, and L. Pierre (2012): "Consistency of maximum-likelihood and variational estimators in the stochastic block model," *Electronic Journal of Statistics*, 6(none), 1847 – 1899.

Dasgupta, A., J. Hopcroft, and F. McSherry (2004): "Spectral analysis of random graphs with skewed degree distributions," in *45th Annual IEEE Symposium on Foundations of Computer Science*, pp. 602–610.

Daudin, J. J., F. Picard, and S. Robin (2008): "A mixture model for random graphs," *Statistics and Computing*, 18(2), 173–183.

Govaert, G., and M. Nadif (2008): "Block clustering with Bernoulli mixture models: Comparison of different approaches," *Computational Statistics & Data Analysis*, 52(6), 3233–3245.

Karrer, B., and M. E. J. Newman (2011): "Stochastic blockmodels and community structure in networks," *Physical Review E*, 83(1).

Kasahara, H., and K. Shimotsu (2014): "Nonparametric identification and estimation of the number of components in multivariate mixtures," *Journal of the Royal Statistical Society, Series B*, 76, 97–111.

Larremore, D. B., A. Clauset, and A. Z. Jacobs (2014): "Efficiently inferring community structure in bipartite networks," *Phys. Rev. E*, 90, 012805.

Lee, C., and D. J. Wilkinson (2019): "A review of stochastic block models and extensions for graph clustering," *Applied Network Science*, 4(1), 122.

Lentz, R., S. Piyapromdee, and J.-M. Robin (2023): "The Anatomy of Sorting - Evidence From Danish Data," *Econometrica*, 91(6), 2409–2455.

Longepierre, L., and C. Matias (2019): "Consistency of the maximum likelihood and variational estimators in a dynamic stochastic block model," *Electronic Journal of Statistics*, 13(2), 4157 – 4223.

Mariadassou, M., and C. Matias (2015): "Convergence of the groups posterior distribution in latent or stochastic block models," *Bernoulli*, 21(1), 537–573.

Rabiner, L. (1989): "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, 77(2), 257–286.

Snijders, T. A. B., and K. Nowicki (1997): "Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure," *Journal of Classification*, 14(1), 75–100.

Zhao, Y., N. Hao, and J. Zhu (2024): "Variational Estimators of the Degree-corrected Latent Block Model for Bipartite Networks," *Journal of Machine Learning Research*, 25(150), 1–42.

APPENDIX

A Proof of the identification Theorem

We prove identification through the following steps.

1. Consider $Q_{\ell\neg} \equiv [\mathbb{P}(\ell, w, \neg, w')]_{w\times w'} = F_\ell \Pi_\ell D_{\ell\neg} A_\ell F_\ell^\top = F_\ell \Pi_\ell D_{\ell\neg} G_\ell^\top$ with $G_\ell = F_\ell A_\ell^\top$. Given the SVD $Q_{\ell\neg} = U_\ell \Lambda_\ell V_\ell^\top$,[4] write

$$U_\ell^\top Q_{\ell\neg} V_\ell \Lambda_\ell^{-1} = U_\ell^\top F_\ell \Pi_\ell D_{\ell\neg} G_\ell^\top V_\ell \Lambda_\ell^{-1} = I_K.$$

   Let $W_\ell = U_\ell^\top F_\ell \Pi_\ell D_{\ell\neg}$. Then $W_\ell^{-1} = G_\ell^\top V_\ell \Lambda_\ell^{-1}$.

2. Consider $Q_{\ell\ell} \equiv [\mathbb{P}(\ell, w, \ell, w')]_{w\times w'} = F_\ell \Pi_\ell D_{\ell\ell} A_\ell F_\ell^\top = F_\ell \Pi_\ell D_{\ell\ell} G_\ell^\top$. It holds that

$$U_\ell^\top Q_{\ell\ell} V_\ell \Lambda_\ell^{-1} = U_\ell^\top F_\ell \Pi_\ell D_{\ell\ell} G_\ell^\top V_\ell \Lambda_\ell^{-1} = W_\ell D_{\ell\neg}^{-1} D_{\ell\ell} W_\ell^{-1}.$$

   The diagonal matrix $D_{\ell\neg}^{-1} D_{\ell\ell}$ is identified as the eigenvectors (up to labeling). And knowing that the entries of $D_{\ell\neg}^{-1} D_{\ell\ell} = \text{diag}[M(\ell|k,\ell)/M(\neg|k,\ell)]$ are all distinct (Condition 5), any matrix of eigenvectors is thus of the form

$$W_\ell = U_\ell^\top F_\ell \Pi_\ell D_{\ell\neg} \Delta_\ell^{-1}, \quad W_\ell^{-1} = \Delta_\ell G_\ell^\top V_\ell \Lambda_\ell^{-1},$$

   for some non singular diagonal matrix Δ_ℓ.

3. Consider W_ℓ^{-1}. The matrix $W_\ell^{-1} \Lambda_\ell V_\ell^\top$ identifies . The rows of F_ℓ and the columns of A_ℓ sum to one. It follows that the rows of G_ℓ also sum to one, which then identifies he diagonal matrix Δ_ℓ from $\Delta_\ell G_\ell^\top$, and G_ℓ is also identified. The matrix W_ℓ then identifies $F_\ell \Pi_\ell D_{\ell\neg}$. And since the rows of F_ℓ sums to one, the diagonal matrix $\Pi_\ell D_{\ell\neg}$ is identified and F_ℓ is identified. Since $G_\ell = F_\ell A_\ell^\top$ with G_ℓ and F_ℓ both identified, we finally also identify A_ℓ. It is quite remarkable that $Q_{\ell\neg}$ and $Q_{\ell\ell}$ (mobility within and between firms of type ℓ) suffice to identify F_ℓ, A_ℓ and the products of diagonal matrices $D_{\ell\neg}^{-1} D_{\ell\ell}$ and $\Pi_\ell D_{\ell\neg}$.

