Labor Market Entry and Earnings Dynamics: Bayesian Inference Using Mixtures-of Experts Markov Chain Clustering

Sylvia Frühwirth-Schnatter* Christoph Pamminger Andrea Weber
Rudolf Winter-Ebmer

October 18, 2010

Abstract

Keywords: labor market, transition data

*Department of Applied Statistics, Johannes Kepler University Linz, Altenbergerstraße 69, A-4040 Austria; Tel: ++43 732 2468 8295; e-mail address: sylvia.fruehwirth-schnatter@jku.at

Contents

1 Introduction

A competitive model of the labor market implies that the development of individual earnings over the life cycle will follow the development of individual marginal productivity. Beside factors such as on-the-job learning and better employer matches, which increase the productivity of workers over time, shocks to aggregate labor demand - such as a major recession - can also have an impact on wage rates. In a spot labor market, however, those temporary changes to labor demand are relately short lived and should not influence wages over prolonged periods of time. This view has been seriously challenged both by studies on cohort size effects (Welch, 1979) and studies on the impact of early career problems on later outcomes. The general approach taken by these studies is to assess the initial wage or employment penalties from entering the labor market in a bad year and to test whether this initial impact persists over time. Raaum and Roed (2006), e.g. show for Norway, that school leavers facing particularly depressed labor market conditions at the start of their career face a higher risk of unemployment both initially and after ten years. Oreopoulos et al. (2008) study careers of Canadian college graduates and find a high initial wage penalty of entering in a recession, but the penalty fades away during the first decade of a worker's career. [1]

In this paper we study a slightly different aspect of the impact labor market entry conditions can have on career development. We depart from the traditional strategy of modeling wage or employment outcomes at a particular point in time, but we focus on mobility throughout the complete career path. Thereby our aim is twofold. First, we want to identify specific career patterns that characterize the earnings development of individuals after entry in the labor market. The idea is to extend the traditional mover/stayer classification to a wider variety of career types. Intuitively, some individuals may be in stable employment relationships throughout their working lives, while others are observed in more volatile jobs; still others could be considered as social climbers with a consistent upward mobility, while others could be characterized as losers with a high tendency of downward mobility. Our second goal is find out whether labor market conditions at the start of one's career have an impact on the type of career pursued over the

[1] Studies for Austria (Brunner and Kuhn, 2009), the UK (Burgess et al., 2003), Japan (Kondo, 2007), Sweden (Kwon and Meyersson-Milgrom, 2007) or the US (Oyer, 2006; Kahn, 2009; Genda et al., 2010) use essentially the same strategy.

lifetime. While entering the labor market in a recession might impose an immediate penalty in the form a lower starting wages, it might also influence the life-time career path; i.e. an individual might be characterized by a different career-type when entering the labor market in a recession as opposed to a boom period.

The statistical problem behind our empirical analysis consists of finding groups of similar time series in a set or panel of time series that are unlabeled a priori. In this paper we introduce new clustering techniques which determine subsets of similar time series within the panel. Compared to cross-sections, distance-based clustering methods are rather difficult to define for time series data. Frühwirth-Schnatter and Kaufmann (2008) demonstrated recently that model-based clustering based on finite mixture models (Banfield and Raftery, 1993; Fraley and Raftery, 2002) extends to time series data in quite a natural way. The crucial point in model-based clustering is to select an appropriate clustering kernel in terms of a sampling density which captures salient features of the observed time series. Various such clustering kernels were suggested for panels with real-valued time series observations by Frühwirth-Schnatter and Kaufmann (2008) and Juárez and Steel (2010). The econometric methods we develop in this paper will also be of interest in other areas of applied statistics like economics, finance or public health where it is often desirable to find groups of similar time series in a set of a-priori unlabeled time series.

For discrete-valued individual level panel data such as the panel considered in this paper, clustering kernels are typically based on first-order time-homogeneous Markov chain models. For discrete-valued time series it is particularly difficult to define distance measures and model-based clustering has been shown to be a useful alternative. Fougère and Kamionka (2003), for instance, considered a mover-stayer model in continuous time which is a constrained mixture of two Markov chains to incorporate a simple form of heterogeneity across individual labor market transition data. Mixtures of time-homogeneous Markov chains both in continuous and discrete time are also considered in Frydman (2005) including an application to bond ratings migration. Pamminger and Frühwirth-Schnatter (2010) construct more general clustering kernels based on first-order time-homogeneous Markov chain models to capture unobserved heterogeneity in the transition behavior within each cluster. In this paper we further extend clustering of Markov chain models based on discrete-valued data panel further by modeling the prior probability to belong to a certain cluster to depend on a set of covariates via a multinomial logit model. The

determinants we consider in our application are individual characteristics, such as the type of skill and occupation, and local labor market characteristics at the time of entry.

For estimation, we pursue a Bayesian approach which offers several advantages compared to EM estimation considered, for instance, in Frydman (2005). In particular, Bayesian inference easily copes with problems that occur with ML estimation if for any cluster no transitions are observed in the data for any cell of the cluster-specific transition matrix. A Bayesian approach to Markov chain clustering has been used earlier by Pamminger and Frühwirth-Schnatter (2010), and by Fougère and Kamionka (2003) for the special case of a mover-stayer model. In the present paper we extend the two-block Markov chain Monte Carlo sampler of Pamminger and Frühwirth-Schnatter (2010) to the mixture-of-experts extension of their method. To estimate the parameters in the multinomial regression model describing group membership we use auxiliary mixture sampling in the dRUM representation (Frühwirth-Schnatter and Frühwirth, 2010). This method turned out to be superior to other MCMC methods such as Frühwirth-Schnatter and Frühwirth (2007), Scott (2009) and Holmes and Held (2006) in term of the effective sampling rate.

2 Data

Our empirical analysis is based on data from the Austrian Social Security Data Base (ASSD), which combines detailed longitudinal information on employment and earnings of all private sector workers in Austria since 1975 (Zweimüller et al., 2009).

The sample we consider consists of $N = 49\,279$ male Austrian workers, who enter the labor market for the first time in the years 1975 to 1985 and are less than 25 years old at entry. We do not consider females in our sample, because hours of work are not observed. For non-Austrian citizens it is not always clear, if we can measure the entry in the labor market correctly. We extract yearly earnings observations measured by gross monthly wages in May of successive years and observe wages for a time span between 2 to 31 years. The the median time an individual is observed in our panel is equal to 22 years. Following Weber (2001), the gross monthly wage is divided into six categories labeled with 0 up to 5. Category zero corresponds to zero-income, i.e. unemployment or out of labor force. The categories one to five correspond to the quintiles

of the income distribution which are calculated for each year from all non-zero wages observed in that year for the population of all male employees in Austria. The use of wage categories has the advantage that no inflation adjustment has to be made and that it circumvents the problem that in Austria recorded wages are right-censored because wages that exceed a social security payroll tax cap are recorded with exactly that limit only. We cut the time series of workers who had zero income for more than five years. For individuals first observed in the data as apprentices, we consider their first "real" wage-income as the point job entry, because the apprenticeship allowance is very low compared to average wages.

As we are interested in characterizing the wage path since the first job, we are including only pre-determined variables, like age, education and type of first job; all other variables, like job mobility or work experience or tenure are endogenous. As education is not directly available in the data, we approximate it with apprenticeship education and age at entry in the first job: We take young men who served more than 2.5 years as apprentice, as baseline. We consider young men entering the labor market before their 18th birthday without having finished apprenticeship as "unskilled". Furthermore, those starting after their 18th birthday without finishing apprenticeship are coded as "skilled", because they are likely to have finished some kind of higher education such as high school or university.

The period from 1975 to 1985 for which we observe labor market entries is characterized by a fair amount of business cycle variation, ranging from a boom period in the mid 1970's to the recession in the early 1980's. The state of the labor market is captured by the unemployment rate across 65 counties, which is measured at the date of entry into the labor market.

3 Method

3.1 Mixtures-of-Experts Markov Chain Models

As for many data sets available for empirical labor market research, the structure of the data introduced in Section 2 takes the form of a discrete-valued panel data. The categorical outcome variable y_{it} assumes one of K states, labeled by $\{1, \ldots, K\}$, and is observed for N individuals $i = 1, \ldots, N$ over T_i discrete time periods, i.e. for $t = 0, \ldots, T_i$. For each individual i, we model the state of y_{it} in period t to depend on past values of the outcome variable, e.g. on the

state of $y_{i,t-1}$ in a first order model. Furthermore, we allow the transition process to depend on observable and unobservable covariates. Subsequently, $\mathbf{y}_i = \{y_{i1}, \ldots, y_{i,T_i}\}$ denotes an individual time series, excluding the initial state y_{i0}.

