Bayesian Clustering of Categorical Time Series Using Finite Mixtures of Markov Chain Models

by

Sylvia FRÜHWIRTH-SCHNATTER, Christoph PAMMINGER

Working Paper No. 0907

July 2009

Supported by the Austrian Science Funds

FWF

The Austrian Center for Labor Economics and the Analysis of the Welfare State

JKU Linz
Department of Applied Statistics
Altenberger Strasse 69
4040 Linz, Austria
www.labornrn.at

silvia.fruehwirth-schnatter@jku.at
phone +43 (0)70 2468 - 8295, - 9846 (fax)

Bayesian Clustering of Categorical Time Series Using Finite Mixtures of Markov Chain Models

Sylvia Frühwirth-Schnatter* Christoph Pamminger

Abstract

Two approaches for model-based clustering of categorical time series based on time-homogeneous first-order Markov chains are discussed. For Markov chain clustering the individual transition probabilities are fixed to a group-specific transition matrix. In a new approach called Dirichlet multinomial clustering the rows of the individual transition matrices deviate from the group mean and follow a Dirichlet distribution with unknown group-specific hyperparameters. Estimation is carried out through Markov chain Monte Carlo. Various well-known clustering criteria are applied to select the number of groups. An application to a panel of Austrian wage mobility data leads to an interesting segmentation of the Austrian labor market.

Keywords: Markov chain Monte Carlo, model-based clustering, panel data, transition matrices, labor market, wage mobility

*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

1 Introduction

In many areas of applied statistics like economics, finance or public health it is often desirable to find groups of similar time series in a set or panel of time series that are unlabeled a priori. To this aim, clustering techniques are required which determine subsets of similar time series within the panel. However, distance-based clustering methods cannot be easily extended to time series data, where an appropriate distance-measure is rather difficult to define, see e.g. the review by Liao (2005).

As opposed to that, 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). Recently, Juárez and Steel (2009) suggested to use skew-t distributions to capture skewness in the cluster-specific sampling density.

In the present paper we are interested in clustering discrete-valued time series which are considered as outcomes of a categorical variable with several states. In our econometric application in Section 5, we will study individual wage mobility in the Austrian labor market. Wage mobility describes chances but also risks of an individual to move between wage categories over time. The analysis is based on a panel reporting for young men entering the labor market between 1975 and 1980 their wage category in May of successive years. To give a more detailed picture of this panel several individual time series showing wage careers for a few employees are presented in Figure 1. The wage career is similar for some of them and quite different for others. The panel contains almost ten thousand of such wage careers and we are interested in searching for clusters of individuals with similar wage mobility behavior.

For such discrete-valued time series it is particularly difficult to define distance measures and model-based clustering appears to be a promising alternative. To capture the dynamics inherit in such data we consider two clustering kernels both of which are based on first-order time-homogeneous Markov chain models.

The first approach, called Markov chain clustering, assumes that for all time series within a cluster the transition behavior could be sufficiently described by the same cluster-specific transition matrix. Several papers found such an approach useful for clustering discrete-valued time series, see for instance Cadez et al. (2003) who clustered users according to their behavior on a web site and Ramoni et al. (2002) who clustered sensor data from mobile robots. Fougère and Kamionka (2003) 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.

Markov chain clustering could be viewed as fitting a dynamic multinomial model with cluster-specific parameters to each time series in the panel. While such a model allows the transition behavior to be different across clusters, it does not account for differences between individuals within a cluster. One way to capture unobserved heterogeneity within a cluster is to consider finite mixtures of random-effects models. Such models turned out to be useful in economic growth analysis, see e.g. Canova (2004) and Frühwirth-Schnatter and Kaufmann (2008), and in marketing research, see e.g. Lenk and DeSarbo (2000), Frühwirth-Schnatter et al. (2004), and Rossi et al. (2005). Our second clustering approach, called Dirichlet multinomial clustering, could be viewed as such a finite mixture of random-effects models, designed specifically to capture unobserved heterogeneity in the transition behavior across time series within a cluster. Such a model may be regarded as a mixture of Markov chain models where within each cluster the individual transition matrix of each time series deviates from an average group-specific transition matrix according to a Dirichlet distribution.

For estimation, we pursue a Bayesian approach which offers several advantages compared to EM estimation considered in Cadez et al. (2003) and 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 Ramoni et al. (2002) who applied a heuristic search method to find a good partition of the data. In the present paper we follow Frühwirth-Schnatter and Kaufmann (2008) and use a two-block Markov chain Monte Carlo sam-

pler for both clustering methods. A similar sampler has been used by Fougère and Kamionka (2003) for the special case of a mover-stayer model.

The remaining paper is organized as follows. Section 2 discusses Markov chain clustering as well as Dirichlet multinomial clustering. Bayesian estimation is considered in Section 3. In Section 4 we give a short review of some well-known criteria for selecting the number of groups. Model-based clustering is applied in Section 5 to a large panel of Austrian wage mobility data extending earlier work by Fougère and Kamionka (2003) for the French labor market. Section 6 concludes.

2 Clustering through Finite Mixtures of Markov Chain Models

2.1 Model-Based Clustering of Categorical Time Series

Let $\{y_{it}\}, t = 0, \ldots, T_i$ be a panel of categorical time series observed for N units $i = 1, \ldots, N$ where the number T_i of individual observations can vary from individual to individual. The observation y_{it} of individual i at time t arises from a categorical variable with K potential states labelled by $k \in \{1, \ldots, K\}$.

Model-based clustering as introduced by Frühwirth-Schnatter and Kaufmann (2008) is based on formulating a clustering kernel for an individual time series $\mathbf{y}_i = \{y_{i0}, \ldots, y_{i,T_i}\}$ in terms of a sampling density $p(\mathbf{y}_i|\boldsymbol{\vartheta})$, where $\boldsymbol{\vartheta}$ is an unknown model parameter. It is assumed that the N time series arise from H hidden groups, whereby within each group, say h, the clustering kernel $p(\mathbf{y}_i|\boldsymbol{\vartheta}_h)$ could be used for describing all time series in this group.

A latent group indicator S_i taking a value in the set $\{1, \ldots, H\}$ is introduced for each time series $\mathbf{y}_i$ to indicate which cluster the time series belongs to:

$$p(\mathbf{y}_i|S_i, \boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_H) = p(\mathbf{y}_i|\boldsymbol{\vartheta}_{S_i}) = \begin{cases} p(\mathbf{y}_i|\boldsymbol{\vartheta}_1), & \text{if } S_i = 1, \\ \vdots & \vdots \\ p(\mathbf{y}_i|\boldsymbol{\vartheta}_H), & \text{if } S_i = H. \end{cases} \quad (1)$$

It is assumed that $S_1, \ldots, S_N$ are a priori independent and $\Pr(S_i = h) = \eta_h$, where η_h is equal to the relative size of cluster h, i.e. $\sum_{h=1}^{H} \eta_h = 1$.

An important aspect of model-based clustering is that we do not assume to know a priori

which time series belong to which group and the group indicators $\mathbf{S} = (S_1, \ldots, S_N)$ are estimated along with the group-specific parameters $(\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_H)$ and the group sizes $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_H)$ from the data, see Section 3 for more details.

2.2 Markov Chain Clustering

An important building block for clustering discrete-valued time series is the first-order time-homogeneous Markov chain model characterized by the transition matrix $\boldsymbol{\xi}$, where

$$\xi_{jk} = \Pr(y_{it} = k | y_{i,t-1} = j), \qquad j, k = 1, \ldots, K \qquad \text{and} \quad \sum_{k=1}^{K} \xi_{jk} = 1. \tag{2}$$

ξ_{jk} represents the probability of the event that y_{it} takes the value k at time t given it took the value j at time $t-1$. Evidently, each row $\boldsymbol{\xi}_{j\cdot} = (\xi_{j1}, \ldots, \xi_{jK})$ of $\boldsymbol{\xi}$ represents a probability distribution over the discrete set $\{1, \ldots, K\}$. An individual time series $\mathbf{y}_i$ is said to be generated by a Markov chain model with transition matrix $\boldsymbol{\xi}$, if the sampling distribution $p(\mathbf{y}_i|\boldsymbol{\xi})$ of $\mathbf{y}_i$ given $\boldsymbol{\xi}$ reads:

$$p(\mathbf{y}_i|\boldsymbol{\xi}) = \prod_{t=1}^{T_i} p(y_{it}|y_{i,t-1}, \boldsymbol{\xi}) = \prod_{t=1}^{T_i} \xi_{y_{i,t-1}, y_{it}} = \prod_{j=1}^{K} \prod_{k=1}^{K} \xi_{jk}^{N_{i,jk}}, \tag{3}$$

where

$$N_{i,jk} = \#\{y_{it} = k, y_{i,t-1} = j\} \tag{4}$$

is the number of transitions from state j to state k observed in time series i. Note that in (3) we condition on the first observation y_{i0}.

Markov chain clustering is based on the assumption that within each cluster such a Markov chain model with group-specific transition matrix $\boldsymbol{\xi}_h$ could be used as clustering kernel. In the notation of Subsection 2.1 the group-specific parameter $\boldsymbol{\vartheta}_h$ is equal to $\boldsymbol{\xi}_h$ and the time series model $p(\mathbf{y}_i|\boldsymbol{\vartheta}_h)$ used for clustering in (1) is equal to the sampling distribution defined in (3):

$$p(\mathbf{y}_i|S_i = h, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_H) = p(\mathbf{y}_i|\boldsymbol{\xi}_h) = \prod_{j=1}^{K} \prod_{k=1}^{K} \xi_{h,jk}^{N_{i,jk}}. \tag{5}$$

A special version of this clustering method has been applied 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 we 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.

2.3 Finite Mixtures of Dynamic Random Coefficient Multinomial Logit Models

To extend Markov chain clustering, we rewrite the first-order time-homogeneous Markov chain model defined in (2) as a dynamic multinomial logit model:

$$\xi_{jk} = \Pr(y_{it} = k | y_{i,t-1} = j) = \frac{\exp(\gamma_{jk})}{\sum_{l=1}^{K} \exp(\gamma_{jl})}, \tag{6}$$

where the $(K \times K)$-dimensional transition matrix $\boldsymbol{\xi}$ is parameterized in terms of the $(K \times K)$-dimensional coefficient matrix $\boldsymbol{\gamma}$ with elements γ_{jk}. For each row of $\boldsymbol{\gamma}$, some normalization is required to achieve identifiability. It could be assumed, for instance, that $\gamma_{j,k_0} = 0$ for some baseline category k_0.

Next, we consider the random utility model (McFadden, 1974) corresponding to model (6):

$$\begin{aligned} u_{kit} &= \sum_{j=1}^{K} \gamma_{jk} I_{\{y_{i,t-1}=j\}} + \epsilon_{kit}, \quad k = 1, \ldots, K, \\ y_{it} = k &\Leftrightarrow u_{kit} = \max_{l \in \{1,\ldots,K\}} u_{lit}, \end{aligned} \tag{7}$$

where $\epsilon_{1it}, \ldots, \epsilon_{Kit}$ are independent random utility shocks each following a type-I extreme value distribution.

Thus the first-order time-homogeneous Markov chain model implies that any two individuals i and i' being in the same state j at time $t-1$ have exactly the same expected utility of moving

to category k. However, such a model appears to be too restrictive, because the expected utility of moving to category k is likely to depend on more factors than just the immediate past.

To account for unobserved heterogeneity in the individual transition behavior a dynamic random coefficient multinomial logit model may be considered where for each individual i the coefficient matrix $\boldsymbol{\gamma}_i^s$ is a random coefficient drawn from some distribution $p(\boldsymbol{\gamma}_i^s)$:

$$u_{kit} = \sum_{j=1}^{K} \gamma_{i,jk}^s I_{\{y_{i,t-1}=j\}} + \epsilon_{kit}, \quad k = 1, \ldots, K, \tag{8}$$

where $\gamma_{i,jk}^s$ is the (j,k)th element of $\boldsymbol{\gamma}_i^s$.