4. Having proceeded independently for each firm type ℓ, how do we know that the worker group k that we have thus labelled for firm type ℓ corresponds to the worker group k' that we have thus labelled for firm type ℓ'? Take matrix $Q_{\ell\ell'} = [\mathbb{P}(\ell, w, \ell', w')]_{w\times w'}$ corresponding to a job mobility from ℓ to ℓ'. We know that

$$Q_{\ell\ell'} = F_\ell \Pi_\ell D_{\ell\ell'} A_\ell F_{\ell'}^\top = (F_\ell \Pi_\ell D_{\ell\neg})(D_{\ell\neg}^{-1} D_{\ell\ell'}) A_\ell F_{\ell'}^\top$$

[4] For a matrix $Q \in \mathbb{R}^{m\times n}$ of rank $K \leq m, n$, the SVD has $Q = U\Lambda V^\top$ where U is $m \times K$ with orthonormal columns, Λ is diagonal $K \times K$ and V is $n \times K$ with orthonormal columns.

with $D_{\ell\neg}^{-1}D_{\ell\ell'}$ diagonal. So given arbitrarily chosen worker-group labels for ℓ and ℓ' one should relabel worker groups for ℓ' (say) by reordering the columns of $F_{\ell'}$ so that $(F_\ell \Pi_\ell D_{\ell\neg})^+ Q_{\ell\ell'}(F_{\ell'}^+)^\top A_\ell = D_{\ell\neg}^{-1}D_{\ell\ell'}$ is a diagonal matrix (denoting $B^+ = (B^\top B)^{-1}B^\top$ for any full column rank matrix B). This procedure also identifies $D_{\ell\neg}^{-1}D_{\ell\ell'}$.

5. Note that $D_{\ell\neg} + \sum_{\ell'} D_{\ell\ell'} = I_K$. Hence

$$D_{\ell\neg} = [I_K + \sum_{\ell'} D_{\ell\neg}^{-1}D_{\ell\ell'}]^{-1}$$

is identified, and so are all $D_{\ell\ell'}$. Hence, we also identify $D_{\ell\neg}$ and Π_ℓ from $D_{\ell\neg}^{-1}D_{\ell\ell}$ and $\Pi_\ell D_{\ell\neg}$.

6. Consider $Q_{\ell 0} = [\mathbb{P}(\ell, w, 0)]_w = F_\ell \Pi_\ell D_{\ell 0} e_K$, for $D_{\ell 0} = \text{diag}[M(0|k,\ell)]_k$ and e_K is the K-vector of ones. Hence $D_{\ell 0}e_K = [M(0|k,\ell)]_k$ is identified by regressing $[\mathbb{P}(\ell, w, 0)]_w$ on the columns of $F_\ell \Pi_\ell$.

7. Consider $Q_{\ell 0\ell'} = [\mathbb{P}(\ell, w, 0, \ell', w')]_{w\times w'} = F_\ell \Pi_\ell D_{\ell 0} A_\ell D_{0\ell'} A_0 F_{\ell'}^\top$, with $A_0 = [A(k'|k,0)]_{k\times k'}$ and $D_{0\ell'} = \text{diag}[M(\ell'|k,0)]_k$. This identifies $D_{0\ell'}A_0 = [A(k'|k,0)M(\ell'|k,0)]$ for all $\ell' = 1, ..., L$. Consider diagonal terms $A(k|k,0)M(\ell'|k,0)$. One can eliminate $A(k|k,0)$(non zero by Condition 6) from these equations and identify $M(\ell'|k,0)/M(1|k,0)$ for all $\ell' = 2, ..., L$. Hence, all $M(\ell'|k,0)$, $\ell' = 1, ..., L$, are identified because $\sum_{\ell'=1}^L M(\ell'|k,0) = 1$. Then the whole matrix A_0 follows.

B MLE for the complete information model

We start by rewriting the log-likelihood as follows:

$$\ln \mathcal{L}\left(Z^f\right) = \sum_{\ell=1}^{L} n_\ell^f \ln \pi_\ell^f,$$
$$\ln \mathcal{L}\left(Z^w \mid Z^f\right) = \sum_{k=1}^{K} n_k^w(1) \ln \pi_k^w + \sum_{\ell=0}^{L}\sum_{k=1}^{K}\sum_{k'=1}^{K} n_{k\ell k'}^w \ln A_{k\ell k'},$$

where

$$n_\ell^f = \sum_{j=1}^{J} Z_{j\ell}^f$$

counts the number of firms of type ℓ, and where $n_k^w(1)$ counts the number of workers initially of type k and $n_{k\ell k'}^w$ the number of workers of type k in a state ℓ in one period,

and switching to type k' in the next period:

$$
\begin{aligned}
n^w_k(1) &= \sum_{i=1}^{I} Z^w_{ik}(1), \\
n^w_{k0k'} &= \sum_{i=1}^{I} \sum_{s=1}^{S_i-1} Z^w_{ik}(s) Y_{i0}(s) Z^w_{ik'}(s+1), \\
n^w_{k\ell k'} &= \sum_{i=1}^{I} \sum_{s=1}^{S_i-1} Z^w_{ik}(s) \left(\sum_{j=1}^{J} Y_{ij}(s) Z^f_{j\ell} \right) Z^w_{ik'}(s+1), \ \ell \geq 1.
\end{aligned}
$$

The MLE of type probabilities are obtained by maximizing $\ln \mathcal{L}\left(Z^f\right)$ and $\ln \mathcal{L}\left(Z^w \mid Z^f\right)$ subject to the constraint that probabilities must add up to one, that is

$$
\widehat{\pi}^f_\ell = \frac{n^f_\ell}{J}, \quad \widehat{\pi}^w_k = \frac{n^w_k(1)}{I}, \quad \widehat{A}_{k\ell k'} = \frac{n^w_{k\ell k'}}{\sum_{k'=1}^{K} n^w_{k\ell k'}},
$$

where $\sum_{k'=1}^{K} n^w_{k\ell k'}$ is the number of matches (k, ℓ) in periods $1 : S_i - 1$.