3.1.1 Markov Chain Clustering

Individual level transition data can be considered as a special case of a panel of discrete-valued time series. To capture the presence of unobserved heterogeneity on the dynamics in a panel of discrete-valued time series, Pamminger and Frühwirth-Schnatter (2010) extended model based clustering as introduced by Frühwirth-Schnatter and Kaufmann (2008) to this type of time series. They assume that H hidden clusters are present in the panel and a clustering kernel $p(\mathbf{y}_i|\boldsymbol{\vartheta}_h)$ with cluster-specific parameter $\boldsymbol{\vartheta}_h$ is used for describing all time series in group h, $h = 1, \ldots, H$, i.e. $p(\mathbf{y}_i|S_i, \boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_H) = p(\mathbf{y}_i|\boldsymbol{\vartheta}_{S_i})$, where $S_i \in \{1, \ldots, H\}$ is a latent group indicator. To capture the discrete nature of the data, Pamminger and Frühwirth-Schnatter (2010) considered various clustering kernel $p(\mathbf{y}_i|\boldsymbol{\vartheta}_h)$ based on Markov chains like Markov chain clustering, Dirichlet multinomial clustering and clustering based on inhomogeneous Markov chains. Markov chain clustering, for instance, is based on modeling separate transition processes for each group through a first-order time-homogeneous Markov chain model with cluster-specific transition matrix $\boldsymbol{\xi}_h$, where $\xi_{h,jk} = \Pr(y_{it} = k|y_{i,t-1} = j, S_i)$, $j, k = 1, \ldots, K$. Hence each row of $\boldsymbol{\xi}_h$ represents a probability distribution over the discrete set $\{1, \ldots, K\}$, i.e. $\sum_{k=1}^{K} \xi_{h,jk} = 1$. The clustering kernel $p(\mathbf{y}_i|\boldsymbol{\xi}_h)$ reads with $\boldsymbol{\vartheta}_h = \boldsymbol{\xi}_h$:

$$p(\mathbf{y}_i|\boldsymbol{\xi}_h) = \prod_{t=1}^{T_i} p(y_{it}|y_{i,t-1}, \boldsymbol{\xi}_h) = \prod_{j=1}^{K} \prod_{k=1}^{K} \xi_{h,jk}^{N_{i,jk}}, \tag{1}$$

where $N_{i,jk} = \#\{y_{it} = k, y_{i,t-1} = j\}$ is the number of transitions from state j to state k observed in time series i. Note that we condition in (1) on the first observation y_{i0} and the actual number of observations is equal to T_i for each time series.

A special version of this Markov chain clustering method has been applied to labor market transition data in Fougère and Kamionka (2003) who considered a mover-stayer model where $H = 2$ and $\boldsymbol{\xi}_1$ is equal to the identity matrix while only $\boldsymbol{\xi}_2$ is unconstrained. Frydman (2005) considered another constrained mixture of Markov chain models where the transition matrices

$\boldsymbol{\xi}_h, h \geq 2$, are related to the transition matrix $\boldsymbol{\xi}_1$ of the first group through $\boldsymbol{\xi}_h = \mathbf{I} - \boldsymbol{\Lambda}_h(\mathbf{I} - \boldsymbol{\xi}_1)$ where $\mathbf{I}$ is the identity matrix and $\boldsymbol{\Lambda}_h = \text{Diag}\,(\lambda_{h,1}, \ldots, \lambda_{h,K})$ with $0 \leq \lambda_{h,j} \leq 1/(1 - \xi_{1,jj})$ for $j = 1, \ldots, K$. In contrast to these approaches, Pamminger and Frühwirth-Schnatter (2010) assume that the transition matrices $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_H$ are entirely unconstrained which leads to more flexibility in capturing differences in the transition behavior between the groups.

3.1.2 Modeling prior group membership

Clustering as in Pamminger and Frühwirth-Schnatter (2010) is based on the standard finite mixture model which assumes that the group indicators $\mathbf{S} = (S_1, \ldots, S_N)$ are a priori independent with $\Pr(S_i = h) = \eta_h$ such that $\sum_{h=1}^{H} \eta_h = 1$. In the present application this assumption implies that each individual has the same prior probability to follow a particular group-specific career dynamic, regardless of the individual's observable characteristics or the circumstances at labor market entry.

To obtain a more meaningful model for the data introduced in Section 2, an extension of model-based clustering for discrete-valued panel data which allows pre-determined variables to impact on group membership is suggested in this subsection. Specifically, we model prior group membership $\Pr(S_i = h)$ through a multinomial logit model (MNL) for $\mathbf{S}$:

$$\Pr(S_i = h|\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_H) = \frac{\exp\left(\mathbf{x}_i\boldsymbol{\beta}_h\right)}{1 + \sum_{l=2}^{H} \exp\left(\mathbf{x}_i\boldsymbol{\beta}_l\right)}, \tag{2}$$

where $\mathbf{x}_i$ is a row vector of regressors, including 1 for the intercept and $\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_H$ are group-specific, unknown parameters. This model is known as mixture-of-experts models, see e.g. Frühwirth-Schnatter (2006) and has been applied in many different areas, see e.g. For identifiability reasons we set $\boldsymbol{\beta}_1 = \mathbf{0}$, which means that $h = 1$ is the baseline group and $\boldsymbol{\beta}_h$ is the effect on log-odds ratio relative to the baseline. Mixture-of-experts models yield important insights into the factors that determine group membership (Frühwirth-Schnatter and Kaufmann, 2008).

Model (2) allows us to capture the influence of individual characteristics, cohort effects, or labor market conditions that are determined at time of the entry in the labor market on group membership and thereby on mobility patterns. As will be demonstrated in Subsection 3.1.3, we deal with the initial condition problem present in discrete-time dynamic panels by adding the

initial wage category to the set of regressors appearing in $\mathbf{x}_i$.

3.1.3 A simple solution to the initial conditions problem

Inference in Pamminger and Frühwirth-Schnatter (2010) is carried out conditional on the initial condition y_{i0}, by treating this variable as exogenous. In our dynamic model with unobserved heterogeneity this assumption implies that the initial period earnings y_{i0} are independent of group membership S_i, which is apparently a very unsatisfactory assumption.

There is a long literature discussing the problem with initial conditions in non-linear dynamic models with unobserved heterogeneity. See Heckman (1981) for an early reference and Wooldridge (2005) for a recent review. These papers focus on models where unobserved heterogeneity is captured through an individual effect S_i following a continuous distribution. However, the initial condition problem has also been addressed also in the case where S_i follows a discrete distribution as for model based clustering in a transition model.

To handle the initial condition problem, we recall that the joint distribution of $y_{i0}, \ldots, y_{i,T_i}$ and S_i may be formulated in a way that separates the choice of the clustering kernel density $p(y_{i1}, \ldots, y_{i,T_i}|y_{i0}, S_i, \boldsymbol{\theta})$ from the choice of a joint model for y_{i0} and S_i:

$$p(y_{i0}, \ldots, y_{i,T_i}, S_i|\boldsymbol{\theta}) = p(y_{i1}, \ldots, y_{i,T_i}|y_{i0}, S_i, \boldsymbol{\theta})p(y_{i0}, S_i|\boldsymbol{\theta}), \tag{3}$$

where $\boldsymbol{\theta}$ contains all unknown model parameter.

Evidently, there exist two ways of formulating a joint distribution $p(y_{i0}, S_i|\boldsymbol{\vartheta})$ for y_{i0} and S_i. The first approach, which closely follows the suggestions discussed in Heckman (1981), specifies a conditional model for the initial condition y_{i0} conditional on unobserved heterogeneity S_i and a marginal model for S_i:

$$p(y_{i0}, S_i|\boldsymbol{\vartheta}) = p(y_{i0}|S_i, \boldsymbol{\vartheta})p(S_i|\boldsymbol{\vartheta}). \tag{4}$$

For the choice of $p(y_{i0}|S_i, \boldsymbol{\vartheta})$ two approaches have been applied in the literature. One of them relies on the existence of a stationary distribution of $p(y_{i0}|\boldsymbol{\vartheta}_2) = \pi_\infty(y_{i0}; \boldsymbol{\vartheta}_2)$. This assumption is unattractive in our case, as starting wages usually are not drawn from a stationary wage distribution. The other approach consists of directly modelling $p(y_{i0}|S_i, \boldsymbol{\vartheta})$ i.e. as a multinomial

logit model where some parameters have to be group-specific (that is switching with S_i) to achieve dependence between y_{i0} and S_i. In our case this would lead to a complicated parametric structure, because in certain groups we may find only very few individuals in certain initial states and thus run into problems with parameter identification.

For this reason, we prefer the second approach, which extends the "simple solution to the initial conditions problem" suggested by Wooldridge (2005) to model-based clustering in dynamic panels. We specify $p(y_{i0}, S_i|\boldsymbol{\theta})$ appearing in (3) by formulating a conditional model for unobserved heterogeneity S_i for a given initial condition y_{i0} and a marginal model for y_{i0}:

$$p(y_{i0}, S_i|\boldsymbol{\vartheta}) = p(S_i|y_{i0}, \boldsymbol{\vartheta})p(y_{i0}|\boldsymbol{\vartheta}). \tag{5}$$

In terms of our clustering procedure this means that the logit model used for modeling S_i has to be extended such that it depends on the initial conditions y_{i0}. This is achieved by adding indicator variables for the initial states to the covariate matrix $\mathbf{x}_i$ of the MNL model introduced for S_i in (2).

