

On Worker and Firm Heterogeneity in Wages and Employment Mobility: Evidence from Danish Register Data

Rasmus Lentz* Suphanit Piyapromdee† Jean-Marc Robin‡

June 13, 2018

Abstract

In this paper, we develop a model of wage dynamics and employment mobility with unrestricted interactions between worker and firm unobserved characteristics in both wages and employment mobility. We adopt the finite mixture approach of Bonhomme et al. (2017). The model is estimated on Danish matched employer-employee data for the period 1985-2013. The estimation includes gender, education, age, tenure and time controls. We find significant sorting on wages and it is stable over the period. Sorting is established early in careers, increasing during the first decade after which it declines steadily. Job-to-job mobility displays a “mean-reverting” pattern that maintains correlations between worker and firm types to a stationary level. Counterfactuals demonstrate that sorting is primarily driven by two channels: First, a “preference” channel whereby higher wage workers are more likely to accept jobs in higher wage firms. Second, a job finding channel where the job destination distribution out of non-employment is stochastically increasing in the wage type of the worker.

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

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

†University College London. Email s.piyapromdee@ucl.ac.uk. I gratefully acknowledge the financial support from the Danish Council for Independent Research Social Sciences, grant DFF – 6109-00037

‡Sciences-Po, Paris and University College London. Email jeanmarc.robin@sciencespo.fr

1 Introduction

What is the assignment of workers to firms, and how is it realized? To answer these two questions, we estimate a flexible parametric model of wages and mobility with two-sided unobserved heterogeneity using Danish matched employer-employee data. First, this allows us to measure how wages depend on worker and firm unobserved heterogeneity, and how worker and firm unobserved types correlate across matches in any cross-section (sorting). Second, because we also model the dynamic matching process between workers and firms, we can explain how the estimated cross-sectional match distribution is produced.

Since the seminal contribution of Abowd et al. (1999) (hereafter AKM) the literature studying how individual wages vary across workers and firms has focussed on estimating a linear model with worker and firm fixed effects, sometimes a match-specific effect, using Ordinary Least Squares.¹ The AKM model, in its version without match effects, has been criticized for being too restrictive. Eeckhout and Kircher (2011) claim that non monotonicity in the way wages depend on worker and firm effects may explain the weak correlation that is found when estimating the AKM additive-effect model. Several recent studies have designed and estimated search-matching models that confirm Eeckhout and Kircher’s hypothesis.² We shall not follow the structural route because any structural model is also bound to severely restrict the way worker and firm heterogeneity determine wages and employment mobility. Our aim in this paper is to document these relationships under the same restrictions as the AKM model regarding job-to-job mobility, but with more flexibility in the link between wages and heterogeneity. We leave to further work such extensions as endogenous mobility (depending on past wages given heterogeneity), or time-varying unobserved types (for example in order to model human capital accumulation beyond deterministic age effects).

We depart from AKM’s linearity by following the recent approach of Bonhomme et al. (2017) (hereafter BLM), who put no shape restrictions on the way wages and job-matching depend on worker and firm heterogeneity.³ This is rendered possible by assuming that worker and firm unobserved types are discrete; that is, workers and firms can be clustered into homogeneous groups of identical workers and firms. Like all finite-mixture models, but unlike fixed-effect models, BLM show that the model is identified using a panel of workers of finite length.

As in BLM’s application, the log wage of a match is assumed to be normally distributed with a mean and variance that depends on both observed and unobserved worker and firm het-

¹See Holzer et al. (2004); Martins (2008); Iranzo et al. (2008); Gruetter and Lalive (2009); Bagger et al. (2013); Card et al. (2013); Woodcock (2015); Song et al. (2015); Card et al. (2016). Recent application correct the naive OLS estimator for finite sample biases using the method in Andrews et al. (2008).

²See Lise et al. (2016); Lise and Robin (2017); Hagedorn et al. (2017); Bagger and Lentz (2014); de Melo (2018).

³As far as we know, BLM and our paper are the only ones which model the dynamic process of wages and employment mobility with both worker and firm heterogeneity and without structural assumptions that tie mobility to wage outcomes. There are many papers with only worker heterogeneity (e.g. Altonji et al., 2013) and one paper with only firm heterogeneity (Abowd et al., 2006).

erogeneity in a non-parametric way. Our way of modeling mobility yet differs. Consider the probability for a worker of type k and observed time-varying characteristics x who is currently employed at a firm of type ℓ to move to a firm of type ℓ' before the end of the period. The collection of these different probabilities for all k, x, ℓ, ℓ' is soon a very large set. BLM do not restrict these probabilities, which makes them easily estimated (given observed types) by empirical frequencies, but very imprecisely. In order to reduce the dimensionality of the estimation problem, we assume that employed workers meet other employment opportunities at a rate that depends on worker characteristics k and x . The type of the employment opportunity ℓ' is drawn from a distribution that is common to all workers and independent of the type of the current firm. Finally, the probability that the worker decides to quit the current job and move to the new firm takes the shape of a binomial logit, each alternative match being assigned a specific utility value (function of k, ℓ and x). Hence, one interpretation of the mobility patterns in the paper is that of a standard on-the-job search model with random utility. By doing so, we make the transition probabilities both easier to identify and estimate, and easier to interpret.

A second difference with BLM is that we use the entire panel of Danish matched employer-employee data for the period 1987-2013 for estimation (instead of just two years). We assume that worker and firm unobserved types do not change over time. However, the way wage and mobility parameters depend on observed and unobserved types is allowed to change over time by periods of three years – some degree of stationarity being required for identification. For example, the complementarity of worker and firm types in match productivity (wage) can thus change along the business cycle and be subject to structural change. Therefore, a type here conditions wage and mobility in a dynamic way. This allows us to run counterfactual exercises where we remove structural changes to the parameters (time variations). This could also allow, in particular, to calculate ex post earnings present values, like in the recent work of Guvenen et al. (2017), conditional on worker types. The downside of this approach to unobserved heterogeneity is that we may need more types to capture the diversity of possible dynamics.⁴

A third difference with BLM relates to the estimation procedure, which is in two steps. First, they classify firms using an automatic classification procedure (k -means) based on the distribution of wages within each firm. In a second step, they estimate wage and mobility parameters with unobserved worker types, but observed firm types using the EM algorithm. Instead, we propose a Classification EM algorithm in the spirit of Celeux and Govaert (1992), which updates the firm classification after the EM step so as to improve the expected likelihood. We make thereby use of mobility data, in addition to wages, to classify firm types. This can be important for measuring inequality as two firms offering identical wages but with different levels of job stability would result in different lifetime earnings for workers. Moreover, the mobility parameters involve substantial non-linearity in the likelihood function. The structure of

⁴We checked that this modeling assumption was not determining our results by alternatively estimating 9 completely separate models, one for each 3-year subpanel. Both our evaluation of the correlation between worker and firm effects and the variance decomposition barely changed.

transition probabilities being related to a Bradley-Terry model, we formulate an MM algorithm (Hunter, 2004; Hunter and Lange, 2004) that is embedded in the M-step and allows a fast solution for mobility parameters – a substantial contribution for the feasibility of the estimator. Monte Carlo simulations show that this CEM algorithm improves over a one-step k -means.

Nevertheless, we do not view our paper’s main contribution as being technical. Our empirical application is rich with interesting lessons. First, we show both in the data and in large-scale simulations (i.e. with a one million workers and 100,000 firms, and using our estimated complex wage and mobility model) that the AKM model is overfitting in finite samples. It tends to underestimate the residual variance and sorting (i.e. the correlation between worker and firm fixed effects, and the contribution of the covariance between worker and firm fixed effects to the log-wage variance), and overestimate the contribution of the worker and firm effects to the overall variance. We find relatively large biases for the OLS estimator of the AKM model.

Second, our estimates show an apparent disagreement in the way unobserved heterogeneity determines conditional mean wages, on one hand, and, on the other hand, the idiosyncratic wage variance and the mobility parameters.⁵ The strongest link is estimated for layoff rates and job finding probabilities for non-employed workers. The parameters governing the way a worker of a given type values job types display a much weaker link, and only at low tenure. As a consequence, the cross sectional distribution of match types that results from the first match drawn when entering the labor market, and from subsequent employment mobility therefore shows evidence of moderate, yet non negligible sorting. We measure sorting as the correlation of the worker and firm components obtained from a linear projection of conditional mean wages on worker and firm type dummies. Our estimates are around 28% and very stable over the all period. This is however a greater correlation than is usually estimated using the AKM model.

Third, our parametric model allows us to run counterfactual and synthetic cohort simulations that help trace key drivers of sorting. Surprisingly, we find that this correlation level is largely already obtained at the first draw of a match in workers’ careers. Cohort simulations show that subsequent job mobility strengthens sorting during the first decade of a cohort’s work life after which sorting steadily weakens. Also surprisingly, we find that a cohort’s peak of sorting is therefore attained quite early, after about 10 years of experience. If we counterfactually start a cohort in non-employment, it quickly (within a year or two) attains the level of sorting implied by our initial match distribution estimates. The strong link between worker types and the estimated job destination distributions of non-employed workers is one primary force for the result. The mean-reverting nature of firm mobility is another important force. Conditional on one worker’s type, the dynamic process of firm types induced by job mobility is estimated strongly stationary. The main parameters determining the degree of (stationary) sorting between worker and firm types are the preference for the job, the sampling distribution of firm types and

⁵For example, we estimate $\mu_{k\ell}$ and $\sigma_{k\ell}$ the conditional mean log-wage and standard deviation for a worker of type k and an employer of type ℓ . The correlation between $\mu_{k\ell}$ and $\sigma_{k\ell}$ across matches (k, ℓ) is estimated around 0.10 for all workers with tenure greater than 100 days.

the job finding rates of unemployed.

The layout of the paper is as follow. Section 2 describes the model, Section 3 the estimation procedure, Section 4 the parametric specification. Then, Section 5 presents the estimation results, and the last two section presents applications. Section 7 is about measuring sorting and Section 8 about the log-wage variance decomposition.

2 The Model

We use the matched employer-employee data from Denmark from 1985-2013. Wages are reported at annual frequency and mobility data of workers are reported at a weekly level. Our analysis panel starts in 1987 and we use information from 1985 and 1986 to distinguish between short and long tenure jobs in the stock of jobs in 1987. We restrict the sample to include only employment spells that start after individuals finished their highest levels of education. We remove any spells that start after the individuals turn 50 years old and treat any spells with less than an average of 25 work hours in a week as a non-employment spell.⁶

Workers are indexed by $i \in \{1, \dots, I\}$ and firms by $j \in \{0, 1, \dots, J\}$, where $j = 0$ reflects non-employment. For each worker i , we observe a set of time-invariant characteristics z_i including gender and education. Firms also differ from each other by a set of observed, fixed characteristics r_j such as public/private status. Each worker i is either drawn from the working age population in 1987, or enters the panel in the first week of the first year following his or her last year of schooling. Individual trajectories $(w_{it}, j_{it}, x_{it})_{t=1}^T$ are recorded at weekly frequency, where $j_{it} \equiv j(i, t) \in \{0, 1, \dots, J\}$ is the employer's ID in the t -th week of observation, x_{it} are observed worker controls including potential experience (age minus age upon leaving school), job tenure, and calendar time, and w_{it} is the worker's log-wage rate at occurrence t . Note that although the number of repeated observations T varies across individuals, we adopt the simplified notation of a balanced panel. Job mobility is measured at a weekly frequency but wages are measured at annual frequency. Thus, wage observations are missing except for the first week of any new match and the first week of the year.⁷

We assume that employers (firms) can be clustered into L different groups indexed by $\ell \in \{1, \dots, L\}$ and that workers can be clustered into K different groups indexed by $k \in \{1, \dots, K\}$. The index ℓ_j is the type of firm j and k_i is the type of worker i . Non-employment is observable and is denoted by $\ell = 0$. Worker and firm type assignments are assumed to be fixed over the duration of the panel. That is, tenure and experience impact a worker's mobility and wages conditional on a fixed type assignment. By implication, a type is characterized also by its

⁶We operate this selection essentially for comparability with other studies keeping only full-time workers.

⁷At the end of the year, employers declare to the tax administration services the cumulative salary paid to each of their employees during the elapsed year. Total salary divided by the total number of hours worked by the worker in that year is the wage rate that we assign to the first week of that year or of the next employment spell if the job started inside the year. Hours worked are inferred by Statistics Denmark from observed mandatory pension contributions that are conditional on weekly hours worked.

dynamic wage and mobility paths. An often used alternative in the literature (see, for example Card et al. (2013)) is to stratify the data by calendar time intervals (a common length seems to be around 7 years) and estimate the model independently for each stratification.

We make two assumptions regarding worker and firm classifications. We omit the individual index i for simplicity and denote as $\mathbb{P}(x)$ the (generic) probability mass or density function of a random variable X , describing the distribution of some trait in the population of workers, at a point x .

Assumption 1 (Initial condition). *1) Initial x_1 does not predict worker type k conditional on z : $\mathbb{P}(k|x_1, z) = \mathbb{P}(k|z)$. 2) Initial employer type ℓ_1 is independent of z given k and x_1 : $\mathbb{P}(\ell_1|z, k, x_1) = \mathbb{P}(\ell_1|k, x_1)$.*

Leaving aside the dependence to x_1 , these assumptions are natural. First, it is well known that the effects of time-invariant controls are not identified in fixed-effect models and are subsumed in the fixed effects. Second, by this way, we have a common scale of heterogeneity to compare workers of different gender or education and jobs in different industries. Third, there is *a priori* no loss of generality in proceeding this way. Suppose for example that wages are constant across all men and across all women, but differ across gender types. Then the best way of classifying wages given gender will be the align k on gender. However, in practice, K is likely to be lower than the number of different values of z . So there may be some loss of information.

The assumption that z predicts k independently of x_1 is more disputable. First, because some workers in 1987 are drawn from the stock. So x_1 contains tenure and potential experience for these workers. If better workers tend to have higher tenure, for example, then tenure predicts worker type. Second, the association between type and gender and education may vary across cohorts of workers.⁸

Our next assumption is a conditional independence assumption giving a Markovian structure to the model.

Assumption 2 (Conditional independence). *The next employer type and wage (w_t, ℓ_{t+1}) depend on the current information: $z, \boldsymbol{\ell}_t = (\ell_t, \ell_{t-1}, \dots)$, $\mathbf{w}_{t-1} = (w_{t-1}, \dots)$ and $\mathbf{x}_t = (x_t, x_{t-1}, \dots)$ as follows*

$$\mathbb{P}(w_t, \ell_{t+1}|z, k, \boldsymbol{\ell}_t, \mathbf{w}_{t-1}, \mathbf{x}_t) = \mathbb{P}(w_t|k, \ell_t, x_t) \mathbb{P}(\ell_{t+1}|k, \ell_t, x_t).$$

We implicitly assume that mobility to ℓ_{t+1} occurs at the end of period t , and hence is conditioned by x_t , like the wage w_t that is realized in period t .

The identification of finite mixture models is now beginning to be well understood (see e.g. Hall and Zhou, 2003; Hu, 2008; Hu and Schennach, 2008; Kasahara and Shimotsu, 2009; Allman et al., 2009; Hu and Shum, 2012; Henry et al., 2014; Bonhomme et al., 2016b,a; Gassiat

⁸A more sophisticated assumption would allow k to depend on z and cohort. Then, for experienced workers in 1987, we would translate the current date 1987 and potential experience into the corresponding cohort. Moreover, instead of conditioning on tenure, we would add a term in the likelihood for the probability of initial tenure given worker type, firm type, initial experience and initial date (1987).

et al., 2016). Recently, Bonhomme et al. (2017) have applied these identification techniques to this model of wages and mobility with two-sided heterogeneity. They show that two consecutive observations of wages and employer’s types are generically sufficient to identify the model. In general, two repeated observations are not enough to identify a finite mixture (Allman et al., 2009). BLM’s proof based on the existence of “cycles”, i.e. workers returning to some initial firm type after a while, is a bit complicated. In Appendix A we discuss two identification arguments that complement their analysis. In particular, we show identification with two periods if individuals can change employer within the same group of firms. In this last case, wage dynamics for stayers and wage dynamics for movers within the same group of firms allows to write the identification problem in the form of a joint eigenvalue decomposition that delivers uniqueness of eigenvectors. In the application, to stay on the safe side of identification and in order to increase the precision of estimates, we shall use three years of observations to estimate the parameters.

3 The estimation procedure

In this section we develop a Classification Expectation Maximization (CEM) algorithm for estimating the mixture model. We shall be treating the unobserved firm types $F = (\ell_1, \dots, \ell_J)$ as fixed effects (i.e. a parameter to be estimated), and worker types $E = (k_1, \dots, k_J)$ as random effects.

3.1 Likelihood

We state the likelihood for a given firm classification. Let $\ell_{it} = \ell_{j(i,t)}$ denote the type of the firm employing worker i in period t . Let also

$$D_{it} = \begin{cases} 1 & \text{if } j_{i,t+1} \neq j_{it}, \\ 0 & \text{if } j_{i,t+1} = j_{it}, \end{cases}$$

indicate an employer change between t and $t + 1$.

For the given firm classification $F = (\ell_1, \dots, \ell_J)$ let

$$q(\ell|r, F) = \frac{\#\{j : r_j = r, \ell_j = \ell\}}{\#\{j : r_j = r\}} \quad \text{and} \quad q(\ell|F) = \frac{\#\{j : \ell_j = \ell\}}{J}$$

denote the share of type- ℓ firms given observed firm type r (e.g. public/private status) and the unconditional share.

Let f, M, π and m denote parametric versions of the wage density function $\mathbb{P}(w|k, \ell, x)$, the transition probability $\mathbb{P}(\ell'|k, \ell, x)$, the worker type probability $\mathbb{P}(k|z)$, and the distribution of initial employer types $\mathbb{P}(\ell_1|k, x_1)$. For a value $\beta = (f, M, \pi, m)$ of the parameters and a

classification F of firms, the likelihood for one worker i – i.e. of $(w_{it}, j_{it}, x_{it})_{t=1}^T$ conditional on (z_i, x_{i1}) – is

$$\sum_{k=1}^K L_i(k|\beta, F),$$

where the complete individual likelihood is

$$L_i(k|\beta, F) = \frac{m(\ell_{i1}|k, x_{i1}) \pi(k|z_i)}{q(\ell_{i1}|F)} \prod_{t=1}^T f(w_{it}|k, \ell_{it}, x_{it}) \times \prod_{t=1}^{T-1} M(\neg|k, \ell_{it}, x_{it})^{1-D_{it}} \left(\frac{M(\ell_{i,t+1}|k, \ell_{it}, x_{it})}{q(\ell_{i,t+1}|F)} \right)^{D_{it}}, \quad (1)$$

where $M(\neg|k, \ell, x) = 1 - \sum_{\ell'=0}^L M(\ell'|k, \ell, x)$ is the probability of staying with the same employer, and assuming that for the last observation period we do not know whether a mobility occurs or not by the end of it. By convention $f(w|k, \ell, x) = 1$ if the wage observation w is missing. The term $m(\ell_{i1}|k, x_{i1}) \pi(k|z_i)$ is the probability of the initial match type (k_i, ℓ_{i1}) , for $k_i = k$, under Assumption 1. We assume that each firm within a group is equally likely to be selected. With that, the ratio $1/q(\ell_{it}|F)$ is proportional to the probability that *this particular* firm j_{it} , which is of type ℓ_{it} , be selected, either initially or upon job-to-job mobility. The rest of the likelihood factors in this way under Assumption 2. Note that the terms $1/q(\ell_{it}|F)$ do not show up in BLM because they do not imbed the estimation of the firm classification F in the same likelihood-maximization framework as the other parameters. Omitting them would make the firm classification step fail.

3.2 The EM algorithm for a given firm classification

The firm classification in the data is unobserved. It is infeasible to evaluate the likelihood function for the formulation of the model where a firm's unobserved type is a latent variable symmetric to the unobservable worker type formulation in equation (1). The difficulty lies with accounting for the co-dependency between a firm's workers resulting from their matches to a common firm type in a setup where workers move between firms. Consequently, we estimate the model for a given firm classification F . We shall explain in the next subsection how we set and update F .

For a given value of $\beta = (f, M, \pi, m)$, the posterior probability of worker i to be of type k given all wages and controls (all the available information) is

$$p_i(k|\beta, F) \equiv \frac{L_i(k|\beta, F)}{\sum_{k=1}^K L_i(k|\beta, F)}. \quad (2)$$

Note that the factors $1/q(\ell|F)$ in the definition of $L_i(k|\beta, F)$ (equation (1)) appear in the numerator and the denominator of the definition of posterior probabilities in the same way and can be simplified out.