We next turn to the log-likelihood of emissions given worker and firm types:

$$
\begin{aligned}
\ln \mathcal{L}\left(X \mid Z^f, Z^w\right) = &\sum_{j=1}^{L} \sum_{\ell=1}^{L} n^w_{j\ell} \ln \theta_{j\ell} + \sum_{k=1}^{K} \sum_{\ell=0}^{L} n^w_{k\ell}(1) \ln m_{k\ell} \\
&+ \sum_{k=1}^{K} \sum_{\ell'=1}^{L} n^w_{k0\ell'} \ln M_{k0\ell'} + \sum_{k=1}^{K} n^w_{k0\neg} \ln M_{k0\neg} \\
&+ \sum_{k=1}^{K} \sum_{\ell=1}^{L} \sum_{\ell'=0}^{L} n^w_{k\ell\ell'} \ln M_{k\ell\ell'} + \sum_{k=1}^{K} \sum_{\ell=1}^{L} n^w_{k\ell\neg} \ln M_{k\ell\neg} \\
&+ \sum_{k=1}^{K} \sum_{\ell=1}^{L} \sum_{i=1}^{I} \sum_{s=1}^{S_i} Z^w_{ik}(s) \left(\sum_{j=1}^{J} Y_{ij}(s) Z^f_{j\ell} \right) \ln \varphi\left(W_i(s), \eta_{k\ell}\right),
\end{aligned}
$$

where

$$
n^w_{j\ell} = \sum_{i=1}^{I} Y_{ij}(1) Z^f_{j\ell} + \sum_{i=1}^{I} \sum_{s=2}^{S_i} \sum_{j_-=0}^{J} \mathbf{1}\{j_- \neq j\} Y_{ij_-}(s-1) Y_{ij}(s) Z^f_{j\ell}, \ j, \ell \geq 1,
$$

is the number of workers initially employed by firm j/type ℓ or who later are observed to

move to such a firm and type. The other worker counts are transparent:

$$
\begin{aligned}
n^w_{k0}(1) &= \sum_{i=1}^{I} Z^w_{ik}(1) Y_{i0}(1), \\
n^w_{k\ell}(1) &= \sum_{i=1}^{I} \sum_{s=1}^{S_i-1} Z^w_{ik}(1) \sum_{j=1}^{J} Y_{ij}(1) Z^f_{j\ell}, \ \ell \geq 1, \\
n^w_{k0\ell'} &= \sum_{i=1}^{I} \sum_{s=1}^{S_i-1} Z^w_{ik}(s) Y_{i0}(s) \sum_{j=1}^{J} Y_{ij}(s+1) Z^f_{j\ell'}, \ \ell' \geq 1, \\
n^w_{k0\neg} &= \sum_{i=1}^{I} \sum_{s=1}^{S_i-1} Z^w_{ik}(s) Y_{i0}(s) D_i(s), \\
n^w_{k\ell 0} &= \sum_{i=1}^{I} \sum_{s=1}^{S_i-1} Z^w_{ik}(s) \sum_{j=1}^{J} Y_{ij}(s) Z^f_{j\ell} Y_{i0}(s+1), \ \ell \geq 1, \\
n^w_{k\ell\ell'} &= \sum_{i=1}^{I} \sum_{s=1}^{S_i-1} Z^w_{ik}(s) \sum_{j=1}^{J} Y_{ij}(s) Z^f_{j\ell} \sum_{j'=1}^{J} Y_{ij'}(s+1) Z^f_{j'\ell'}, \ \ell, \ell' \geq 1, \\
n^w_{k\ell\neg} &= \sum_{i=1}^{I} \sum_{s=1}^{S_i-1} Z^w_{ik}(s) \sum_{j=1}^{J} Y_{ij}(s) Z^f_{j\ell} D_i(s), \ \ell \geq 1,
\end{aligned}
$$

We may also set $n^w_{k00} = 0$ (no transition from unemployment to unemployment as they are counted in $n^w_{k0\neg}$).

Maximizing $\ln \mathcal{L}\left(X \mid Z^f, Z^w\right)$ subject to appropriate constraints yields the following simple moment estimators:

$$
\begin{aligned}
\widehat{\theta}_{j\ell} &= \frac{n^w_{j\ell}}{\sum_{j=1}^{J} n^w_{j\ell}}, \\
\widehat{m}_{k\ell} &= \frac{n^w_{k\ell}(1)}{\sum_{\ell=0}^{L} n^w_{k\ell}(1)}, \ \ell \geq 0, \\
\widehat{M}_{k0\neg} &= \frac{n^w_{k0\neg}}{n^w_{k0\neg} + \sum_{\ell'=1}^{L} n^w_{k0\ell'}}, \quad \widehat{M}_{k0\ell'} = \frac{n^w_{k0\ell'}}{n^w_{k0\neg} + \sum_{\ell'=1}^{L} n^w_{k0\ell'}}, \ \ell' \geq 1, \\
\widehat{M}_{k\ell\neg} &= \frac{n^w_{k\ell\neg}}{n^w_{k\ell\neg} + \sum_{\ell'=0}^{L} n^w_{k\ell\ell'}}, \quad \widehat{M}_{k\ell\ell'} = \frac{n^w_{k\ell\ell'}}{n^w_{k\ell\neg} + \sum_{\ell'=0}^{L} n^w_{k\ell\ell'}}, \ \ell \geq 1, \ell' \geq 0,
\end{aligned}
$$

and

$$
\begin{aligned}
\psi_1'(\widehat{\eta}_{k\ell}) = \widehat{\mu}_{k\ell} = \widehat{W}_{k\ell} &= \frac{1}{n^w_{k\ell}} \sum_{i=1}^{I} \sum_{s=1}^{S_i} Z^w_{ik}(s) \sum_{j=1}^{J} Y_{ij}(s) Z^f_{j\ell} W_i(s), \\
\psi_2'(\widehat{\eta}_{k\ell}) = \widehat{\mu}^2_{k\ell} + \widehat{\sigma^2}_{k\ell} = \widehat{W^2}_{k\ell} &= \frac{1}{n^w_{k\ell}} \sum_{i=1}^{I} \sum_{s=1}^{S_i} Z^w_{ik}(s) \sum_{j=1}^{J} Y_{ij}(s) Z^f_{j\ell} W_i^2(s).
\end{aligned}
$$