A crucial point in this model is the choice of the distribution of heterogeneity $p(\boldsymbol{\gamma}_i^s)$. Markov chain clustering as discussed in Subsection 2.2 corresponds to the assumption that $p(\boldsymbol{\gamma}_i^s)$ is a discrete distribution with H support points. To obtain more flexibility, one could follow Rossi et al. (2005) and assume that $p(\boldsymbol{\gamma}_i^s)$ may be described by a multivariate finite mixture of normal distributions, i.e.:

$$p(\boldsymbol{\gamma}_i^s) = \sum_{h=1}^{H} \eta_h p(\boldsymbol{\gamma}_i^s | \boldsymbol{\vartheta}_h), \tag{9}$$

where $p(\boldsymbol{\gamma}_i^s | \boldsymbol{\vartheta}_h)$ is the density of a multivariate normal distribution and the group-specific parameter $\boldsymbol{\vartheta}_h$ contains the unknown mean vector and all distinct parameters of the unknown variance-covariance matrix.

It is, in principle, possible to cluster panels of categorical time-series using such a multinomial model with random-effects as clustering kernel. However, in this general form the model involves the estimation of the covariance matrix of the distribution of random effects for each cluster and for this reason might be intractable for the purpose of clustering short individual time series with possibly many categories. To obtain a more parsimonious clustering kernel one could use constrained covariance matrices in the random effects distribution (9), like diagonal matrices. However, a general drawback of choosing a normal distribution as clustering kernel, either with an arbitrary or a constrained covariance matrix, is that the marginal distribution $p(\mathbf{y}_i | \boldsymbol{\vartheta}_h)$ is not available in closed form, because the integral $\int p(\mathbf{y}_i | \boldsymbol{\gamma}_i^s) p(\boldsymbol{\gamma}_i^s) d\boldsymbol{\gamma}_i^s$ cannot be solved analytically.

Subsequently, we consider a distribution of heterogeneity $p(\boldsymbol{\gamma}_i^s)$ which is a finite mixture

distribution as in (9), however, the group-specific distribution $p(\boldsymbol{\gamma}_i^s|\boldsymbol{\vartheta}_h)$ is different from Rossi et al. (2005) and is defined as follows. We assume that within each cluster h all elements $\gamma_{i,jk}^s$ of $\boldsymbol{\gamma}_i^s$ are independent random coefficients each following a log-Gamma distribution with a cluster and element-specific shape parameter $e_{h,jk}$ and common scaling parameter equal to 1:

$$\gamma_{i,jk}^s|S_i = h \sim \log \mathcal{G}\left(e_{h,jk}, 1\right), \quad j,k = 1,\ldots,K. \tag{10}$$

The mean and the variance of $\gamma_{i,jk}^s$ are given by

$$\mathrm{E}(\gamma_{i,jk}^s|S_i = h) = -\psi(e_{h,jk}), \quad \mathrm{Var}(\gamma_{i,jk}^s|S_i = h) = \psi'(e_{h,jk}), \tag{11}$$

where $\psi(s) = \Gamma'(s)/\Gamma(s)$ and $\Gamma(s)$ is the Gamma distribution, see e.g. Balakrishnan (1992, Appendix 18.2.A) or Frühwirth-Schnatter et al. (2009, Appendix A.1).

Note that we do not force identifiability, but leave all elements $\gamma_{i,jk}^s$ of the $(K \times K)$ matrix $\boldsymbol{\gamma}_i^s$ unconstrained. This distribution is, to a certain extent, related to a finite mixture of spherical normal distributions, however, the distribution of heterogeneity is skewed rather than symmetric. Choosing this group-specific distribution of heterogeneity has two distinct advantages compared to choosing a normal distribution as in Rossi et al. (2005). As will be discussed subsequently in Subsection 2.4, it leads to a closed form for the heterogeneity distribution of the individual transition matrices $\boldsymbol{\xi}_i^s$ corresponding to the coefficient matrix $\boldsymbol{\gamma}_i^s$ and allows a closed form expression for the clustering kernel, i.e. the distribution $p(\mathbf{y}_i|\boldsymbol{\vartheta}_h)$ of the individual times $\mathbf{y}_i$ given the group-specific parameter $\boldsymbol{\vartheta}_h$.

2.4 Dirichlet Multinomial Clustering

It is possible to rewrite the finite mixture of random coefficient multinomial logit models introduced in (8), (9) and (10) in the following way. Each individual time series $\mathbf{y}_i$ is generated by a Markov chain model with individual transition matrix $\boldsymbol{\xi}_i^s$ where the element $\xi_{i,jk}^s$ is determined by the jth row of the random coefficient matrix $\boldsymbol{\gamma}_i^s$:

$$\xi_{i,jk}^s = \frac{\exp(\gamma_{i,jk}^s)}{\sum_{l=1}^K \exp(\gamma_{i,jl}^s)}. \tag{12}$$

The sampling distribution of $\mathbf{y}_i$ given $\boldsymbol{\xi}_i^s$ is obtained from (3):

$$p(\mathbf{y}_i|\boldsymbol{\xi}_i^s) = \prod_{j=1}^{K}\prod_{k=1}^{K}(\xi_{i,jk}^s)^{N_{i,jk}}. \tag{13}$$

Furthermore, within each cluster h the heterogeneity distribution of the individual transition matrix $\boldsymbol{\xi}_i^s$ is available in closed form. The independence assumption for the elements of $\boldsymbol{\gamma}_i^s$ implies that in each cluster the rows $\boldsymbol{\xi}_{i,j\cdot}^s$ are independent apriori. From (10) we obtain for all elements in row $\boldsymbol{\xi}_{i,j\cdot}^s$ that $\exp(\gamma_{i,jk}^s)|S_i = h \sim \mathcal{G}(e_{h,jk}, 1)$. Hence it follows immediately from (12) that each row $\boldsymbol{\xi}_{i,j\cdot}^s$ follows a Dirichlet distribution with cluster-specific parameter $\mathbf{e}_{h,j\cdot} = (e_{h,j1}, \ldots, e_{h,jK})$:

$$\boldsymbol{\xi}_{i,j\cdot}^s|(S_i = h) \sim \mathcal{D}(e_{h,j1}, \ldots, e_{h,jK}), \qquad j = 1, \ldots, K. \tag{14}$$

For $H = 1$, this model is closely related to the Dirichlet multinomial model as for each row $\boldsymbol{\xi}_{i,j\cdot}^s$ of $\boldsymbol{\xi}_i^s$ the multinomial distribution for the number of transitions starting from state j is combined with a Dirichlet prior on the cell probabilities. For $H > 1$, such a Dirichlet multinomial model is used as clustering kernel, hence the method is called Dirichlet multinomial clustering. The group-specific parameter $\boldsymbol{\vartheta}_h$ is identical with the $(K \times K)$-dimensional parameter matrix $\mathbf{e}_h = \{\mathbf{e}_{h,j\cdot}, j = 1, \ldots, K\}$ appearing in (14).

A distinctive advantage of modeling the distribution of heterogeneity in this way is that the clustering kernel $p(\mathbf{y}_i|S_i = h, \mathbf{e}_1, \ldots, \mathbf{e}_H) = p(\mathbf{y}_i|\mathbf{e}_h)$ where $\boldsymbol{\xi}_i^s$ is integrated out is available in closed form. Hence, the clustering kernel may be entirely characterized by the group-specific parameter $\mathbf{e}_h$. This is easily verified by combining (13) and (14):

$$\begin{aligned} p(\mathbf{y}_i|\mathbf{e}_h) &= \int p(\mathbf{y}_i|\boldsymbol{\xi}_i^s)p(\boldsymbol{\xi}_i^s|\mathbf{e}_h)d\boldsymbol{\xi}_i^s = \\ &= \frac{\prod_{j=1}^{K}\Gamma(\sum_{k=1}^{K}e_{h,jk})}{\prod_{j=1}^{K}\prod_{k=1}^{K}\Gamma(e_{h,jk})}\int\prod_{k=1}^{K}\prod_{j=1}^{K}(\xi_{i,jk}^s)^{N_{i,jk}+e_{h,jk}-1}d\xi_{i,jk}^s = \\ &= \frac{\prod_{j=1}^{K}\Gamma(\sum_{k=1}^{K}e_{h,jk})}{\prod_{j=1}^{K}\prod_{k=1}^{K}\Gamma(e_{h,jk})}\frac{\prod_{j=1}^{K}\prod_{k=1}^{K}\Gamma(N_{i,jk}+e_{h,jk})}{\prod_{j=1}^{K}\Gamma(\sum_{k=1}^{K}(N_{i,jk}+e_{h,jk}))}. \end{aligned} \tag{15}$$

It is evident from (15) that this clustering kernel no longer is a first-order Markov process but allows for higher order dependence.

Next, we study the group-specific transition behavior implied by the parameter $\mathbf{e}_h$ in more detail. Each group may be characterized by the average group-specific transition matrix $\boldsymbol{\xi}_h$ given by the expected value of the individual transition matrix $\boldsymbol{\xi}_i^s$ in group h:

$$\xi_{h,jk} = \mathrm{E}(\xi_{i,jk}^s | S_i = h, \mathbf{e}_h) = \frac{e_{h,jk}}{\sum_{l=1}^K e_{h,jl}}. \tag{16}$$

From this formula it follows that each row of $\mathbf{e}_h$ determines the corresponding row in the group-specific transition matrix $\boldsymbol{\xi}_h$. The matrices $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_H$ may be compared to the corresponding matrices in the Markov chain clustering approach studied in Subsection 2.2.

While for Markov chain clustering the individual transition matrix $\boldsymbol{\xi}_i^s$ is equal to the group-specific transition matrix $\boldsymbol{\xi}_h$ for all individuals in group h, $\boldsymbol{\xi}_i^s$ is allowed to be different from $\boldsymbol{\xi}_h$ for Dirichlet multinomial clustering. The variability of $\boldsymbol{\xi}_i^s$ within each group is given by the variance of the individual transition probabilities $\xi_{i,jk}^s$:

$$\mathrm{Var}(\xi_{i,jk}^s | S_i = h, \mathbf{e}_h) = \frac{e_{h,jk} \sum_{l \neq k} e_{h,jl}}{\left(\sum_{l=1}^K e_{h,jl}\right)^2 \left(1 + \sum_{l=1}^K e_{h,jl}\right)}. \tag{17}$$

It can easily be shown that

$$\frac{\mathrm{Var}(\xi_{i,jk}^s | S_i = h, \mathbf{e}_h)}{\mathrm{E}(\xi_{i,jk}^s | S_i = h, \mathbf{e}_h)\ (1 - \mathrm{E}(\xi_{i,jk}^s | S_i = h, \mathbf{e}_h))} = \frac{1}{1 + \sum_{l=1}^K e_{h,jl}}. \tag{18}$$

As the right hand side of (18) is the same for all elements of row $\boldsymbol{\xi}_{i,j\cdot}^s$, a single parameter depending only on the row sum $\Sigma_{hj} = \sum_{l=1}^K e_{h,jl}$ controls variability for all elements in the jth row of group h. Thus the row sums of $\mathbf{e}_h$ are a measure of heterogeneity in the corresponding rows of $\boldsymbol{\xi}_i^s$ in group h. The smaller Σ_{hj}, the more variable are the individual transition probabilities and the larger deviations of $\boldsymbol{\xi}_{i,j\cdot}^s$ from the group mean $\boldsymbol{\xi}_{h,j\cdot}$ are to be expected. On the other hand, if Σ_{hj} is very large, then variability in row j is very small meaning that the individual transition probabilities are nearly equal to the group mean $\boldsymbol{\xi}_{h,j\cdot}$. If this is the case for all rows in all groups, Dirichlet multinomial clustering reduces to Markov chain clustering.

Note that Dirichlet multinomial clustering provides a very parsimonious way of introducing group-specific unobserved heterogeneity in individual transition matrices. While the dimension of the group-specific parameter $\boldsymbol{\vartheta}_h = \boldsymbol{\xi}_h$ is equal to $K(K-1)$ for Markov chain clustering,

the dimension of $\boldsymbol{\vartheta}_h = \mathbf{e}_h$ is equal to K^2 for Dirichlet multinomial clustering, introducing only K additional parameters for each group. Each of these K parameters controls group-specific unobserved heterogeneity in exactly one row of $\boldsymbol{\xi}_i^s$.