Then, define

$$Q_i(f|\beta^{(m)}, F) = \sum_{k=1}^K p_i(k|\beta^{(m)}, F) \left[\sum_{t=1}^T \ln f(w_{it}|k, \ell_{it}, x_{it}) \right] \quad (3)$$

as the expected log-likelihood of worker i 's wages for a given value $\beta^{(m)}$ of the parameter. The worker posteriors are determined by the model parameters and firm classification $(\beta^{(m)}, F)$, where the superscript m is used to denote the given EM-algorithm iteration. Also, let

$$H_i(M|\beta^{(m)}, F) = \sum_{k=1}^K p_i(k|\beta^{(m)}, F) \left[\sum_{t=1}^{T-1} \left\{ (1 - D_{it}) \ln M(\cdot|k, \ell_{it}, x_{it}) + D_{it} \ln M(\ell_{i,t+1}|k, \ell_{it}, x_{it}) \right\} \right] \quad (4)$$

be the expected log-likelihood of worker i 's employment history conditional on the first state ℓ_{i1} .

The EM algorithm iterates the following steps:

E-step For $\beta^{(m)} = (f^{(m)}, M^{(m)}, \pi^{(m)}, m^{(m)})$ and F , calculate posterior probabilities $p_i(k|\beta^{(m)}, F)$.

M-step Update $\beta^{(m)}$ by maximizing $\sum_i p_i(k|\beta^{(m)}, F) \ln L_i(k|\beta, F)$ subject to $\sum_k \pi(k|z) = 1$ for all z and $\sum_{\ell_1} m(\ell_1|k, x_1) = 1$ for all k, x_1 , that is

$$f^{(m+1)} = \arg \max_f \sum_{i=1}^I Q_i(f|\beta^{(m)}, F), \quad (5)$$

$$M^{(m+1)} = \arg \max_M \sum_{i=1}^I H_i(M|\beta^{(m)}, F), \quad (6)$$

$$\pi^{(m+1)}(k|z) = \frac{\sum_{i=1}^I p_i(k|\beta^{(m)}, F) \mathbf{1}\{z_i = z\}}{\#\{i : z_i = z\}}, \quad (7)$$

$$m^{(m+1)}(\ell|k, x_1) = \frac{\sum_{i=1}^I p_i(k|\beta^{(m)}, F) \mathbf{1}\{x_{i1} = x_1, \ell_{i1} = \ell\}}{\sum_{i=1}^I p_i(k|\beta^{(m)}, F) \mathbf{1}\{x_{i1} = x_1\}}. \quad (8)$$

Note that $m(\ell_1|k, x_1)$ is identified for all levels of experience and tenure only in the first survey year. For all subsequent years, all workers entering the survey are also entering the labor market and hence have zero experience and tenure.

3.3 Firm re-classification given other parameters

Given an initial value $(\hat{\beta}^{(s)}, F^{(s)})$, where $\hat{\beta}^{(s)}$ can be obtained given $F^{(s)}$ using the previous EM algorithm, we update $F^{(s)}$ as

$$F^{(s+1)} = \arg \max_F \sum_{i=1}^I \sum_{k=1}^K p_i(k|\hat{\beta}^{(s)}, F^{(s)}) \ln L_i(k; \hat{\beta}^{(s)}, F). \quad (9)$$

In practice we only search for a firm reclassification that increases the likelihood. To do that we order firms by size from the largest to the smallest. We find $\ell_1^{(s+1)}$ that maximizes equation (9) keeping all other firm types equal to their values in $F^{(s)}$. Then we find $\ell_2^{(s+1)}$ given $\ell_1^{(s+1)}$ and $\ell_3^{(s)}, \dots, \ell_J^{(s)}$ and so on until $\ell_J^{(s+1)}$. We then return to the EM iterations with the updated $F^{(s+1)}$. We call this algorithm a Classification EM algorithm as it resembles the eponym algorithm proposed by Celeux and Govaert (1992) as a variant of the EM algorithm of Dempster et al. (1977). See Appendix B for a detailed exposition of our CEM algorithm and why it is working.

This leaves the question of initialization of the firm classification, $F^{(0)}$. For this we opt for simplicity: We rank firms by average wage per worker in the firm, and cluster firms equally into L groups based on that sorting. Alternatively, we could use the k -means algorithm as in Bonhomme et al. (2017).

4 Empirical specification

In this section, we provide the details of parametric specifications of f and M , and on the number of groups and their labelling.

4.1 Wage distribution

Wages are assumed lognormal given match type. Specifically,

$$f(w|k, \ell, x) = \frac{1}{\sigma_{k\ell}(x)} \varphi\left(\frac{w - \mu_{k\ell}(x)}{\sigma_{k\ell}(x)}\right), \quad (10)$$

with $\varphi(u) = (2\pi)^{-1/2} e^{-u^2/2}$. This specification of the log-wage mean allows for a match-specific mean $\mu_{k\ell}$ and variance $\sigma_{k\ell}^2$.⁹

The M-step update 5 takes the following form:

$$\begin{aligned} \mu_{k\ell}^{(m+1)}(x) &= \frac{\sum_{i=1}^I p_i(k|\beta^{(m)}, F) \sum_{t=1}^T \mathbf{1}\{\ell_{it} = \ell, x_{it} = x\} w_{it}}{\sum_{i=1}^I p_i(k|\beta^{(m)}, F) \sum_{t=1}^T \mathbf{1}\{\ell_{it} = \ell, x_{it} = x\}}, \\ \sigma_{k\ell}^{(m+1)}(x) &= \sqrt{\frac{\sum_{i=1}^I p_i(k|\beta^{(m)}, F) \sum_{t=1}^T \mathbf{1}\{\ell_{it} = \ell, x_{it} = x\} [w_{it} - \mu_{k\ell}^{(m+1)}(x)]^2}{\sum_{i=1}^I p_i(k|\beta^{(m)}, F) \sum_{t=1}^T \mathbf{1}\{\ell_{it} = \ell, x_{it} = x\}}} \end{aligned}$$

(simple averages weighted by posterior type-probabilities).

⁹The distribution f is the distribution of wages for matches with given characteristics. One may interpret these wages as productivity if there is no selection into employment conditional on types.

4.2 Transition probabilities

We omit conditioning on x_{it} to simplify the notations. The probability for a worker of type k of a transition from a firm of type $\ell = 1, \dots, L$ to a firm of type $\ell' = 1, \dots, L$ at time t is specified as

$$M(\ell'|k, \ell, x) = \lambda_k(x) v_{\ell'}(x) P_{k\ell\ell'}(x). \quad (11)$$

Parameter $\lambda_k \in [0, 1]$ is the worker k conditional probability of a meeting with an outside employer. Parameter $v_{\ell'} \geq 0$, with $\sum_{\ell'=1}^L v_{\ell'} = 1$, is the probability that the outside draw is a job of type ℓ' .

The parameter $P_{k\ell\ell'}$ is the probability that the transition from ℓ to ℓ' becomes effective. We assume a Bradley-Terry specification for $P_{k\ell\ell'}$ (see e.g. Agresti, 2003; Hunter, 2004). That is,

$$P_{k\ell\ell'}(x) = \frac{\gamma_{k\ell'}(x)}{\gamma_{k\ell}(x) + \gamma_{k\ell'}(x)}. \quad (12)$$

Parameter $\gamma_{k\ell}$, with $\sum_{\ell=1}^L \gamma_{k\ell} = 1$, measures the quality of the match (k, ℓ) . If the worker draws a same-type job, with little loss of generality, we assume that the worker moves with probability $1/2$.¹⁰

We also model unemployment-employment transitions in a completely unrestricted way:

$$M(\ell'|k, 0) = \psi_{k\ell'}, \quad M(0|k, \ell) = \delta_{k\ell}.$$

By convention, $M(0|k, 0) = 0$. There is no transition from unemployment to unemployment.

With this it follows that the probability of staying non-employed is

$$M(-|k, 0) = 1 - \sum_{\ell'=1}^L M(\ell'|k, 0) = 1 - \sum_{\ell'=1}^L \psi_{k\ell'},$$

and for $\ell \geq 1$, the probability of staying with the same employer is

$$M(-|k, \ell) = 1 - \sum_{\ell'=0}^L M(\ell'|k, \ell) = 1 - \delta_{k\ell} - \lambda_{k\ell} \sum_{\ell'=1}^L v_{\ell'} P_{k\ell\ell'} = 1 - \delta_{k\ell} - \lambda_{k\ell} + \lambda_{k\ell} \sum_{\ell'=1}^L v_{\ell'} (1 - P_{k\ell\ell'}).$$

We prove in Appendix C that the more flexible specification $\lambda_{k\ell}, v_{k\ell'}, \gamma_{k\ell}$ is not identified given knowledge of unrestricted transition probabilities $M(\ell'|k, \ell) = \lambda_{k\ell} v_{k\ell'} P_{k\ell\ell'}$. In this case, for every choice of $M_{k\ell\ell'}$ for all k and $\ell, \ell' \geq 1$ there are two and only two solutions. Let $(\lambda_{k\ell}, v_{k\ell'}, \gamma_{k\ell})$ be such that

$$M(\ell'|k, \ell) = \lambda_{k\ell} v_{k\ell'} \frac{\gamma_{k\ell'}}{\gamma_{k\ell} + \gamma_{k\ell'}}.$$

¹⁰We experimented the specification $P_{k\ell\ell'} = \frac{\gamma_{k\ell'}}{\theta \gamma_{k\ell} + \gamma_{k\ell'}}$, where $\theta > 0$ measures the incumbent's advantage and parametrizes mobility within the same group of firms. However it appeared difficult to disentangle θ from λ .

Then we also have

$$M(\ell'|k, \ell) = \frac{\lambda_{k\ell}}{\gamma_{k\ell}} v_{k\ell'} \gamma_{k\ell'} \frac{1/\gamma_{k\ell'}}{1/\gamma_{k\ell} + 1/\gamma_{k\ell'}}.$$

Hence $\left(\frac{\lambda_{k\ell}}{\gamma_{k\ell}}, v_{k\ell'} \gamma_{k\ell'}, \frac{1}{\gamma_{k\ell}}\right)$ is also a solution (with the appropriate normalizations). We thus need to restrict the parametrization further. We opt for making $\lambda_{k\ell}$ independent of ℓ and $v_{k\ell'}$ independent of k (but dependent on x). An employee will draw an alternative offer with probability λ_k , this offer is from a firm of type ℓ' with probability $v_{\ell'}$ and ℓ' beats ℓ with probability $P_{k\ell\ell'} = \frac{\gamma_{k\ell'}}{\gamma_{k\ell} + \gamma_{k\ell'}}$.

This parametric restriction on transition probabilities $M(\ell'|k, \ell, x)$ is a considerable reduction of dimensionality with respect to letting transition probabilities unrestricted as in BLM. We believe that the loss of generality is amply compensated by the gains in efficiency and intelligibility. However, we also lose in simplicity. Transition probability estimates for the M-step of the EM algorithm are simple frequencies in the unrestricted case. Obtaining estimates in the the parametric restriction is another challenge. In Appendix D we develop an MM algorithm (Hunter, 2004; Hunter and Lange, 2004) that allows to maximize $H(M|\beta^{(m)})$ subject to the parametric restriction on M very rapidly.¹¹

4.3 The number of groups and group labeling

The numbers of latent groups, K and L , are set by experimentation. We tried automatic selection techniques to select K, L such as *BIC* penalization. For a reason that we do not fully understand, such likelihood penalization seems to work for K (workers), but not for L (firms). This is likely due to our treatment of each ℓ_j as a fixed effect. A deeper theoretical analysis of the CEM algorithm (on a simpler specification) is necessary to understand it better.

For each individual i present in the panel at any time, we can calculate $p_i(k)$ the posterior probability of being of type k using equation (2) at the estimated parameters (β and firm classification F), and for any given worker and employment-firm types k, ℓ and for any x indicating a particular tenure, experience and calendar time period, let

$$p(k, \ell, x) \propto \sum_{i,t} p_i(k) \sum_{t=1}^T \mathbf{1}\{\ell_{it} = \ell, x_{it} = x\} \quad (13)$$

¹¹The MM algorithm works by finding a function that minorizes the objective function and that is more easily maximized. Let $f(\theta)$ be the objective concave function to be maximized. At the m step of the algorithm, the constructed function $g(\theta|\theta_m)$ will be called the minorized version of the objective function at θ_m if

$$g(\theta|\theta_m) \leq f(\theta), \forall \theta, \text{ and } g(\theta_m|\theta_m) = f(\theta_m).$$

Then, maximize $g(\theta|\theta_m)$ instead of $f(\theta)$, and let $\theta_{m+1} = \arg \max_{\theta} g(\theta|\theta_m)$. The above iterative method guarantees that $f(\theta_m)$ converges to a local optimum or a saddle point as m goes to infinity because

$$f(\theta_{m+1}) \geq g(\theta_{m+1}|\theta_m) \geq g(\theta_m|\theta_m) = f(\theta_m).$$

be the estimated proportion of observations with such k, ℓ, x values in the panel.¹²

The labeling of the groups is arbitrary. We use wage observations to give a cardinal labeling to the groups. For a given time period, let

$$\mu_{k\ell}(x) = \bar{\mu}(x) + a_k + b_\ell + \tilde{\mu}_{k\ell}(x), \quad (14)$$

be the linear projection of $\mu_{k\ell}(x)$ on all tenure and experience interactions and worker and firm indicators. The term $\tilde{\mu}_{k\ell}(x)$ denotes the residual, and we call it the “match effect”. The projection is obtained by regressing $\mu_{k\ell}(x)$ on tenure*experience dummies, worker dummies and firm dummies, weighing each (k, ℓ, x) by $p(k, \ell, x)$. The estimation algorithm will assign labels arbitrarily. We relabel k and ℓ so that a_k and b_ℓ , given in 5, are now increasing in k, ℓ . Hence, by construction $\mu_{k\ell}$ will on average be increasing in the k and ℓ indices.

5 Estimation results

This section presents the results. First on the distributions of latent types, then on how conditional mean wages and other parameters depend on latent types.

The time-invariant worker characteristics z_i include gender and education. Education level is based on the normed number of years of education associated with the worker’s highest completed degree. The low education group comprises all degrees normed to less than 12 years of education. The medium education group has a norm of exactly 12 years, and the high education group is any education level with a norm greater than 12 years. The time-invariant firm characteristics r_j include the public/private status. The worker’s time variant characteristics x_i include the short/long tenure status, potential experience (time since graduation), and calendar year.¹³ Short tenure is defined as less than 100 weeks of employment, and 26 weeks for non-employment. We divide experience into four groups: less than 5 years; 5-10 years; 11-15 years; and more than 15 years. Additionally, we allow the wage and mobility parameters to vary by 3-year time intervals. This leads to 9 different calendar time groups between 1987-2013.¹⁴ Thus, x_{it} is one of 72 different groups.

5.1 Distributions of worker and firm types

We set the number of worker types to $K = 14$ and the number of firm types to $L = 24$. We indeed obtained greater likelihood gains by categorizing firms more finely than workers.

¹²The proportionality symbol \propto means that the right-hand side of the “equality” needs to be normalized for the left-hand side to be a proper probability.

¹³We can calculate actual experience only for the workers entering the labor market after 1987. This is why we stick to potential experience, or age.

¹⁴We shall interpret calendar time as reflecting the macro environment. It could also refer to cohort effects. There is obviously no way to separate cohort, time and age.

Figure 1: Cross-sectional distributions of types (10-15 years of experience, long tenure, 1999-01)

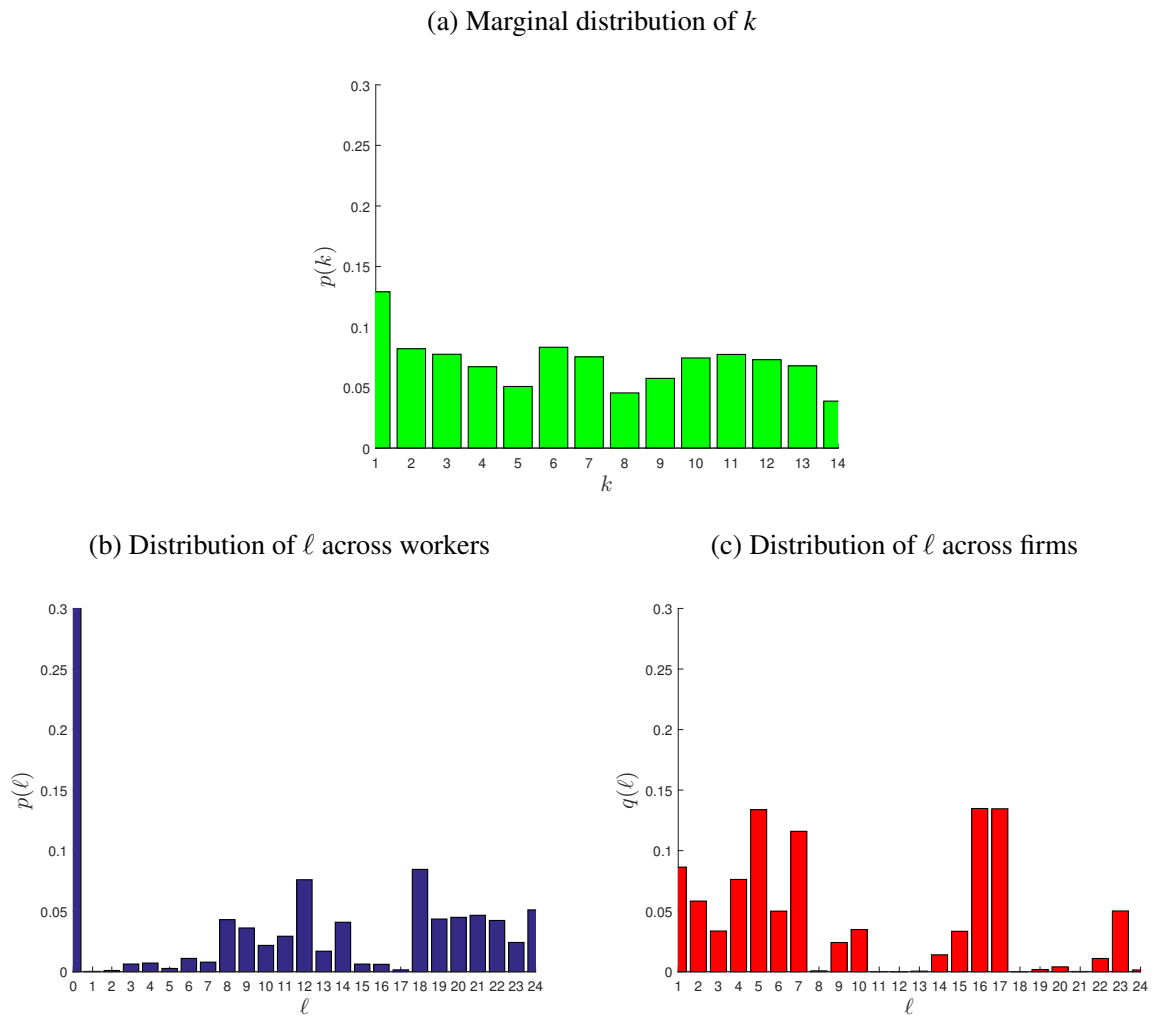


Figure 2: Gender and education conditional average wages

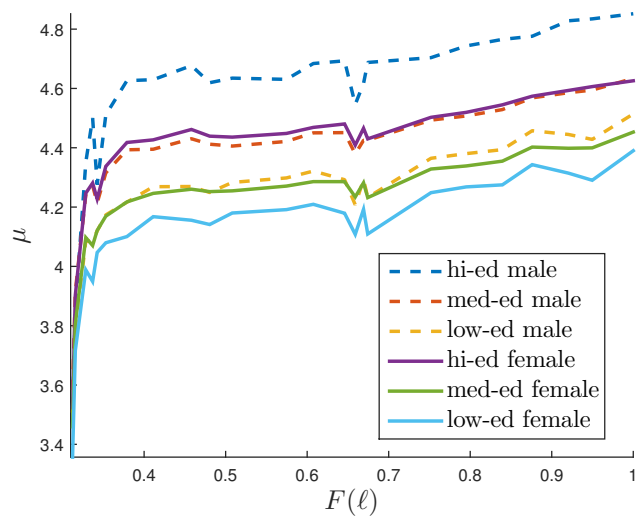


Table 1: Proportion of gender and education by worker type

Worker type k	Gender		Education		
	Male	Female	Low	Medium	High
1	0.48	0.52	0.64	0.31	0.05
2	0.13	0.87	0.40	0.52	0.08
3	0.31	0.69	0.28	0.55	0.17
4	0.42	0.58	0.23	0.57	0.20
5	0.45	0.55	0.30	0.45	0.25
6	0.51	0.49	0.18	0.55	0.27
7	0.40	0.60	0.13	0.42	0.45
8	0.62	0.38	0.24	0.44	0.32
9	0.77	0.23	0.27	0.58	0.15
10	0.76	0.24	0.14	0.66	0.20
11	0.55	0.45	0.07	0.33	0.60
12	0.68	0.32	0.07	0.37	0.56
13	0.78	0.22	0.04	0.28	0.68
14	0.83	0.17	0.03	0.20	0.77

Table 2: Firm Characteristics

Firm type ℓ	Sector			Group size	Avg no. spells	%no. spells
	No cat.	Public	Private			
1	0.17	0.09	0.73	31638	3	0.00
2	0.05	0.04	0.91	21363	5	0.01
3	0.09	0.04	0.87	12307	29	0.03
4	0.07	0.06	0.87	27911	8	0.00
5	0.17	0.05	0.79	49002	2	0.01
6	0.02	0.06	0.92	18328	11	0.19
7	0.37	0.05	0.57	42465	13	0.02
8	0.05	0.71	0.24	262	1736	0.02
9	0.17	0.12	0.71	8829	99	0.05
10	0.05	0.11	0.84	12745	34	2.32
11	0.33	0.43	0.24	58	7209	9.82
12	0.65	0.21	0.15	34	26491	0.00
13	0.41	0.20	0.39	176	3183	30.16
14	0.02	0.18	0.80	5113	122	3.97
15	0.03	0.06	0.91	12222	24	0.00
16	0.04	0.06	0.90	49364	4	39.24
17	0.18	0.04	0.78	49284	2	0.11
18	0.22	0.49	0.30	37	27776	0.58
19	0.09	0.14	0.77	636	1063	0.02
20	0.07	0.11	0.83	1471	491	1.74
21	0.10	0.14	0.76	78	6300	9.73
22	0.12	0.09	0.79	4000	181	0.26
23	0.14	0.06	0.80	18370	27	0.03
24	0.08	0.10	0.82	525	1413	1.67

For conciseness, we only present graphical results for workers who have 10-15 years of experience and more than 100 days of tenure in the 1999-01 period. The online appendix contains graphs for all periods across different characteristics of workers. There are variations over time without any pattern obviously emerging. We calculate

$$p(k) \propto \sum_i p_i(k), \quad p(\ell) \propto \sum_{i,t} \mathbf{1}\{\ell_{it} = \ell\},$$

the estimated marginal cross-sectional distributions of k and ℓ across workers (the marginals of $p(k, \ell, x)$). Figure 1 displays these distributions. The marginal CDFs of k and ℓ (we shall use $F(k), F(\ell)$ to denote them) offer a different labeling for the plots that we are going to analyze. The advantage of plotting for example $\mu_{k\ell}(x)$ against $F(k), F(\ell)$ for a given x (year, tenure and experience) over using k and ℓ for the axes is that we thus compress the axis-scale where there are few workers (as in the case of lower firm types, $\ell \leq 5$) and stretch it where there are many. Interestingly, some firm types are relatively rare in the population of firms (i.e. $\ell = 1, 2, 5, 15, 16, 17$) but frequent in the population of employed workers because these firms have larger sizes. The firm classification respects the assumption that firms within a group are equally likely to be sampled. Hence, to the extent that firm size and vacancy posting are related (as one would expect), the classification will group firms by size, in addition to wage and mobility patterns. Moreover, $F(k)$ is invariant to the labelling of worker types, whether $k = 1, 2, 3$ or $k = 1, 5, 7$, both choices deliver the same $F(k)$.