Our approach is directly related to Wooldridge (2005)'s treatment of the Maximum Likelihood case, where he models the mean of the random intercept distribution as being dependent on the initial state. Under the assumption that $p(S_i|\mathbf{x}_i, \boldsymbol{\vartheta}_1)$ and $p(y_{i0}|\boldsymbol{\vartheta}_2)$ have no common parameters, the marginal distribution $p(y_{i0}|\boldsymbol{\vartheta}_2)$ need not be specified explicitly, because it cancels from all posterior distributions.

3.2 Model Specification

We specify the model for earnings dynamics of labor market entrants as a fist order Markov model with group specific transition parameters:

$$\Pr(y_{it} = k|y_{i,t-1} = j, S_i = h) = \xi_{h,jk}. \tag{6}$$

The estimated parameters are $\xi_{h,jk}$ with $j, k \in \{1, \ldots, K\}$ and $h = 1, \ldots, H$. Our model treats the group membership indicator S_i and the number of different groups H as latent parameters. See the next subsection 3.4 for the procedure used to determine H.

Group membership, or $\Pr(S_i = h)$, is modeled by the multinomial logit model given by

equation (2). To address the initial condition problem we model $\Pr(S_i = h|\mathbf{x}_i, y_{i0})$ as outlined in Subsection 3.1.3 we extend the list of covariates by variables z that capture the relationship of unobserved heterogeneity with to the initial earnings categories

$$\Pr(S_i = h|\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_H) = \frac{\exp\left(\mathbf{x}_i\boldsymbol{\beta}_h + z_i\gamma_h\right)}{1 + \sum_{l=2}^{H} \exp\left(\mathbf{x}_i\boldsymbol{\beta}_l + z_i\gamma_l\right)}, \tag{7}$$

The estimated parameters are $\boldsymbol{\beta}_h$ and γ_h. Our choice of variables x includes factors that are fixed at the time of labor market entry and which we assume to be relevant for the determination of earnings mobility. We therefore include individual characteristics such as education and the type of occupation as well as cohort effects, expressed by a set of dummies for the year of labor market entry. The central variable measuring labor market characteristics at the time of entry is the unemployment rate in the region and the year of labor market entry.

To allow for correlation of the unobserved group membership with initial earnings, the variables z are chosen to include a set of indicators for the initial wage category. Our model specification implies that the only way that covariates impact on earnings trajectories is via their effect on group membership. To allow for additional flexibility in the relationship between covariates and initial earnings we include interaction terms between the regional unemployment rate and earnings categories in the initial period in z_i. We experimented with even more flexible specifications, such as interactions of the initial earnings categories with education or leads and lags or the unemployment rate. But they did not improve the fit of the model and are thus not reported here.

Pamminger and Frühwirth-Schnatter (2010) performed an illustrative comparison of two clustering kernels for discrete-valued time series, namely Markov chain clustering and Dirichlet multinomial clustering, for a smaller and less well specified version of the panel data set introduced in Section 2. Since this comparison revealed that both methods yielded comparable results, we decided to focus subsequently on Markov chain clustering, because Bayesian inference is computationally less demanding, see Subsection 3.3.

3.3 Bayesian Inference for a Fixed Number of Clusters

In this paper we pursue a Bayesian approach toward estimation for fixed H. $\mathbf{S}$ is estimated along with the group-specific parameters $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_H$ and $\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_H$ from the data.

3.3.1 Prior distributions

We assume prior independence between $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_H$ and $\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_H$. All regression coefficients β_{hj} are assumed to be independent a priori, each following a standard normal prior distribution. The rows of $\boldsymbol{\xi}_h$ are independent a priori each following a Dirichlet distribution, i.e. $\boldsymbol{\xi}_{h,j\cdot} \sim \mathcal{D}(e_{0,j1}, \ldots, e_{0,jK})$ for $j = 1, \ldots, K$ with prior parameters $\mathbf{e}_{0,j\cdot} = (e_{0,j1}, \ldots, e_{0,jK}) = N_0 \times \boldsymbol{\xi}^*_{j\cdot}$ where $N_0 = 10$ and

$$\boldsymbol{\xi}^* = \begin{pmatrix} 0.7 & 0.2 & 0.025 & 0.025 & 0.025 & 0.025 \\ 0.15 & 0.6 & 0.15 & 0.03\dot{} & 0.03\dot{} & 0.03\dot{} \\ 0.03\dot{} & 0.15 & 0.6 & 0.15 & 0.03\dot{} & 0.03\dot{} \\ 0.03\dot{} & 0.03\dot{} & 0.15 & 0.6 & 0.15 & 0.03\dot{} \\ 0.03\dot{} & 0.03\dot{} & 0.03\dot{} & 0.15 & 0.6 & 0.15 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.2 & 0.7 \end{pmatrix}.$$

The choice of this prior takes into account that to stay in the same wage category is much more likely than a transition to another wage category and transitions into adjacent categories are more likely than into the other categories.

3.3.2 MCMC estimation

For practical Bayesian estimation we extend the Markov chain Monte Carlo (MCMC) sampler discussed by Pamminger and Frühwirth-Schnatter (2010) to the mixtures-of-experts formulation introduced in (2). First, a step has to be added to sample the parameters appearing in (2) conditional on knowing $\mathbf{S}$. Second, model (2) acts as prior group membership in the classification step:

(a) Sample the cluster-specific transition matrices $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_H$ given $\mathbf{S}$. The various rows $\boldsymbol{\xi}_{h,j\cdot}$ of the transition matrices $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_H$ are conditionally independent and may be sampled

line-by-line from a total of KH Dirichlet distributions:

$$\boldsymbol{\xi}_{h,j\cdot}|\mathbf{S},\mathbf{y} \sim \mathcal{D}\left(e_{0,j1}+N_{j1}^h(\mathbf{S}),\ldots,e_{0,jK}+N_{jK}^h(\mathbf{S})\right) \quad j=1,\ldots,K,\ h=1,\ldots,H, \tag{8}$$

where $N_{j1}^h(\mathbf{S}) = \sum_{i:S_i=h} N_{i,jk}$ is the total number of transitions from j to k observed in group h and is determined from the transitions $N_{i,jk}$ for all individuals falling into that particular group.

(b) Sample parameters $\boldsymbol{\beta}_2,\ldots,\boldsymbol{\beta}_H$ given $\mathbf{S}$: draw $\boldsymbol{\beta}$ from the multinomial logit model (2) using auxiliary mixture sampling in the dRUM representation (Frühwirth-Schnatter and Frühwirth, 2010).

(c) Bayes' classification for each individual i: draw $S_i, i=1,\ldots,N$ from the following discrete probability distribution which combines the likelihood $p(\mathbf{y}_i|\boldsymbol{\xi}_h)$ and the prior (2))

$$\Pr(S_i=h|\mathbf{y}_i,\mathbf{x}_i,\boldsymbol{\beta},\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_H) \propto p(\mathbf{y}_i|\boldsymbol{\xi}_h)\ \frac{\exp\left(\mathbf{x}_i\boldsymbol{\beta}_h\right)}{1+\sum_{l=2}^H \exp\left(\mathbf{x}_i\boldsymbol{\beta}_l\right)}, \quad h=1,\ldots,H. \tag{9}$$

For details on MCMC inference in general, we refer to standard monographs like Geweke (2005) and Gamerman and Lopes (2006). In step (b), we apply a very efficient method of auxiliary mixture sampling introduced by Frühwirth-Schnatter and Frühwirth (2010), see Appendix A for details.

We start MCMC estimation by choosing initial values for the group-indicators $\mathbf{S}$ in one of the following ways: non-random initial clustering such as $\mathbf{S}=(1,\ldots,H,1,\ldots,H,\ldots)$, random initial clustering by sampling S_i from $(1,...,H)$ with replacement, or k-means clustering (as implemented in R) of the transition frequencies observed for each individual.

3.3.3 Dealing with Label Switching

Like for any finite mixture model, label switching may occur during MCMC sampling, see Jasra et al. (2005) or Frühwirth-Schnatter (2006, Section 3.5) for a recent review. Pamminger and Frühwirth-Schnatter (2010) followed Frühwirth-Schnatter (2006, p. 96f) to identify the finite mixture model, by applying k-means clustering to all MH posterior draws of the vector $\boldsymbol{z}_{m,h} = (\xi_{h,11}^{(m)},\ldots,\xi_{h,KK}^{(m)})^T$ containing the posterior draws of the group-specific persistence probabilities.

Provided that the simulation clusters in the point process representation of the MCMC draws are well-separated, the classification sequence $(d_1^{(m)}, \ldots, d_H^{(m)})$ corresponding to $(\boldsymbol{z}_{m,1}, \ldots, \boldsymbol{z}_{m,H})$ is a permutation of the labels $\{1, \ldots, H\}$. This classification sequence is used for each $m = 1, \ldots, M$ to relabel the H MCMC draws $(\boldsymbol{\vartheta}_1, \eta_1)^{(m)}, \ldots, (\boldsymbol{\vartheta}_H, \eta_H)^{(m)}$. The same permutation is used to relabel the MCMC draws $\mathbf{S}^{(m)} = (S_1^{(m)}, \ldots, S_N^{(m)})$ of the hidden group indicators. Since this method is unaffected by the mixture-of-expert extension, it may be applied without modifications to our extension.