3 Bayesian Inference for a Fixed Number of Clusters

In this paper we pursue a Bayesian approach toward estimation for fixed H. We assume prior independence between $\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_H$ and $\boldsymbol{\eta}$. We apply the Dirichlet prior $\boldsymbol{\eta} \sim \mathcal{D}(\alpha_0, \ldots, \alpha_0)$ which is commonly used in mixture modeling, see e.g. Frühwirth-Schnatter (2006), and choose specific priors $p(\boldsymbol{\vartheta}_h)$ for $\boldsymbol{\vartheta}_h$, depending on the clustering kernel. For practical Bayesian estimation we use a Markov chain Monte Carlo (MCMC) sampler based on data augmentation as in Frühwirth-Schnatter and Kaufmann (2008) which is described in Algorithm 1.

Algorithm 1.

1. Bayes' classification for each individual i: draw $S_i, i = 1, \ldots, N$ from the discrete probability distribution

$$\Pr(S_i = h|\mathbf{y}_i, \boldsymbol{\eta}, \boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_H) \propto p(\mathbf{y}_i|\boldsymbol{\vartheta}_h)\eta_h, \quad h = 1, \ldots, H. \tag{19}$$

2. Sample mixing proportions $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_H)$: draw $\boldsymbol{\eta}$ from the Dirichlet distribution $\mathcal{D}(\alpha_1, \ldots, \alpha_H)$ where $\alpha_h = \#\{S_i = h\} + \alpha_0$.

3. Sample component parameters $\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_H$: draw $\boldsymbol{\vartheta}_h$ from $p(\boldsymbol{\vartheta}_h|\mathbf{S}, \mathbf{y}), \quad h = 1, \ldots, H$.

This algorithm has been applied by Fougère and Kamionka (2003) for the special case of a mover-stayer model and has been mentioned in a short note by Ridgeway and Altschuler (1998). An alternative Bayesian approach has been used by Ramoni et al. (2002) who apply a heuristic search method for finding a good partition $\mathbf{S}$ of the data based on the marginal likelihood function $p(\mathbf{y}|\mathbf{S})$ where $\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_H$ are integrated out.

3.1 Bayesian Inference for Markov Chain Clustering

As the likelihood function of the Markov chain model given in (5) factors into K independent terms each depending only on the j-th row of the transition matrix we assume that the rows of

$\boldsymbol{\xi}_h$ are a priori independent each following a Dirichlet distribution, i.e. $\boldsymbol{\xi}_{h,j\cdot} \sim \mathcal{D}(e_{0,j1}, \ldots, e_{0,jK})$ with prior parameters $\mathbf{e}_{0,j\cdot} = (e_{0,j1}, \ldots, e_{0,jK})$ for $j = 1, \ldots, K$. This prior is conjugate to the complete data likelihood and allows straightforward implementation of Markov chain Monte Carlo estimation as in Algorithm 1 with $\boldsymbol{\vartheta}_h = \boldsymbol{\xi}_h, h = 1, \ldots, H$.

Classification in step 1 is based on the Markov chain model $p(\mathbf{y}_i|\boldsymbol{\vartheta}_h) = p(\mathbf{y}_i|\boldsymbol{\xi}_h)$ defined in (5). The complete data posterior distribution $p(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_H|\mathbf{S}, \mathbf{y})$ appearing in the third step where classifications $\mathbf{S}$ are considered to be known is of closed form due to conjugacy:

$$\begin{aligned} p(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_H|\mathbf{S}, \mathbf{y}) &\propto \prod_{i=1}^{N} p(\mathbf{y}_i|\boldsymbol{\xi}_{S_i}) \prod_{h=1}^{H} p(\boldsymbol{\xi}_h) = \prod_{i=1}^{N} \prod_{j=1}^{K} \prod_{k=1}^{K} (\xi_{S_i,jk})^{N_{i,jk}} \prod_{h=1}^{H} p(\boldsymbol{\xi}_h) \\ &\propto \prod_{h=1}^{H} \prod_{j=1}^{K} \left(\prod_{k=1}^{K} (\xi_{h,jk})^{N_{jk}^h + e_{0,jk} - 1} \right), \end{aligned}$$

where $N_{jk}^h = \sum_{i:S_i=h} N_{i,jk}$ 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.

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:

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

The Bayesian approach offers several advantages in the context of Markov chain clustering compared to EM estimation as in Cadez et al. (2003) or Frydman (2005). First, in many applications the diagonal elements in the transition matrices are expected to be rather high whereas the off-diagonal probabilities are comparatively low and the Bayesian approach allows to incorporate this information by setting the prior parameters adequately.

Second, the Bayesian approach based on a Dirichlet prior $\mathcal{D}(e_{0,j1}, \ldots, e_{0,jK})$ where $e_{0,jk} > 0$ is able to handle problems that occur under zero transitions when applying the EM algorithm to Markov chain clustering. The EM algorithm breaks down, if no transitions starting from j are observed in group h, i.e. $\sum_{k=1}^{K} N_{jk}^h = 0$ for some j. Then the complete data likelihood function

$p(\mathbf{y}|\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_H, \mathbf{S})$ is independent of the jth row of $\boldsymbol{\xi}_h$, $\boldsymbol{\xi}_{h,j\cdot}$:

$$p(\mathbf{y}|\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_H, \mathbf{S}) = \prod_{h=1}^{H} \prod_{l=1}^{K} \prod_{k=1}^{K} (\xi_{h,lk})^{N_{lk}^h} = \prod_{h=1}^{H} \prod_{l=1, l\neq j}^{K} \prod_{k=1}^{K} (\xi_{h,lk})^{N_{lk}^h},$$

and no estimator for $\boldsymbol{\xi}_{h,j\cdot}$ exists in the M-step. Additionally, the EM algorithm fails if not a single transition from j to k is observed for the whole panel. In this case $N_{jk}^h = 0$ for all $h = 1, \ldots, H$ and the M-step leads to an estimator of $\xi_{h,jk}$ that lies on the boundary of the parameter space, i.e. $\hat{\xi}_{h,jk} = 0$ for $h = 1, \ldots, H$. This causes difficulties with the computation of $\Pr(S_i = h|\mathbf{y}_i, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_H)$ for all observations in all groups in the subsequent E-step.

To avoid these problems one could follow the rule of thumbs discussed e.g. in Agresti (1990) and add a small constant $e_{0,jk}$, e.g. $e_{0,jk} = 0.5$ to the number of observed transitions. It is easy to verify that this is equivalent to combining the complete data likelihood $p(\mathbf{y}|\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_H, \mathbf{S})$ with a Dirichlet prior $\mathcal{D}(e_{0,j1}, \ldots, e_{0,jK})$ for each row $\boldsymbol{\xi}_{h,j\cdot}$ within our Bayesian approach.

3.2 Bayesian Inference for Dirichlet Multinomial Clustering

3.2.1 Prior Distributions

For Bayesian estimation a prior has to be chosen for each group-specific parameter $\mathbf{e}_h, h = 1, \ldots, H$ which is a matrix of size $(K \times K)$. In contrast to Subsection 3.1 no conjugate prior allowing straightforward MCMC estimation is available, however, the structure of the complete data likelihood to be discussed in Subsection 3.2.2 still suggests to assume that all rows $\mathbf{e}_{h,j\cdot}$ are independent within and across each group.

To avoid all problems with empty transitions that have been discussed in Subsection 3.1 we assume that $\mathbf{e}_{h,j\cdot} \geq 1$ for all rows in all groups. To take dependencies between the elements of $\mathbf{e}_{h,j\cdot}$ into account we assume that $\mathbf{e}_{h,j\cdot} - 1$ is a discrete-valued multivariate random variable following a negative multinomial distribution, $\mathbf{e}_{h,j\cdot} - 1 \sim \text{NegMulNom}(p_{j1}, \ldots, p_{jK}, \beta)$, where

$$p_{jk} = \frac{N_0 \cdot \hat{\xi}_{jk}}{\alpha + N_0}.$$

This prior depends on the hyperparameters N_0, β, α and $\hat{\xi}_{jk}$, the choice of which is discussed

below. The density of this prior reads:

$$p(\mathbf{e}_{h,j\cdot}) = \frac{\Gamma(\beta - K + \sum_{k=1}^{K} e_{h,jk})}{\Gamma(\beta)\prod_{k=1}^{K}(e_{h,jk}-1)!} p_{j0}^{\beta} \prod_{k=1}^{K} p_{jk}^{e_{h,jk}-1},$$

where $p_{j0} = 1 - \sum_{k=1}^{K} p_{jk}$, while expectation and variance are given by:

$$\begin{aligned}
\mathrm{E}(e_{h,jk}) &= 1 + \frac{\beta\, p_{jk}}{p_{j0}} = \frac{\beta}{\alpha} N_0 \hat{\xi}_{jk}, \\
\mathrm{Var}(e_{h,jk}) &= \frac{\beta\, p_{jk}(p_{jk}+p_{j0})}{p_{j0}^2} = \frac{\beta \cdot N_0\hat{\xi}_{jk}(N_0\hat{\xi}_{jk}+\alpha)}{\alpha^2} \\
&= \mathrm{E}(e_{h,jk}-1)\left(\frac{\mathrm{E}(e_{h,jk}-1)}{\beta}+1\right).
\end{aligned}$$

The negative multinomial distribution arises as a mixture distribution, if the K elements of $\mathbf{e}_{h,j\cdot}$ are independent random variables from the following Poisson distribution: $e_{h,jk} - 1 \sim \mathcal{P}(\gamma\lambda_{jk})$ with $\gamma \sim \mathcal{G}(\alpha, \beta)$. After integrating over γ, the marginal distribution reads $\mathbf{e}_{h,j\cdot} - 1 \sim$ NegMulNom $(p_{j1}, \ldots, p_{jK}, \beta)$-distribution with $p_{jk} = \lambda_{jk}/(\alpha + \sum_{l=1}^{K}\lambda_{jl})$.

This representation suggests choosing following hyperparameters: $\lambda_{jk} = N_0\hat{\xi}_{jk}$, where N_0 is the size of an imaginary experiment, e.g. $N_0 = 10$, and $\hat{\boldsymbol{\xi}}$ is a prior guess of the transition matrix, while α and β are small integers, e.g. $\alpha = \beta = 1$.

Alternatively, it is possible to assume that each element of $\mathbf{e}_{h,j\cdot} - 1$ is a continuous random variable following independently some prior distribution, for instance, the Gamma distribution $e_{h,jk} - 1 \sim \mathcal{G}(b_{jk}, 1)$ where $b_{jk} = N_0\hat{\xi}_{jk}$. However, we do not pursue this form of a prior distribution in the present paper.

3.2.2 MCMC Estimation

The parameters $\mathbf{e}_1, \ldots, \mathbf{e}_H$, $\boldsymbol{\eta}$ and the hidden indicators $\mathbf{S}$ are jointly estimated by MCMC sampling using Algorithm 1 where $\boldsymbol{\vartheta}_h = \mathbf{e}_h$. Classification in the first step of Algorithm 1 is based on the marginal time series model $p(\mathbf{y}_i|\boldsymbol{\vartheta}_h) = p(\mathbf{y}_i|\mathbf{e}_h)$ defined in (15).