Table 1 shows how observed characteristics z , gender and education, correlate with unobserved worker types (mean posterior probabilities by observed type). There is a strong correlation between types and observed characteristics. It is seen that high worker types tend to be male and highly educated workers.¹⁵ Figure 2 shows average wage realizations conditional on gender and education and firm type. Although we do not specifically make wages depend on gender and education, the number of worker types K is large enough to deliver the expect link.

Table 2 repeats the operation for firms. Strikingly, most groups gather many firms that hire very few workers, and a few groups gather a small number of firms that hire many workers. For example, the 262 firms in group $\ell = 8$ are mostly public and account for 0.02% of all spells. Then, the 49,364 firms in group $\ell = 16$ and the 6,300 firms in group $\ell = 21$ are mostly private and account for 39.2% and 9.7% of all spells. Although firm size and sector seem to be important factors to explain the estimated classification, it would be obviously difficult to explain why groups 11 and 18 are thus split.

Summing up, although we estimate a relatively small number of groups ($K = 14$ groups of workers and $L = 24$ groups of firms), this classification contains information that differs from the observed characteristics.

¹⁵In reality, types may reflect a combination of heterogeneity in skills, productivity and bargaining power. A more detailed analysis controlling for observed characteristics, such as occupation, and further analyzing sources of heterogeneity in types would be interesting future work.

5.2 Conditional mean wages and other parameters

In Figure 3, we show the time-aggregated $\mu_{k\ell}$ for one experience and tenure status (10-15 years of experience and more than 100 days of tenure). Wages are generally increasing in both tenure and experience, but apart from that the links between expected wages and match types by experience and tenure are similar and we do not show them.

Wages vary significantly more with worker than firm types. It is furthermore notable that $\mu_{k\ell}$ is broadly speaking monotone in the ranked firm and worker types. Worker types agree on the wage ranks of firms, and vice versa. Moreover the lines $\ell \mapsto \mu_{k\ell}$ for all k (left panel) are essentially parallel. These results are broadly consistent with the assumption of linear wage in the AKM model. We will later perform a variance decomposition exercise that will allow us to make this assertion more precise.

In order to avoid an inflation of figures (which are relegated to an online appendix), we only show (in Table 3) how the other parameters correlate with mean wage $\mu_{k\ell}$ with respect to the cross-sectional match distribution $p(k, \ell) \propto \sum_{i,t} p_i(k) \mathbf{1}\{\ell_{it} = \ell\}$ is generally negative and low.

The correlation between idiosyncratic dispersion (as measured by standard deviation $\sigma_{k\ell}$) and mean wages is sizable and negative only for the short-tenure workers. However, looking at the estimated values of $\sigma_{k\ell}(x)$ more closely, idiosyncratic dispersion seems generally quite concentrated with a few outliers (mainly $k = 2$).

Overall, search intensity λ_k is increasing in worker type k (with noise), and this relationship is more pronounced for younger workers who are long-tenure (less than 15 years of labor market experience). It is worth noting that the core sorting mechanism in Bagger and Lentz (2014) implies that more skilled workers search more intensely, consistent with this finding. The sampling distribution v_{ℓ} shows evidence that better firms are more likely to be sampled by higher-tenure workers with little variation across workers of different experience levels. This relationship, however, only exists for long-tenure workers.

Layoff rates $\delta_{k\ell}$ are strongly decreasing in mean wage, especially so for low-tenure workers. The negative link between layoff rates and firm types is also emphasized in Bagger and Lentz (2014) and Jarosh (2015). Note, however, that for higher-tenure workers at least the layoff rate is quite flat and homogeneous. For job finding rates $\psi_{k\ell}$ we find a positive link with mean wage, less strong than for layoff rates and again stronger for low-tenure workers.

Finally, the link between mean wage and the preference for the job (or match value), $\gamma_{k\ell}$, is slightly negative for longer tenure. It is weakly positive (correlation of the order of 0.3) at lower tenure. This may indicate that some workers, for example unemployed, may be forced to accept lower quality jobs that they will sooner than later quit when better opportunities show up. Now, the overall very low correlation between $\gamma_{k\ell}$ and $\mu_{k\ell}$ shows that the drivers of mobility across firm types are to a significant extent unrelated to wages, particularly for long tenure workers. Many other factors such as job amenities, family shocks, and job specific idiosyncratic shocks may explain mobility in ways that is not necessarily optimal in terms of income.

Figure 3: Mean wage μ_{kl} (10-15 years of experience, long tenure, 1999-01)

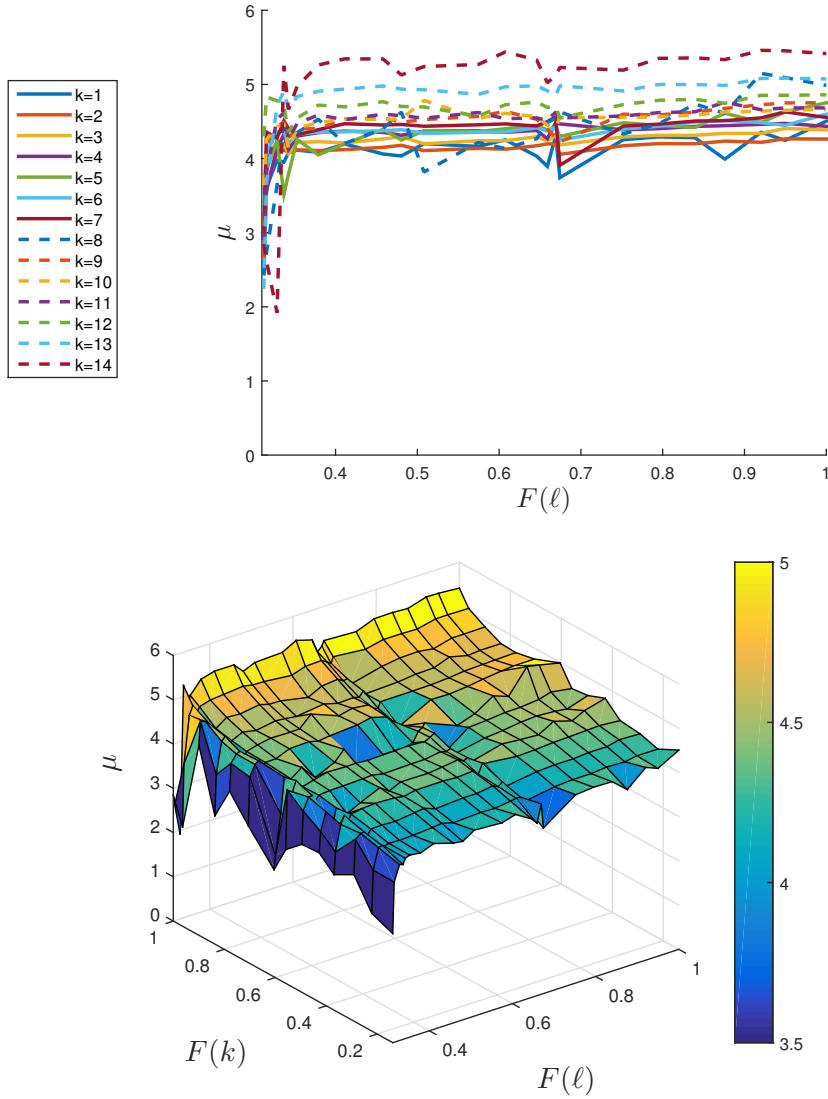


Table 3: Correlation with μ_{kl} in 1999-01

	σ_{kl}	δ_{kl}	ψ_{kl}	γ_{kl}	λ_k	$\nu_{\ell'}$
experience	tenure < 100 weeks					
<5 years	-0.50	-0.79	0.51	0.31	0.33	-0.11
5-10 years	-0.41	-0.72	0.56	0.36	0.40	-0.07
10-15 years	-0.31	-0.66	0.57	0.41	0.46	-0.08
> 15 years	-0.35	-0.61	0.55	0.36	0.24	-0.17
	tenure > 100 weeks					
<5 years	0.01	-0.19	0.28	0.06	0.63	0.45
5-10 years	0.04	-0.25	0.11	-0.10	0.65	0.49
10-15 years	0.06	-0.21	-0.06	0.07	0.58	0.49
> 15 years	0.08	-0.14	-0.26	-0.10	0.22	0.20

6 Is there a ladder effect?

To explore the ladder structure further and the mobility implied by the $\lambda_k, \nu_\ell, \gamma_{k\ell}$ estimates, Figure 4 presents the expected firm type destination conditional on the current firm type and a job-to-job move. There is surprisingly little effect of age or experience. Tenure seems to matter. At low tenure the curves diverge. For low skill workers (say $k \leq 7$; solid lines) the curves are essentially flat, except for experienced workers with long tenure. For workers of higher type ($k > 7$) then the expected destination firm rank is increasing in current firm rank. At higher tenure, we see a reverse phenomenon. Low-skilled workers tend to catch up with high-skill workers. Furthermore, catching-up seems stronger later in the life cycle (the main experience effect).

In Figure 5 we repeat the exercise but we now consider the expected firm type destination not in the next year, but five years ahead. We make two important changes though. First, we reintroduce non-employment either in the origin state ($\ell = 0$), the destination state ($\ell' = 0$), or any intermediate year. Second, we run simulations allowing for tenure to change. However we proceed to the simulation separately for each experience category. The results are striking. There is a much bigger dispersion in destination states across worker groups for younger workers than for older ones. We can see also that starting non-employed is not as bad as starting in a job from one of the worst types. The third amazing result is that the average firm type does not seem to vary much with experience. Finally, we see that mobility is mean-reverting. Starting from a low firm type, mobility will improve the firm location; but starting from a high firm type, mobility reduces the firm quality. It is thus hard to maintain a position at the top of the job ranking if, for some reason, one has to move.

To illustrate the interaction between experience on mobility patterns, we construct a synthetic cohort for the 1999-01 period, a mid point of the longitudinal dimension of our data. The online appendix contains graphs for all periods. The patterns we highlight are invariant across the time periods in the data. The cohort is initialized with zero experience and tenure. Its initial allocation is drawn according to the estimated $m(\ell|k, x)$ where the x category is the one corresponding to zero experience, zero tenure and the time period in question. The cohort is then simulated forward 30 years with experience and tenure evolving endogenously. Calendar time is held counterfactually constant to avoid conflation of experience effects with time effects.

Figure 6a shows that non-employment risk is strongly associated with worker type. Low wage workers have significantly higher non-employment rates early in life than high wage workers. The evolution with experience is such that the sorting on the extensive employment margin is less pronounced for high experience workers. Note that the relatively high non-employment rates are in part a result of our focus on more stable and close to full time jobs.

Figure 6b shows the average firm location, by experience and worker type, for the employed. Higher worker types are on average matched with higher wage firms. On average the age profile of the firm location is flat and even decreasing for the high wage workers, whereas low wage

Figure 4: Conditional mobility across firm types, 1999-01

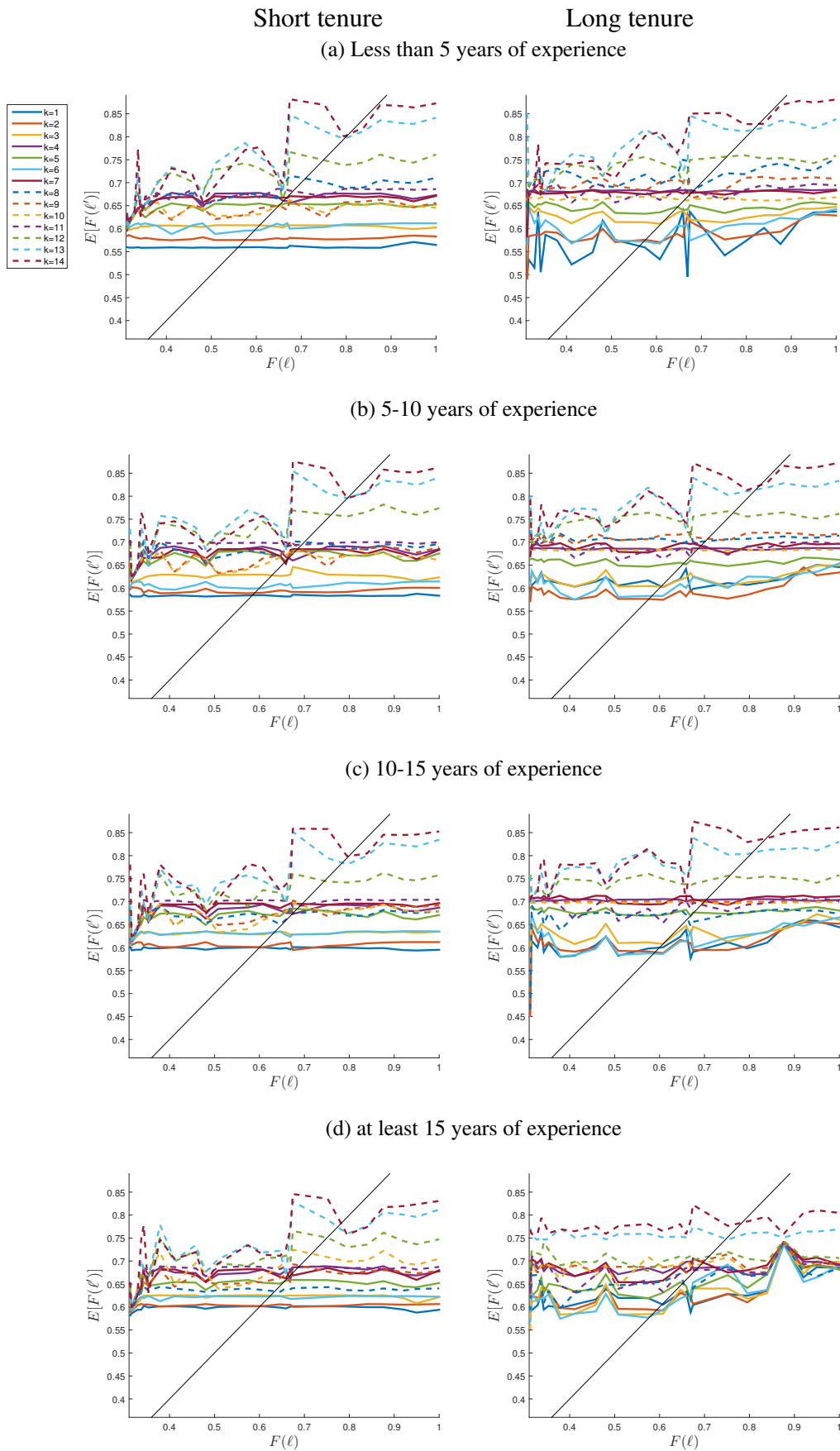


Figure 5: Conditional mobility across firm types (5-year ahead and unconditional on tenure)

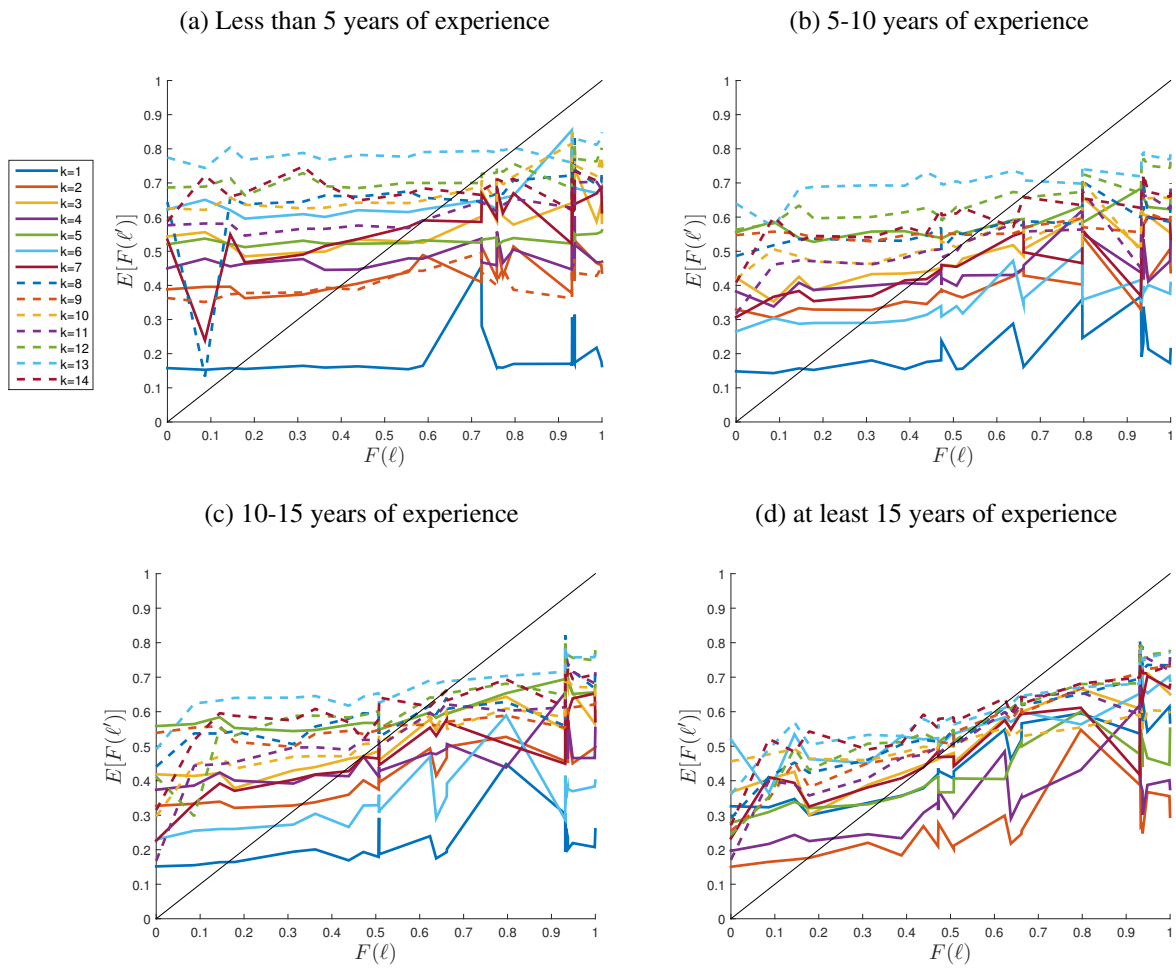


Figure 6: Non-employment and average firm location over the life-cycle, 1999-01 cohort.