3.4 Selecting the Number of Clusters

Despite much research effort, it is still an open issue how to select the number H of clusters in an optimal manner. The difficulties with identifying H are particularly well-documented for the BIC criterion (Schwarz, 1978) $BIC(H) = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_H) + d_H \log n$, where $\hat{\boldsymbol{\theta}}_H$ is the ML estimator of $\boldsymbol{\theta}_H = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_H, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_H)$, $p(\mathbf{y}|\boldsymbol{\theta}_H)$ denotes the likelihood function, $\hat{\boldsymbol{\theta}}_H$ is the ML estimator, and d_H is the number of parameters in a model with H clusters. Since the mixture-of-experts model is applied to panel data it is not obvious how to choose the sample size n (Kass and Raftery, Jun., 1995). As each time series is modeled independently, the number N of time series is a natural choice for the sample size, i.e. $n = N$. On the other hand, since multiple observations are available for each time series, one might prefer the total number of observations as sample size, i.e. $n = \sum_{i=1}^{N} T_i$.

The AIC criterion (Akaike, 1974) defined by $AIC(H) = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_H) + 2\, d_H$ is independent of the sample size, but is well-known to be inconsistent and leads to overfitting the number of clusters H. $BIC(H)$ is known to be consistent for the number of components, if the component density is correctly specified (Keribin, 2000), although in small data sets it tends to choose models with too few components (Biernacki et al., 2000). On the other hand, simulation studies reported in Biernacki and Govaert (1997), Biernacki et al. (2000), and McLachlan and Peel (2000, Section 6.11) show that $BIC(H)$ will overrate the number of clusters under misspecification of the component density.

Since $BIC(H)$ is an asymptotic approximation to minus twice the marginal likelihood $-2 \log p(\mathbf{y}|H)$, see e.g. Kass and Raftery (Jun., 1995), it is not surprising that selecting H as to maximize the marginal likelihood $p(\mathbf{y}|H)$ or the posterior probability distribution

$p(H|\mathbf{y}) \propto p(\mathbf{y}|H)p(H)$ may not be adequate either, as demonstrated in various applications of model-based clustering, see e.g. Frühwirth-Schnatter and Pyne (2010).

A criterion that was found to be able to identify the correct number of clusters even when the component densities are misspecified is the approximate weight of evidence $AWE(H)$ (Banfield and Raftery, 1993). Biernacki and Govaert (1997) expressed $AWE(H)$ as a criterion which penalizes the complete data log-likelihood function $p(\mathbf{y}, \mathbf{S}|\boldsymbol{\theta}_H)$ with model complexity, i.e $AWE(H) = -2 \log p(\mathbf{y}, \hat{\mathbf{S}}|\hat{\boldsymbol{\theta}}_H^C) + 2\, d_H(\frac{3}{2} + \log n)$, where $(\hat{\boldsymbol{\theta}}_H^C, \hat{\mathbf{S}})$ maximizes $\log p(\mathbf{y}, \mathbf{S}|\boldsymbol{\theta}_H)$.

Various criteria involve the entropy $EN(H, \boldsymbol{\theta}_H) = -\sum_{h=1}^{H}\sum_{i=1}^{N} t_{ih}(\boldsymbol{\theta}_H) \log t_{ih}(\boldsymbol{\theta}_H)$, where $t_{ih}(\boldsymbol{\theta}_H) = \Pr(S_i = h|\mathbf{y}_i, \boldsymbol{\theta}_H)$ is the posterior classification probability defined in (9). The entropy is close to 0 if the resulting clusters are well-separated and increases with increasing overlap of the clusters. The CLC criterion (Biernacki and Govaert, 1997), for instance, penalizes the log likelihood function by the entropy rather than by model complexity, i.e. $CLC(H) = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_H) + 2\, EN(H, \hat{\boldsymbol{\theta}}_H)$. However, the CLC criterion works well only for well-separated clusters with a fixed weight distribution, hence its properties are not known for the more general mixture-of-experts model.

The $ICL\text{-}BIC$ criterion (McLachlan and Peel, 2000) penalizes the log likelihood function both by model complexity and the entropy, i.e. $ICL\text{-}BIC(H) = BIC(H) + 2\, EN(H, \hat{\boldsymbol{\theta}}_H)$. Simulation studies in McLachlan and Peel (2000, Section 6.11) indicate that $ICL\text{-}BIC$ may identify the correct number of clusters for (multivariate) continuous data even under a misspecified multivariate normal clustering kernel. However, simulation studies in Biernacki et al. (2008) show that this criterion tends to fail for discrete-valued data, even if the true model is used as clustering kernel.

For discrete-valued data, Biernacki et al. (2008) recommend to use the (exact) integrated classification likelihood (ICL) which is defined as $ICL(H) = \int p(\mathbf{y}, \hat{\mathbf{S}}|\boldsymbol{\theta}_H) p(\boldsymbol{\theta}_H|\mathbf{y}) d\,\boldsymbol{\theta}_H$, where $p(\mathbf{y}, \mathbf{S}|\boldsymbol{\theta}_H)$ is the complete-data likelihood function and $\hat{\mathbf{S}}$ This criterion showed good performance for latent class models. For Markov chain clustering with the mixture-of-expert extension the $ICL(H)$ reads:

$$ICL(H) = p(\hat{\mathbf{S}}) \prod_{j=1}^{K} \left(\frac{\Gamma(\sum_{k=1}^{K} e_{0,jk})}{\prod_{k=1}^{K} \Gamma(e_{0,jk})} \right)^{H} \prod_{h=1}^{H} \frac{\prod_{k=1}^{K} \Gamma(N_{jk}^{h}(\hat{\mathbf{S}}) + e_{0,jk})}{\Gamma(\sum_{k=1}^{K}(N_{jk}^{h}(\hat{\mathbf{S}}) + e_{0,jk}))}, \tag{10}$$

where the integral $p(\hat{\mathbf{S}}) = \int p(\hat{\mathbf{S}}|\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_H) p(\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_H) d\,\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_H$ is approximated by importance sampling where for each $h = 2, \ldots, H$ a multivariate normal distribution is used as a proposal density for $\boldsymbol{\beta}_h$ where the mean and covariance matrix are set to the corresponding MCMC sample estimates.

4 Results

To identify groups of individuals with similar wage career, we applied Markov chain clustering for 2 up to 5 groups. For each number H of groups we simulated 10 000 MCMC draws after a burn-in of 5 000 draws with a thinning parameter equal to 5.

4.1 Model Selection and Clustering

The model selection criteria described in Section 3.4 are applied to select the number H of clusters, see Figure 1.

AIC and BIC decrease with increasing H and suggests at least 5 components. However, as outlined in Section 3.4, we cannot expect that the Markov chain model is a perfect description of the component-specific distribution for time series in a real data panel. Thus it is likely that BIC is overfitting and that two or even more components in the mixture model correspond to a single cluster with rather similar transition behavior.

This hypothesis is supported by the other criteria; all of which suggest a smaller number of clusters. As described in Section 3.4, the evaluation of these criteria is based on approximate ML/MCL-estimators $\hat{\boldsymbol{\theta}}_H$ and $(\hat{\boldsymbol{\theta}}_H^C, \hat{\mathbf{S}})$ derived from all available MCMC draws. To check the stability of model choice we repeated several independent MCMC runs (see Figure 1). CLC, $ICL\text{-}BIC$ and particularly the (exact) ICL indicate three clusters for different MCMC runs. However, the AWE refers to a four-group solution which has also more importance from an economic point of view. We can easily interpret four different wage-mobility groups, which are characterized by the trend over time and the variability of earnings: an upward, a downward group as well as a static and a mobile group.

In the following, we concentrate on the four-cluster solution in more detail because this solution led to more sensible interpretations from an economic point of view. The model is

identified as described in Subsection 3.3.3 by applying k-means clustering to the MCMC draws. All classification sequences resulting from k-means clustering turned out to be permutations of $\{1, \ldots, 4\}$ and allowed straightforward identification of the four-components model.

Individuals are assigned to the four wage mobility groups using the posterior classification probabilities $t_{ih}(\boldsymbol{\theta}_H) = \Pr(S_i = h|\mathbf{y}_i, \boldsymbol{\theta}_H)$ for $H = 4$. The posterior expectation $\hat{t}_{ih} = \mathrm{E}(t_{ih}(\boldsymbol{\theta}_4)|\mathbf{y})$ of these probabilities is estimated by evaluating and averaging $t_{ih}(\boldsymbol{\theta}_4)$ over the last 10 000 MCMC draws of $\boldsymbol{\theta}_4$ with a thinning parameter equal to 5 (with effectively 2000 draws remaining). Each employee is then allocated to that cluster which exhibits the maximum posterior probability, i.e. $\hat{S}_i$ is defined in such a way that $\hat{t}_{i,\hat{S}_i} = \max_h \hat{t}_{i,h}$. The closer $\hat{t}_{i,\hat{S}_i}$ is to 1, the higher is the segmentation power for individual i.