The third step of Algorithm 1 is the only step which is essentially different from the corresponding step for Markov chain clustering. To implement this step the complete data posterior distribution $p(\mathbf{e}_1, \ldots, \mathbf{e}_H|\mathbf{S}, \mathbf{y})$ where the classifications $\mathbf{S}$ are considered to be known for each

individual is derived:

$$p(\mathbf{e}_1,\ldots,\mathbf{e}_H|\mathbf{S},\mathbf{y}) \propto \prod_{i=1}^{N} p(\mathbf{y}_i|\mathbf{e}_{S_i}) \prod_{h=1}^{H} p(\mathbf{e}_h) = \prod_{h=1}^{H} p(\mathbf{e}_h) \prod_{i:S_i=h} \prod_{j=1}^{K} p(\mathbf{y}_i|\mathbf{e}_{h,j\cdot})$$
$$\propto \prod_{h=1}^{H}\prod_{j=1}^{K} p(\mathbf{e}_{h,j\cdot}) \frac{\Gamma(\sum_{k=1}^{K} e_{h,jk})^{N_h}}{\left(\prod_{k=1}^{K}\Gamma(e_{h,jk})\right)^{N_h}} \left(\prod_{i:S_i=h} \frac{\prod_{k=1}^{K}\Gamma(N_{i,jk}+e_{h,jk})}{\Gamma(\sum_{k=1}^{K}(N_{i,jk}+e_{h,jk}))}\right), \tag{20}$$

where N_h is the number of time series in group h. Note that the KH rows $\mathbf{e}_{h,j\cdot}$ of $\mathbf{e}_1,\ldots,\mathbf{e}_H$ are independent, however, the conditional posterior $p(\mathbf{e}_{h,j\cdot}|\mathbf{y},\mathbf{S})$ given by

$$p(\mathbf{e}_{h,j\cdot}|\mathbf{y},\mathbf{S}) \quad \propto \quad p(\mathbf{e}_{h,j\cdot}) \frac{\Gamma(\sum_{k=1}^{K} e_{h,jk})^{N_h}}{\left(\prod_{k=1}^{K}\Gamma(e_{h,jk})\right)^{N_h}} \left(\prod_{i:S_i=h} \frac{\prod_{k=1}^{K}\Gamma(N_{i,jk}+e_{h,jk})}{\Gamma(\sum_{k=1}^{K}(N_{i,jk}+e_{h,jk}))}\right)$$

is no longer of closed form. Thus the group-specific parameters $\mathbf{e}_1,\ldots,\mathbf{e}_H$ are sampled line-by-line by drawing each row $\mathbf{e}_{h,j\cdot}$ from $p(\mathbf{e}_{h,j\cdot}|\mathbf{y},\mathbf{S})$ by means of a Metropolis-Hastings algorithm.

As the computation of $p(\mathbf{e}_{h,j\cdot}|\mathbf{y},\mathbf{S})$ is rather time-consuming we decided to update only $l \leq K$ elements per row simultaneously while the other elements remained unchanged. As these elements are randomly chosen, this is a valid updating strategy to reduce computation time which comes at the cost of possibly higher autocorrelations than updating all elements.

We propose each element $e_{h,jk}$ to be updated independently from a discrete random walk proposal density $q(e_{h,jk}|e_{h,jk}^{(m-1)})$ since the support of $e_{h,jk}$ are the natural numbers according to our prior assumption. If $e_{h,jk}^{(m-1)} \geq 2$ we add with equal probability $-1, 0$ or 1, if $e_{h,jk}^{(m-1)} = 1$ we add 0 or 1. This proposal is equivalent to a uniform distribution on $[\max(1, e_{h,jk}^{(m-1)} - 1), e_{h,jk}^{(m-1)} + 1]$. We accept the proposed value $\mathbf{e}_{h,j\cdot}^{new}$ with probability $\min(1, r)$ where

$$r = \frac{p(\mathbf{e}_{h,j\cdot}^{new}|\mathbf{y},\mathbf{S})\; q(\mathbf{e}_{h,j\cdot}^{(m-1)}|\mathbf{e}_{h,j\cdot}^{new})}{p(\mathbf{e}_{h,j\cdot}^{(m-1)}|\mathbf{y},\mathbf{S})\; q(\mathbf{e}_{h,j\cdot}^{new}|\mathbf{e}_{h,j\cdot}^{(m-1)})}.$$

Note that our MCMC implementation avoids the expensive generation of the individual transition matrices $\boldsymbol{\xi}_1^s,\ldots,\boldsymbol{\xi}_N^s$ during each iteration. Such a step would require drawing all K rows $\boldsymbol{\xi}_{i,j\cdot}^s$ of $\boldsymbol{\xi}_i^s$ for each $i = 1,\ldots,N$ from

$$\boldsymbol{\xi}_{i,j\cdot}^s|(S_i = h, \mathbf{e}_h, \mathbf{y}) \sim \mathcal{D}\left(e_{h,j1} + N_{i,j1},\ldots,e_{h,jK} + N_{i,jK}\right),$$

where $N_{i,jk}$ is the number of transitions from state j to k of individual i, see (4).

In our labor market application in Section 5, for instance, where we are dealing with nearly 10 000 time series and $K = 6$ categories, this would require sampling from about 60 000 Dirichlet distributions which in turn means sampling about 360 000 random variables from a Gamma distribution for each MCMC sweep. This expensive step can be avoided under the special structure of the distribution of heterogeneity underlying Dirichlet multinomial clustering, because the density $p(\mathbf{y}_i|\mathbf{e}_{S_i})$ is available in closed form.

4 Selecting the Number of Clusters

If a finite mixture model is applied to model the distribution of the data in a flexible way, selecting the number of components H reduces to a model selection problem which could be solved by computing marginal likelihoods or running some model space methods, see e.g. Frühwirth-Schnatter (2006, Chapter 4 and 5).

In a clustering context, however, it is not so clear how to select an optimal number of components. Most clustering criteria are based on measuring model fit through some kind of likelihood function which is then penalized in an appropriate way to avoid overfitting. For any of these criteria the optimal number H of clusters is defined as that value of H which minimizes the criterion. Subsequently, $\boldsymbol{\theta}_H$ denotes the model parameter in a finite mixture model with H components, i.e. $\boldsymbol{\theta}_H = (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_H, \eta_1, \ldots, \eta_H)$.

The most popular model selection criteria are AIC (Akaike, 1974) and BIC (Schwarz, 1978) which penalize the log mixture likelihood by model complexity defined as the total number d_H of independent parameters to be estimated in a mixture model with H components:

$$AIC(H) = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_H) + 2\, d_H, \tag{21}$$

$$BIC(H) = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_H) + d_H \log N, \tag{22}$$

where $\hat{\boldsymbol{\theta}}_H$ is the ML estimator maximizing the log mixture likelihood $\log p(\mathbf{y}|\boldsymbol{\theta}_H)$ given by:

$$\log p(\mathbf{y}|\boldsymbol{\theta}_H) = \sum_{i=1}^{N} \log \left(\sum_{h=1}^{H} \eta_h\, p(\mathbf{y}_i|\boldsymbol{\vartheta}_h) \right). \tag{23}$$

Because the ML estimator is not available within the framework of MCMC estimation, $\hat{\boldsymbol{\theta}}_H$ is chosen as that posterior draw which maximizes the log mixture likelihood $\log p(\mathbf{y}|\boldsymbol{\theta}_H)$.

AIC is generally known to be overfitting and tends to select too many components also when fitting a finite mixture model with unknown number of components to the data. This happens even if the clustering kernel is correctly specified.

BIC is an asymptotic approximation to minus twice the marginal likelihood $-2\log p(\mathbf{y}|H)$, see e.g. Kass and Raftery (1995). Because the posterior probability of a model is the higher the smaller BIC, this criterion could be used to compare various clustering kernels for the same number H of components. However, one should be cautious when using BIC to select the number of clusters in the data. Only if the data form H well-separated clusters and the clustering kernels appearing in the finite mixture model are chosen from the true cluster-specific distribution, then the number of components in the mixture selected by BIC is asymptotically equal to the number of clusters due to a consistency result proven by Keribin (2000).

However, the number of components in the mixture selected by BIC need not be identical with the number of clusters in the data, if at least one of these assumptions is violated. First of all, BIC is not good a estimator for the number of *distinct* clusters, if the component densities are strongly overlapping. Furthermore, simulation studies reported in Biernacki et al. (2000) show that BIC typically overrates the number of clusters if the distribution underlying the clustering kernel is not identical with the true cluster-specific distribution.

Approximate weight of evidence (AWE) which is derived in Banfield and Raftery (1993) as another approximation to minus twice the log Bayes factor is expressed in Biernacki and Govaert (1997) as a criterion which penalizes the complete data log-likelihood function with model complexity:

$$AWE(H) = -2\ \log\ p(\mathbf{y}, \hat{\mathbf{S}}|\hat{\boldsymbol{\theta}}_H^C) + 2\, d_H(\frac{3}{2} + \log N), \tag{24}$$

where $\hat{\boldsymbol{\theta}}_H^C$ and $\hat{\mathbf{S}}$ are determined jointly as that combination of parameters and allocations that maximize the complete data log-likelihood $\log\ p(\mathbf{y}, \mathbf{S}|\boldsymbol{\theta}_H)$ given by

$$\log\ p(\mathbf{y}, \mathbf{S}|\boldsymbol{\theta}_H) = \sum_{i=1}^{N} \log\left(\eta_{S_i} p(\mathbf{y}_i|\boldsymbol{\vartheta}_{S_i})\right). \tag{25}$$

Again, approximate estimators $\hat{\boldsymbol{\theta}}_H^C$ and $\hat{\mathbf{S}}$ are obtained by choosing the posterior draw maximizing the complete data log-likelihood function.

None of these criteria directly takes into account that in a clustering context a finite mixture model is fitted with the hope of finding a good partition of the data. For this reason various criteria were developed which involve the quality of the resulting partition measured through the entropy $EN(H, \boldsymbol{\theta}_H)$ which is given by

$$EN(H, \boldsymbol{\theta}_H) = -\sum_{h=1}^{H}\sum_{i=1}^{N} t_{ih}(\boldsymbol{\theta}_H) \log t_{ih}(\boldsymbol{\theta}_H) \geq 0, \tag{26}$$

where $t_{ih}(\boldsymbol{\theta}_H) = \Pr(S_i = h|\mathbf{y}_i, \boldsymbol{\theta}_H)$ is the posterior classification probability defined in (19). The entropy is a measure of how well the data are classified given the finite mixture distribution defined by $\boldsymbol{\theta}_H$. It is close to 0 if the resulting clusters are well-separated and increases with increasing overlap of the mixture components.

The CLC criterion (Biernacki and Govaert, 1997) penalizes the log mixture likelihood by the entropy rather than by model complexity as in AIC or BIC:

$$CLC(H) = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_H) + 2\, EN(H, \hat{\boldsymbol{\theta}}_H), \tag{27}$$

where $\hat{\boldsymbol{\theta}}_H$ is again the (approximate) ML estimator.

Since CLC works well only for well-separated clusters with a fixed weight distribution Biernacki et al. (2000) proposed the integrated classification likelihood (ICL) criterion. A special approximation to this criterion is the $ICL\text{-}BIC$ criterion (McLachlan and Peel, 2000) which penalizes not only model complexity, but also the failure of the mixture model to provide a classification of the data into well-separated clusters:

$$ICL\text{-}BIC(H) = BIC(H) + 2\, EN(H, \hat{\boldsymbol{\theta}}_H). \tag{28}$$

Simulation studies reported by McLachlan and Peel (2000, Section 6.11) showed that $ICL\text{-}BIC$ is able to identify the correct number of clusters in the context of multivariate mixtures of normals even when the component densities are misspecified.

5 Application to Austrian Wage Mobility Data

In this section we consider wage mobility in the Austrian labor market. Wage mobility describes chances but also risks of an individual to move between wage categories over time, see also Raferzeder and Winter-Ebmer (2007). In the present paper, the moves and transitions between the categories are expressed in terms of transition matrices which determine the income career and career progressions for an individual. It is sensible to assume that the income careers and career progressions are different between the employees. Our goal is to find meaningful groups of employees with similar wage mobility behavior using both Markov chain clustering as well as Dirichlet multinomial clustering.

5.1 Data Description

The data set has been provided by the Austrian social security authority who collects detailed data for all workers in Austria and has been taken from the ASSD (Austrian Social Security Data Base), see Zweimüller et al. (2009).

The panel consists of time series observations for $N = 9\,809$ men entering the labor market in the years 1975 to 1980 at an age of at most 25 years. The time series represent gross monthly wages in May of successive years and exhibit individual lengths ranging from 2 to 27 years with the median length being equal to 23. Following Weber (2001), the gross monthly wage is divided into six categories labeled with 0 up to 5. Category zero corresponds to zero-income or non-employment which is not equivalent to be out of labor force. The categories one to five correspond to the quintiles of the income distribution which are determined 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 circumvents the problem that in Austria recorded wages are right-censored because wages that exceed a social security payroll tax cap which is an upper limit of the assessment base for the contribution fee are recorded with exactly that limit.