Finally, we deduce the following value of the log-likelihood of emissions given types:

$$\begin{aligned}\ln \mathcal{L}\left(X \mid Z^f, Z^w\right) = &\sum_{\ell=1}^{L} n_\ell^f \sum_{j=1}^{J} \widehat{\theta}_{j\ell} \ln \theta_{j\ell} + \sum_{k=1}^{K} n_k^w(1) \sum_{\ell=0}^{L} \widehat{m}_{k\ell} \ln m_{k\ell}^w \\ &+ \sum_{k=1}^{K} \sum_{\ell=0}^{L} n_{k\ell}^w \left(\sum_{\ell'=0}^{L} \widehat{M}_{k\ell\ell'} \ln M_{k\ell\ell'} + \widehat{M}_{k\ell\neg} \ln M_{k\ell\neg} \right) \\ &+ \sum_{k=1}^{K} \sum_{\ell=1}^{L} n_{k\ell}^w \ln \varphi\left(\widehat{W}_{k\ell}, \widehat{W^2}_{k\ell}, \eta_{k\ell}\right),\end{aligned}$$

where

$$n_\ell^f = \sum_{j=1}^{J} n_{j\ell}^w, \quad n_k^w(1) = \sum_{\ell=0}^{L} n_{k\ell}^w(1), \quad n_{k\ell}^w = n_{k\ell\neg}^w + \sum_{\ell'=0}^{L} n_{k\ell\ell'}^w.$$

C Details on the VEM estimation algorithm

C.1 Expected log-likelihood

The expected log-likelihood is

$$\mathbb{E}_R \ln \mathcal{L}(X, Z^f, Z^w) = \sum_{Z^f} R^f(Z^f) \sum_{Z^w} R^w(Z^w) \ln \mathcal{L}(X, Z^f, Z^w)$$

where

$$\begin{aligned}\ln \mathcal{L}(X, Z^f, Z^w) &= \ln \mathcal{L}(Z^f) + \ln \mathcal{L}(X, Z^w \mid Z^f) \\ &= \ln \mathcal{L}(Z^f) + \sum_{i=1}^{I} \ln \mathcal{L}_i(X, Z^w \mid Z^f)\end{aligned}$$

with

$$\mathbb{E}_R \ln \mathcal{L}(Z^f) = \sum_{j=1}^{J} \sum_{\ell=1}^{L} \tau_{j\ell} \ln \pi_\ell^f$$

and where the individual likelihoods $\ln \mathcal{L}_i(X, Z^w \mid Z^f)$ can be integrated iteratively exactly like the complete log-likelihood as we now explain.

The initial worker type probability is still

$$\alpha_k(1) = \pi_k^w,$$

but the worker type transition probabilities for $s = 2, ..., S_i$ replace the certain Z^f by the

probability distribution $\tau = (\tau_{j\ell})$:

$$\alpha_{kk'}(s|\tau) = \exp\left[\sum_{j=1}^{J} Y_{ij}(s-1) \sum_{\ell=0}^{L} \tau_{j\ell} \ln A_{k\ell k'}\right].$$

The emission probabilities also change in the same way:

- for $X_i(s)$:

$$\beta_k(1|\tau) = \exp\Bigg[\sum_{j=0}^{J} Y_{ij}(1) \sum_{\ell=0}^{L} \tau_{j\ell} \Bigg(\ln m_{k\ell} + \ln \theta_{j\ell} + \mathbf{1}\{j \neq 0\} \ln\left[\frac{1}{\sigma}\varphi\left(\frac{W_i(1) - \mu_{k\ell}}{\sigma}\right)\right] + D_i(1) \ln M_{k\ell\neg} + \sum_{j'=0}^{J} Y_{ij'}(2) \sum_{\ell'=0}^{L} \tau_{j'\ell'} \mathbf{1}\{j' \neq j\} [\ln M_{k\ell\ell'} + \ln \theta_{j'\ell'}]\Bigg)\Bigg],$$

- for $X_i(s),\ s = 2, ..., S_i - 1$:

$$\beta_k(s|\tau) = \exp\Bigg[\sum_{j=0}^{J} Y_{ij}(s) \sum_{\ell=1}^{L} \tau_{j\ell} \Bigg(\mathbf{1}\{j \neq 0\} \ln\left[\frac{1}{\sigma}\varphi\left(\frac{W_i(s) - \mu_{k\ell}}{\sigma}\right)\right] + D_i(s) \ln M_{k\ell}^{\neg} + \sum_{j'=0}^{J} Y_{ij'}(s+1) \sum_{\ell'=0}^{L} \tau_{j'\ell'} \mathbf{1}\{j' \neq j\} [\ln M_{k\ell\ell'} + \ln \theta_{j'\ell'}]\Bigg)\Bigg],$$

- for $X_i(S_i)$:

$$\beta_k(S_i|\tau) = \exp\left[\sum_{j=0}^{J} Y_{ij}(S_i) \sum_{\ell=1}^{L} \tau_{j\ell} \left(\mathbf{1}\{j \neq 0\} \ln\left[\frac{1}{\sigma}\varphi\left(\frac{W_i(S_i) - \mu_{k\ell}}{\sigma}\right)\right] + D_i(s) \ln M_{k\ell}^{\neg}\right)\right].$$