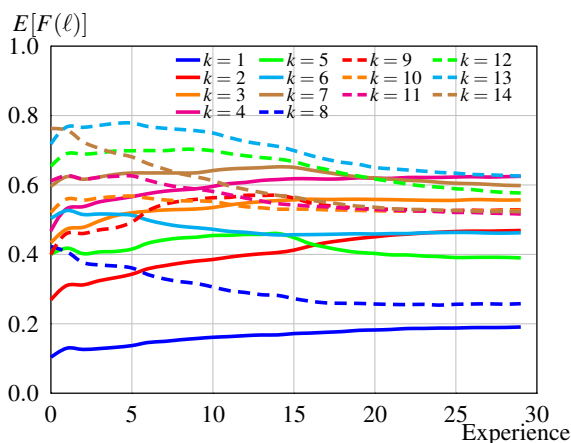
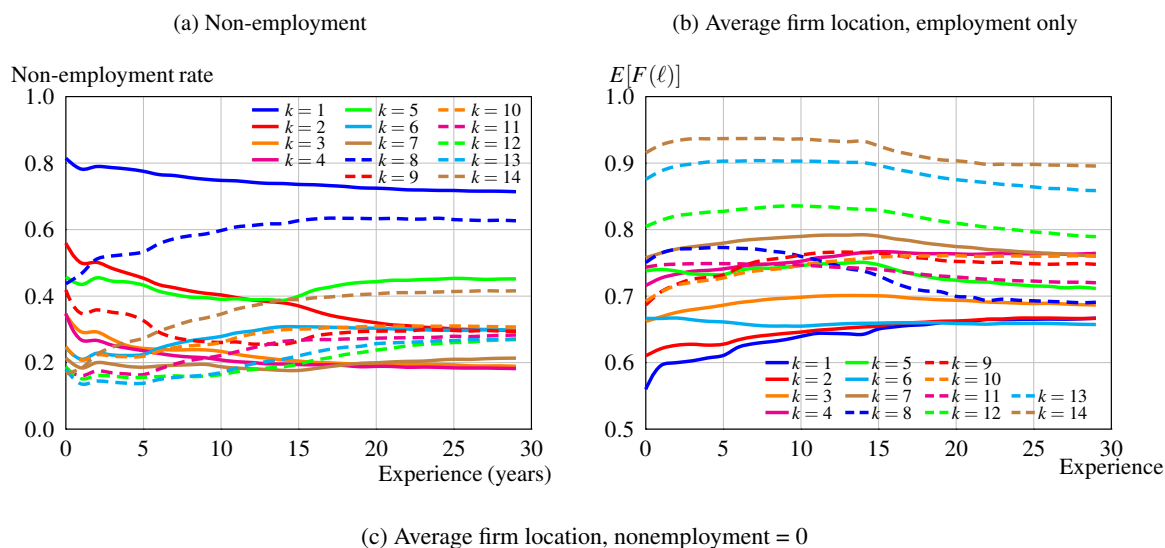
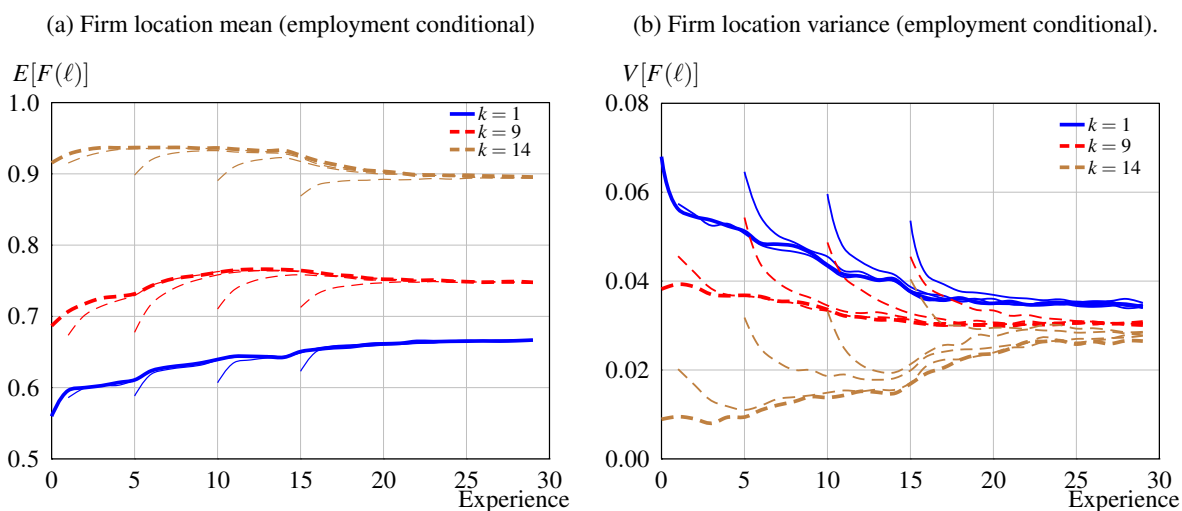


Figure 7: Firm location mean and variance after nonemployment shocks, 1999-01 cohort.



Note: Cohort initialized by the estimated initial distribution in heavy line. In thin lines, cohort initialized to be nonemployed in either week 1, 260, 520, or 780 and is simulated forward according to the estimated mobility model after that. Lines show employment conditional firm location means and variances for given worker types. For exposition purposes, only select types are shown.

workers show some tendency to improve their position on the ladder over time. The ladder position variation in age is however considerably less than that across worker types. Figure 6c includes non-employed workers by assigning a value 0 to $F(\ell)$ in this case. The extensive employment margin tends to add to the picture that ladder position deteriorates in age for high wage workers and something of an opposite trend for low wage workers. This is the mirror image of the non-employment risk patterns previously described.

Finally, we also show the mean and variance of firm location by worker type along the life cycle in Figure 7 subject to initial non-employment. Specifically, a cohort is initialized to be unemployed at one of four different experience levels, 1 week, 260 weeks, 520 weeks, or 780 weeks. The cohort is then simulated forward according to the estimated mobility model. By this, we demonstrate the impact of non-employment at different levels of experience. These results are drawn in thin lines in Figure 7. They are compared to a baseline where the cohort is initialized by the estimated initial distribution.

The first observation of a cohort is one year after the non-employment spell and ladder position is a result both of the match distribution directly out of nonemployment and subsequent mobility within the year. First of all, conditional on employment, it is seen that a cohort recovers ladder position quite quickly. That said, there is a prolonged subsequent period of some 5 years where the cohort regains additional ladder position after the non-employment spell, which demonstrates that job-to-job mobility does result in improved ladder position. The move out of non-employment is associated with initially higher variance in ladder position which declines during the subsequent recovery period and job-to-job moves.

7 Measuring sorting

The preceding section produces a rather depressing view of mobility. There is an extremely weak ladder mechanism. Workers draw a different firm match given their type when they enter the labor market, and essentially stay there. The aim of this section is first to provide a measurement of sorting, i.e. how worker and firm types correlate, and second to understand how it is built.

7.1 How much sorting there is?

To illustrate the estimated sorting, Figure 8 shows the joint distribution of 10-15 years of experience and long tenure workers and firms relative to the matching probability under the assumption of independence, $c[F(k), F(\ell)] \equiv \frac{p(k,\ell)}{p(k)p(\ell)}$ in 1999-01. This is a copula density. There is evidence of positive assortative matching. The entries in the first diagonal of the $c()$ mapping is significantly higher than off-diagonal entries. Low type workers tend to match with low type firms, high type workers with high type firms.

There is significant noise in the relationship, in particular because it is unconditional on

Figure 8: Copula $c[F(k), F(\ell)] \equiv \frac{p(k,\ell)}{p(k)p(\ell)}$ (10-15 years of experience, long tenure, 1999-01)

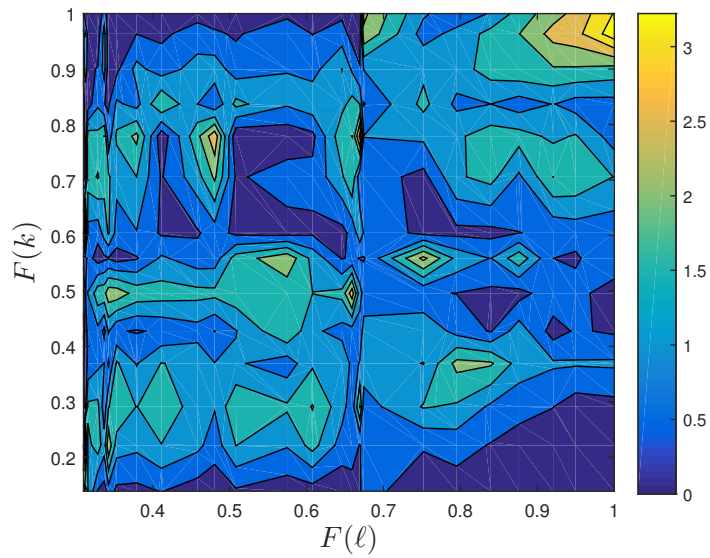


Figure 9: Sorting by tenure and experience

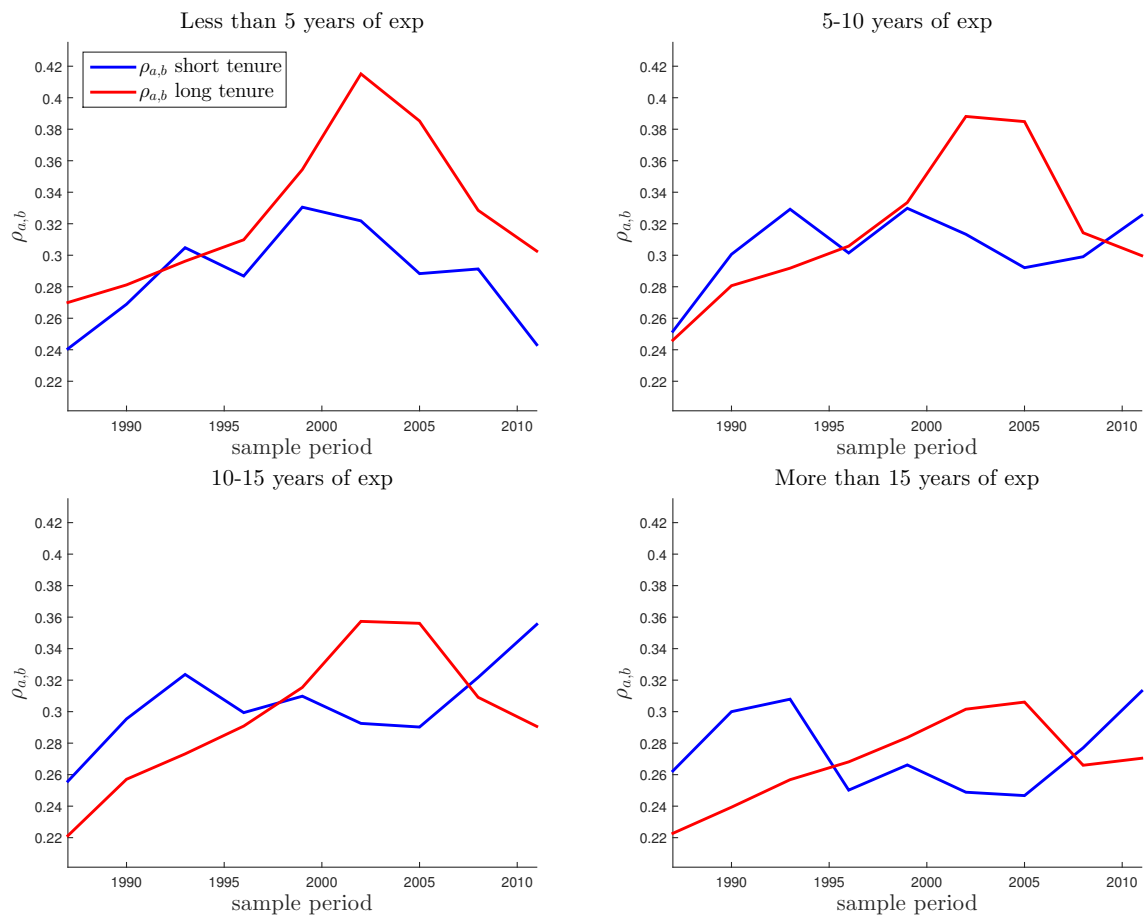
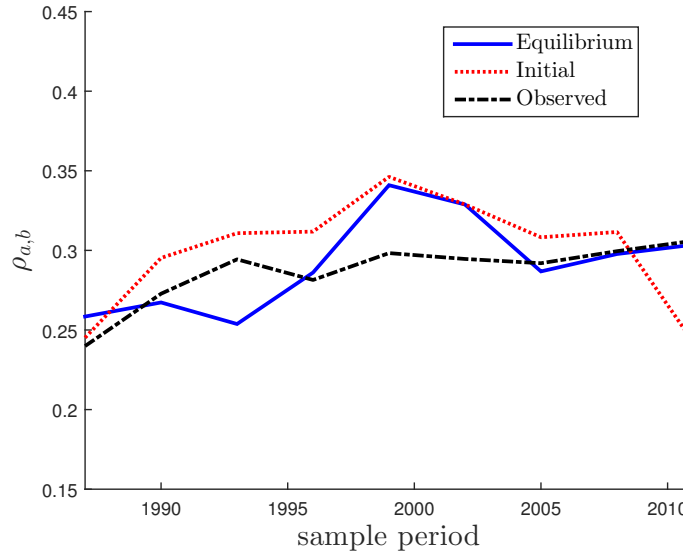


Figure 10: Observed, initial and equilibrium sorting



calendar time, experience and tenure. To quantify the strength of the sorting conditional on worker characteristics, we calculate the correlation coefficient between the cardinal measures of worker and firm types we obtained in the wage projection. We do that in two ways, allowing tenure and experience to interact with worker and firm types, or not (like in the literature).

Figure 9 shows the correlation coefficient between types conditional on time, experience, and tenure. That is, the worker type measure is $a_k(x)$ and the firm type measure is $b_\ell(x)$, obtained by projecting $\mu_{k\ell}(x)$ on k and ℓ separately for each x , each case (k, ℓ) being weighted by

$$p(k, \ell | x) = \frac{\sum_{i,t} p_i(k) \mathbf{1}\{\ell_{it} = \ell, x_{it} = x\}}{\sum_{i,t} \mathbf{1}\{x_{it} = x\}}.$$

As can be seen the correlation coefficients tend to take values between 0.2 and 0.4, confirming the positive sorting between firm and worker types. Furthermore, the positive sorting tends to get stronger from the beginning of the sample period until the early 2000's. The increasing pattern is particularly pronounced for long tenure matches. Also, sorting seems a bit stronger for younger workers, which is consistent with the search ladder being steeper for younger workers.

With the more usual projection obtained without interacting k and ℓ with tenure and experience, we find a fairly modest increase in the correlation coefficient over time from 0.25 to 0.3 (see Figure 10, line “Observed”). These estimates are considerably larger than those in Bagger et al. (2013) using the AKM model (from a low -.07 in 1981 to a high .14 in 2001). Overall, our sorting pattern is thus stronger and more stable. Recently, Borovičková and Shimer (2017) have proposed a different way of calculating this correlation based on mean wage per worker and per firm. Using their estimator, we find a correlation of 47%, much larger than ours. Understanding the sources of discrepancy is beyond the aim of this paper, but it is certain that this estimator measures something different from the correlation between worker and firm linear effects.

Figure 11: Sorting over the life cycle

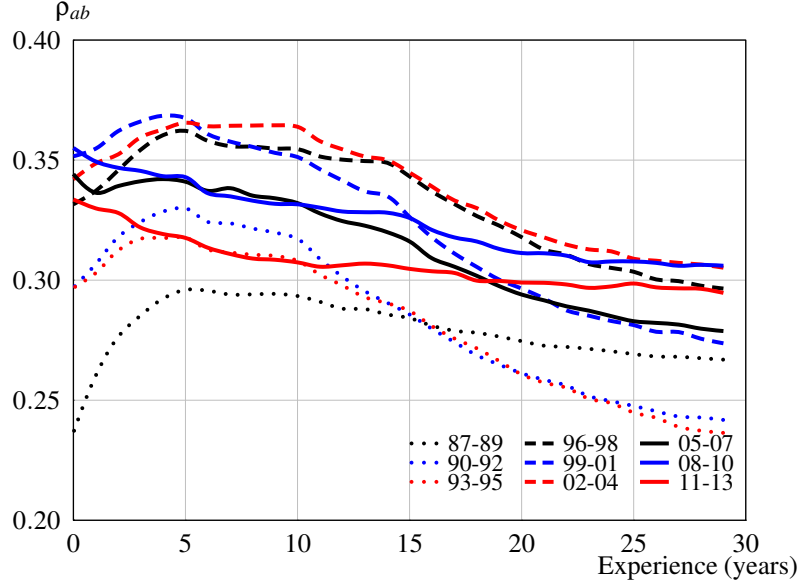


Figure 10 also compares three correlations between a_k and b_ℓ (without interactions with tenure and experience) obtained from three different match distributions: the observed cross-sectional match distribution $p(k, \ell, x)$ calculated for a given year; the stationary (or equilibrium) distribution $p^\infty(k, \ell, x)$ that is obtained from the estimated transition probabilities $M_{k\ell\ell'}(x)$,¹⁶ and lastly, the initial distribution of matches $p^0(k, \ell|x_1) = p(k)m(\ell|k, x_1)$. We find that the correlations calculated with these alternative distributions are close to each other. It may be the initial correlation is bigger than the observed and equilibrium ones. This confirms that mobility is not a mechanism that increases sorting, instead mobility just maintains sorting close to its initial level.

To study the interaction between experience and tenure on mobility patterns, we construct a synthetic cohort for each of our 9 time periods, as explained in the previous section. Figure 11 shows the correlation between worker and firm wage effects among employed workers for the 9 different cohorts by experience. Already at the outset, the cohorts are significantly sorted. It is then a general pattern that during the first 5 to 10 years of a cohort's life, mobility patterns strengthen sorting, and the cohort's maximal sorting level is attained already after about 5 or 10 years. Sorting subsequently weakens as the cohort ages further. Interestingly, this life-cycle pattern seems to have disappeared with the 2008 crisis showing a steady decline in sorting throughout the life of a cohort. Strikingly, for these cohorts maximal sorting is attained already at the outset of the cohort's life.

¹⁶From $M_{k\ell\ell'}(x)$ calculate $m^\infty(\ell|k, x)$ as the ergodic distribution associated with the Markov chain where we also allow tenure status to reset. Then, $p^\infty(k, \ell|x) = p(k)m^\infty(\ell|k, x)$, where $p(k) = \frac{1}{N} \sum_i p_i(k)$. Then average over x .

Table 4: Average counterfactual sorting over time

Benchmark	$\rho_{a,b}$				
	Average	<5 exp	5-10 exp	10-15 exp	15> exp
	0.29	0.31	0.32	0.30	0.27
Counterfactual	% change				
	Average	<5 exp	5-10 exp	10-15 exp	15> exp
No k variation in γ_{kl}	-57.74	-52.05	-59.60	-60.26	-57.03
No k, ℓ variation in γ_{kl}	-64.55	-66.68	-63.70	-63.38	-61.44
No k variation in ψ_{kl}	-39.57	-36.42	-37.61	-38.90	-45.53
No ℓ variation in ψ_{kl}	-44.88	-39.27	-47.87	-44.10	-45.52
No ℓ variation in v_ℓ	-50.79	-39.16	-53.31	-52.05	-51.96
No k variation in λ_k	-3.11	-3.68	-3.48	-3.16	-1.84
No E-E transition ($\lambda_k = 0$)	-33.97	-39.21	-35.68	-32.86	-22.11
No k variation in δ_{kl}	1.74	-5.42	0.37	2.60	5.35
No ℓ variation in δ_{kl}	-1.19	5.70	1.73	-2.09	-2.85
No k variations in δ_{kl} given no k variation in ψ_{kl}	-45.72	-49.64	-45.38	-43.75	-47.57
No k variation in γ_{kl} given no ℓ variation in v_ℓ	-93.59	-84.44	-91.42	-99.96	-92.81
No tenure effect in γ_{kl}	-5.72	-8.29	-4.43	-5.71	-3.69
No tenure effect in λ_k	-0.79	-0.39	-1.14	-1.43	-0.70
No tenure effect in v_ℓ	-1.80	-2.82	-0.68	-0.59	-2.61
No experience effect in γ_{kl}	-4.31	-12.32	-11.03	-2.44	4.81
No experience effect in λ_k	-1.78	-1.50	-2.06	-1.54	-2.09
No experience effect in v_ℓ	-1.15	1.91	-1.73	-2.67	-2.25
No E-E transition ($\lambda_k = 0$)	-33.97	-39.21	-35.68	-32.86	-22.11
No E-U transition ($\delta_{kl} = 0$)	-21.69	0.32	-15.91	-22.80	-18.77

7.2 The anatomy of equilibrium sorting

We finally perform a series of counterfactuals to assess the relative strength of the different mobility channels that may impact sorting in the model. Each counterfactual changes a particular mobility variable by removing one dimension of variation (such as worker-type, firm-type, tenure, experience and time variation) and setting the variable to its weighted average in a given dimension while holding the others parameters constant. We simulate 1,000,000,000 workers to calculate the counterfactual steady state match probability $p_{alt}^\infty(k, \ell)$. We then use $p_{alt}^\infty(k, \ell)$ to re-calculate the correlations between the benchmark cardinal measures of worker a_k and firm b_ℓ types, obtained from the original steady-state distribution.¹⁷ Table 4 displays the percentage change of counterfactual correlations, relative to the benchmark case, for both unconditional and conditional on experience.