5.2 Running Model-Based Clustering

To identify groups of individuals with similar wage mobility behavior, we apply both Markov chain clustering as well as Dirichlet multinomial clustering for 2 up to 6 groups.

For the Dirichlet prior of the weight distribution $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_H)$ we choose $\alpha_0 = 4$ as recommended by Frühwirth-Schnatter (2006). For Markov chain clustering the prior for each row for each matrix $\boldsymbol{\xi}_h$ is based on a Dirichlet prior where $e_{0,jj} = 2$ and $e_{0,jk} = 1$, if $j \neq k$. For Dirichlet multinomial clustering, the prior for each row for each matrix $\mathbf{e}_h$ is based on the negative multinomial distribution with $\alpha = \beta = 1$, $N_0 = 70$ and $\hat{\boldsymbol{\xi}}_h = \hat{\boldsymbol{\xi}}$, where $\hat{\xi}_{jj} = 0.7$ and $\hat{\xi}_{jk} = 0.06$, if $j \neq k$. Alternative hyperparameters were considered but showed negligible differences in the results.

We start MCMC estimation by choosing initial values for the group-specific parameters. Initial values for the weights are $\eta_h^{(0)} = 1/H$, $h = 1, \ldots, H$, both for Markov chain clustering as well as for Dirichlet multinomial clustering. To choose initial values for the remaining parameters, we define a transition matrix $\hat{\boldsymbol{\xi}}$ with diagonal elements $\hat{\xi}_{jj} = 0.7$ and $\hat{\xi}_{jk} = 0.06$, if $j \neq k$. For Markov chain clustering we choose $\boldsymbol{\xi}_h^{(0)} = \hat{\boldsymbol{\xi}}$, while for Dirichlet multinomial clustering we set $\mathbf{e}_h^{(0)} = N_0 \hat{\boldsymbol{\xi}}$ with N_0 being the hyperparameter appearing in the prior.

For each number H of groups we simulated 10 000 MCMC draws after a burn-in of 10 000 draws for Markov chain clustering and a burn-in of 15 000 draws for Dirichlet multinomial clustering. To update the elements of $\mathbf{e}_h$ in Dirichlet multinomial clustering we choose $l = 2$ elements per row randomly and apply the Metropolis-Hastings algorithm described in Subsection 3.2.2, leading to an average acceptance rate of 0.245.

5.3 Selecting the Number of Clusters

The model selection criteria described in Section 4 are applied to select the number H of clusters both under Dirichlet multinomial as well as under Markov chain clustering, see Figure 2.

For both clustering kernels, AIC and BIC decrease with increasing H and suggests at least 6 components. However, as outlined in Section 4, we cannot expect that the Markov chain model or even the more flexible Dirichlet multinomial 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. For Dirichlet multinomial clustering AWE takes a minimum at $H = 4$, while, somewhat surprisingly, CLC and $ICL\text{-}BIC$ show a non-monotonic behavior with two local minima at $H = 2$ and $H = 4$. For Markov chain clustering all criteria suggest the presence of 5 clusters. As described in Section 4, the evaluation of these criteria is based on approximate 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. While model choice was stable for Dirichlet multinomial clustering, CLC and $ICL\text{-}BIC$ sometimes indicated only 4 clusters under Markov chain clustering for different MCMC runs. To sum up, these criteria provide evidence for 4 clusters under Dirichlet multinomial clustering and 4 or 5 clusters for Markov chain clustering.

When we compare Dirichlet multinomial clustering with Markov chain clustering for a fixed number H of clusters using BIC, we find that Dirichlet multinomial clustering has in general a higher posterior probability than Markov chain clustering. First, this indicates that some unobserved heterogeneity is present in the cluster even after accounting for differences in the cluster-specific transition behavior. Second, Dirichlet multinomial clustering is expected to exhibit a higher robustness to untypical group members. It should be noted that this difference gets smaller with increasing H, because adding components reduces the within-cluster unobserved heterogeneity and allows to introduce small components containing untypical wage careers.

When $ICL\text{-}BIC$ which penalizes BIC by entropy is used to compare the clustering methods we find that Dirichlet multinomial clustering dominates Markov chain clustering up to 4 clusters. For 5 and 6 clusters Dirichlet multinomial clustering is outperformed by Markov chain clustering although giving a higher posterior probability for the observed data, mainly because the entropy of the resulting classification of the time series is larger than for Markov chain clustering.

To provide a more profound comparison of Dirichlet multinomial clustering versus Markov chain clustering we decided to discuss the four-cluster solution for both clustering methods in more details. These solutions also led to sensible interpretations from an economic point of view.

5.4 Empirical Results

For reasons discussed in Subsection 5.3, we discuss in more detail Bayesian inference for the four-group solution both for Dirichlet multinomial as well as for Markov chain clustering.

As common for modern Bayesian inference, we approximate the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$ of any quantity $\boldsymbol{\theta}$ of interest by MCMC draws from the posterior distribution. We use the posterior mean $\mathrm{E}(\boldsymbol{\theta}|\mathbf{y})$ which is approximated by the average of the corresponding MCMC draws as a point estimator for $\boldsymbol{\theta}$. To evaluate the uncertainty associated with estimating $\boldsymbol{\theta}$, we consider for each component θ_j of $\boldsymbol{\theta}$ the posterior standard deviation $\mathrm{SD}\,(\theta_j|\mathbf{y}) = \sqrt{\mathrm{Var}(\theta_j|\mathbf{y})}$ which is approximated by the standard deviation of the corresponding MCMC draws. Confidence regions are derived from the corresponding percentiles of the MCMC draws. For more details on MCMC inference we refer to standard monographs like Geweke (2005) and Gamerman and Lopes (2006).

For finite mixture models parameter estimation based on MCMC draws is possible only, if no label switching is present meaning that the labeling of the clusters did not change during MCMC sampling, see e.g. Frühwirth-Schnatter (2006, Section 3.5) for an exhaustive review of the label switching problem. Label switching typically occurs if the finite mixture model is overfitting the number of components. However, as indicated by BIC reported in Figure 2 it is very unlikely that a mixture with 4 components overfits the data under investigation. This is supported by the visual inspection of the MCMC draws (not reported to save space) of the cluster sizes $\boldsymbol{\eta}$ and the cluster-specific parameter $(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_4)$ and $(\mathbf{e}_1, \ldots, \mathbf{e}_4)$ which did not reveal any signs of label switching.

5.4.1 Analyzing Wage Mobility

To analyze wage mobility in the different clusters we investigate for each $h = 1, \ldots, 4$ the posterior distribution of the group-specific transition matrix $\boldsymbol{\xi}_h$. For Markov chain clustering, MCMC draws for $\boldsymbol{\xi}_h$ are directly available. For Dirichlet multinomial clustering, posterior draws for $\boldsymbol{\xi}_h$ are obtained by applying the nonlinear transformation (16) to each MCMC draw of $\mathbf{e}_h$.

MCMC based posterior inference for Dirichlet multinomial clustering is summarized in Table 1 by reporting the posterior expectation $\mathrm{E}(\xi_{h,jk}|\mathbf{y})$ as well as the standard deviation $\mathrm{SD}\,(\xi_{h,jk}|\mathbf{y})$ for each cell of the group-specific transition matrices $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_4$. In addition, the

posterior expectations are visualized in Figure 3 using "balloon plots" generated by means of function `balloonplot()` from the R package `gplots` (Jain and Warnes, 2006). These plots also show the relative size of each group.

Based on the transition matrices reported in Table 1 and in Figure 3, we assign a labeling to each cluster, namely "low wage", "flexible", "unemployed", and "climbers" which will be further corroborated by the long-run wage distribution to be discussed later in this subsection as well as by the wage careers of typical group members to be discussed in Subsection 5.4.3.

A remarkable difference in the transition behavior of individuals belonging to different clusters is evident from Figure 3. 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 "unemployed" and the "flexible" cluster than for the other clusters.

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

For all matrices, the main diagonal refers to the probabilities to remain in the various wage categories. Persistence is pretty high except for the "flexible" cluster. Members of this 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 much higher in the "climbers" cluster than in the other clusters.

These obvious differences in the one-step ahead transition matrix between the clusters have a strong impact on the wage mobility and the long-run wage career of the group members, as shown by Figure 4. This figure shows for each cluster h the 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. Additionally, Figure 4 shows the posterior expectation $\mathrm{E}(\boldsymbol{\pi}_{h,t}|\mathbf{y}, \boldsymbol{\pi}_{h,0})$ of the cluster-specific wage distribution $\boldsymbol{\pi}_{h,t}$ after a period of t years, which is defined by $\boldsymbol{\pi}_{h,t} = \boldsymbol{\pi}_{h,0}\boldsymbol{\xi}_h^t$. 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 100 MCMC draws of $\boldsymbol{\xi}_h$.

For $t = 100$, the wage distribution is practically equal to the equilibrium distribution $\boldsymbol{\pi}_{h,\infty}$ of the transition matrix $\boldsymbol{\xi}_h$, i.e. $\boldsymbol{\pi}_{h,\infty} = \boldsymbol{\pi}_{h,\infty}\boldsymbol{\xi}_h$. In the "unemployed" and the "flexible" cluster the equilibrium distribution is reached after only a few years, whereas in the other two clusters this distribution is reached after about two decades.

The wage distributions shown in Figure 4 provide further evidence for the labeling of the clusters we introduced earlier. Young men belonging to the "unemployed" cluster have a much higher risk to start in the no-income category then young men belonging to the other clusters. Furthermore, about 60 % of the members of this group have no income in the long-run.

For young men belonging either to the "low wage", the "flexible", or the "climbers" cluster there is little difference between the initial wage distribution when they enter the labor market. However, in the long run considerable differences in the wage distribution become evident due to the observed differences in wage mobility. Members of the "flexible" cluster have a much higher risk to end up in the no-income category than members of the "low wage" or the "climbers" cluster. In the long-run, however, members of the "low wage" cluster are disadvantaged and end up in lower wage categories while members of the "climbers" cluster move into the highest wage categories.

5.4.2 Analyzing Unobserved Heterogeneity

To analyze how much unobserved heterogeneity is present in the various clusters, we report in Table 2 the posterior expectation of the variance of the individual transition probabilities $\xi^s_{i,jk}$ within each cluster which has been defined in (17). These variances are multiplied by 10^4 to obtain the variance of the individual transition probabilities in percent. In addition, we show the posterior expectation and the posterior standard deviation of the group-specific unobserved heterogeneity in row j as defined in (18). All expectations are estimated as the average of MCMC draws obtained by applying, respectively, transformation (17) and (18) to each MCMC draw of $\mathbf{e}_h$.

The variances of the individual transition probabilities as well as the unobserved heterogeneity measure varies considerably between the clusters as well as between the rows within each cluster. Unobserved heterogeneity is highest in the "flexible" cluster and lowest in the "unem-

ployment" cluster. Consequently, several high variances for individual transition matrices are observed in the "flexible" cluster, while the "unemployment" cluster typically has smaller variances. In general, persistence probabilities have higher variances than the off-diagonal elements.

Apart from a few cells with high individual variance, the amount of unobserved heterogeneity is rather moderate for most of the cells. Thus it is to be expected that the cluster-specific transition matrices obtained by Dirichlet multinomial clustering are similar to the ones obtained by Markov chain clustering. Indeed, when we studied the transition matrices and the long-run wage distributions of the four-group solution obtained through Markov chain clustering we were able to identify clusters with a similar meaning, namely a "low wage", a "flexible", an "unemployed", and a "climbers" cluster.