4.2 Estimation Results

4.2.1 Analyzing Wage Mobility

To analyze wage mobility in the different clusters we investigate for each $h = 1, \ldots, 4$ the posterior expectation of the group-specific transition matrix $\boldsymbol{\xi}_h$. The four group-specific transition matrices are best visualized in Figure 2 using "balloon plots" [2]. The circles are proportional to the size of the corresponding entry in the transition matrix. Based on these transition matrices, we assign a labeling to each cluster, namely "upward", "static" "downward", and "mobile".

A remarkable difference in the transition behavior of individuals belonging to different clusters is evident from Figure 2. Consider, for instance, the first column of each matrix containing the risk for an individual in income category j to drop into the no-income category in the next year. This risk is much higher for the "downward" cluster than for the other clusters.

The probability to remain in the no-income category is located in the top left cell and is again higher in the "downward" cluster than in the other ones. The remaining probabilities in the first row correspond to the chance to move out of the no-income category. These chances are smaller for the "downward" cluster than for the other clusters. In the "upward" cluster chances are high to move into any wage category while in the "static" cluster only the chance to move in wage category one is comparatively high.

[2] They are generated with the function `balloonplot()` from the R package `gplots` (Jain and Warnes, 2006). Full numerical results together with standard deviations are in the Appendix.

For all matrices, the main diagonal refers to the probabilities to remain in the various wage categories. Persistence is highest in the "static" cluster. Members of the "mobile" cluster move quickly between the various wage categories. The upper secondary diagonal represents the chance to move forward into the next higher wage category, which is higher in the "upward" and "mobile" cluster than in the other clusters. On the other hand, the lower secondary diagonal - representing the risk to move into the next lower wage category - is stronger in the "downward" cluster.

Based on the posterior classification probabilities we can also calculate the size of the clusters: 29% of persons belong to the "static" cluster, 27% to the "upward" group and 25% to the "mobile" cluster; only 20% of male workers starting a career fall in the "downward" trap.

In Figure 3 these "balloon plots" are adjusted to show relative transition frequencies within groups: the entries in each matrix sum to one. We can see easily that the mass of individuals' transitions in the "upward" cluster lies in the bottom left corner, the reverse is true for the "downward" cluster. For the "static" group most individuals are located in the center and the lower quintiles, whereas in the "mobile" group the pattern is more diverse, but concentrated in the upper quintiles.

These differences in the transition matrices between the clusters have a strong impact on the long-run wage career of the group members, as shown by Figure 4. This figure starts for each cluster h with an initial wage distribution $\boldsymbol{\pi}_{h,0}$ at $t = 0$ which is estimated from the initial wage category y_{i0} observed for all individuals i being classified to group h. The posterior expectations $\mathrm{E}(\boldsymbol{\pi}_{h,t}|\mathbf{y}, \boldsymbol{\pi}_{h,0})$ of the cluster-specific wage distribution $\boldsymbol{\pi}_{h,t}$ after t years ($\boldsymbol{\pi}_{h,t} = \boldsymbol{\pi}_{h,0}\boldsymbol{\xi}_h^t$) are shown for several periods as well as the steady state. [3]

For $t = 100$, the wage distribution is already practically equal to the steady state $\boldsymbol{\pi}_{h,\infty}$ of the transition matrix $\boldsymbol{\xi}_h$, i.e. $\boldsymbol{\pi}_{h,\infty} = \boldsymbol{\pi}_{h,\infty}\boldsymbol{\xi}_h$. In the "downward" cluster the steady state is reached after only a few years, whereas in the other three clusters it takes one to two decades.

The wage distributions shown in Figure 4 are consistent with our labeling of the clusters introduced earlier. Young men belonging to the "downward" cluster have a much higher risk to start in the no-income category then any other young men. Furthermore, about 40 % of the

[3]The posterior expectation is estimated by averaging MCMC draws of $\boldsymbol{\pi}_{h,t}$ obtained by computing $\boldsymbol{\pi}_{h,t}$ for $t = 1, \ldots, 100$ for the last 10000 MCMC draws with a thinning parameter equal to 5 (with effectively 2000 draws remaining) of $\boldsymbol{\xi}_h$.

members of this group have no income in the long-run. For young men belonging either to the "mobile" or the "upward" cluster there is little difference between the initial wage distribution when they enter the labor market. However, in the long run the pattern diverges considerably: while the members of the "upward" cluster gather themselves in the upmost quintiles, those from the "mobile" cluster are to be seen in the middle of the wage distribution. Members from the "static" cluster end up in a very balanced steady state.

4.2.2 Posterior Classification

Table 2 analyzes the segmentation power for the clustering method by reporting the quartiles and the median of classification probabilities $\hat{t}_{i,\hat{S}_i}$ defined in Subsection 3.3.3 within the various groups as well as for all individuals. We find that the overall segmentation power is rather high. 3 out of 4 individuals are assigned with at least 63.8% to their respective groups. For 1 out of 4 individuals assignment probability amounts to at least 97.5%. Segmentation power varies between the clusters and is the highest for the "upward" cluster and the lowest for the "mobile" cluster.

4.2.3 The Impact of Observables on Group Membership

The previous clustering analysis was more descriptive, specifying common mobility patterns of certain groups in the labor market. From an economic point of view, it is interesting to understand, what characteristics of a particular person makes him more prone to fall into one or the other cluster. Moreover, our main question is: do random differences in the labor market situation at the time of entry in the labor force have a long-run impact on mobility behavior of workers? We model the prior probability of an individual to belong to a certain cluster by the multinomial logit model specified in equation (7). The estimation results are presented using the "upward" cluster as baseline.

As discussed above, we capture the general labor market situation at the time of entry into the labor market by the unemployment rate in the district together with a set of yearly time dummies to control for unspecified time trends. Further we allow for impacts of educational categories and the type of occupation on mobility patterns. To model the correlation between group membership and initial earnings categories in period zero, we add dummies for the wage

category at entry with non-employment or zero income serving as baseline. Correlation between labor market entry conditiona and entyr wages are captured by interaction terms between these dummies and the unemployment rate.

Bayesian inference for the regression parameters in this multinomial logit model is summarized in Table 3, which reports the posterior expectations and the posterior standard deviations of all regression parameters. The results show that, indeed, bad economic conditions at the time of entry reduce the probability of an individual to end up in the favorable "upward" cluster. Individuals are almost equally shifted towards one of the three other clusters. This result is remarkable because other studies were primarily concerned with short-run impacts of a bad start, whereas differen mobility patterns are a typical long-run phenomenon.

The other results are mostly according to expectations: individuals starting in white-collar jobs are most likely to end up in "upward" clusters and least likely in "downward" clusters. The picture is less clear for our skill categories: while skilled workers are most likely to be classified in the "upward" cluster, the unskilled are most likely to be in the "static" cluster and least likely to be in the "upward" and in particular the "mobile" cluster.

We include dummy variables to indicate in which wage quintile the worker started his first job to control for initial conditions. The initial earnings category is an important determinant of group membership, which implies that there is substantial correlation between unobserved heterogeneity and initial conditions. The coefficients are fairly consistent in the sense that starting in a high wage quintile makes it much less likely to end up in the "downward" or the "static " cluster; there is no consistent pattern relating the starting wage with either being in the "mobile " or the "upward" cluster, though. No clear pattern emerges from the interaction terms between unemployment rate and initial earnings categories. Those terms are included mainly to allow for arbitrary correlations between the initial conditions and the covariates influencing group membership, therefore we do not give them any interpretation. We note, however, that the inclusion of the interaction terms has a significant impact.

5 Conclusions

In this paper we discussed an approach to model-based clustering of categorical time series based on time-homogeneous first-order Markov chains with unknown transition matrices. In the Markov chain clustering approach the individual transition probabilities are fixed to a group-specific transition matrix.

We discussed in detail an application of this approach to modeling and clustering a panel of Austrian wage mobility data describing the wage career of nearly 50 000 young men entering the labor market between 1975 and 1985. Model choice indicated in terms of posterior probability (approximated by BIC) that for this cohort the labor market should be segmented into three/four groups. The group-specific transition behavior turned out to be very different across the clusters and led to a meaningful interpretation from an economic point of view showing four types of wage careers, namely "mobile" , "downward", "upward", and "static".

We investigated the segmentation power of the four-group solution and found that it is rather high. 3 out of 4 individuals are assigned with at least 63.8% probability to their respective cluster.

We conclude from our investigation that this clustering kernel is a sensible tool for model-based clustering of discrete-valued panel data.

For other panels of discrete-valued time series other clustering kernels might be sensible. More complex clustering kernels could involve the use of kth order Markov chains in order to extend the memory of the clustering kernel to the past k observations, see e.g. Saul and Jordan (1999). MCMC estimation as discussed in this paper is easily extended to this case. Another promising alternative is to use inhomogeneous Markov chains as clustering kernels. This method could be based on modeling each row of the transition matrix through a dynamic multinomial logit model with random effects.

Using a dynamic multinomial logit model with random effects as clustering kernel has the advantage that it allows to include subject-specific as well as aggregate economic covariates and, at the same time, is able to capture first or even higher order dependence by including past observations of the time series as covariates.

However, MCMC estimation of a model where the clustering kernel is a dynamic multinomial logit model with random effects is much more involved, because no explicit expression for the

marginal distribution is available, and we leave this for future research.