For any distribution $R^w(Z^w)$ with

$$\mathbb{P}_{R^w}\{Z_i^w(s) = k\} = \zeta_{ik}(s), \quad \mathbb{P}_{R^w}\{Z_i^w(s) = k, Z_i^w(s+1) = k'\} = \xi_{ikk'}(s+1),$$

we finally obtain

$$\mathbb{E}_R \ln \mathcal{L}_i\left(X_i, Z_i^w \mid Z^f\right) = \sum_{k=1}^{K} \zeta_{ik}(1) \ln[\alpha_k(1)\beta_k(1|\tau)] + \sum_{s=2}^{S_i} \sum_{k=1}^{K} \sum_{k'=1}^{K} \xi_{ikk'}(s) \ln \alpha_{kk'}(s|\tau) + \sum_{s=2}^{S_i} \sum_{k=1}^{K} \zeta_{ik}(s) \ln \beta_k(s|\tau).$$

C.2 M-step

The expected log-likelihood has exactly the same structure as the complete log-likelihood. It suffices to replace $Z^f_{j\ell}$ by $\tau_{j\ell}$, $Z^w_{ik}(s)$ by $\zeta_{ik}(s)$ and $Z^w_{ik}Z^w_{ik'}$ by $\xi_{ikk'}$. The VEM formulas for the emission parameters are therefore

$$\widetilde{\pi}^f_\ell = \frac{\widetilde{n}^f_\ell}{J}, \quad \widetilde{\pi}^w_k = \frac{\widetilde{n}^w_k(1)}{I}, \quad \widetilde{A}_{k\ell k'} = \frac{\widetilde{n}^w_{k\ell k'}}{\sum_{k'=1}^K \widetilde{n}^w_{k\ell k'}},$$

and

$$\begin{aligned}
\widetilde{\theta}_{j\ell} &= \frac{\widetilde{n}^w_{j\ell}}{\sum_{j=1}^J \widetilde{n}^w_{j\ell}}, \\
\widetilde{m}_{k\ell} &= \frac{\widetilde{n}^w_{k\ell}(1)}{\sum_{\ell=0}^L \widetilde{n}^w_{k\ell}(1)}, \ \ell \geq 0, \\
\widetilde{M}_{k0\neg} &= \frac{\widetilde{n}^w_{k0\neg}}{\widetilde{n}^w_{k0\neg} + \sum_{\ell'=1}^L \widetilde{n}^w_{k0\ell'}}, \quad \widetilde{M}_{k0\ell'} = \frac{\widetilde{n}^w_{k0\ell'}}{\widetilde{n}^w_{k0\neg} + \sum_{\ell'=1}^L \widetilde{n}^w_{k0\ell'}}, \ \ell' \geq 1, \\
\widehat{M}_{k\ell\neg} &= \frac{n^w_{k\ell\neg}}{\widetilde{n}^w_{k\ell\neg} + \sum_{\ell'=0}^L \widetilde{n}^w_{k\ell\ell'}}, \quad \widehat{M}_{k\ell\ell'} = \frac{n^w_{k\ell\ell'}}{\widetilde{n}^w_{k\ell\neg} + \sum_{\ell'=0}^L \widetilde{n}^w_{k\ell\ell'}}, \ \ell \geq 1, \ell' \geq 0,
\end{aligned}$$

and

$$\begin{aligned}
\psi_1'(\widetilde{\eta}_{k\ell}) = \widetilde{\mu}_{k\ell} = \widetilde{W}_{k\ell} &= \frac{1}{\widetilde{n}^w_{k\ell}} \sum_{i=1}^I \sum_{s=1}^{S_i} \zeta_{ik}(s) \sum_{j=1}^J Y_{ij}(s) \tau_{j\ell} W_i(s), \\
\psi_2'(\widetilde{\eta}_{k\ell}) = \widetilde{\mu}^2_{k\ell} + \widetilde{\sigma^2}_{k\ell} = \widetilde{W^2}_{k\ell} &= \frac{1}{\widetilde{n}^w_{k\ell}} \sum_{i=1}^I \sum_{s=1}^{S_i} \zeta_{ik}(s) \sum_{j=1}^J Y_{ij}(s) \tau_{j\ell} W_i^2(s).
\end{aligned}$$

with

$$
\begin{aligned}
\widetilde{n}^f_{\ell} &= \sum_{j=1}^{J} \tau_{j\ell} \\
\widetilde{n}^w_{k}(1) &= \sum_{i=1}^{I} \zeta_{ik}(1), \\
\widetilde{n}^w_{k0k'} &= \sum_{i=1}^{I} \sum_{s=1}^{S_i-1} \sum_{i=1}^{I} Y_{i0}(s)\xi_{ikk'}(s+1), \\
\widetilde{n}^w_{k\ell k'} &= \sum_{i=1}^{I} \sum_{s=1}^{S_i-1} \left(\sum_{j=1}^{J} Y_{ij}(s) Z^f_{j\ell} \right) \xi_{ikk'}(s+1), \ \ell \geq 1. \\
\widetilde{n}^w_{j\ell} &= \sum_{i=1}^{I} Y_{ij}(1)\tau_{j\ell} + \sum_{i=1}^{I} \sum_{s=2}^{S_i} \sum_{j'=0}^{J} \mathbf{1}\{j' \neq j\} Y_{ij'}(s-1) Y_{ij}(s) \tau_{j\ell}, \ j, \ell \geq 1, \\
\widetilde{n}^w_{k0}(1) &= \sum_{i=1}^{I} \zeta_{ik}(s) Y_{i0}(s), \\
\widetilde{n}^w_{k\ell}(1) &= \sum_{i=1}^{I} \sum_{s=1}^{S_i-1} \zeta_{ik}(s) \sum_{j=1}^{J} Y_{ij}(s)\tau_{j\ell}, \ \ell \geq 1, \\
\widetilde{n}^w_{k0\ell'} &= \sum_{i=1}^{I} \sum_{s=1}^{S_i-1} \zeta_{ik}(s) Y_{i0}(s) \sum_{j=1}^{J} Y_{ij}(s+1)\tau_{j\ell}, \ \ell' \geq 1, \\
\widetilde{n}^w_{k0\neg} &= \sum_{i=1}^{I} \sum_{s=1}^{S_i-1} \zeta_{ik}(s) Y_{i0}(s) D_i(s), \\
\widetilde{n}^w_{k\ell 0} &= \sum_{i=1}^{I} \sum_{s=1}^{S_i-1} \zeta_{ik}(s) \sum_{j=1}^{J} Y_{ij}(s)\tau_{j\ell} Y_{i0}(s+1), \ \ell \geq 1, \\
\widetilde{n}^w_{k\ell\ell'} &= \sum_{i=1}^{I} \sum_{s=1}^{S_i-1} \zeta_{ik}(s) \sum_{j=1}^{J} Y_{ij}(s)\tau_{j\ell} \sum_{j'=1}^{J} Y_{ij'}(s+1)\tau_{j'\ell'}, \ \ell, \ell' \geq 1, \\
\widetilde{n}^w_{k\ell\neg} &= \sum_{i=1}^{I} \sum_{s=1}^{S_i-1} \zeta_{ik}(s) \sum_{j=1}^{J} Y_{ij}(s)\tau_{j\ell} D_i(s), \ \ell \geq 1,
\end{aligned}
$$