¹⁷We also tried with re-estimating a_k and b_ℓ with the counterfactual distribution. Results are similar.

If all workers agree on their preference for firms ($\gamma_{k\ell} = \bar{\gamma}_\ell$) then the only sources of sorting are that better workers search more intensely and climb the common ladder faster, and that layoff rates may vary across worker types. This counterfactual removes all homophilic source of sorting. As can be seen, the impact on sorting is substantial. We observe a similar impact on sorting due to variation across worker types in the firm type destination distribution out of unemployment ($\psi_{k\ell} = \bar{\psi}_\ell$). We see this channel to be a significant source of sorting in the estimated model. It is a notable feature of the estimation since it is a channel that is typically absent in standard random search models.

Removing the heterogeneity in search intensity λ_k has little effect on sorting. However, when we set λ_k to zero — so that workers keep the first jobs they draw out of unemployment until the next layoff — sorting is reduced by almost 40 percent for the youngest group of workers, and to a lesser extent for the older groups.

The removal of heterogeneity in the sampling probability of firm types $v_{\ell'}$ also has a significant effect on sorting, similar in magnitude to the effect obtained by removing heterogeneity in the job finding rates for unemployed. Hence an heterogeneous firm sampling probability does amplify sorting driven by heterogeneous preferences $\gamma_{k\ell}$, but not heterogeneous search intensity. This is likely due to the fact that search intensity does not vary enough across workers.

There is limited impact on sorting from the layoff channel, $\delta_{k\ell} = \bar{\delta}_\ell$ and $\delta_{k\ell} = \bar{\delta}_k$. Even though low-quality matches are more likely to be sent into unemployment than high-quality matches, heterogeneous layoff risk does not significantly alter overall sorting. It is conceivable that the lack of impact on sorting could be a consequence of heterogeneous job finding rates for unemployed workers, $\psi_{k\ell}$. The estimated destination distribution out of unemployment goes a long way to directly place a worker type back to the position implied by the job-to-job mobility. Hence, layoffs could have minor sorting implications in our setting. Therefore, differential layoff differences across worker types would not affect sorting either.

Finally, we assess the extent to which the transition from employment to unemployment impacts sorting by setting $\delta_{k\ell}$ to zero. When workers only move from job to job, sorting rises among young workers, while sorting reduces for workers in other age groups. This is consistent with our finding that sorting is largely built early on in one's career and become less strong as workers age.

8 Decomposing the variance of log-wages and wage growth

In this last section, we measure the contribution of sorting to log-wage inequality and wage growth dispersion.

8.1 Log-wage levels

For a given time period, we project $\mu_{k\ell}(x)$ on all tenure and experience interactions and worker and firm indicators as in equation (14). We then decompose the log-wage variance as follows (omitting the conditioning on calendar time):

$$\begin{aligned} V(w_{it}) &= E[V(w_{it}|k_i, \ell_{it}, x_{it})] + V[E(w_{it}|k_i, \ell_{it}, x_{it})] \\ &= E[\sigma^2(k_i, \ell_{it}, x_{it})] + V[\mu(k_i, \ell_{it}, x_{it})], \end{aligned}$$

and

$$\begin{aligned} V[\mu(k_i, \ell_{it}, x_{it})] &= V[\bar{\mu}(x_{it})] + V[a(k_i)] + V[b(\ell_{it})] \\ &\quad + V[\tilde{\mu}(k_i, \ell_{it}, x_{it})] + 2\text{Cov}[\bar{\mu}(x_{it}), a(k_i)] \\ &\quad + 2\text{Cov}[\bar{\mu}(x_{it}), b(\ell_{it})] + 2\text{Cov}[a(k_i), b(\ell_{it})], \end{aligned}$$

where expectation operators (and variance and covariance) are with respect to distribution $p(k, \ell, x)$. For example,

$$E[\sigma^2(k_i, \ell_{it}, x_{it})] = \frac{\sum_{k, \ell \neq 0, x} p(k, \ell, x) \sigma_{k\ell}^2(x)}{\sum_{k, \ell \neq 0, x} p(k, \ell, x)}.$$

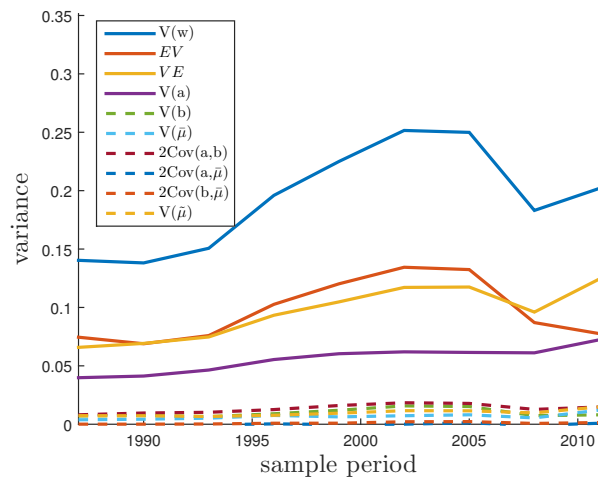
Figure 12 shows the log-wage variance decomposition over time. In panel (a), it is seen that overall log wage variance, $V(w)$, is increasing over time. Panel (b) shows the relative importance of within match wage variance, $E\sigma^2$, and between match wage variance, $V\mu$, respectively. As can be seen, the between match wage variance is rising proportionately over time so that by the end of the sample, in the early 2010s, the within match wage variance accounts for 62.1% of overall wage variance, from 47.6% at the end of the 1980s. The residual variance, the average of the idiosyncratic component $\sigma_{k\ell}(x)$, conversely decreased from around 52.4% to 37.9%.

Note that we estimate considerably more idiosyncratic wage dispersion than is usually the case using the AKM model (Table 5). Along with the increased importance of between match wage variance follows an increase in the variance across worker types $V(a)$. Wage dispersion across worker types is the dominant source of dispersion across matches (around 29% between 1987 and 2007, and 35% after 2007), but its importance as a fraction of between match variation is decreasing over time (from 61.2% to 58.5%). Since we estimate so much more idiosyncratic variance, our estimates of person effect contribution is sizably smaller than in the AKM-based literature.

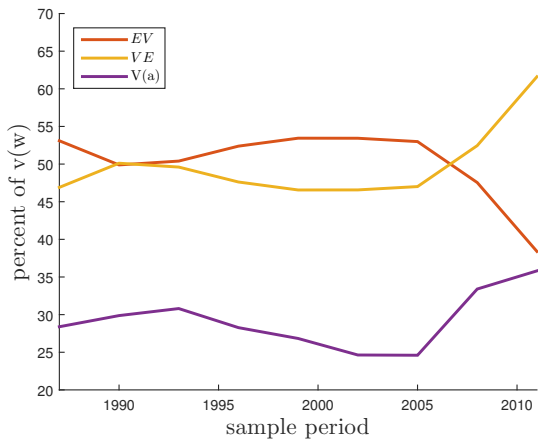
The effect of sorting $2\text{Cov}(a, b)$ comes next, which explains about 7% of the overall log-wage variance. We confirm the increasing trend that has been observed elsewhere (Card et al., 2013; Song et al., 2015), although in a much more subdued way. The firm effect has remained third in contribution after the worker effect and what we conventionally call “sorting” until the

Figure 12: Variance decomposition

(a) Levels



(b) Main components



(c) Firm effect and sorting

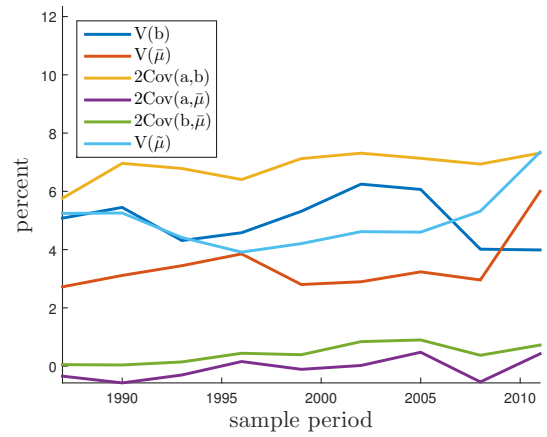


Table 5: Log-wage variance decompositions (percents)

	AKM														BLM		Our estimator		
	FR	US1	AU	US2	IT	BR	DE	US3	DK	SW	87-89	99-01	11-13	average					
	76-87	84-93	90-97	90-99	81-97	95-05	85-91	80-86	87-13	02-04									
Residual	15.8	9.3	5.0	9.3	15.1	7.0	7.7	20.3	39.0	25.2	53.1	53.4	38.4	52.7					
Person effect	76.9	81.6	66.3	63.7	43.9	60.0	61.0	47.5	42.9	60.1	28.4	26.8	35.8	29.2					
Firm effect	30.2	19.2	37.0	15.4	13.1	26.9	18.5	16.0	11.6	2.5	5.1	5.3	4.0	5.2					
Cross effect	-27.2	-2.0	-22.4	0.62	2.1	3.2	2.3	1.6	3.3	12.2	5.8	7.1	7.3	7.1					
Match effect				5.1			2.6				5.2	4.2	7.3	6.4					
Observed	6.8	52.0	3.1	4.0	7.5	3.0	10.7	7.6	1.8		2.7	2.8	6.0	4.0					
heterogeneity	-3.1	-69.0	9.4	0.64	15.5		-2.6		0.88		-0.3	-0.1	0.4	-0.1					
Sorting	0.7	9.0	1.7	1.25	2.6		1.6		0.52		0.1	0.4	0.7	0.41					
	-28.3	-2.5	-22.7	1.0	4.4	4.0	3.4	24.9	7.4	49.1	24.0	29.8	30.6	28.9					

Notes: All decompositions under the label AKM refer to estimations of the AKM model with or without a match-specific component on various matched employer-employee data using the method of Abowd et al. (2002). Columns FR (France) and US1 come from Table 2 in Abowd et al. (2002). Column AU (Austria) comes from Table 4 in Gruetter and Lalive (2009). Column US2 results from Tables 5 and 6 in Woodcock (2015). They use the LEHD data discussed in Abowd et al. (2009). Column IT (Italy) refers to Table 4 in Iranzo et al. (2008). Column BR (Brazil) refers to Table 1 in de Melo (2018). Columns DE (Germany) are calculated from Table III in Card et al. (2013). Columns US3 result from our own calculations from Table 2 in Song et al. (2015). The worker effect is the sum of the “mean worker effect across firms” and the “difference of worker effect from mean worker effect across firms”, and the effect of observed heterogeneity is the sum of the “mean xb across firms” and the “difference of xb from mean xb across firms”. Column DK, under the AKM label, corresponds to our own estimation of the AKM model on entire 10987-2013 sample. Column BLM refers to Bonhomme et al. (2017)’s estimation for Sweden, 2002-04 (Table 2).

mid-2000, explaining about 5% of the log wage variance. We observe a big drop after 2005 benefiting to the match-specific effect $V(\tilde{\mu})$ and the effect of age and tenure $V(\bar{\mu})$. It is difficult to say if this corresponds to a structural change or if it is just some anecdotic phenomenon. Finally, there is evidence that worker and firm heterogeneity impact mean wages in a non additive way. The match-specific effect $V(\tilde{\mu})$ has doubled in size between 1995 and 2011. Its contribution to log wage variance is now comparable to that of the worker effect.

Summing up, we estimate much more residual variance than with the AKM model, and therefore a smaller contribution of worker effects. Sorting comes next and its contribution is increasing over the period. The firm effect and the match effect have similar, smaller contributions. The relative shares of worker, firm, sorting and residual effects are different from the estimates in Bonhomme et al. (2017) using Swedish matched employer-employee data. They estimate a greater correlation between worker and firm types, less residual variance, a much greater contribution of the person effect and a lower contribution of the firm effect. However, note that they estimate their more flexible model (as far as mobility is concerned) on just two years of data, whereas we use 26 years of data to identify worker and firm heterogeneity.

8.2 Monte Carlo simulations

In order to both check the accuracy of our estimation procedure and evaluate the biases of AKM, we ran a Monte Carlo simulation as close as possible from the true data. The simulations involve 1,000,000 workers and 100,000 firms where $K = 14, L = 24$. The length of the panel is 7 years. We use the parameters estimated for a given time period and age class, and we simulation worker trajectories, allowing for tenure to change. We hold age fixed in the simulations so as not to worry about workers' entry into and exit from the labor market. We draw initial conditions from the steady-state distribution.

The results can be found in Table 6. Each horizontal section of the table separated by an horizontal line corresponds to a different simulation with a different time period and a different experience group. Within each horizontal panel, the row labeled "true" displays the actual variance decomposition and correlation between worker and firm effects for employed workers. The row "LPR" shows the variance decomposition and correlation after first estimating the parameters using our model and estimation procedure. It is a test of the capacity of our CEM algorithm to recover the true parameters when the model is well specified. As can be seen, our algorithm works very well. The row labeled "AKM" shows the variance decomposition and correlation that are obtained by estimating an AKM model on the simulated data. The simulations confirm the overfitting tendency of AKM, overestimating the contributions of worker and firm effects, and underestimating the contribution of sorting as well as the residual variance.

As part of the AKM decomposition, we also calculate the correlation between a worker's wage effect and the average of the worker's co-workers' wage fixed effects as proposed in de Melo (2018). As seen, our simulations produce significant correlations between co-workers

Table 6: Simulated log-wage variance decomposition

Date	Experience	Model	Residual variance	Explained variance	Components of explained variance							Corr. btw pe and co-worker pe	
					Person effect	Firm effect	Sorting	Match effect	Obs. hetero.		Correlation btw pe and fe		
								var	cov				
1987-89	<5 yrs	true	68.7	31.4	15.9	5.4	5.3	3.0	1.1	0.2	0.4	28.6	23.6
		LPR	68.5	31.5	16.3	6.0	4.9	2.5	1.3	-0.1	0.5	24.6	
		AKM	57.4	42.6	29.6	8.9	2.2		1.2	0.1	0.5	6.9	
		BS										34.0	
	5-10 yrs	true	60.2	40.0	24.4	5.9	6.8	2.5	0.4	0.1	-0.1	28.5	23.6
		LPR	60.2	39.8	24.1	6.2	6.5	2.7	0.5	-0.1	-0.1	26.5	
		AKM	49.8	50.2	36.9	9.3	3.7		0.4	0.0	-0.1	10.0	
		BS										37.6	
	10-15 yrs	true	49.9	50.2	32.9	6.4	7.2	3.6	0.2	0.1	-0.2	24.7	22.1
		LPR	49.8	50.2	32.6	7.1	7.2	3.1	0.3	0.0	-0.2	23.6	
		AKM	41.0	59.0	45.2	10.2	3.4		0.3	0.1	-0.2	7.9	
		BS										37.4	
>15 yrs	true	40.0	60.0	41.4	5.9	7.3	5.6	0.1	-0.1	-0.2	23.3	21.9	
	LPR	40.4	59.6	41.1	7.1	7.4	4.1	0.1	-0.1	-0.2	21.7		
	AKM	33.2	66.8	54.3	11.3	1.4		0.2	-0.2	-0.3	2.8		
	BS										37.0		
1999-01	<5 yrs	true	60.1	39.9	19.4	6.1	7.4	3.5	1.4	1.6	0.7	34.0	28.8
		LPR	60.6	39.4	18.1	6.6	6.6	4.1	2.0	1.0	1.0	30.0	
		AKM	51.1	48.9	31.4	8.6	5.2		1.3	1.5	0.7	15.8	
		BS										37.5	
	5-10 yrs	true	57.2	42.9	22.6	6.1	7.8	3.3	1.3	1.1	0.6	33.2	28.8
		LPR	57.6	42.4	22.1	6.3	7.2	3.6	1.7	0.8	0.7	30.5	
		AKM	48.4	48.9	31.4	8.6	5.2		1.3	1.5	0.7	15.8	
		BS										37.7	
	10-15 yrs	true	53.5	46.4	27.2	5.9	7.9	3.1	1.2	0.8	0.5	31.0	29.4
		LPR	53.9	46.1	26.9	6.1	7.4	3.1	1.4	0.6	0.5	28.8	
		AKM	44.5	55.5	39.3	9.2	4.6		1.0	0.8	0.5	12.1	
		BS										38.0	
>15 yrs	true	50.0	50.1	31.7	5.4	6.7	3.8	1.4	0.8	0.4	25.7	28.2	
	LPR	50.0	50.0	31.8	5.6	6.5	3.5	1.5	0.6	0.5	24.2		
	AKM	40.9	59.1	45.6	10.6	0.4		1.2	0.9	0.4	1.0		
	BS										35.0		
2011-13	<5 yrs	true	33.9	66.2	43.7	3.1	5.4	8.8	4.1	0.3	0.7	23.0	25.2
		LPR	35.5	64.5	44.5	3.5	4.5	6.8	4.1	0.3	0.8	18.1	
		AKM	31.8	68.2	53.9	6.6	2.5		4.5	-0.5	1.2	6.6	
		BS										33.8	
	5-10 yrs	true	40.5	59.5	39.4	4.1	7.6	5.5	1.8	0.8	0.4	29.7	30.6
		LPR	41.5	58.5	38.8	4.6	7.8	4.4	1.8	0.8	0.3	29.0	
		AKM	33.9	66.1	52.0	8.0	3.1		2.3	0.3	0.3	7.6	
		BS										39.1	
	10-15 yrs	true	41.8	58.3	36.5	6.2	8.4	5.4	1.3	0.2	0.3	28.0	32.3
		LPR	42.7	57.3	36.5	6.7	8.3	4.1	1.4	0.0	0.3	26.7	
		AKM	34.6	65.4	51.3	10.6	1.8		1.6	-0.1	0.2	3.8	
		BS										42.7	
>15 yrs	true	42.9	57.1	37.2	5.9	8.6	4.1	1.2	-0.2	0.3	29.1	33.5	
	LPR	43.3	56.7	37.3	6.1	8.6	3.3	1.3	-0.3	0.4	28.7		
	AKM	34.8	65.2	51.6	11.6	0.6		1.3	-0.2	0.3	1.3		
	BS										43.9		

Notes: The simulations involve 1,000,000 workers and 100,000 firms where $K = 14, L = 24$. The length of the panel is 7 years. The simulation is initialized in steady state. “Model = true” is the infeasible variance decomposition as if heterogeneity were observed. “Model = AKM” refers to (therefore misspecified) AKM estimates. “Model = LPR” refers to our model. “Model = BS” refers to Borovičková and Shimer (2017)’s calculation of the degree of sorting. We show the mean simulation. With such a big sample size, there is very little variance across simulations.

as they sort on firm types. As a reference, the co-worker correlation is 0.38 when AKM is performed on our full data set.

Lastly, the row labeled “BS” shows Borovičková and Shimer (2017)’s calculation of the degree of sorting. The simulations confirm the tendency of the correlation between mean wage per worker and mean wage per firm (with all the adjustments to the naive formula proposed by BS) to overestimate (sometimes quite significantly) the correlation between worker and firm linear effects. This statistic captures more than just the correlation between linear projections of log wages on worker and firm dummies. The role of the match-specific effect, for example, must be better understood.

9 Conclusion

In this paper we use the finite mixture framework of Bonhomme et al. (2017) to estimate a model of wages and employment mobility on Danish panel of matched employer-employee data over the period 1987-2013. Our model allows for structural changes in the parameters and we propose a new parametrization for state transition probabilities. We develop a Classification Expectation Maximization algorithm allowing to estimate a random component model for $K = 14$ worker types and a classification of firms into $L = 24$ discrete groups. Our estimation algorithm works well and is fast, despite the nonlinear specification of state transition probabilities.

Our estimates of unobserved worker and firm heterogeneity and structural parameters shows an apparent disagreement in the way unobserved heterogeneity determines conditional mean wages, on one hand, and, on the other hand, the idiosyncratic wage variance and the mobility parameters. The strongest link is estimated for layoff rates and job finding probabilities for unemployed workers. The parameters governing the way a worker of a given type values job types display a much weaker link, and only at low tenure.

The joint distribution of match types shows evidence of moderate sorting. We measure sorting as the correlation of the worker and firm components obtained from a linear projection of conditional mean wages on worker and firm type dummies. Our estimates are around 28% and very stable over the all period. We find that this correlation level is obtained at the first draw of a match in workers’ careers. The moderate relationship between conditional mean wages and mobility parameters is only sufficient to maintain sorting at its initial level. There is no evidence of a strong ladder effect. The main parameters here are the preference for the job, the sampling distribution of firm types and the job finding rates of unemployed.