Acknowledgements

This research is supported by the Austrian Science Foundation (FWF) under the grant S 10309-G14 (NRN "The Austrian Center for Labor Economics and the Analysis of the Welfare State", Subproject "Bayesian Econometrics") as well as the Austrian Marshallplan Jubiläumsstiftung.

References

Akaike, H., 1974. A new look at statistical model identification. IEEE Transactions on Automatic Control 19, 716–723.

Banfield, J. D., Raftery, A. D., 1993. Model-based gaussian and non-gaussian clustering. Biometrics 49, 803–821.

Biernacki, C., Celeux, G., Govaert, G., 2000. Assessing a mixture model for clustering with the integrated classification likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 719–725.

Biernacki, C., Celeux, G., Govaert, G., 2008. Exact and monte carlo calculations of integrated likelihoods for the latent class model. Tech. rep., Research report No 6609, INRIA, Sophia Antipolis, France.

Biernacki, C., Govaert, G., 1997. Using the classification likelihood to choose the number of clusters. Computing Science and Statistics 29, 451–457.

Brunner, B., Kuhn, A., 2009. To shape the future: How labor market entry conditions affect individuals' long-run wage profiles, working Paper 0929, NRN: The Austrian Center for Labor Economics and the Analysis of the Welfare State.

Burgess, S., Propper, C., Rees, H., Shearer, A., 2003. The class of 1981: the effects of early career unemployment on subsequent unemployment experiences. Labour Economics 10 (3), 291–309.

Fougère, D., Kamionka, T., 2003. Bayesian inference for the mover-stayer model in continuous time with an application to labour market transition data. Journal of Applied Econometrics 18 (6), 697–723.

Fraley, C., Raftery, A. E., 2002. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97 (458), 611–631.

Frühwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer Series in Statistics. New York/Berlin/Heidelberg: Springer.

Frühwirth-Schnatter, S., Frühwirth, R., 2007. Auxiliary mixture sampling with applications to logistic models. Computational Statistics and Data Analysis 51, 3509–3528.

Frühwirth-Schnatter, S., Frühwirth, R., 2010. Data augmentation and MCMC for binary and multinomial logit models. In: Kneib, T., Tutz, G. (Eds.), Statistical Modelling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir. Physica-Verlag, Heidelberg, pp. 111–132, also available at `http://www.ifas.jku.at/e2550/e2756/index_ger.html`, IFAS Research Paper Series 2010-48.

Frühwirth-Schnatter, S., Kaufmann, S., 2008. Model-based clustering of multiple time series. Journal of Business & Economic Statistics 26, 78–89.

Frühwirth-Schnatter, S., Pyne, S., 2010. Bayesian Inference for Finite Mixtures of Univariate and Multivariate Skew Normal and Skew-t Distributions. Biostatistics 11, 317 – 336, doi: 10.1093/biostatistics/kxp062.

Frydman, H., 2005. Estimation in the mixture of markov chains moving with different speeds. Journal of the American Statistical Association 100, 1046–1053.

Gamerman, D., Lopes, H. F., 2006. Markov Chain Monte Carlo. Stochastic Simulation for Bayesian Inference, 2nd Edition. Boca Raton, FL: Chapman & Hall/CRC.

Genda, Y., Kondo, A., Otha, S., 2010. Long-term effects of a recession at labor market entry in japan and the united states. Journal of Human Resources 45, 157–196.

Geweke, J., 2005. Contemporary Bayesian Econometrics and Statistics. Wiley.

Heckman, J., 1981. The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process. In: Manski, C. F., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge MA, pp. 179–195.

Holmes, C. C., Held, L., 2006. Bayesian auxiliary variable models for binary and multinomial regression. Bayesian Analysis 1, 145–168.

Jain, N., Warnes, G. R., 2006. Balloon plot. R News 6 (2), 35–38.

Jasra, A., Holmes, C. C., Stephens, D. A., 2005. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling. Statistical Science 20, 50–67.

Juárez, M. A., Steel, M. F. J., 2010. Model-based clustering of non-gaussian panel data based on skew-t distributions. Journal of Business & Economic Statistics 28 (1), 52–66.

Kahn, L. B., 2009. The long-term labor market consequences of graduating from college in a bad economy. Labour Economics 17 (2), 303–316.

Kass, R. E., Raftery, A. E., Jun., 1995. Bayes factors. Journal of the American Statistical Association 90 (430), 773–795.

Keribin, C., 2000. Consistent estimation of the order of mixture models. Sankhya A 62, 49–66.

Kondo, A., 2007. Does the first job really matter? state dependency in employment status in japan. Journal of the Japanese and International Economies 21 (3), 379–402.

Kwon, I., Meyersson-Milgrom, E. M., 2007. Cohort effects in wages and promotions. Tech. Rep. 07-25, SIEPR Discussion Paper.

McFadden, D., 1974. Frontiers of Econometrics. New York: Academic, Ch. Conditional logit analysis of qualitative choice behaviour, pp. 105–142.

McLachlan, G. J., Peel, D., 2000. Finite Mixture Models. Wiley Series in Probability and Statistics. New York: Wiley.

Oreopoulos, P., von Wachter, T., Heisz, A., June 2008. The short- and long-term career effects of graduating in a recession: Hysteresis and heterogeneity in the market for college graduates, nBER Working Paper 12159.

Oyer, P., Summer 2006. Initial labor market conditions and long-term outcomes for economists. Journal of Economic Perspectives 20 (3), 143160.

Pamminger, C., Frühwirth-Schnatter, S., 2010. Model-based Clustering of Categorical Time Series. Bayesian Analysis 5, 345–368.

Raaum, O., Roed, K., 2006. Do business cycle conditions at the time of labor market entry affect future employment prospects? The Review of Economics and Statistics 88(2), 193–210.

Saul, L. K., Jordan, M. I., 1999. Mixed memory markov models: Decomposing complex stochastic processes as mixture of simpler ones. Machine Learning 37, 75–87.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461–464.

Scott, S. L., 2009. Data augmentation and the Bayesian analysis of multinomial logit models. Statistical Papers XX, accepted for publication.

Weber, A., 2001. State dependence and wage dynamics: A heterogeneous markov chain model for wage mobility in austria. Research report, Institute for Advanced Studies, Vienna.

Welch, F., 1979. Effects of cohort size on earnings: The baby boom babies' financial bust. The Journal of Political Economy 87 (5), S65–S97.

Wooldridge, J. M., 2005. Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. Journal of Applied Econometrics 20, 39–54.

Zweimüller, J., Winter-Ebmer, R., Lalive, R., Kuhn, A., Wuellrich, J.-P., Ruf, O., Büchi, S., 2009. The austrian social security database (ASSD). Tech. rep., Working Paper 0903, NRN: The Austrian Center for Labor Economics and the Analysis of the Welfare State, Linz, Austria.

6 Tables

			"upward"			
	0	1	2	3	4	5
0	0.3594(0.736)	0.1364(0.696)	0.0970(0.353)	0.1105(0.318)	0.1481(0.363)	0.1486(0.364)
1	0.1255(0.574)	0.5970(1.330)	0.1570(0.623)	0.0644(0.289)	0.0396(0.198)	0.0166(0.109)
2	0.0768(0.283)	0.0596(0.281)	0.4318(0.706)	0.3179(0.520)	0.0922(0.274)	0.0217(0.120)
3	0.0610(0.195)	0.0172(0.108)	0.0567(0.187)	0.4744(0.445)	0.3482(0.396)	0.0424(0.152)
4	0.0490(0.123)	0.0075(0.047)	0.0093(0.051)	0.0482(0.125)	0.6419(0.305)	0.2441(0.265)
5	0.0481(0.065)	0.0026(0.015)	0.0014(0.012)	0.0032(0.018)	0.0381(0.068)	0.9065(0.097)
			"static"			
	0	1	2	3	4	5
0	0.3548(0.938)	0.4191(0.783)	0.1595(0.523)	0.0540(0.282)	0.0123(0.129)	0.0003(0.021)
1	0.1120(0.241)	0.7421(0.334)	0.1278(0.232)	0.0143(0.065)	0.0035(0.029)	0.0004(0.010)
2	0.0518(0.128)	0.0745(0.149)	0.7318(0.265)	0.1341(0.204)	0.0075(0.049)	0.0004(0.012)
3	0.0361(0.116)	0.0144(0.074)	0.0822(0.197)	0.7554(0.298)	0.1105(0.232)	0.0013(0.026)
4	0.0362(0.138)	0.0052(0.055)	0.0062(0.062)	0.0556(0.253)	0.8456(0.318)	0.0512(0.218)
5	0.0430(0.247)	0.0015(0.051)	0.0015(0.054)	0.0012(0.055)	0.0308(0.365)	0.9219(0.474)
			"downward"			
	0	1	2	3	4	5
0	0.5749(0.334)	0.2456(0.320)	0.1027(0.183)	0.0527(0.134)	0.0209(0.088)	0.0032(0.034)
1	0.3523(0.509)	0.4834(0.643)	0.1161(0.290)	0.0344(0.125)	0.0126(0.068)	0.0011(0.020)
2	0.2699(0.454)	0.1678(0.348)	0.4084(0.611)	0.1263(0.311)	0.0253(0.137)	0.0024(0.039)
3	0.2406(0.521)	0.0794(0.287)	0.1746(0.444)	0.3804(0.660)	0.1172(0.410)	0.0077(0.093)
4	0.2196(0.701)	0.0580(0.373)	0.0607(0.372)	0.2167(0.687)	0.3967(1.078)	0.0483(0.396)
5	0.2551(1.884)	0.0275(0.625)	0.0367(0.711)	0.0805(1.039)	0.2365(1.825)	0.3638(2.740)
			"mobile"			
	0	1	2	3	4	5
0	0.2914(0.809)	0.1469(0.509)	0.2381(0.535)	0.2078(0.423)	0.1074(0.308)	0.0084(0.087)
1	0.2143(0.639)	0.3524(1.054)	0.2901(0.631)	0.1048(0.341)	0.0356(0.179)	0.0027(0.046)
2	0.1006(0.242)	0.0797(0.205)	0.5315(0.530)	0.2478(0.350)	0.0381(0.134)	0.0023(0.030)
3	0.0666(0.148)	0.0198(0.072)	0.1037(0.168)	0.6153(0.352)	0.1907(0.232)	0.0039(0.032)
4	0.0531(0.120)	0.0080(0.043)	0.0154(0.061)	0.1233(0.192)	0.7341(0.277)	0.0662(0.155)
5	0.0453(0.285)	0.0042(0.079)	0.0103(0.123)	0.0215(0.179)	0.3432(0.983)	0.5755(1.111)