We may also set $\widetilde{n}^w_{k00} = 0$ (no transition from unemployment to unemployment as they are counted in $n^w_{k0\neg}$).

Note that we also obtain the following expression for

$$\begin{aligned}\mathbb{E}_R \ln \mathcal{L}\left(X, Z^f, Z^w\right) = J \sum_{\ell=1}^{L} \widetilde{\pi}_\ell^f \ln \pi_\ell^f + I \sum_{k=1}^{K} \widetilde{\pi}_k^w \ln \pi_k^w \\ + \sum_{\ell=1}^{L} \widetilde{n}_\ell^f \sum_{j=1}^{J} \widetilde{\theta}_{j\ell} \ln \theta_{j\ell} + \sum_{k=1}^{K} \widetilde{n}_k^w(1) \sum_{\ell=0}^{L} \widetilde{m}_{k\ell} \ln m_{k\ell}^w \\ + \sum_{k=1}^{K} \sum_{\ell=0}^{L} \widetilde{n}_{k\ell}^w \left(\sum_{\ell'=0}^{L} \widetilde{M}_{k\ell\ell'} \ln M_{k\ell\ell'} + \widetilde{M}_{k\ell\neg} \ln M_{k\ell\neg} \right) \\ + \sum_{k=1}^{K} \sum_{\ell=1}^{L} \widetilde{n}_{k\ell}^w \ln \varphi\left(\widetilde{W}_{k\ell}, \widetilde{W^2}_{k\ell}, \eta_{k\ell}\right).\end{aligned}$$

C.3 Worker E-step

We update $\zeta_{ik}(s)$ and $\xi_{ikk'}(s)$ in the worker E-step as follows.

The worker E-step problem is

$$\begin{aligned}&\max_{R^w} \sum_{Z^f} R^f(Z^f) \sum_{Z^w} R^w(Z^w) \ln \mathcal{L}(X, Z^w | Z^f) - \sum_{Z^w} R^w(Z^w) \ln R^w(Z^w) \\ &= \max_{R^w} \sum_{Z^w} R^w(Z^w) \ln \left[\exp\left(\sum_{Z^f} R^f(Z^f) \ln \mathcal{L}(X, Z^w \mid Z^f) \right) \right] - \sum_{Z^w} R^w(Z^w) \ln R^w(Z^w)\end{aligned}$$

subject to the restriction $\sum_{Z^w} R^w(Z^w) = 1$. Hence

$$\begin{aligned}R^w(Z^w) &\propto \exp\left(\sum_{Z^f} R^f(Z^f) \ln \mathcal{L}(X, Z^w \mid Z^f) \right) \\ &= \prod_{i=1}^{I} \frac{\exp \sum_{Z^f} R^f(Z^f) \ln \mathcal{L}_i(X_i, Z_i^w \mid Z^f)}{\sum_{Z_i^w} \exp \sum_{Z^f} R^f(Z^f) \ln \mathcal{L}_i(X_i, Z_i^w \mid Z^f)} \\ &= \prod_{i=1}^{I} R_i^w(Z_i^w).\end{aligned}$$

This means that the optimal worker posterior distribution $R^w(Z^w)$ is indeed independent across workers.

We then obtain the marginal posterior probabilities as:

$$\begin{aligned}\zeta_{ik}(s) &= \mathbb{P}_{R^w}\left(Z_i^w(s) = k\right) \\ &= \frac{\sum_{Z_i^w : Z_i^w(s) = k} \overline{\mathcal{L}}_{i,\tau}\left(X_i, Z_i^w\right)}{\sum_{Z_i^w} \overline{\mathcal{L}}_{i,\tau}\left(X_i, Z_i^w\right)}\end{aligned}$$

and

$$\begin{aligned}\xi_{ikk'}(s+1) &= \mathbb{P}_{R^w}\left(Z_{ik}^w(s)=1, Z_{ik'}^w(s+1)=1\right)\\ &= \frac{\sum_{Z_i^w: Z_i^w(s)=k, Z_i^w(s+1)=k'} \overline{\mathcal{L}}_{i,\tau}\left(X_i, Z_i^w\right)}{\sum_{Z_i^w} \overline{\mathcal{L}}_{i,\tau}\left(X_i, Z_i^w\right)},\end{aligned}$$

where

$$\begin{aligned}\ln \overline{\mathcal{L}}_{i,\tau}\left(X_i, Z_i^w\right) &:= \sum_{Z^f} R^f(Z^f) \ln \mathcal{L}_i\left(X_i, Z_i^w \mid Z^f\right)\\ &= \sum_{k=1}^{K} Z_{ik}^w(1) \ln\left[\alpha_k(1)\beta_k(1|\tau)\right]\\ &+ \sum_{s=2}^{S_i-1}\sum_{k=1}^{K} Z_{ik}^w(s-1) \sum_{k'=1}^{K} Z_{ik'}^w(s) \ln \alpha_{kk'}(s|\tau) + \sum_{s=2}^{S_i}\sum_{k=1}^{K} Z_{ik}^w(s) \ln \beta_k(s|\tau)\\ &= \ln\left[\alpha_{k_1}(1)\beta_{k_1}(1|\tau)\alpha_{k_1k_2}(2|\tau)\beta_{k_2}(2|\tau)\alpha_{k_2k_3}(3|\tau)\beta_{k_3}(3|\tau)...\right],\end{aligned}$$

with $Z_i^w = (k_1, k_2, k_3, ...)$. The fact that Z^f has been integrated out does not alter the structure of this likelihood. It is like the complete likelihood, with Z^f replaced by τ in the definition of the α, β state and emission probabilities.