Finally, we estimate the log-wage decomposition. We estimate much more residual variance than with the AKM model, and therefore a smaller contribution of worker effects. Sorting comes next and its contribution is increasing over the period. The firm effect and the match effect have similar, smaller contributions. The relative shares of worker, firm and sorting effects are consistent with the estimates in Bonhomme et al. (2017) using Swedish matched employer-

employee data.

What all this means is that the more precise description of wage and mobility that BLM's approach facilitates does not fundamentally change the estimation of sorting that the AKM model delivers. A 28% correlation is certainly not large. We could understand such a small correlation for young workers, but most search-matching models would predict that a stronger correlation is built after a while. Our estimates of the preference for a job (determining which of the current job and an alternative one a worker selects) are very weakly correlated with mean wages (across match types), in particular for older workers. This implies that workers tend to change employment for reasons that relate very little to productivity. Amenities, following one's spouse, skill obsolescence may explain mobility at older age better than wage prospects.

However, it may also be that some of the basic assumptions of our model, which we maintained from the AKM model, are not satisfied by the data. For example, wages and mobility may depend on current and past wages conditional on worker and firm heterogeneity (endogenous mobility). BLM consider an extension with autoregressive wages, and find it too weak to make a difference on the measurement of sorting. Or, it could be that worker and firm types are not fixed over time. Developing a Hidden Markov model with time-varying worker and firm types is an interesting project for future research.

In any case, one conclusion seems clear to us. It is certainly important to understand how worker and firm heterogeneity determine wages, but it is at least as important to understand how worker and firm heterogeneity determine employment mobility. The finite mixture approach of Bonhomme, Lamadon and Manresa offers a tractable way to do that, and to go beyond the wage model of Abowd, Kramarz and Margolis. We are just beginning to mine this seam.

APPENDIX

A Non-parametric identification

The basic linear algebraic tool for identification is the following lemma. Let there be two $m \times n$ matrices A_0 and A_1 such that

$$A_0 = G\Delta_0F^\top, \quad A_1 = G\Delta_1F^\top,$$

for two matrices G, F of dimensions $m \times k$ and $n \times k$, and two diagonal $k \times k$ matrices Δ_0, Δ_1 . Suppose that $\text{rank}(A_0) = k \leq \min(m, n)$. This implies that G and F have linearly independent columns and that Δ_0 is invertible. Let the Singular Value Decomposition (SVD) of A_0 be $A_0 = U\Lambda V^\top$, for two orthogonal matrices U, V of dimensions $m \times k$ and $n \times k$, and an invertible $k \times k$ diagonal matrix Λ .¹⁸

Two interesting results follow. Firstly,

$$U^\top A_0 V \Lambda^{-1} = U^\top G \Delta_0 F^\top V \Lambda^{-1} = I_k \quad (\text{identity matrix}).$$

This implies that if we let $W = U^\top G$, then $W^{-1} = \Delta_0 F^\top V \Lambda^{-1}$. Secondly,

$$U^\top A_1 V \Lambda^{-1} = U^\top G \Delta_1 F^\top V \Lambda^{-1} = W \Delta_1 \Delta_0^{-1} W^{-1}.$$

This shows that $\Delta_1 \Delta_0^{-1}$ is the diagonal matrix of the eigenvalues of $U^\top A_1 V \Lambda^{-1}$, and that W is one particular matrix of eigenvectors. If there are no multiple eigenvalues (all associated eigenspaces have dimension one), W is unique up to a multiplication of its columns by non zero numbers. If there are several such A_1 matrices, the second step's diagonalization becomes a simultaneous diagonalization problem.

A.1 Identification of remaining parameters with observed firm types

We first consider the case where the firm type is observed.

A.1.1 One instrument, one wage, one job mobility

Assume that for each worker, we observe a set of characteristics z (gender, education, etc), the employer's type ℓ and the current wage w , as well as employment mobility at the end of the period (either no mobility, or mobility to a firm of type ℓ'). The worker type k is not observed. For simplicity, we neglect the control x and we assume that all variables are discrete.

¹⁸The standard SVD has $A_0 = U\Lambda V^\top$ where U is $m \times m$ orthogonal, Λ is rectangular diagonal $m \times n$ and V is $n \times n$ orthogonal. If $\text{rank}(A_0) = k$, then we extract from U and V the first k columns, and from Λ the first k rows and columns (assuming this block contains the non zero singular values). Orthogonality of U and V means that $U^\top U = V^\top V = I_k$ (the identity matrix of size k).

Identifying matrices

We start by building identifying matrices using the following assumption specializing Assumptions 1 and 2 to the current case.

Assumption 3. 1) $\mathbb{P}(\ell|z, k) = \mathbb{P}(\ell|k)$; 2) $\mathbb{P}(w|z, k, \ell) = \mathbb{P}(w|k, \ell)$; 3) $\mathbb{P}(\ell'|z, k, \ell, w) = \mathbb{P}(\ell'|k, \ell)$.

The first condition states that z helps measuring k but does not predict the employer's type ℓ given k . The second condition states the same for the wage w given the match type (k, ℓ) and z . The third condition says that job-to-job mobility only depends on the current match type (k, ℓ) and not on z or the current wage w .

Under Assumption 3,

$$\mathbb{P}(k, z, \ell, w) = \mathbb{P}(k) \mathbb{P}(z|k) \mathbb{P}(\ell|z, k) \mathbb{P}(w|z, k, \ell) = \mathbb{P}(k) \mathbb{P}(z|k) \mathbb{P}(\ell|k) \mathbb{P}(w|k, \ell).$$

Then summing $\mathbb{P}(k, z, \ell, w)$ over the unobserved worker type k ,

$$\mathbb{P}(z, \ell, w) = \sum_k \mathbb{P}(z|k) \mathbb{P}(k, \ell) \mathbb{P}(w|k, \ell), \quad (15)$$

where $\mathbb{P}(k, \ell) = \mathbb{P}(k) \mathbb{P}(\ell|k)$. Define the matrices

$$Q_\ell = [\mathbb{P}(z, \ell, w)]_{z \times w}, \quad G = [\mathbb{P}(z|k)]_{z \times k}, \quad F_\ell = [\mathbb{P}(w|k, \ell)]_{w \times k}, \quad \Delta_\ell = \text{diag}[\mathbb{P}(k, \ell)],$$

where the subscript indicates which variable indexes rows and which variable indexes columns.

It follows from (15) that $Q_\ell = G \Delta_\ell F_\ell^\top$.

Next, consider the probability of z, w, ℓ and of moving to a firm of any type ℓ' at the end of the period:

$$\mathbb{P}(z, \ell, w, \ell') = \sum_k \mathbb{P}(z|k) \mathbb{P}(k, \ell) \mathbb{P}(w|k, \ell) \mathbb{P}(\ell'|k, \ell), \quad (16)$$

where the probability of moving to a new job of type ℓ' is $\mathbb{P}(\ell'|z, k, \ell, w) = \mathbb{P}(\ell'|k, \ell)$ by Assumption 3.

We can similarly define the matrices

$$Q_{\ell\ell'} = [\mathbb{P}(z, \ell, w, \ell')]_{z \times w}, \quad \Delta_{\ell\ell'} = \text{diag}[\mathbb{P}(k, \ell, \ell')],$$

with $\mathbb{P}(k, \ell, \ell') = \mathbb{P}(k, \ell) \mathbb{P}(\ell'|k, \ell)$. Equation (16) implies that $Q_{\ell\ell'} = G \Delta_{\ell\ell'} F_{\ell'}^\top$.

Identification

We can then apply the lemma to our setup under the following assumption.

Assumption 4. 1) All matches form with positive probability: $\mathbb{P}(k, \ell) \neq 0$ for all (k, ℓ) . 2) $\mathbb{P}(z|k)$ and $\mathbb{P}(w|k, \ell)$ are linearly independent with respect to k given ℓ . 3) No two groups of workers k

and k' have the same transition probabilities: $\mathbb{P}(\ell'|k, \ell) \neq \mathbb{P}(\ell'|k', \ell)$ for some (ℓ, ℓ') . 4) Workers of different type k have different distributions $\mathbb{P}(z|k)$ of observed characteristics z .

Condition 1 implies that matrix Δ_ℓ is non singular. It is a stronger assumption than necessary which is made to simplify the argument. We could allow for different sets of matching k 's for each ℓ . Condition 2 guarantees that matrices G and F_ℓ are full-column rank. Both conditions together guaranty that the matrices \mathbb{P}_ℓ are full column rank.

We can then apply the identification lemma with $A_0 = Q_\ell$ and $A_1 = Q_{\ell 1}, \dots, Q_{\ell L}$, separately for each initial employer type ℓ . Let $Q_\ell = U_\ell \Lambda_\ell V_\ell^\top$ be the SVD of Q_ℓ . The lemma shows that the matrices $U_\ell^\top Q_{\ell \ell'} V_\ell \Lambda_\ell^{-1}$, $\ell' = 1, \dots, L$, can be simultaneously diagonalized. The matrices of eigenvalues are $\Delta_{\ell \ell'} \Delta_\ell^{-1} = \text{diag}[\mathbb{P}(\ell'|k, \ell)]$, which identifies transition probabilities. Under Condition 3 of Assumption 4 the entries of $\Delta_{\ell \ell'} \Delta_\ell^{-1}$ are all distinct. The common matrix of eigenvectors W_ℓ is such that

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

for some non singular diagonal matrix D_ℓ . Hence, $G D_\ell$ is identified, and as the rows of G sum to one because $\mathbb{P}(z|k)$ is a probability, it follows that the matrix D_ℓ is identified from the sum of the rows of $U_\ell W_\ell$. In the same way, $\Delta_\ell D_\ell^{-1}$ is also identified, hence Δ_ℓ , that is $\mathbb{P}(k, \ell)$ for all k . The identification of G and F_ℓ immediately follows.

Finally, note that we have proceeded separately for workers employed in different sectors ℓ . This means that the classification of workers is contingent on the type of the first period's employer. One can identify worker groups across employer types ℓ if the distributions of observed types z vary across worker groups k (Assumption 4, Condition 4).

Wrapping up, if the distributions of observed worker characteristics and wages are sufficiently "rich" to differentiate worker types, and if the worker classification is minimal in the sense that no two groups are observationally identical and no group is degenerate (probability zero), then one wage observation and one job mobility are enough to identify the model when firm types are observed. If there are too few observed characteristics, say only gender, then the matrix G will not be full column rank for $K > 2$ and we shall need more information, such as two wage observations.

A.1.2 One job mobility, two wages

If there are no or insufficiently many observed worker characteristics, let us consider two consecutive periods of time, in which we observe the worker's wages and both employers' types. We now offer a simple proof of identification for this case that differs from BLM's in that we use both stayers and movers, and we consider the possibility of a mobility in the same job class (from ℓ to ℓ).¹⁹

¹⁹BLM's proof only uses job-to-job transitions and relies on the existence of some special, "alternating" job paths.

Identifying matrices

We can proceed as in the previous subsection and build identifying matrices under a similar specialization of Assumptions 1 and 2.

Assumption 5. 1) $\mathbb{P}(\ell|z, k) = \mathbb{P}(\ell|k)$; 2) $\mathbb{P}(w|z, k, \ell) = \mathbb{P}(w|k, \ell)$; 3) $\mathbb{P}(\ell'|z, k, \ell, w) = \mathbb{P}(\ell'|k, \ell)$; 4) $\mathbb{P}(w'|z, k, \ell, w, \ell') = \mathbb{P}(w'|k, \ell')$.

Assumption 5 adds to Assumption 3 the Condition 4, which is a conditional independence assumption stating that wages only depend on the current match type.

First, consider the probability of wages w and w' in periods 1 and 2 in the same firm of type ℓ (no job mobility is denoted as \neg):

$$\mathbb{P}(\ell, w, \neg, w') = \sum_k \mathbb{P}(k, \ell) \mathbb{P}(w|k, \ell) \mathbb{P}(\neg|k, \ell) \mathbb{P}(w'|k, \ell),$$

where $\mathbb{P}(\neg|k, \ell) = 1 - \sum_{\ell'} \mathbb{P}(\ell'|k, \ell)$ is the probability of staying with the same employer given (k, ℓ) . Let

$$F_\ell = [\mathbb{P}(w|k, \ell)]_{w \times k}, \quad \bar{\Delta}_\ell = \text{diag}[\mathbb{P}(k, \ell, \neg)],$$

with $\mathbb{P}(k, \ell, \neg) = \mathbb{P}(k, \ell) \mathbb{P}(\neg|k, \ell)$. We have $\bar{\mathbb{P}}_\ell = [\mathbb{P}(\ell, w, \neg, w')]_{w \times w'} = F_\ell \bar{\Delta}_\ell F_\ell^\top$.

Second, consider the probability of wages w and w' in periods 1 and 2 with a job mobility from ℓ to ℓ' :

$$\mathbb{P}(\ell, w, \ell', w') = \sum_k \mathbb{P}(k, \ell) \mathbb{P}(w|k, \ell) \mathbb{P}(\ell'|k, \ell) \mathbb{P}(w'|k, \ell').$$

Denoting $\Delta_{\ell\ell'} = \text{diag}[\mathbb{P}(k, \ell, \ell')]$, with $\mathbb{P}(k, \ell, \ell') = \mathbb{P}(\ell'|k, \ell) \mathbb{P}(k, \ell)$, we have

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

Identification

Now let us update Assumption 4 as follows.

Assumption 6. 1) $\mathbb{P}(k, \ell) \neq 0$ for all (k, ℓ) . 2) $\mathbb{P}(w|k, \ell)$ is linearly independent with respect to k given ℓ . 3) For all (k, ℓ) , no mobility is always possible: $\mathbb{P}(\neg|k, \ell) \neq 0$. 4) For all (k, k') there exists ℓ such that $\frac{\mathbb{P}(\ell|k, \ell)}{\mathbb{P}(\neg|k, \ell)} \neq \frac{\mathbb{P}(\ell|k', \ell)}{\mathbb{P}(\neg|k', \ell)}$.

Under Conditions 1 and 3, $\bar{\Delta}_\ell$ is non singular for all ℓ . Under Condition 2, F_ℓ is full-column rank. For the SVD $\bar{\mathbb{P}}_\ell = U_\ell \Lambda_\ell U_\ell^\top$, let $W_\ell = \Lambda_\ell^{-1/2} U_\ell F_\ell \bar{\Delta}_\ell^{-1/2}$. Then

$$W_\ell^{-1} = \bar{\Delta}_\ell^{-1/2} F_\ell^\top U_\ell^\top \Lambda_\ell^{-1/2}.$$

It also holds that

$$\Lambda_\ell^{-1/2} U_\ell \mathbb{P}_{\ell\ell} U_\ell^\top \Lambda_\ell^{-1/2} = W_\ell \bar{\Delta}_\ell^{-1} \Delta_{\ell\ell} W_\ell^{-1}.$$

Assuming that the entries of $\bar{\Delta}_\ell^{-1} \Delta_{\ell\ell} = \text{diag}[\mathbb{P}(\ell|k, \ell)/\mathbb{P}(-|k, \ell)]$ are all distinct (Condition 4), any matrix of eigenvectors is thus of the form

$$W_\ell = \Lambda_\ell^{-1/2} U_\ell F_\ell \bar{\Delta}_\ell^{-1/2} D_\ell, \quad W_\ell^{-1} = D_\ell^{-1} \bar{\Delta}_\ell^{1/2} F_\ell^\top U_\ell^\top \Lambda_\ell^{-1/2},$$

for a non singular diagonal matrix D_ℓ . Since the rows of F_ℓ sum to one, then $\bar{\Delta}_\ell^{-1/2} D_\ell$ and $\bar{\Delta}_\ell^{1/2} D_\ell^{-1}$ are both identified. Hence $\bar{\Delta}_\ell$ is identified, and so are D_ℓ and F_ℓ .

Now, 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 $Q_{\ell\ell'}$ corresponding to a job mobility from ℓ to ℓ' . We know that $Q_{\ell\ell'} = F_\ell \Delta_{\ell\ell'} F_{\ell'}^\top$ with $\Delta_{\ell\ell'}$ 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 $F_{\ell'}^+ Q_{\ell\ell'} F_{\ell'}^+ = \Delta_{\ell\ell'}$ is a diagonal matrix (denoting $A^+ = (A^\top A)^{-1} A^\top$ for any full column rank matrix A).

Finally $\text{diag}[\mathbb{P}(k, \ell)] = \bar{\Delta}_\ell + \sum_{\ell'} \Delta_{\ell\ell'}$ is also identified.

A.2 Identification of firm types from a cross-section of wages

Suppose that for each firm j we observe a set of characteristics r_j and two independent wages w_{j1}, w_{j2} . Moreover, suppose that for firms as for workers wages are independent of independent of observed types r_j given ℓ_j . 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.

More precisely let us make the following assumption. Let q denote probability distributions across firms.

Assumption 7. 1) Within a given firm of type (ℓ, r) , the distribution of wages given (ℓ, r) is independent of r . 2) Wages are mutually independent given ℓ . 3) The distribution of ℓ is not degenerate ($q(\ell) \neq 0$ for all ℓ).

Under this assumption, we can write the probability of (w_1, w_2) as

$$q(w_1, w_2) = \sum_{\ell} q(\ell) q(w_1|\ell) q(w_2|\ell)$$

and the probability of (w_1, w_2, r) as

$$q(w_1, w_2, r) = \sum_{\ell} q(\ell, r) q(w_1|\ell) q(w_2|\ell).$$

Per se r does not identify the firm classification. For example, suppose that we have manufacturing and services and public and private firms. We can classify firms into two groups by clustering them by their public/private status, or by industry. However, observable firm

heterogeneity helps building matrices of moments with a common algebraic structure. Specifically, $\mathbb{P}_r = [q(w_1, w_2, r)]$, w_1 in rows and w_2 in columns, can be factored as $\mathbb{P}_r = F\Delta_r F^\top$, for $F = [q(w|\ell)]$ and $\Delta_r = \text{diag}[q(\ell, r)]$. Because $q(\ell) = \sum_r q(\ell, r) \neq 0$ by assumption, the identification lemma identifies $\Delta_r \Delta_{r_0}^{-1} = \text{diag}[q(r|\ell)]$. Finally, using a similar argument as above, the matrix of eigenvectors, together with the fact that $q(w|\ell)$ is a probability and sums to one, identifies both conditional wage distributions F and type probabilities $q(\ell)$.

If we do not observe firm characteristics, a third wage observation is necessary to identify $q(\ell)$ and $q(w|\ell)$. And if we observe many wages per firm, then the posterior probability of firm type will precisely estimate the unobserved firm type. In this paper, we thus follow BLM's idea of classifying firms into distinct groups, using the employees' wages and other observed firm characteristics, while leaving worker types random. However, we shall iterate an algorithm that estimates the model first given a current firm classification, and then proceeds to an improved firm re-classification.

B The CEM principle

Given the statistical model which generates a set X of observed data, a set of two unobserved latent data or missing values E, F , and a vector of unknown parameters β , along with a likelihood function $\mathbb{P}(X, E, F|\beta)$, the maximum likelihood estimate (MLE) of the unknown parameters is determined by the marginal likelihood of the observed data

$$\mathbb{P}(X|\beta) = \sum_{E, F} \mathbb{P}(X, E, F|\beta).$$

In our application E and F are two discrete variables: E is the set of worker types and F is the set of employer types (E for *employee* and F for *firm*). However, because different workers can be in the same firm at different periods of time, the summation over F can be extremely costly to compute and not even feasible. So instead we treat F as a parameter (a set of firm fixed effects) and we seek to maximize the conditional likelihood

$$\mathbb{P}(X|F, \beta) = \sum_E \mathbb{P}(X, E|F, \beta),$$

with respect to β and F . For this we use the following Classification Expectation Maximization (CEM) algorithm (not quite like the one in Celeux and Govaert, 1992, but related).