For further comparison, Figure 5 shows the difference between the posterior expectation of the cluster-specific transition matrices $\boldsymbol{\xi}_h$ obtained, respectively, by Dirichlet multinomial clustering (DMC) and Markov chain clustering (MCC), i.e. $\mathrm{E}(\xi_{h,jk}|\mathbf{y}, \mathrm{DMC}) - \mathrm{E}(\xi_{h,jk}|\mathbf{y}, \mathrm{MCC})$ for each cell (j, k) in each cluster. We observe the biggest differences in the last row of the transition matrix in the "low wage" cluster. This row concerns those (rare) members in that cluster who manage to move to the highest wage category. Under DMC, the expected chance to remain in the highest wage category is 76.4%. For MCC, this chance decreases to 47.3%, while the risk to drop back to wage category one, which is the lowest non-zero wage, increases by 17.1% and is 26.8% instead of 9.7%.

Less pronounced differences are present in the last row of the transition matrix in the "flexible" cluster, where the persistence chance is by 9.6% smaller for DMC than for MCC (61.6%), while the risk to move back to wage category four is by 4.2% larger than for MCC (15.8%). For the remaining cells differences occur mainly for the persistence probabilities with MCC overrating persistence in relation to DMC by up to about 5%. This phenomenon is well-known in the analysis of dynamic panels, see e.g. Hsiao (2003), where it is often observed that ignoring unobserved heterogeneity leads to overrating persistence, see also Frühwirth-Schnatter and Frühwirth (2007) for a related panel data analysis for the Austrian labor market.

5.4.3 Posterior Classification

Next we study for both clustering methods how 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 $H = 4$, see e.g. Frühwirth-Schnatter (2006, pp.221) for various ways of clustering observations based on finite mixture models. 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 5000 MCMC draws of $\boldsymbol{\theta}_4$ with a thinning parameter equal to 20. 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.

Table 3 analyzes the segmentation power for both clustering methods by reporting the quartiles and the median of $\hat{t}_{i,\hat{S}_i}$ 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 74.7% (MCC) and 72.3% (DMC) to their respective groups. For 1 out of 4 individuals assignment probability amounts to at least 99.2% (MCC) and 97.6% (DMC). Segmentation power varies between the clusters and is the highest for the "unemployed" cluster and the lowest for the "flexible" cluster. We find that Markov chain clustering has a slightly higher segmentation power than Dirichlet multinomial clustering in particular for the "low wage" cluster where we found the largest differences in the estimated transition matrices.

To get an even better understanding of the various wage mobility groups typical group members are selected for each cluster and their individual time series are plotted in Figure 6 and 7. Figure 6 shows for both clustering methods the five members having the highest classification probability to belong to a particular cluster, while Figure 7 shows five individuals selected from ranks between 10 and 200.

These figures further emphasize the interpretation of the wage mobility groups given above and is surprisingly robust to the clustering method. The "flexible" cluster obviously represents the more flexible and fluctuating employees. Typical members of the "low wage" cluster stay mainly in the lowest wage category. The "unemployment" cluster contains the employees who fall into the no-income category more often and remain there much longer than members of the other clusters. Finally, the "climbers" cluster comprises of employees who get out of the no-income

category more easily and make rather straight career advancements. Such huge differences in the wage mobility in the Austrian labor market have never been documented before.

5.4.4 Analyzing Group Membership

To learn more about the factors that effect the probability of an individual to belong to a certain cluster we use the classifications $\hat{S}_i$ obtained for each person $i = 1, \ldots, 9\,809$ under Dirichlet multinomial clustering as input for a multinomial logit regression model. We select the "unemployed" cluster as baseline and use several covariates to model the odds of belonging to any of the other clusters.

To capture the general economic situation at time of entry into the labor market we introduce time dummies for each year 1976 to 1980 with 1975 serving as baseline and add the unemployment rate in the district to capture the regional economic situation.

Unfortunately, little individual information about the employees is available. We only know whether a person started as blue or white collar worker, the age at entry as well as the days a person served as an apprentice. We use the last two variables to define a proxy for the education of a person which is not observed directly. We take young men who finished apprenticeship, meaning that they 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, we consider young men starting after their 18th birthday without finishing apprenticeship as "skilled", because they are likely to have finished some kind of higher education such as high school or university. Finally, we add dummies for the wage category at entry with zero income serving as baseline.

We perform Bayesian inference for the resulting multinomial regression model based on a standard normal prior for all regression parameters. The posterior expectations and the posterior standard deviations of all regression parameters are reported in Table 4. These results are based on 20 000 MCMC draws (after discarding 5000 draws as burn-in) obtained by auxiliary mixture sampling (Frühwirth-Schnatter and Frühwirth, 2007).

From Table 4 we find that having a non-zero income in the initial year increases the odds significantly to belong to any cluster but the "unemployment" cluster. For young men starting with relatively high wages the odds of belonging to the "climbers" rather than to any other

cluster are high. The same is true for young men starting in 1976 to 1978 which was a period of high real GDP growth rate while the real GDP growth rate was negative in 1975, see also Table 5. An increasing unemployment rate in the district increases the odds of belonging to the "low wage" cluster. For "unskilled" young men the odds of belonging to any but the "unemployment" cluster are negative.

6 Concluding Remarks

In this paper we discussed two approaches 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. In a new approach called Dirichlet multinomial clustering it is assumed that within each group unobserved heterogeneity is still existent and is captured by allowing the individual transition matrices to deviate from the group means by describing this variation for each row through a Dirichlet distribution with unknown hyperparameters.

We discussed in detail an application of these two approaches to modeling and clustering a panel of Austrian wage mobility data describing the wage career of nearly 10 000 young men entering the labor market during the second half of the 1970s. Model choice indicated that Dirichlet multinomial clustering outperforms Markov chain clustering in terms of posterior probability (approximated by BIC) and that for this cohort the labor market should be segmented into 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 "unemployed", "low wage", "flexible" and "climbers". When further analyzing the results obtained by Dirichlet multinomial clustering, we found that unobserved heterogeneity is present in the various clusters and, as expected from previous investigations, ignoring it would lead to overrating the persistence probability.

However, the amount of unobserved heterogeneity within each cluster is small compared to the differences between the clusters. Thus it is not surprising that the types discovered under Dirichlet multinomial clustering turned out to be robust to the choice of clustering kernel and were more less the same the types as obtained by Markov chain clustering under a four-group

solution.

We investigated the segmentation power of the four-group solution for both clustering methods and found that it is rather high. 3 out of 4 individuals are assigned with at least 74.7% (Markov chain clustering) and 72.3% (Dirichlet multinomial clustering) probability to their respective cluster.

We conclude from our investigation that both clustering kernels are a sensible tool for model-based clustering of discrete-valued panel data. For our case study, Dirichlet multinomial clustering indicated the presence of unobserved heterogeneity and, consequently, outperformed Markov chain clustering in terms of BIC. However, the clusters we discovered had a similar meaning for both methods and Markov chain clustering showed a slightly higher segmentation power.

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. As discussed in detail in Subsection 2.4, Dirichlet multinomial clustering is a restricted variant of this model with a different parameterization.

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. Furthermore, such a model is able to capture more general correlation patterns in the distribution of unobserved heterogeneity than the restricted version corresponding to Dirichlet multinomial clustering.

Under Dirichlet multinomial clustering, individual transition probabilities $\xi^s_{i,jk}$ and $\xi^s_{i,j'l}$ appearing in different rows of $\boldsymbol{\xi}^s_i$ are independent, while for transition probabilities $\xi^s_{i,jk}$ and $\xi^s_{i,jl}$ appearing in the same row of $\boldsymbol{\xi}^s_i$ the following holds:

$$\frac{\mathrm{Cov}(\xi^s_{i,jk}\,\xi^s_{i,jl}|S_i=h,\mathbf{e}_h)}{\mathrm{E}(\xi^s_{i,jk}|S_i=h,\mathbf{e}_h)\mathrm{E}(\xi^s_{i,jl}|S_i=h,\mathbf{e}_h)} = -\frac{1}{1+\sum_{k'=1}^{K} e_{h,jk'}}.$$

Thus the dependence structure within each row is rather restricted and, apart from the sign, is controlled by the same expression which controls the total amount of unobserved heterogeneity in that row, see also (18).

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

We thank Andrea Weber and Rudolf Winter-Ebmer for numerous remarks as well as comments on this research. Special thanks go to Helga Wagner and other members of our department for helpful comments and discussions. The second author's research is supported by the Austrian Science Foundation (FWF) under the grants P 17 959 ("Gibbs sampling for discrete data") and S 10309-G14 (NRN "The Austrian Center for Labor Economics and the Analysis of the Welfare State", Subproject "Bayesian Econometrics").

References

Agresti, A. (1990). *Categorical Data Analysis*. Chichester: Wiley.

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

Balakrishnan, N. (Ed.) (1992). *Handbook of the Logistic Distribution*. New York: Marcel Dekker.

Banfield, J. D. and A. E. Raftery (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics 49*, 803–821.

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

Cadez, I., D. Heckerman, C. Meek, P. Smyth, and S. White (2003). Model-based clustering and visualization of navigation patterns on a web site. *Data Mining and Knowledge Discovery 7*(4), 399–424.

Canova, F. (2004). Testing for convergence clubs in income per-capita: A predictive density approach. *International Economic Review 45*, 49–77.

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

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

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

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

Frühwirth-Schnatter, S., R. Frühwirth, L. Held, and H. Rue (2009). Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. *Statistics and Computing 19*, forthcoming.

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

Frühwirth-Schnatter, S., R. Tüchler, and T. Otter (2004). Bayesian analysis of the heterogeneity model. *Journal of Business & Economic Statistics 22*, 2–15.

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. and H. F. Lopes (2006). *Markov Chain Monte Carlo. Stochastic Simulation for Bayesian Inference* (2 ed.). Boca Raton, FL: Chapman & Hall/CRC.

Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*. Wiley.

Hsiao, C. (2003). *Analysis of Panel Data* (2 ed.). New York: Cambridge University Press.

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

Juárez, M. A. and M. F. J. Steel (2009). Model-based clustering of non-Gaussian panel data based on skew-t distributions. *Journal of Business & Economic Statistics 27*, to appear.

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

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

Lenk, P. J. and W. S. DeSarbo (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika 65*, 93–119.

Liao, T. W. (2005). Clustering of time series data – a survey. *Pattern Recognition 38*, 1857–1874.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour. In P. Zarembka (Ed.), *Frontiers of Econometrics*, pp. 105–142. New York: Academic.

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

Raferzeder, T. and R. Winter-Ebmer (2007). Who is on the rise in Austria: Wage mobility and mobility risk. *Journal of Economic Inequality 5*(1), 39–51.

Ramoni, M., P. Sebastiani, and P. Cohen (2002). Bayesian clustering by dynamics. *Machine Learning 47*, 91–121.

Ridgeway, G. and S. Altschuler (1998). Clustering finite discrete Markov chains. In *Proceedings of the Section on Physical and Engineering Sciences*, pp. 228–229. American Statistical Association.

Rossi, P. E., G. M. Allenby, and R. McCulloch (2005). *Bayesian Statistics and Marketing*. Chichester: Wiley.

Saul, L. K. and M. I. Jordan (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.

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.