The Forward-Backward algorithm can therefore applied to this non standard context, yielding the following simple formulas for the posterior worker type probabilities:

$$\zeta_{ik}(s) = \frac{F_{ik}(s)B_{ik}(s)}{L_i}, \quad \xi_{ikk'}(s+1) = \frac{F_{ik}(s)\alpha_{kk'}(s+1|\tau)\beta_{k'}(s+1|\tau)B_{ik'}(s+1)}{L_i},$$

where $F_{ik}(s)$ and $B_{ik}(s)$ are the Forward and Backward operators (defined just below) and with

$$L_i = \sum_{k=1}^{K} F_{ik}(s)B_{ik}(s).$$

The Forward algorithm. The forward step defines the variables,

$$\begin{aligned}F_{ik}(s) &= \sum_{Z_i^w(1:s-1)} \exp \sum_{Z^f} R^f(Z^f) \ln \mathcal{L}_i\left(X_i(1:s), Z_i^w(1:s-1), Z_i^w(s)=k \mid Z^f\right)\\ &= \sum_{Z_i^w(1:s-1)} \overline{\mathcal{L}}_{i,\tau}\left(X_i(1:s), Z_i^w(1:s-1), Z_i^w(s)=k\right).\end{aligned}$$

Up to the averaged out firm types, this is the likelihood of the observed emissions for periods 1 to s, and of the worker being in state k at s. The forward variables can be

calculated iteratively as

$$F_{ik}(1) = \alpha_k(1)\beta_k(1),$$
$$F_{ik}(s) = \sum_{\kappa=1}^{K} F_{i\kappa}(s-1)\alpha_{\kappa k}(s|\tau)\beta_\kappa(s|\tau), \quad s = 2, ..., S_i.$$

Remark. For S_i large enough, the multiplication of many values lower than one can run into numerical underflow problems. To deal with this, we can use the "log-sum-exp" trick. First, we use a log coding:

$$\ln F_{ik'}(s) = \ln\left(\sum_{k=1}^{K} \exp\left[\ln F_{ik}(s-1) + \ln \alpha_{kk'}(s) + \ln \beta_{k'}(s)\right]\right).$$

Then, to avoid numerical underflows, we write the first log as

$$\ln \sum_{k=1}^{K} e^{x_k} = c + \ln \sum_{k=1}^{K} e^{x_k - c}, \quad c = \max_k x_k.$$

This way, we make sure that $e^{x_k - c} = 1$ for some $k = 1, ..., K$, and therefore that the log is positive. Alternatively, we can use the method in Rabiner (1989).

The Backward algorithm. The backward step defines, for $s = 1, ..., S_i - 1$,

$$\begin{aligned} B_{ik}(s) &= \sum_{Z_i^w(s+1:S_i)} \exp \sum_{Z^f} R^f(Z^f) \ln \mathcal{L}_i \left(X_i(s+1:S_i), Z_i^w(s+1:S_i) \mid Z^f, Z_i^w(s) = k\right) \\ &= \sum_{Z_i^w(s+1:S_i)} \overline{\mathcal{L}}_{i,\tau} \left(X_i(s+1:S_i), Z_i^w(s+1:S_i) \mid Z_i^w(s) = k\right). \end{aligned}$$

This is the likelihood of the emissions after $s+1$ given the worker type at s. Here again, the backward variables can be calculated iteratively. For

$$B_{ik}(s) = \sum_{k'=1}^{K} \alpha_{kk'}(s+1|\tau)\beta_{k'}(s+1|\tau)B_{ik'}(s+1), \;\; s = 1, ..., S_i - 1,$$
$$B_{ik}(S_i) = 1.$$

We finally obtain, for $s = 1, ..., S_i$,

$$\sum_{Z_i^w(1:s-1), Z_i^w(s+1:S_i)} \overline{\mathcal{L}}_{i,\tau} \left(X_i, Z_i^w(1:s-1), Z_i^w(s) = k, Z_i^w(s+1:S_i)\right) = F_{ik}(s)B_{ik}(s),$$

and for $s = 1, ..., S_i - 1$,

$$\sum_{Z_i^w(1:s-1), Z_i^w(s+2:S_i)} \overline{\mathcal{L}}_{i,\tau}\left(X_i, Z_i^w(1:s-1), Z_i^w(s) = k, Z_i^w(s+1) = k', Z_i^w(s+1:S_i)\right)$$
$$= F_{ik}(s)\alpha_{kk'}(s+1|\tau)\beta_{k'}(s+1|\tau)B_{ik'}(s+1).$$

C.4 Firm E-step

Here we directly solve the problem

$$\max_{R^f} \sum_{Z^f} R^f(Z^f) \sum_{Z^w} R^w(Z^w) \ln \mathcal{L}(X, Z^f, Z^w) - \sum_{Z^f} R^f(Z^f) \ln R^f(Z^f),$$

subject to

$$R^f(Z^f) = \prod_{j=1}^{J} \prod_{\ell=1}^{L} \tau_{j\ell}^{Z^f_{j\ell}}, \quad \sum_{\ell=1}^{L} \tau_{j\ell} = 1.$$

For workers, this direct approach was not as straightforward because $R_i^w(Z_i^w)$ did not factorize as the product of marginals for $Z_i^w(s)$. With the independence restriction

$$\sum_{Z^f} R^f(Z^f) \ln R^f(Z^f) = \sum_{j=1}^{J} \sum_{\ell=1}^{L} \tau_{j\ell} \ln \tau_{j\ell}.$$