Table 1: Posterior expectation $\mathrm{E}(\boldsymbol{\xi}_h|\mathbf{y})$ and, in parenthesis, posterior standard deviations $\mathrm{SD}(\boldsymbol{\xi}_h|\mathbf{y})$ (multiplied by 100) of the average transition matrix $\boldsymbol{\xi}_h$ in the various clusters.

	Markov chain clustering		
	1st Qu.	Median	3rd Qu.
"upward"	0.7751	0.9552	0.9940
"static"	0.6009	0.7977	0.9558
"downward"	0.6272	0.8538	0.9727
"mobile"	0.6042	0.7851	0.9337
overall	0.6378	0.8532	0.9746

Table 2: Segmentation power of Markov chain clustering; reported are the lower quartile, the median and the upper quartile of the individual posterior classification probabilities $\hat{t}_{i,\hat{S}_i}$ for all individuals within a certain cluster as well as for all individuals.

	"static"	"downward"	"mobile"
Intercept	1.08723 (0.11849)	0.80076 (0.11773)	1.10707 (0.13555)
Unemployment rate in district	0.14118 (0.02449)	0.13051 (0.02434)	0.12481 (0.02743)
Unskilled	0.27972 (0.07014)	0.95308 (0.06825)	-0.79275 (0.09425)
Skilled	-1.30045 (0.04716)	-1.05160 (0.04638)	-1.98995 (0.05165)
White collar	-1.63902 (0.04242)	-2.25963 (0.04712)	-2.27425 (0.05522)
Start in wage category 1	0.79487 (0.10522)	0.24602 (0.10373)	0.74447 (0.12897)
Start in wage category 2	-0.05383 (0.12918)	-0.12639 (0.12353)	0.72716 (0.14039)
Start in wage category 3	-0.85094 (0.17030)	-0.80229 (0.16030)	0.46321 (0.15584)
Start in wage category 4	-0.93842 (0.33328)	-0.80421 (0.21876)	0.21289 (0.19899)
Start in wage category 5	-0.80603 (0.69196)	-0.72659 (0.44669)	0.65145 (0.38215)
Start in year 1976	-0.49488 (0.09674)	-0.26228 (0.10153)	-0.56900 (0.10585)
Start in year 1977	-0.24680 (0.09231)	-0.07513 (0.09935)	-0.39870 (0.09816)
Start in year 1978	-0.26623 (0.09484)	-0.03402 (0.10175)	-0.29910 (0.10375)
Start in year 1979	-0.19094 (0.09744)	0.02542 (0.10636)	-0.34746 (0.11084)
Start in year 1980	-0.07144 (0.09559)	0.19426 (0.10330)	-0.24927 (0.10355)
Start in year 1981	-0.21170 (0.10248)	0.16996 (0.10996)	-0.47084 (0.11714)
Start in year 1982	-0.44602 (0.12702)	-0.08256 (0.13638)	-0.63841 (0.14005)
Start in year 1983	-0.62936 (0.14637)	-0.18905 (0.15402)	-0.69356 (0.15943)
Start in year 1984	-0.40915 (0.15154)	0.00586 (0.15988)	-0.67097 (0.16587)
Start in year 1985	-0.56454 (0.15508)	0.00686 (0.16393)	-0.54190 (0.16968)
U rate * Wage C 1	-0.07307 (0.01972)	-0.08631 (0.01881)	-0.05062 (0.02365)
U rate * Wage C 2	-0.12715 (0.02388)	-0.16707 (0.02276)	-0.08512 (0.02563)
U rate * Wage C 3	-0.11664 (0.03391)	-0.13907 (0.03131)	-0.09320 (0.03023)
U rate * Wage C 4	-0.46343 (0.12115)	-0.22873 (0.04621)	-0.17130 (0.04005)
U rate * Wage C 5	-1.02326 (0.44505)	-0.39920 (0.11977)	-0.44577 (0.10191)

Table 3: Multinomial logit model to explain group membership in a particular cluster (baseline: "upward" cluster); the numbers are the posterior expectation and, in parenthesis, the posterior standard deviation of the various regression coefficients.

	upward	static	downward	mobile
0	15942	19369	57258	20060
1	18446	82325	43848	19575
2	15578	62597	23428	43057
3	22503	43974	14124	65905
4	47775	30495	6542	69959
5	125286	8840	987	9688
sum	245530	247600	146187	228244

Table 4: Marginal Distribution: Row sums of absolute transition frequencies within each group.

	upward	static	downward	mobile
0	6.49	7.82	39.17	8.79
1	7.51	33.25	29.99	8.58
2	6.34	25.28	16.03	18.86
3	9.17	17.76	9.66	28.87
4	19.46	12.32	4.48	30.65
5	51.03	3.57	0.68	4.24
sum	100.00	100.00	100.00	100.00

Table 5: Marginal Distribution: 'Relative' row sums of absolute transition frequencies within each group.

Figures

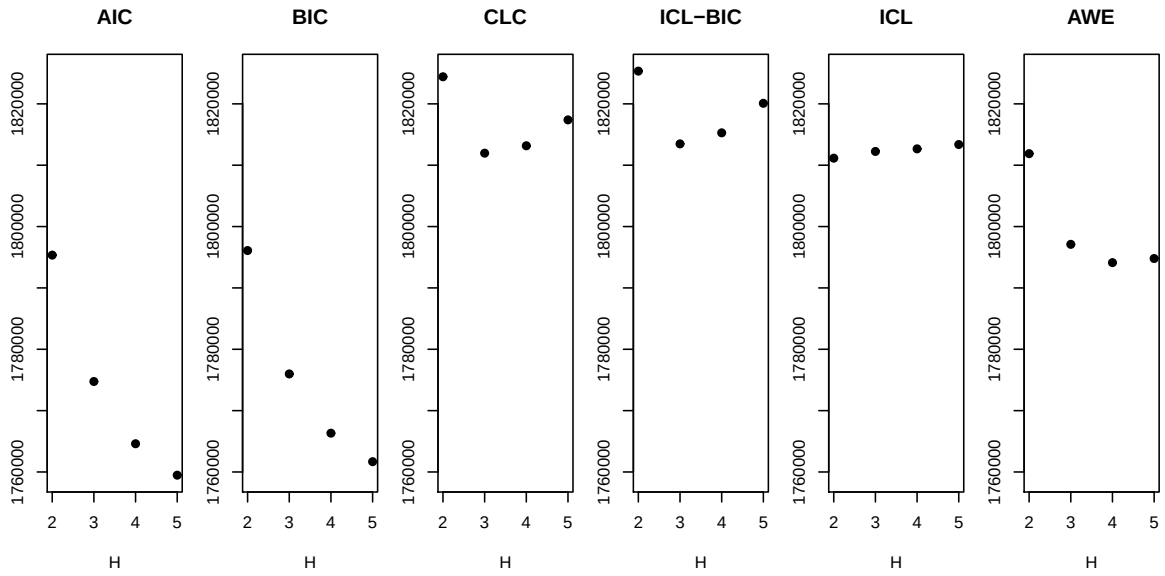

Figure 1: Model selection criteria for various numbers H of clusters.