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

			"unemployed"			
	0	1	2	3	4	5
0	0.914(0.238)	0.047(0.130)	0.016(0.043)	0.008(0.022)	0.008(0.022)	0.008(0.022)
1	0.217(0.637)	0.604(0.830)	0.135(0.508)	0.022(0.131)	0.011(0.065)	0.011(0.065)
2	0.189(0.748)	0.097(0.560)	0.545(0.936)	0.133(0.679)	0.024(0.230)	0.012(0.109)
3	0.177(0.909)	0.037(0.323)	0.115(1.038)	0.457(1.511)	0.178(0.963)	0.037(0.323)
4	0.118(0.913)	0.022(0.277)	0.022(0.277)	0.087(0.942)	0.577(1.571)	0.174(1.212)
5	0.048(0.476)	0.008(0.047)	0.008(0.043)	0.008(0.043)	0.020(0.375)	0.910(0.623)
			"climbers"			
	0	1	2	3	4	5
0	0.136(0.000)	0.227(0.000)	0.227(0.000)	0.182(0.000)	0.136(0.000)	0.091(0.000)
1	0.152(0.491)	0.510(0.863)	0.243(0.589)	0.063(0.358)	0.022(0.129)	0.011(0.065)
2	0.058(0.288)	0.061(0.299)	0.582(0.721)	0.261(0.562)	0.030(0.210)	0.008(0.065)
3	0.037(0.164)	0.012(0.055)	0.091(0.414)	0.644(0.569)	0.205(0.459)	0.012(0.055)
4	0.026(0.085)	0.009(0.028)	0.009(0.028)	0.077(0.288)	0.781(0.352)	0.010(0.252)
5	0.027(0.144)	0.004(0.016)	0.004(0.016)	0.004(0.016)	0.063(0.250)	0.897(0.307)
			"flexible"			
	0	1	2	3	4	5
0	0.560(0.808)	0.240(0.548)	0.088(0.377)	0.048(0.601)	0.043(0.294)	0.021(0.147)
1	0.255(0.578)	0.517(0.852)	0.121(0.533)	0.054(0.216)	0.036(0.144)	0.018(0.072)
2	0.200(0.697)	0.198(0.614)	0.348(1.166)	0.168(0.638)	0.057(0.437)	0.029(0.219)
3	0.139(0.620)	0.095(0.636)	0.136(0.527)	0.408(1.187)	0.180(0.635)	0.043(0.480)
4	0.132(0.431)	0.066(0.216)	0.066(0.216)	0.135(1.031)	0.470(1.079)	0.132(0.431)
5	0.120(0.634)	0.080(0.430)	0.040(0.211)	0.040(0.211)	0.200(1.084)	0.520(1.946)
			"low wage"			
	0	1	2	3	4	5
0	0.247(1.474)	0.478(1.268)	0.180(1.037)	0.054(0.828)	0.021(0.204)	0.021(0.204)
1	0.069(0.237)	0.822(0.411)	0.092(0.338)	0.006(0.023)	0.006(0.023)	0.006(0.023)
2	0.043(0.240)	0.086(0.360)	0.774(0.540)	0.088(0.454)	0.005(0.035)	0.005(0.035)
3	0.025(0.232)	0.025(0.147)	0.105(1.016)	0.763(0.971)	0.070(0.782)	0.012(0.072)
4	0.055(0.823)	0.057(1.062)	0.055(0.822)	0.344(3.322)	0.434(4.752)	0.055(0.823)
5	0.040(1.341)	0.097(2.191)	0.021(0.460)	0.023(0.595)	0.056(2.346)	0.764(4.541)

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

"unemployed"							
row j	0	1	2	3	4	5	$100/(1+\Sigma_{hj})$
0	6.610	3.767	1.299	0.655	0.655	0.655	0.809(0.012)
1	18.880	26.245	12.900	2.350	1.188	1.188	1.090(0.012)
2	15.736	9.332	25.330	11.695	2.067	1.044	1.022(0.010)
3	44.603	8.833	31.974	73.562	43.562	8.833	2.972(0.111)
4	23.925	5.047	5.047	18.209	55.211	33.104	2.247(0.054)
5	3.706	0.554	0.554	0.554	1.981	6.843	0.745(0.013)
"climbers"							
row j	0	1	2	3	4	5	$100/(1+\Sigma_{hj})$
0	51.204	76.356	76.356	64.678	51.204	35.932	4.348(0.000)
1	13.173	25.325	18.752	5.848	2.042	1.027	1.013(0.011)
2	3.670	3.979	16.628	13.219	1.839	0.470	0.685(0.003)
3	4.007	1.368	9.782	26.717	18.836	1.368	1.170(0.016)
4	1.982	0.672	0.672	5.725	13.882	7.322	0.820(0.009)
5	1.072	0.171	0.171	0.171	2.372	3.729	0.414(0.003)
"flexible"							
row j	0	1	2	3	4	5	$100/(1+\Sigma_{hj})$
0	56.775	41.845	19.506	10.259	10.259	5.256	2.293(0.042)
1	32.809	43.730	19.133	8.868	6.023	3.067	1.752(0.009)
2	48.777	48.777	71.540	42.274	18.861	9.755	3.125(0.000)
3	51.738	36.321	51.738	105.353	65.328	19.075	4.372(0.065)
4	69.147	37.164	37.164	69.147	152.242	69.147	6.107(0.180)
5	39.610	27.589	14.386	14.386	60.103	94.577	3.795(0.075)
"low wage"							
row j	0	1	2	3	4	5	$100/(1+\Sigma_{hj})$
0	36.556	49.657	30.005	10.540	3.971	3.971	1.992(0.066)
1	3.442	7.721	4.383	0.283	0.283	0.283	0.532(0.002)
2	1.899	3.579	8.017	3.681	0.210	0.210	0.458(0.006)
3	2.940	2.281	8.735	17.859	6.702	1.000	1.000(0.009)
4	25.384	25.384	25.384	117.125	122.506	25.384	5.041(0.209)
5	5.118	10.858	2.395	2.458	4.052	21.698	1.548(0.034)

Table 2: Posterior expectation of the variance of the individual transition probabilities $100\xi^s_{i,jk}$ (in percent) in the various clusters as defined in (17); last column: posterior expectation and, in parenthesis, posterior standard deviation of the amount of unobserved heterogeneity in row j defined in (18) as $1/(1+\Sigma_{hj})$ and multiplied by a factor 100

	Markov chain clustering			Dirichlet multinomial clustering		
	1st Qu.	Median	3rd Qu.	1st Qu.	Median	3rd Qu.
"unemployed"	0.8480	0.9915	0.9994	0.8494	0.9850	0.9981
"climbers"	0.7452	0.9279	0.9854	0.7456	0.9216	0.9801
"low wage"	0.7432	0.9134	0.9861	0.6546	0.8680	0.9728
"flexible"	0.6749	0.8795	0.9812	0.6540	0.8525	0.9650
overall	0.7465	0.9368	0.9921	0.7226	0.9213	0.9863

Table 3: Segmentation power of Markov chain clustering (left hand side) and Dirichlet multinomial clustering (right hand side); 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.

	"low wage"	"flexible"	"climbers"
Intercept	-0.963(0.137)	-0.975(0.129)	-0.875(0.120)
Unemployment rate in district	0.057(0.017)	0.026(0.016)	-0.017(0.017)
Unskilled	-0.129(0.100)	-0.262(0.098)	-0.713(0.091)
Skilled	-0.382(0.070)	-0.523(0.076)	-0.591(0.061)
White collar	-0.983(0.081)	-1.305(0.083)	-0.293(0.063)
Start in wage category 1	1.389(0.095)	1.880(0.096)	2.103(0.089)
Start in wage category 2	1.510(0.125)	1.433(0.129)	2.537(0.117)
Start in wage category 3	1.208(0.156)	0.841(0.167)	2.347(0.138)
Start in wage category 4	1.210(0.185)	0.717(0.191)	2.077(0.149)
Start in wage category 5	0.730(0.318)	0.568(0.433)	2.205(0.220)
Start in year 1976	-0.024(0.130)	0.054(0.135)	0.184(0.114)
Start in year 1977	-0.146(0.132)	0.072(0.124)	0.217(0.104)
Start in year 1978	0.097(0.128)	0.009(0.122)	0.222(0.104)
Start in year 1979	0.031(0.124)	0.062(0.126)	0.129(0.103)
Start in year 1980	-0.175(0.120)	-0.059(0.123)	-0.020(0.104)

Table 4: Multinomial logit model to explain group membership in a particular cluster (baseline: "unemployment" cluster); the numbers are the posterior expectation and, in parenthesis, the posterior standard deviation of the various regression coefficients.

year	real GDP-growth
1975	$-0.4\,\%$
1976	$4.6\,\%$
1977	$5.0\,\%$
1978	$-0.1\,\%$
1979	$5.5\,\%$
1980	$1.8\,\%$

Table 5: Real GDP-growth in Austria in the years 1975 – 1980 (Source: Statistik Austria)

Figure 1: Individual wage mobility time series of nine randomly selected employees; x-axis: time t (in years); y-axis: income class k (k ranging from 0 to 5).

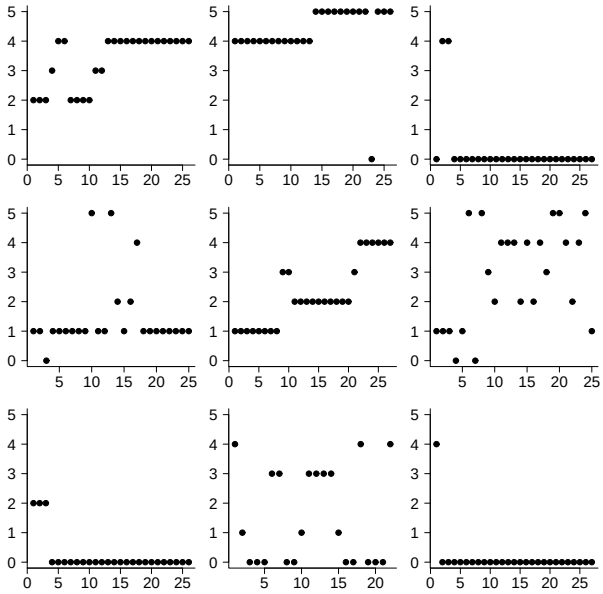
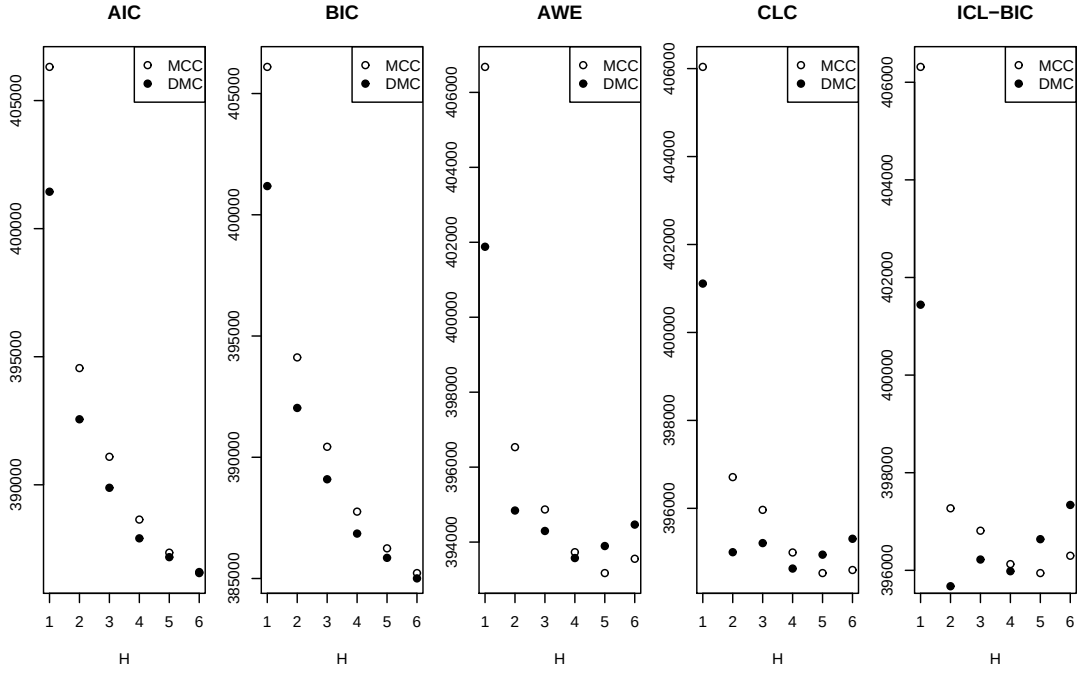


Figure 2: Model selection criteria for various numbers H of clusters for Markov chain clustering (MCC) and Dirichlet multinomial clustering (DMC)