For a given value $F^{(s)}$, we run a standard EM algorithm treating $F^{(s)}$ as data. The EM algorithm consists of applying the following two steps iteratively:

Expectation step (E step): Calculate the expected value of the log likelihood function, with respect to the conditional distribution of E given $X, F^{(s)}$ under the current estimate of the pa-

parameters $\beta^{(m)}$:

$$Q(\beta|\beta^{(m)}, X, F^{(s)}) = \sum_E \mathbb{P}(E|X, F^{(s)}, \beta^{(m)}) \ln \mathbb{P}(X, E|\beta, F^{(s)}),$$

where

$$\mathbb{P}(E|X, F, \beta) = \frac{\mathbb{P}(X, E|F, \beta)}{\sum_E \mathbb{P}(X, E|F, \beta)}.$$

Maximization (M step): Find

$$\beta^{(m+1)} = \arg \max_{\beta} Q(\beta|\beta^{(m)}, X, F^{(s)}).$$

The EM algorithm delivers a sequence of parameter values that increases the likelihood

$$\mathbb{P}(X|F^{(s)}, \beta) = \sum_E \mathbb{P}(X, E|F^{(s)}, \beta).$$

To see this, write the log likelihood of the data by,

$$\ln \mathbb{P}(X|F, \beta) = \ln \mathbb{P}(X, E|F, \beta) - \ln \mathbb{P}(E|X, F, \beta).$$

Then multiplying both sides by $\mathbb{P}(E|X, F^{(s)}, \beta^{(m)})$ and summing over E ,

$$\begin{aligned} \ln \mathbb{P}(X|F^{(s)}, \beta) &= \sum_E \mathbb{P}(E|X, F^{(s)}, \beta^{(m)}) \ln \mathbb{P}(X, E|F^{(s)}, \beta) \\ &\quad - \sum_E \mathbb{P}(E|X, F^{(s)}, \beta^{(m)}) \ln \mathbb{P}(E|X, F^{(s)}, \beta) \\ &= Q(\beta|\beta^{(m)}, X, F^{(s)}) + H(\beta|\beta^{(m)}, X, F^{(s)}). \end{aligned}$$

By Gibbs' inequality $H(\beta|\beta^{(m)}, X, F) \geq H(\beta^{(m)}|\beta^{(m)}, X, F)$. Hence,

$$Q(\beta|\beta^{(m)}, X, F^{(s)}) + H(\beta^{(m)}|\beta^{(m)}, X, F^{(s)})$$

is a minorization of $\ln \mathbb{P}(X|F^{(s)}, \beta)$ in the point $\beta^{(m)}$, and improving $Q(\beta|\beta^{(m)}, X, F^{(s)})$ over $Q(\beta^{(m)}|\beta^{(m)}, X, F^{(s)})$ improves $\mathbb{P}(X, F^{(s)}|\beta)$ relative to $\mathbb{P}(X, F^{(s)}|\beta^{(m)})$.

This is well known. Let $\hat{\beta}^{(s)}$ denote the estimate obtained from $F^{(s)}$. One can let the EM algorithm increase the likelihood $L(\beta, F^{(s)}|X)$ until convergence or stop after any number of iterations. We then update $F^{(s)}$ as follows.

Classification step (C step): Find

$$F^{(s+1)} = \arg \max \sum_E \mathbb{P}(E|X, F^{(s)}, \hat{\beta}^{(s)}) \ln \mathbb{P}(X, E|F, \hat{\beta}^{(s)}).$$

By the same argument as before

$$\ln \mathbb{P}(X|F, \beta) = \sum_E \mathbb{P}(E|X, F^{(s)}, \beta) \ln \mathbb{P}(X, E|F, \beta) - \sum_E \mathbb{P}(E|X, F^{(s)}, \beta) \ln \mathbb{P}(E|X, F, \beta).$$

Therefore, improving $\sum_E \mathbb{P}(E|X, F^{(s)}, \widehat{\beta}^{(s)}) \ln \mathbb{P}(X, E|F, \widehat{\beta}^{(s)})$ delivers $F^{(s+1)}$ such that

$$\mathbb{P}(X|F^{(s+1)}, \widehat{\beta}^{(s)}) > \mathbb{P}(X|F^{(s)}, \widehat{\beta}^{(s)}) > \mathbb{P}(X|F^{(s)}, \widehat{\beta}^{(s-1)}).$$

The sequence $(\beta^{(s)}, F^{(s)})$ delivers an increasing sequence of $(\mathbb{P}(X|F^{(s)}, \beta^{(s)}))$ that converges to a local maximum of $\mathbb{P}(X|F, \beta)$. This CEM algorithm is essentially like a sequential EM algorithm where β is updated given F and then F given β .

C Parametric identification of transition probabilities

We show that the parameters $\lambda_{k\ell}, \mathbf{v}_{k\ell'}, \gamma_{k\ell}$ are not identified given $M_{k\ell\ell'} = \lambda_{k\ell} \mathbf{v}_{k\ell'} \frac{\gamma_{k\ell'}}{\gamma_{k\ell} + \gamma_{k\ell'}}$ for $\ell, \ell' = 1, \dots, K$. Let us omit the index k to simplify the notation.

We have $M_{\ell\ell} = \frac{1}{2} \lambda_{\ell} \mathbf{v}_{\ell}$, which identifies λ_{ℓ} given \mathbf{v}_{ℓ} . Moreover,

$$\frac{M_{\ell\ell'}}{2M_{\ell\ell}} = \frac{\mathbf{v}_{\ell'} \gamma_{\ell'}}{\mathbf{v}_{\ell} \gamma_{\ell} + \gamma_{\ell'}}, \quad \frac{M_{\ell'\ell}}{2M_{\ell'\ell'}} = \frac{\mathbf{v}_{\ell} \gamma_{\ell}}{\mathbf{v}_{\ell'} \gamma_{\ell} + \gamma_{\ell'}},$$

which identifies the parameter ratios $\frac{\gamma_{\ell}}{\gamma_{\ell'}}$ and $\frac{\mathbf{v}_{\ell}}{\mathbf{v}_{\ell'}}$ from the two ratios of transition probabilities $\frac{M_{\ell\ell'}}{2M_{\ell\ell}}$ and $\frac{M_{\ell'\ell}}{2M_{\ell'\ell'}}$. Note that

$$r_{\ell\ell'} \equiv \frac{1 - \sqrt{1 - \frac{M_{\ell\ell'} M_{\ell'\ell}}{M_{\ell\ell} M_{\ell'\ell'}}}}{1 + \sqrt{1 - \frac{M_{\ell\ell'} M_{\ell'\ell}}{M_{\ell\ell} M_{\ell'\ell'}}}} = \frac{\gamma_{\ell} + \gamma_{\ell'} - |\gamma_{\ell} - \gamma_{\ell'}|}{\gamma_{\ell} + \gamma_{\ell'} + |\gamma_{\ell} - \gamma_{\ell'}|} = \frac{\min\{\gamma_{\ell}, \gamma_{\ell'}\}}{\max\{\gamma_{\ell}, \gamma_{\ell'}\}}.$$

We would like to know which of $\gamma_{\ell}, \gamma_{\ell'}$ is the min and which is the max. If we knew the ordering of (γ_{ℓ}) then we could transform $r_{\ell\ell'}$ so as $r_{\ell\ell'} = \frac{\gamma_{\ell}}{\gamma_{\ell'}}$, and the normalization $\sum_{\ell} \gamma_{\ell} = 1$ would identify each γ_{ℓ} precisely. Unfortunately, we do not know a priori the ordering of (γ_{ℓ}) .

Note first that $r_{\ell\ell'} = \frac{\gamma_{\ell}}{\gamma_{\ell'}}$ for all ℓ, ℓ' implies that $r_{\ell\ell''} = r_{\ell\ell'}/r_{\ell'\ell''}$ for all ℓ, ℓ', ℓ'' . However, suppose that $\gamma_2 < \gamma_1 < \gamma_3$. Then $r_{12} = \frac{\gamma_2}{\gamma_1}, r_{13} = \frac{\gamma_1}{\gamma_3}$ and $r_{23} = \frac{\gamma_2}{\gamma_3}$. Hence, $r_{23} = r_{13} \times r_{12}$. If we change r_{13} into $1/r_{13}$ then the ratios are again consistent. The following algorithm does that for all ℓ .

For $i = 2 : L$

For $j = 3 : L$

If $\frac{r_{1j}}{r_{1i}} \notin \left\{ r_{ij}, \frac{1}{r_{ij}} \right\}$ then $r_{1j} \leftarrow \frac{1}{r_{1j}}$

If $\frac{r_{1j}}{r_{1i}} = \frac{1}{r_{ij}}$ then $r_{ij} \leftarrow \frac{1}{r_{ij}}$

end

end

end

This algorithm guaranties that $r_{\ell\ell'} = \frac{\gamma_\ell}{\gamma_{\ell'}}$ for all ℓ, ℓ' or $r_{\ell\ell'} = \frac{\gamma_{\ell'}}{\gamma_\ell}$ for all ℓ, ℓ' .

Second, assume that $r_{\ell\ell'} = \frac{\gamma_\ell}{\gamma_{\ell'}}$. Then

$$\sum_{\ell'=1}^L \frac{1}{r_{\ell\ell'}} = \frac{1}{\gamma_\ell},$$

and $\gamma_\ell = \left(\sum_{\ell'=1}^L \frac{1}{r_{\ell\ell'}} \right)^{-1}$. Moreover,

$$v_\ell = \left(\sum_{\ell'=1}^L \frac{M_{\ell\ell'}}{2M_{\ell\ell}} \frac{\gamma_\ell + \gamma_{\ell'}}{\gamma_{\ell'}} \right)^{-1}.$$

If instead $r_{\ell\ell'} = \frac{\gamma_{\ell'}}{\gamma_\ell}$ then

$$\tilde{\gamma}_\ell = \left(\sum_{\ell'=1}^L \frac{1}{r_{\ell\ell'}} \right)^{-1} \propto \frac{1}{\gamma_\ell}$$

and

$$\tilde{v}_\ell = \left(\sum_{\ell'=1}^L \frac{M_{\ell\ell'}}{2M_{\ell\ell}} \frac{\tilde{\gamma}_\ell + \tilde{\gamma}_{\ell'}}{\tilde{\gamma}_{\ell'}} \right)^{-1} \propto v_\ell \gamma_\ell.$$

Hence, if $\frac{M_{\ell\ell'}}{2M_{\ell\ell}} = \frac{v_{\ell'}}{v_\ell} \frac{\gamma_{\ell'}}{\gamma_\ell + \gamma_{\ell'}}$, then there is an equivalent parametrization $\frac{M_{\ell\ell'}}{2M_{\ell\ell}} = \frac{\tilde{v}_{\ell'}}{\tilde{v}_\ell} \frac{\tilde{\gamma}_{\ell'}}{\tilde{\gamma}_\ell + \tilde{\gamma}_{\ell'}}$ with $\tilde{\gamma}_\ell \propto \frac{1}{\gamma_\ell}$ and $\tilde{v}_\ell \propto v_\ell \gamma_\ell$. For each solution $(\lambda_\ell, v_\ell, \gamma_\ell)$ there is an observationally equivalent one $(\tilde{\lambda}_\ell, \tilde{v}_\ell, \tilde{\gamma}_\ell) \propto \left(\frac{\lambda_\ell}{\gamma_\ell}, v_\ell \gamma_\ell, \frac{1}{\gamma_\ell} \right)$, where \propto means that second and third components must be normalized so as to sum to one. One additional normalization is required to determine which of the two solutions is the right one. For example, one may prefer ex ante that γ_ℓ be increasing.

D An MM algorithm for the M-step update of transition probabilities

In the M-step of the EM algorithm, we maximize the part of the expected likelihood that refers to transitions, i.e.

$$H(M|\beta^{(m)}) \equiv \sum_{k=1}^K \sum_{\ell=0}^L \left\{ n_{k\ell-}(\beta^{(m)}) \ln M_{k\ell-} + \sum_{\ell'=0}^L n_{k\ell\ell'}(\beta^{(m)}) \ln M_{k\ell\ell'} \right\},$$

where $M_{k\ell^-} \equiv M(\neg|k, \ell)$, $M_{k\ell\ell'} \equiv M(\ell'|k, \ell)$, and

$$n_{k\ell^-}(\beta^{(m)}) = \sum_i p_i(k|\beta^{(m)}) \#\{t : D_{it} = 0, \ell_{it} = \ell, x_{it} = x\},$$

$$n_{k\ell\ell'}(\beta^{(m)}) = \sum_i p_i(k|\beta^{(m)}) \#\{t : D_{it} = 1, \ell_{it} = \ell, \ell_{i,t+1} = \ell', x_{it} = x\},$$

where $\#\{\}$ denotes the cardinality of a set and where we reintroduce the control $x_{it} = x$ to remind that we are estimating different parameters for all different control values x .

Parameters $\psi_{k\ell}$ (job finding rate for unemployed) are thus updated as

$$\psi_{k\ell}^{(m+1)} = \frac{n_{k0\ell}(\beta^{(m)})}{n_{k\ell^-}(\beta^{(m)}) + \sum_{\ell'=1}^L n_{k0\ell'}(\beta^{(m)})}.$$

The rest of the likelihood is similar to the likelihood of a Bradley-Terry model except that when the incumbent firm ℓ wins we do not know against which ℓ' . The likelihood is thus rendered more nonlinear by the presence of the term in $\ln \bar{M}_{k\ell}$. An MM algorithm can still be developed as follows.²⁰

Because the logarithm is concave, we can minorize $M(\neg|k, \ell) \equiv M_{k\ell^-}$ as follows. With obvious notations, for $\ell = 1, \dots, L$,

$$\begin{aligned} \ln M_{k\ell^-} &= \ln \left(1 - \delta_{k\ell} - \lambda_k + \lambda_k \sum_{\ell'=1}^L v_{\ell'} (1 - P_{k\ell\ell'}) \right) \\ &\geq \frac{1 - \delta_{k\ell}^{(s)} - \lambda_k^{(s)}}{M_{k\ell^-}^{(s)}} \ln \left(\frac{1 - \delta_{k\ell} - \lambda_k}{1 - \delta_{k\ell}^{(s)} - \lambda_k^{(s)}} M_{k\ell^-}^{(s)} \right) \\ &\quad + \sum_{\ell'=1}^L \frac{\lambda_k^{(s)} v_{\ell'}^{(s)} (1 - P_{k\ell\ell'})}{M_{k\ell^-}^{(s)}} \ln \left(\frac{\lambda_k v_{\ell'} (1 - P_{k\ell\ell'})}{\lambda_k^{(s)} v_{\ell'}^{(s)} (1 - P_{k\ell\ell'})} M_{k\ell^-}^{(s)} \right). \end{aligned}$$

Note that both sides of the inequality are equal if $(\lambda_k, v_{\ell'}, \gamma_{k\ell}) = (\lambda_k^{(s)}, v_{\ell'}^{(s)}, \gamma_{k\ell}^{(s)})$ (no parameter change).

Let

$$\tilde{n}_{k\ell\ell'}^{(s)} = n_{k\ell^-} \frac{\lambda_k^{(s)} v_{\ell'}^{(s)} (1 - P_{k\ell\ell'})}{M_{k\ell^-}^{(s)}},$$

²⁰The MM algorithm works by finding a function that minorizes the objective function and that is more easily maximized. Let $f(\theta)$ be the objective concave function to be maximized. At the m step of the algorithm, the constructed function $g(\theta|\theta_m)$ will be called the minorized version of the objective function at θ_m if

$$g(\theta|\theta_m) \leq f(\theta), \forall \theta, \text{ and } g(\theta_m|\theta_m) = f(\theta_m).$$

Then, maximize $g(\theta|\theta_m)$ instead of $f(\theta)$, and let $\theta_{m+1} = \arg \max_{\theta} g(\theta|\theta_m)$. The above iterative method guarantees that $f(\theta_m)$ converges to a local optimum or a saddle point as m goes to infinity because

$$f(\theta_{m+1}) \geq g(\theta_{m+1}|\theta_m) \geq g(\theta_m|\theta_m) = f(\theta_m).$$

where $n_{k\ell\gamma}$ is implicitly a function of $\beta^{(m)}$. We omit this reference in the rest of this subsection. This is the predicted fraction of stayers such as home beats visitor ℓ' . Given initial values $\lambda^{(s)}, \mathbf{v}^{(s)}$ one can update $\gamma^{(s)}$ so as to maximize

$$\sum_{k=1}^K \sum_{\ell=1}^L \sum_{\ell'=1}^L \left\{ \tilde{n}_{k\ell\ell'}^{(s)} \ln \frac{\gamma_{k\ell}}{\gamma_{k\ell} + \gamma_{k\ell'}} + n_{k\ell\ell'} \ln \frac{\gamma_{k\ell'}}{\gamma_{k\ell} + \gamma_{k\ell'}} \right\},$$

subject to the normalization $\sum_{\ell=1}^L \gamma_{k\ell} = 1$.²¹ Now, because

$$-\ln(\gamma_{k\ell} + \gamma_{k\ell'}) \geq 1 - \ln(\gamma_{k\ell}^{(s)} + \gamma_{k\ell'}^{(s)}) - \frac{\gamma_{k\ell} + \gamma_{k\ell'}}{\gamma_{k\ell}^{(s)} + \gamma_{k\ell'}^{(s)}}$$

(see Hunter, 2004), we can instead maximize

$$\sum_{k=1}^K \sum_{\ell=1}^L \left(\sum_{\ell'=1}^L (\tilde{n}_{k\ell\ell'}^{(s)} + n_{k\ell'\ell}) \right) \ln \gamma_{k\ell} - \sum_{k=1}^K \sum_{\ell=1}^L \sum_{\ell'=1}^L \left((\tilde{n}_{k\ell\ell'}^{(s)} + n_{k\ell'\ell}) \frac{\gamma_{k\ell} + \gamma_{k\ell'}}{\gamma_{k\ell}^{(s)} + \gamma_{k\ell'}^{(s)}} \right).$$

That is (taking special care to indices), for $\ell = 1, \dots, L$,

$$\gamma_{k\ell}^{(s+1)} \propto \left(\sum_{\ell'=1}^L (\tilde{n}_{k\ell\ell'}^{(s)} + n_{k\ell'\ell}) \right) \left[\sum_{\ell'=1}^L \frac{\tilde{n}_{k\ell\ell'}^{(s)} + n_{k\ell\ell'} + \tilde{n}_{k\ell'\ell} + n_{k\ell'\ell}}{\gamma_{k\ell}^{(s)} + \gamma_{k\ell'}^{(s)}} \right]^{-1},$$

where $X_\ell \propto Y_\ell$ means $X_\ell = Y_\ell / \sum_\ell Y_\ell$, that is $\gamma_{k\ell}^{(s+1)}$ should sum to one over $\ell = 1, \dots, L$.

Update $\delta^{(s)}, \lambda^{(s)}$ by maximizing

$$\sum_{k=1}^K \sum_{\ell=1}^L \left(\left(n_{k\ell\gamma} \frac{1 - \delta_{k\ell}^{(s)} - \lambda_k^{(s)}}{M_{k\ell\gamma}^{(s)}} \right) \ln(1 - \delta_{k\ell} - \lambda_k) + n_{k\ell 0} \ln \delta_{k\ell} + \sum_{\ell'=1}^L (\tilde{n}_{k\ell\ell'}^{(s)} + n_{k\ell\ell'}) \ln \lambda_k \right).$$

Let, for $k = 1, \dots, K$,

$$\lambda_k^{(s+1)} = \left[\sum_{\ell=1}^L \sum_{\ell'=1}^L (\tilde{n}_{k\ell\ell'}^{(s)} + n_{k\ell\ell'}) \right] \left[\sum_{\ell=1}^L \left[n_{k\ell 0} + n_{k\ell\gamma} \frac{1 - \delta_{k\ell}^{(s)} - \lambda_k^{(s)}}{M_{k\ell\gamma}^{(s)}} + \sum_{\ell'=1}^L (\tilde{n}_{k\ell\ell'}^{(s)} + n_{k\ell\ell'}) \right] \right]^{-1},$$

$$\delta_{k\ell}^{(s+1)} = \left(1 - \lambda_k^{(s+1)} \right) n_{k\ell 0} \left[n_{k\ell\gamma} \frac{1 - \delta_{k\ell}^{(s)} - \lambda_k^{(s)}}{M_{k\ell\gamma}^{(s)}} + n_{k\ell 0} \right]^{-1}.$$

²¹Notice that for $\ell = \ell' = 0$, we have an extra contribution of $(\tilde{n}_{k00}^{(s)} + n_{k00}) \ln \frac{1}{2}$, but it does not matter because it is independent of parameters.

Finally update $\mathbf{v}^{(s)}$ by maximizing

$$\sum_{\ell'=1}^L \left[\sum_{k=1}^K \sum_{\ell=1}^L \left(\tilde{n}_{k\ell\ell'}^{(s)} + n_{k\ell\ell'} \right) \right] \ln v_{\ell'} \quad \text{s.t.} \quad \sum_{\ell'=1}^L v_{\ell'} = 1.$$

That is

$$v_{\ell'}^{(s+1)} \propto \sum_{k=1}^K \sum_{\ell=1}^L \left[\tilde{n}_{k\ell\ell'}^{(s)} + n_{k\ell\ell'} \right], \quad \ell' = 1, \dots, L.$$

For a given value of $\beta^{(m)}$, the sequence $(H(M^{(s)}|\beta^{(m)}))$, driven by the sequence $(\boldsymbol{\psi}^{(s)}, \boldsymbol{\delta}^{(s)}, \boldsymbol{\lambda}^{(s)}, \mathbf{v}^{(s)}, \boldsymbol{\gamma}^{(s)})$, is increasing. The MM algorithm can thus be stopped at any time, not only after convergence, to deliver the updated values of transition parameters, $(\boldsymbol{\psi}^{(m+1)}, \boldsymbol{\delta}^{(m+1)}, \boldsymbol{\lambda}^{(m+1)}, \mathbf{v}^{(m+1)}, \boldsymbol{\gamma}^{(m+1)})$.