The first-order condition for $\tau_{j\ell}$ $(j, \ell \neq 0)$ is easily obtained from the expected log-likelihood as

$$\begin{aligned}
\ln \pi_\ell^f &+ \sum_{i=1}^{I} \sum_{k=1}^{K} \zeta_{ik}(1) Y_{ij}(1) \left[\ln m_{k\ell} + \ln \theta_{j\ell}\right] \\
&+ \sum_{i=1}^{I} \sum_{s=1}^{S_i - 1} \sum_{k=1}^{K} \sum_{k'=1}^{K} \xi_{ikk'}(s) Y_{ij}(s) \ln A_{k\ell k'} \\
&+ \sum_{i=1}^{I} \sum_{s=1}^{S_i} \sum_{k=1}^{K} \zeta_{ik}(s) Y_{ij}(s) \ln \left[\frac{1}{\sigma} \varphi\left(\frac{W_i(s) - \mu_{k\ell}}{\sigma}\right) + D_i(s) \ln M_{k\ell}^{\neg}\right] \\
&+ \sum_{i=1}^{I} \sum_{s=1}^{S_i - 1} \sum_{k=1}^{K} \zeta_{ik}(s) Y_{ij}(s) \sum_{j'=0}^{J} \mathbf{1}\{j' \neq j\} Y_{ij'}(s+1) \sum_{\ell'=0}^{L} \tau_{j'\ell'} \left[\ln M_{k\ell\ell'} + \ln \theta_{j'\ell'}\right] \\
&+ \sum_{i=1}^{I} \sum_{s=1}^{S_i - 1} \sum_{k=1}^{K} \zeta_{ik}(s) \sum_{j'=0}^{L} \mathbf{1}\{j' \neq j\} Y_{ij'}(s) \sum_{\ell'=0}^{L} \tau_{j'\ell'} Y_{ij}(s+1) \left[\ln M_{k\ell'\ell} + \ln \theta_{j\ell}\right] \\
&- 1 - \ln \tau_{j\ell} + \ln \lambda_j = 0,
\end{aligned}$$

where λ_j is determined by the constraint $\sum_{\ell=1}^{L} \tau_{j\ell} = 1$.

This system is of the form $g(\tau, \lambda) = 0$ with $g(\tau) = (g_j(\tau, \lambda))$ and

$$g_j(\tau, \lambda) = \begin{bmatrix} \tau_j - \lambda_j \exp(a_j + B_j \tau) \\ \mathbf{1}_L^\top \tau_j \end{bmatrix},$$

where $a_j = (a_{j\ell}) \in \mathbb{R}^{L \times 1}$ with

$$\begin{aligned} a_{j\ell} = \ln \pi_\ell^f &+ \sum_{i=1}^{I} \sum_{k=1}^{K} \zeta_{ik}(1) Y_{ij}(1) \left[\ln m_{k\ell} + \ln \theta_{j\ell}\right] \\ &+ \sum_{i=1}^{I} \sum_{s=1}^{S_i - 1} \sum_{k=1}^{K} \sum_{k'=1}^{K} \xi_{ikk'}(s) Y_{ij}(s) \ln A_{k\ell k'} \\ &+ \sum_{i=1}^{I} \sum_{s=1}^{S_i} \sum_{k=1}^{K} \zeta_{ik}(s) Y_{ij}(s) \ln \left[\frac{1}{\sigma} \varphi \left(\frac{W_i(s) - \mu_{k\ell}(s)}{\sigma} \right) + D_i(s) \ln M_{k\ell}^{\neg} \right] - 1 \end{aligned}$$

and $B_j = (B_{j1}, ..., B_{jJ})$ where $B_{jj} = 0_{LL}$ and, for $j \neq j'$,

$$B_{jj'} = \sum_{i=1}^{I} \sum_{s=1}^{S_i - 1} \sum_{k=1}^{K} \zeta_{ik}(s) \left(Y_{ij}(s) Y_{ij'}(s+1) \ln M_k + Y_{ij'}(s) Y_{ij}(s+1) \ln M_k^\top \right)$$

are L^2-matrices, where $\ln M_k$ is the matrix with $\ln M_{k\ell\ell'}$ in position (ℓ, ℓ').

To update the τ_j's, one can either iterate the fixed point system:

$$\tau_j \propto \exp(a_j + B_j \tau), j = 1, ..., J,$$

perhaps sequentially (first the largest firm, then the second largest, etc), without any guarantee of numerical convergence, or use a gradient optimization algorithm exploiting the system's sparsity.

The Jacobian $J(\tau, \lambda) = \frac{\partial g(\tau, \lambda)}{\partial (\tau^\top, \lambda^\top)}$ follows with

$$\begin{aligned} &\frac{\partial g_j(\tau, \lambda)}{\partial \tau_{j'}^\top} = \begin{bmatrix} B_{jj'} \odot \exp(a_j + B_j \tau) \\ 0_{1,L} \end{bmatrix} \ (j \neq j'), \quad \frac{\partial g_j(\tau, \lambda)}{\partial \tau_j} = \begin{bmatrix} I_L \\ 1_{1,L} \end{bmatrix}, \\ &\frac{\partial g_j(\tau, \lambda)}{\partial \lambda_{j'}} = 0_{(L+1),1} \ (j \neq j'), \quad \frac{\partial g_j(\tau, \lambda)}{\partial \lambda_j} = \begin{bmatrix} \exp(a_j + B_j \tau) \\ 0 \end{bmatrix}. \end{aligned}$$

Note also that many couples of firms (j, j') are disconnected because they share no worker though a direct transition. It follows that matrix $B = (B_{jj'})$ is sparse. A Newton-Raphson solver for sparse systems can then be used to solve this problem. For example, one can use a conjugate gradient method for solving the linear systems

$$J(x_k)(x_{k+1} - x_k) = g(x_k), \quad x = (\tau, \lambda),$$

iteratively.