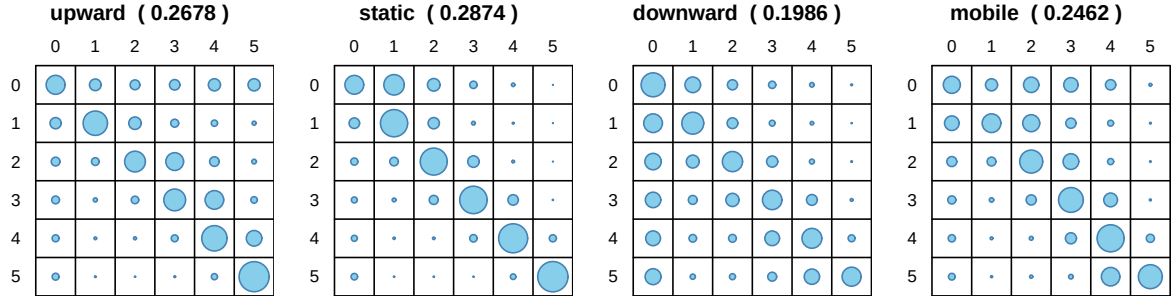

Figure 2: Visualization of posterior expectation of the transition matrices $\boldsymbol{\xi}_1$, $\boldsymbol{\xi}_2$, $\boldsymbol{\xi}_3$, and $\boldsymbol{\xi}_4$ obtained by Markov chain clustering. The circular areas are proportional to the size of the corresponding entry in the transition matrix. The corresponding group sizes are calculated based on the posterior classification probabilities and are indicated in the parenthesis.

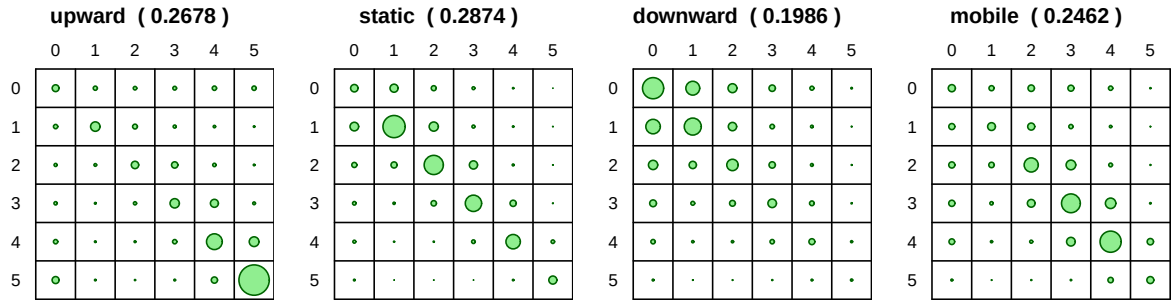

Figure 3: Balloonplots of relative transition frequencies (relative contingency table) within groups (each matrix sums to one!) to visualize the *relevance* of each single transition within each group.

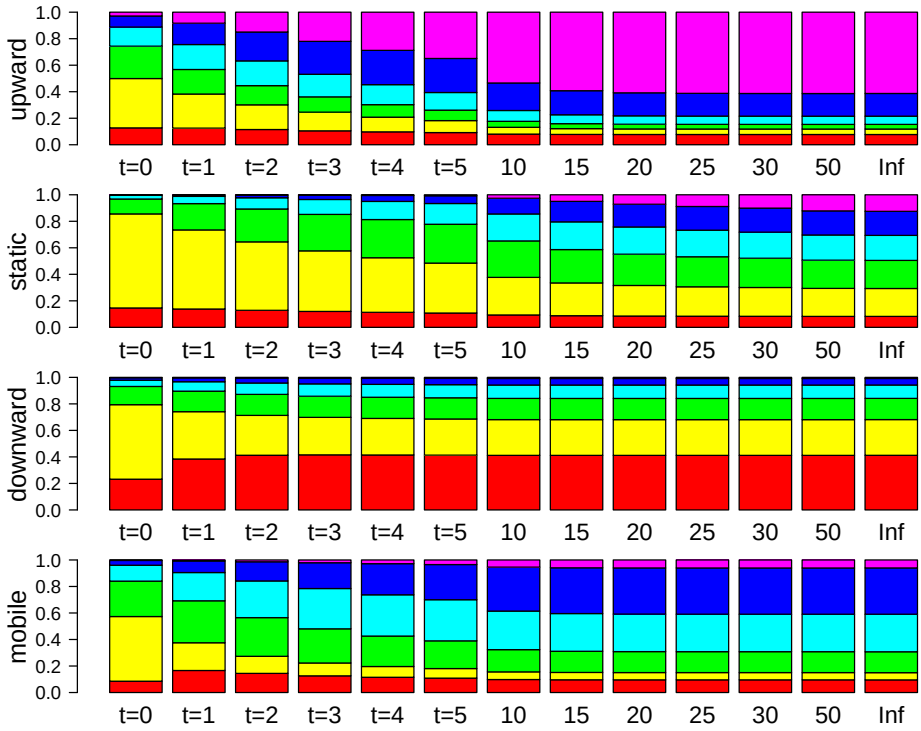

Figure 4: Posterior expectation of the wage distribution $\boldsymbol{\pi}_{h,t}$ over the wage categories 0 to 5 after a period of t years in the various clusters.

A Details on MCMC Estimation of the Mixture-of-Experts Models

A.1 Writing the MNL as Random Utility Model

The interpretation of an MNL as a random utility model (RUM) was introduced by McFadden (1974). Let y_{hi}^u be the utility of choosing category/group h, which is assumed to depend on

covariates $\mathbf{x}_i$:

$$y_{hi}^u = \mathbf{x}_i \boldsymbol{\beta}_h + \delta_{hi}, \tag{11}$$

$$S_i = h \Leftrightarrow y_{hi}^u = \max_{l \in 1,\ldots,H} y_{li}^u. \tag{12}$$

If $\delta_{1i}, \ldots, \delta_{Hi}$ are i.i.d. following a type I extreme value distribution, then the MNL (2) results as the marginal distribution of S_i.

An alternative way to write the MNL as an augmented model involving random utilities is as a differenced RUM (dRUM), which is obtained by choosing a baseline category (here $h_0 = 1$) and considering the model involving the differences of the utilities: $z_{hi} = \mathbf{x}_i \boldsymbol{\beta}_h + \varepsilon_{hi}$, where $z_{hi} = y_{hi}^u - y_{1i}^u$. Marginally, the errors $\varepsilon_{hi} = \delta_{hi} - \delta_{1i}$ follow a logistic distribution but are no longer independent across categories.

It has been shown by Frühwirth-Schnatter and Frühwirth (2010) that for each h, the MNL has the following representation as partial (binary) dRUM:

$$z_{hi} = \mathbf{x}_i \boldsymbol{\beta}_h - \log(\sum_{l \neq h} \lambda_{li}) + \varepsilon_{hi}, \tag{13}$$

where $\varepsilon_{hi}, h \neq 1$ are now i.i.d. following a logistic distribution.

A.2 2-Block Auxiliary Mixture Sampling

The logistic distribution can be approximated by a finite scale mixture of normal distributions with zero means and parameters (s_r^2, w_r). Using this approximation and conditinal on the latent utilities $\mathbf{z} = \{z_{2i}, \ldots, z_{Hi}, i = 1, \ldots, N\}$ and indicators $\mathbf{R} = \{r_{2i}, \ldots, r_{Hi}, i = 1, \ldots, N\}$ the dRUM (13) reduces to a Gaussian regression model:

$$z_{hi} = \mathbf{x}_i \boldsymbol{\beta}_h - \log(\sum_{l \neq h} \lambda_{li}) + \varepsilon_i, \quad \varepsilon_i | r_{hi} \sim \mathcal{N}(0, s_{r_{hi}}^2). \tag{14}$$

Based on this representation, step (b) of the MCMC scheme introduced in Subsection 3.3.2 is implemented in the following way:

(b-1) Sample the regression coefficients $\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_H$ conditional on $\mathbf{z}$ and $\mathbf{R}$ based on the normal

regression model (14). Using a normal prior (with known hyperparameters) the conditional posterior of $\boldsymbol{\beta}_h$ is also given by a multivariate normal density.

(b-2) Sample the latent variables z_{hi} and r_{hi} conditional on $\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_H$ and $\mathbf{S}$ for $i = 1, \ldots, N$ and $h = 2, \ldots, H$ with $\lambda_{hi} = \exp(\mathbf{x}_i \boldsymbol{\beta}_h)$:

(b-2-1) Sample all utilities $z_{2i}, \ldots, z_{Hi}$ simultaneously for each i from:

$$z_{hi} = \log(\lambda^*_{hi} U_{hi} + I\{S_i = h\}) - \log(1 - U_{hi} + \lambda^*_{hi} I\{S_i \neq h\})$$

where $U_{ih} \sim \mathcal{U}[0,1]$ and $\lambda^*_{hi} = \lambda_{hi} / (\sum_{l \neq h} \lambda_{li})$.

(b-2-2) Sample the component indicators r_{hi} conditional on z_{hi} from:

$$\Pr(r_{hi} = j | z_{hi}, \boldsymbol{\beta}_h) \propto \frac{w_j}{s_j} \exp\left\{ -\frac{1}{2} \left(\frac{z_{hi} - \mathbf{x}_i \boldsymbol{\beta}_h + \log(\sum_{l \neq h} \lambda_{li})}{s_j} \right)^2 \right\}$$

To start the MCMC scheme, one has to select starting values for $\mathbf{z}$ and $\mathbf{R}$.