E Numerical implementation

The implementation of the estimation allows the estimation to be scaled up to larger data sets by expansion of the number of CPUs in the computing cluster. The following describes how the storage and computation requirements of the estimation are delegated across CPUs in a parallel computing environment. The coding is done in Fortran and parallelization is performed with OpenMPI.

E.1 Data structure

The Danish Matched Employer-Employee (MEE) data comprise $I = 4,000,000$ workers and $J = 400,000$ firms observed at a weekly frequency from 1985 to 2013. The fundamental observation in the data is a spell (either employment or non-employment).

A worker history consists of a series of employment and unemployment spells. It is stored as a linked list. Each object in the list is a spell. The spell object contains,

- Start and end weeks of the spell.
- ID's of the worker and firm (unemployment has firm ID 0).
- A vector of wage observations for each year of the spell.
- Pointers to the previous and next spell in the worker's history.
- Pointers to the previous and next spell in the firm's spell list (unlike the worker's linked list, the firm list is not necessarily chronological).

In addition, the data structure holds the observable characteristics of each worker and firm separate from the list of spells. The worker i object holds the worker's observable characteristics (gender, education, birth year, year of entry into labor market, etc) as well as pointers to the first and last spells in the worker's labor history. The firm j object holds observable characteristics

(public-private) and pointers to the first and last spell in its list of spells. The firm $j = 0$ list holds all the non-employment spells in the data.

The data storage is divided across CPUs so that each CPU holds its own subset of worker histories. Denote by ι_c the set of worker IDs assigned to CPU c . Each CPU holds the entire set of firms, but CPU c 's list of employment spells in firm j consists only of those that are contributed by workers in the subset ι_c .

The Danish MEE data set is relatively small by international comparison (by the small size of the Danish population). Nevertheless, it does place significant demands on computer memory. Needless to say, this issue only becomes more acute for MEE data from larger countries. It is a virtue of the code that the memory requirement associated with each CPU is roughly $1/C$ of the total size of the data given a total of C CPUs. Thus, the memory pool available to the estimation is the combined memory of the nodes in the cluster, which is trivially scaled up by adding more nodes. This opposed to a data structure where each CPU holds the entire data set, which would place heavy memory requirements on multi-CPU nodes.

E.2 E-step

E.2.1 Likelihood evaluation for a given (β, \mathcal{L}) .

Each CPU holds its own copy of the firm classification, \mathcal{L} . With this, CPU c evaluates $L_i(\beta, \mathcal{L}) = \sum_{k=1}^K L_i(k; \beta, \mathcal{L})$ for any $i \in \iota_c$ by walking through the worker i linked list of spells. CPU c calculates $L^c = \sum_{i \in \iota_c} \ln L_i(\beta, \mathcal{L})$. The likelihood of the data is then found by summing L^c across CPUs, $L(\beta, \mathcal{L}) = \exp(\sum_{c=1}^C L^c)$. This is a modest communication of a single double precision number across the C CPUs. The calculation of the overall likelihood is not necessary for the execution of the E-step, but serves as useful check that the algorithm is indeed proceeding to increase the likelihood in each iteration.

E.2.2 Worker posterior update for a given (β, \mathcal{L}) .

CPU c updates worker posteriors for all $i \in \iota_c$ by, $p_i(k; \beta, \mathcal{L}) = L_k(k; \beta, \mathcal{L})/L_i(\beta, \mathcal{L})$. No communication across CPUs is necessary for this and CPU c knows only the posteriors for workers $i \in \iota_c$. Nowhere in the CEM algorithm does CPU c need to know the worker posterior for workers outside ι_c . This is a significant savings in communication which would otherwise involve a communication of $I \times K$ double precision numbers across the C CPUs in each E step.

E.3 M step

The M step uses the updated posterior $p_i(k; \beta, \mathcal{L})$ from the E step. Each part of the M step requires only modest communication between nodes.

E.3.1 $\pi_k(z)$ update for given (β, \mathcal{L}) .

With the worker posteriors in hand CPU c calculates $\pi_{k,c}(z) = \sum_{i \in \mathcal{I}_c} p_i(k; \beta, \mathcal{L}) \mathbf{1}\{z_i = z\}$, which is communicated across the CPUs. This is a $K \times Z$ dimension double precision array communication across C CPUs where each CPU receives $\sum_{c=1}^C \pi_{k,c}(z)$.²² Each CPU then calculates $\pi_k(z) = \sum_{c=1}^C \pi_{k,c}(z) / [\sum_{k=1}^K \sum_{c=1}^C \pi_{k,c}(z)]$.

E.3.2 $m_{kl}(x)$ update for given (β, \mathcal{L}) .

CPU c calculates $m_{kl,c}(x) = \sum_{i \in \mathcal{I}_c} p_i(k; \beta, \mathcal{L}) \mathbf{1}\{x_{i1} = x, \ell_{i1} = \ell\}$, which is communicated across the CPUs with each CPU receiving $\sum_{c=1}^C m_{kl,c}(x)$. This is a $K \times L \times X_{ini}$ double precision array where X_{ini} is the number of x categories in the initial distribution. Each CPU then calculates $m_{kl}(x) = \sum_{c=1}^C m_{kl,c}(x) / [\sum_{\ell=1}^L \sum_{n=1}^N m_{kl,c}(x)]$.

E.3.3 Wage parameters for given (β, \mathcal{L}) .

CPU c calculates $\mu_{kl,c}(x) = \sum_{i \in \mathcal{I}_c} p_i(k; \beta, \mathcal{L}) \sum_{t=1}^T \mathbf{1}\{\ell_{it} = \ell, x_{it} = x\} w_{it}$ and $d_{kl,c}(x) = \sum_{i \in \mathcal{I}_c} p_i(k; \beta, \mathcal{L}) \sum_{t=1}^T \mathbf{1}\{\ell_{it} = \ell, x_{it} = x\}$. These $2 K \times L \times X$ arrays are communicated across CPUs to form $\sum_{c=1}^C \mu_{kl,c}(x)$ and $\sum_{c=1}^C d_{kl,c}(x)$, where X is the number of relevant x categories for the wage parameters as well as γ and λ mobility parameters. Each CPU proceeds to calculate $\mu_{kl}(x) = \sum_{c=1}^C \mu_{kl,c}(x) / \sum_{c=1}^C d_{kl,c}(x)$.

Moving to the variance, CPU c calculates $\sigma_{kl,c}(x) = \sum_{i=1}^I p_i(k; \beta, \mathcal{L}) \sum_{t=1}^T \mathbf{1}\{\ell_{it} = \ell, x_{it} = x\} [w_{it} - \mu_{kl}(x)]^2$. The $K \times L \times X$ array is communicated across CPUs to form $\sum_{c=1}^C \sigma_{kl,c}(x)$. Each CPU calculates $\sigma_{kl}(x) = \sqrt{\sum_{c=1}^C \sigma_{kl,c}(x) / \sum_{c=1}^C d_{kl,c}(x)}$.

E.3.4 Mobility parameters for given (β, \mathcal{L}) .

Running through worker spell lists, each CPU calculates mobility counts,

$$\bar{n}_{kl,c}(x) = \sum_{i \in \mathcal{I}_c} p_i(k; \beta, \mathcal{L}) \#\{t : D_{it} = 0, \ell_{it} = \ell, x_{it} = x\}$$

and

$$n_{k\ell\ell',c}(x) = \sum_{i \in \mathcal{I}_c} p_i(k; \beta, \mathcal{L}) \#\{t : D_{it} = 1, \ell_{it} = \ell, \ell_{i(t+1)} = \ell', x_{it} = x\}.$$

These two integer arrays (of size $K \times (L+1) \times X$ and $K \times (L+1)^2 \times X$, respectively) are communicated across CPUs to form $\bar{n}_{kl}(x) = \sum_{c=1}^C \bar{n}_{kl,c}(x)$ and $n_{k\ell\ell'}(x) = \sum_{c=1}^C n_{k\ell\ell',c}(x)$. With these counts each CPU updates $\gamma_{kl}(x)$, $\lambda_\ell(x)$ and $\nu_\ell(x)$ according to section D.

E.4 C step

The C-step reassigns firm types in such a way as to increase the value of the expected log likelihood function, thereby increasing the likelihood of the data. The C step can be viewed as

²²Using `mpi_allreduce`.

a simple extension of the M step where the firm classification is just another set of parameters to be chosen so as to improve on the expected log likelihood. While the M step requires very modest communication, the C-step does involve J separate communications of size L arrays within the cluster. This is a significant communication load and consequently, it is advantageous to do multiple EM iterations between C steps.

The firm IDs have been chosen so that firms are ordered by size ($j = 1$ is the largest firm where size is the number of wage observations throughout the panel). The algorithm reassigns firm type j by,

$$\ell_j^{(s+1)} = \arg \max_{\ell} \sum_{i=1}^I \sum_{k=1}^K p_i(k; \hat{\beta}^{(s)}, \mathcal{L}^{(s)}) \ln L_i(k; \hat{\beta}^{(s)}, \mathcal{L}_{-j}^{(s)}(\ell)), \quad (17)$$

where $\mathcal{L}_{-j}^{(s)}(\ell)$ is the firm classification that is obtained by taking the $\mathcal{L}^{(s)}$ classification where all firm types $j' = 1, \dots, j-1$ have already been reassigned, and furthermore replace the j 'th element with ℓ . Do the reassignment in order. This step increases the expected log likelihood.

Done naively, the step is expensive since it involves $L \times J$ expected likelihood evaluations of the data. But the expected log likelihood varies with firm j 's type only through the spells that directly involve firm j and through firm j 's type's impact on the $q(\ell, \mathcal{L}_{-j}^{(s)}(\ell))$ distribution. The latter does involve all spells but in a way that allows simplification. Define by $\Omega(\mathcal{L})$, the contribution to the expected log likelihood from the $q(\cdot|\mathcal{L})$ related terms,

$$\Omega(\beta, \mathcal{L}) = - \sum_{i=1}^I \sum_{k=1}^K p_i(k; \beta, \mathcal{L}) \left[\ln q(\ell_{i1}|\mathcal{L}) + \sum_{t=1}^T D_{it} \ln q(\ell_{i(t+1)}|\mathcal{L}) \right]$$

Define,

$$n_{\ell}^q(\mathcal{L}) = \sum_{i=1}^I \left[\mathbf{1}\{\ell_{i1} = \ell\} + \#\{t : D_{it} = 1, \ell_{i(t+1)} = \ell\} \right]$$

with which we can write,

$$\Omega(\beta, \mathcal{L}) = - \sum_{\ell=1}^L \ln q(\ell|\mathcal{L}) n_{\ell}^q(\mathcal{L}).$$

It is worth noting that another way of calculating Ω is by adding up spells at the firm level. Denote by \hat{n}_j the number of employment spells in firm j ,

$$\hat{n}_j = \sum_{i=1}^I \left[\mathbf{1}[j(i, 1) = j] + \sum_{t=1}^T \mathbf{1}[D_{it} = 1, j(i, t+1) = j] \right].$$

with this, Ω can be written as,

$$\Omega(\beta, \mathcal{L}) = - \sum_{\ell=1}^L \ln q(\ell|\mathcal{L}) \sum_{j=1}^J \hat{n}_j \mathbf{1}[\ell_j = \ell] = - \sum_{\ell=1}^L \ln q(\ell|\mathcal{L}) \hat{n}(\ell|\mathcal{L}),$$

where the number of spells in type ℓ firms is,

$$\hat{n}(\ell|\mathcal{L}) \equiv \sum_{j=1}^J \hat{n}_j \mathbf{1}[\ell_j = \ell]. \quad (18)$$

This firm-centric formulation of Ω is the preferable one for the firm reclassification algorithm.

Continuing the firm-centric formulation of the log-likelihood, denote by $\iota(j) = \{(i, t) \mid j(i, t) = j\}$, that is, all worker-time pairs with firm j . We can then write the the firm j classification update as,

$$\ell_j^{(s+1)} = \arg \max_{\ell} \left[\sum_{(i,t) \in \iota(j)} \sum_{k=1}^K p_i(k; \hat{\beta}^{(s)}, \mathcal{L}^{(s)}) \times \left[f_{k\ell}(w_{it}|x_{it}) + \right. \right. \\ \left. \left. (1 - D_{it}) \ln \bar{M}_{k\ell_{it}}(x_{it}) + D_{i(t-1)} \ln M_{k\ell_{i(t-1)}} + D_{it} \ln M_{k\ell_{i(t+1)}} \right] + \Omega(\beta, \mathcal{L}_{-j}^{(s)}(\ell)) \right]. \quad (19)$$

The algorithm is then as follows:

1. The firm j spell counts, \hat{n}_j , are determined at the outset of the overall estimation where all processors count how many spells they each have for each given firm j . \hat{n}_j is then found by a communication of a size J integer vector across all processors. Furthermore, the firm IDs $j = 1, \dots, J$, are ordered by firm size - specifically the size of $\iota(j)$. These steps are not done in the C-step but rather just once at the outset of the full CEM algorithm.
2. The firm classification at the outset of the C-step is $\mathcal{L}^{(s)}$. Denote by $\mathcal{L}^{(s),0} = \mathcal{L}^{(s)}$, where $\mathcal{L}^{(s),j}$ is the firm classification in the j th substep of the C-step. Initialize the C-step by the determination of $\hat{n}(\ell|\mathcal{L}^{(s)})$ by equation (18).
3. Take firm $j = 1$. Find the optimal firm type for firm j according to equation (19) and firm classification $\mathcal{L}^{(s),j-1}$. The (i, t) pairs in $\iota(j)$ are by the data delegation spread out across different CPUs. Each CPU evaluates the summation in equation (19) for its own (i, t) pairs for each firm type $\ell = 1, \dots, L$. The data structure has for each firm defined a linked list of its spells held by CPU c , which allows quick within CPU evaluation of each CPU's contribution to equation (19). The full sum for each ℓ is then obtained by a summation across all CPUs to the master process. This is a communication of an L size array from each node to the master node. The master process resolves the maximization problem in equation (19), and communicates the optimal firm type $\ell_j^{(s+1)}$ to all CPUs, a single integer.

4. Update the firm classification $\mathcal{L}^{(s),j} = \mathcal{L}_{-j}^{(s),(j-1)}(\ell_j^{(s+1)})$. Thus, as the algorithm steps through $j = 1, \dots, J$, the firm classification is updated sequentially with a new firm type for firm j . Also, update $\hat{n}^j(\ell) = \hat{n}(\ell | \mathcal{L}^{(s),j})$ and the type frequencies $q(\ell | \mathcal{L}^{(s),j})$. This is done by the simple algorithm (stated just for \hat{n}^j)
 - (a) If $\ell_j^{(s+1)} = \ell_j^{(s)}$ then $\hat{n}^j(\ell) = \hat{n}^{(j-1)}(\ell)$ for all ℓ .
 - (b) Else, $\hat{n}^j(\ell_j^{(s+1)}) = \hat{n}^{(j-1)}(\ell_j^{(s+1)}) + \hat{n}_j$ and $\hat{n}^j(\ell_j^{(s)}) = \hat{n}^{(j-1)}(\ell_j^{(s)}) - \hat{n}_j$. For all other firm types, $\hat{n}^j(\ell) = \hat{n}^{(j-1)}(\ell)$.
5. loop back to step 3 for next j . Exit when $j = J$ is completed. Denote by $\mathcal{L}^{(s+1)} = \mathcal{L}^{(s),J}$.

References

- Abowd, J. M., R. H. Creecy, and F. Kramarz (2002, March). Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data. Longitudinal Employer-Household Dynamics Technical Papers 2002-06, Center for Economic Studies, U.S. Census Bureau.
- Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High wage workers and high wage firms. *Econometrica* 67(2), 251–334.
- Abowd, J. M., F. Kramarz, and S. Roux (2006, June). Wages, Mobility and Firm Performance: Advantages and Insights from Using Matched Worker-Firm Data. *Economic Journal* 116(512), 245–285.
- Abowd, J. M., B. E. Stephens, L. Vilhuber, F. Andersson, K. L. McKinney, M. Roemer, and S. Woodcock (2009, June). The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators. In *Producer Dynamics: New Evidence from Micro Data*, NBER Chapters, pp. 149–230. National Bureau of Economic Research, Inc.
- Agresti, A. (2003). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley.
- Allman, E. S., C. Matias, and J. A. Rhodes (2009). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics* 37, 3099–3132.
- Altonji, J. G., A. A. Smith, and I. Vidangos (2013). Modeling earnings dynamics. *Econometrica* 81(4), 1395–1454.
- Andrews, M. J., L. Gill, T. Schank, and R. Upward (2008). High wage workers and low wage firms: negative assortative matching or limited mobility bias? *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171(3), 673–697.
- Bagger, J. and R. Lentz (2014, April). An Empirical Model of Wage Dispersion with Sorting. NBER Working Papers 20031, National Bureau of Economic Research, Inc.

- Bagger, J., K. L. Soerensen, and R. Vejlin (2013). Wage sorting trends. *Economics Letters* 118(1), 63–67.
- Bonhomme, S., K. Jochmans, and J.-M. Robin (2016a). Estimating multivariate latent-structure models. *Annals of Statistics* 44, 540–563.
- Bonhomme, S., K. Jochmans, and J.-M. Robin (2016b). Non-parametric estimation of finite mixtures from repeated measurements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78, 211–229.
- Bonhomme, S., T. Lamadon, and E. Manresa (2017). A Distributional Framework for Matched Employer Employee Data. mimeo, University of Chicago.
- Borovičková, K. and R. Shimer (2017, November). High Wage Workers Work for High Wage Firms. NBER Working Papers 24074, National Bureau of Economic Research, Inc.
- Card, D., A. R. Cardoso, and P. Kline (2016). Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women *. *The Quarterly Journal of Economics* 131(2), 633–686.
- Card, D., J. Heining, and P. Kline (2013). Workplace Heterogeneity and the Rise of West German Wage Inequality. *The Quarterly Journal of Economics* 128(3), 967–1015.
- Celeux, G. and G. Govaert (1992). A classification {EM} algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis* 14(3), 315 – 332.
- de Melo, R. L. (2018). Firm wage differentials and labor market sorting: Reconciling theory and evidence. *Journal of Political Economy* 126(1), 313–346.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1), pp. 1–38.
- Eeckhout, J. and P. Kircher (2011). Identifying Sorting–In Theory. *Review of Economic Studies* 78(3), 872–906.
- Gassiat, E., A. Cleyenen, and S. Robin (2016). Finite state space non parametric hidden Markov models are in general identifiable. *Statistics and Computing*.
- Gruetter, M. and R. Lalive (2009). The importance of firms in wage determination. *Labour Economics* 16(2), 149 – 160.
- Guvenen, F., G. Kaplan, J. Song, and J. Weidner (2017, April). Lifetime Incomes in the United States over Six Decades. NBER Working Papers 23371, National Bureau of Economic Research, Inc.

- Hagedorn, M., T. H. Law, and I. Manovskii (2017, January). Identifying Equilibrium Models of Labor Market Sorting. *Econometrica* 85, 29–65.
- Hall, P. and X.-H. Zhou (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics* 31, 201–224.
- Henry, M., Y. Kitamura, and B. Salanié (2014). Partial identification of finite mixtures in econometric models. *Quantitative Economics* 5(1), 123–144.
- Holzer, H. J., J. I. Lane, and L. Vilhuber (2004). Escaping low earnings: The role of employer characteristics and changes. *Industrial and Labor Relations Review* 57(4), 560–578.
- Hu, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics* 144, 27–61.
- Hu, Y. and S. M. Schennach (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica* 76, 195–216.
- Hu, Y. H. and M. Shum (2012). Nonparametric identification of dynamic models with unobserved state variables. *Journal of Econometrics* 171, 32–44.
- Hunter, D. R. (2004, 02). Mm algorithms for generalized bradley-terry models. *Ann. Statist.* 32(1), 384–406.
- Hunter, D. R. and K. Lange (2004). A tutorial on mm algorithms. *Amer. Statist.*, 30–37.
- Iranzo, S., F. Schivardi, and E. Tosetti (2008). Skill dispersion and firm productivity: An analysis with employer-employee matched data. *Journal of Labor Economics* 26(2), 247–285.
- Jarosh, G. (2015). Searching for job security and the consequences of job loss. Technical report.
- Kasahara, H. and K. Shimotsu (2009). Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica* 77, 135–175.
- Lise, J., C. Meghir, and J.-M. Robin (2016). Matching, sorting and wages. *Review of Economic Dynamics* 19, 63 – 87. Special Issue in Honor of Dale Mortensen.
- Lise, J. and J.-M. Robin (2017, April). The macrodynamics of sorting between workers and firms. *American Economic Review* 107(4), 1104–35.
- Martins, P. S. (2008). Dispersion in wage premiums and firm performance. *Economics Letters* 101(1), 63 – 65.
- Song, J., D. J. Price, F. Guvenen, N. Bloom, and T. von Wachter (2015, May). Firming Up Inequality. NBER Working Papers 21199, National Bureau of Economic Research, Inc.
- Woodcock, S. D. (2015). Match effects. *Research in Economics* 69(1), 100–121.