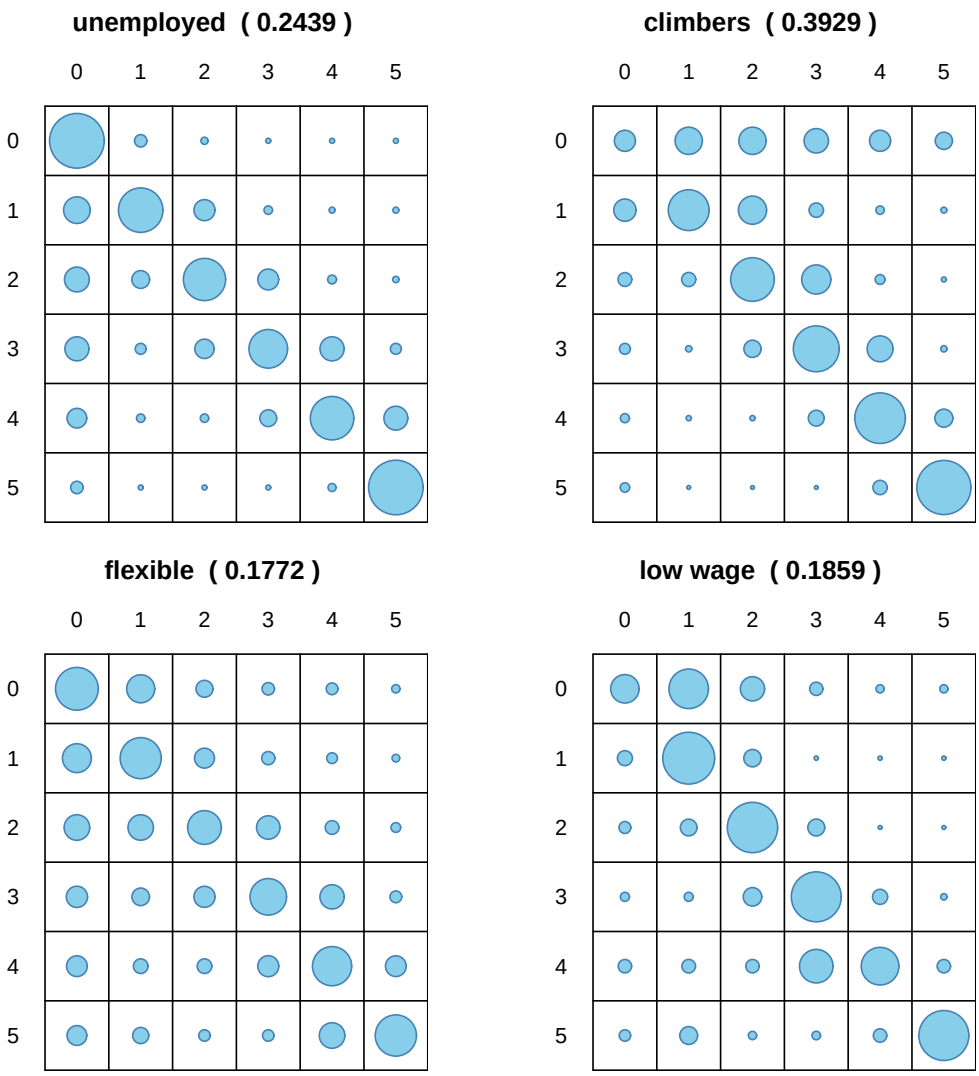


Figure 3: Visualization of posterior expectation of the transition matrices $\boldsymbol{\xi}_1$, $\boldsymbol{\xi}_2$, $\boldsymbol{\xi}_3$, and $\boldsymbol{\xi}_4$ obtained by Dirichlet multinomial clustering. The circular areas are proportional to the size of the corresponding entry in the transition matrix. Posterior expectations of the corresponding group sizes η_1, η_2, η_3 and η_4 are indicated in the parenthesis.

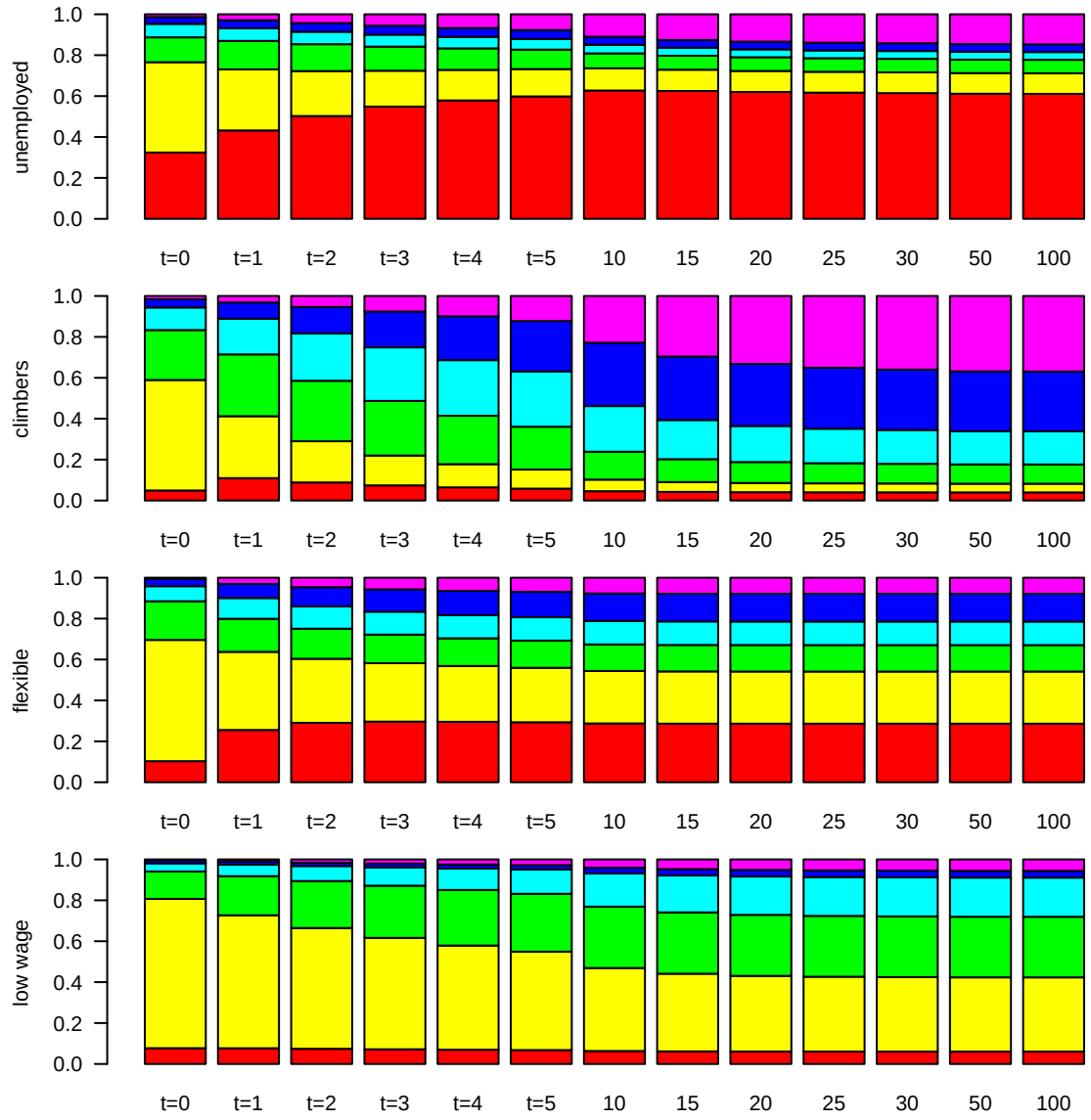


Figure 4: Posterior expectation of the wage distribution $\boldsymbol{\pi}_{h,t}$ over the wage categories 0 to 5 after a period of t years in the various clusters.

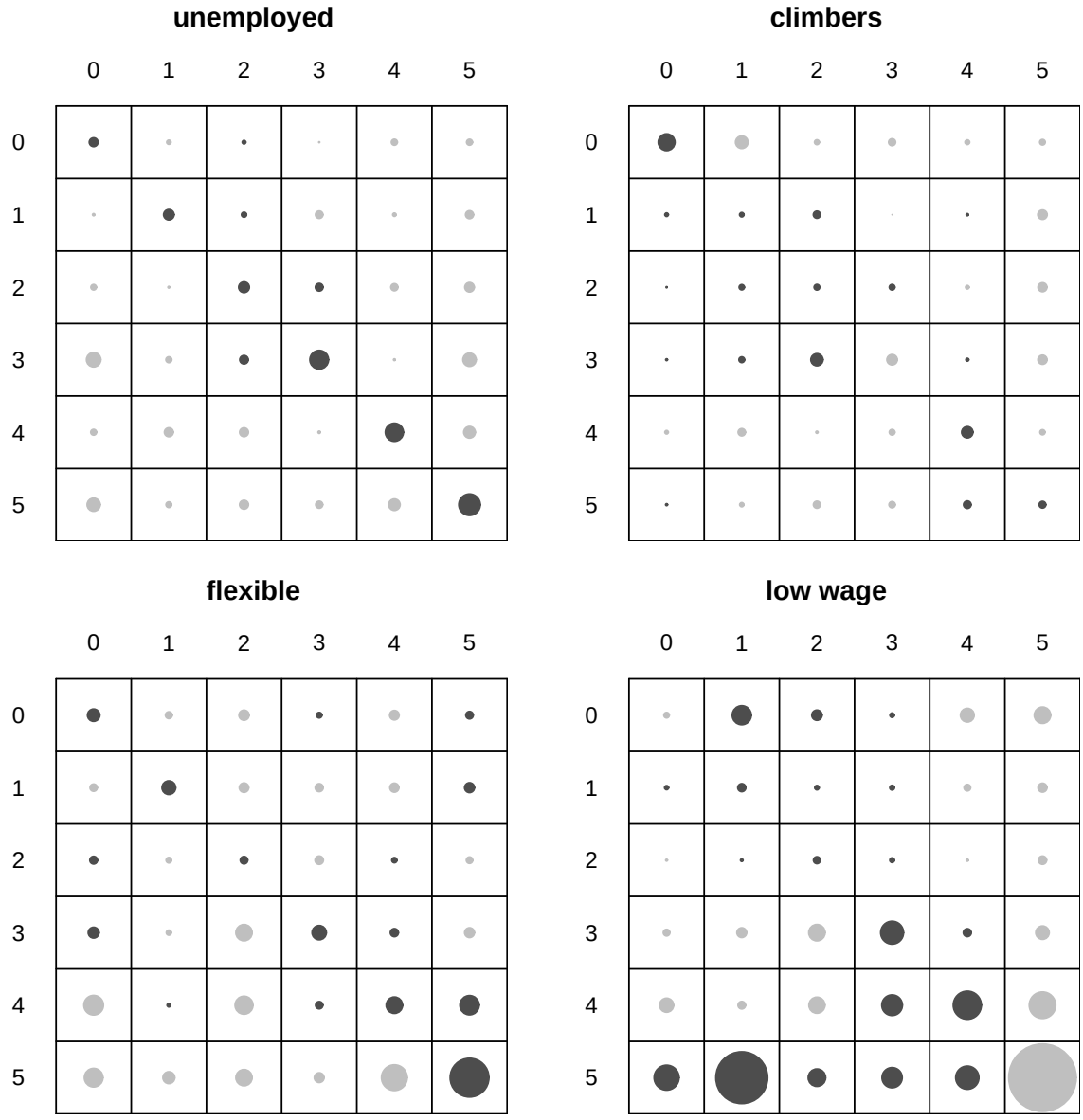


Figure 5: Difference between the posterior expectation of the cluster-specific transition matrices $\boldsymbol{\xi}_h$ obtained by Dirichlet multinomial clustering (DMC) and Markov chain clustering (MCC); each cell shows the difference $\mathrm{E}(\xi_{h,jk}|\mathbf{y}, \mathrm{DMC}) - \mathrm{E}(\xi_{h,jk}|\mathbf{y}, \mathrm{MCC})$; dark gray: difference negative, i.e. $\mathrm{E}(\xi_{h,jk}|\mathbf{y}, \mathrm{MCC}) > \mathrm{E}(\xi_{h,jk}|\mathbf{y}, \mathrm{DMC})$, light gray: difference positive, i.e. $\mathrm{E}(\xi_{h,jk}|\mathbf{y}, \mathrm{DMC}) > \mathrm{E}(\xi_{h,jk}|\mathbf{y}, \mathrm{MCC})$; minimal difference equals -0.1712, maximal difference equals 0.2904.

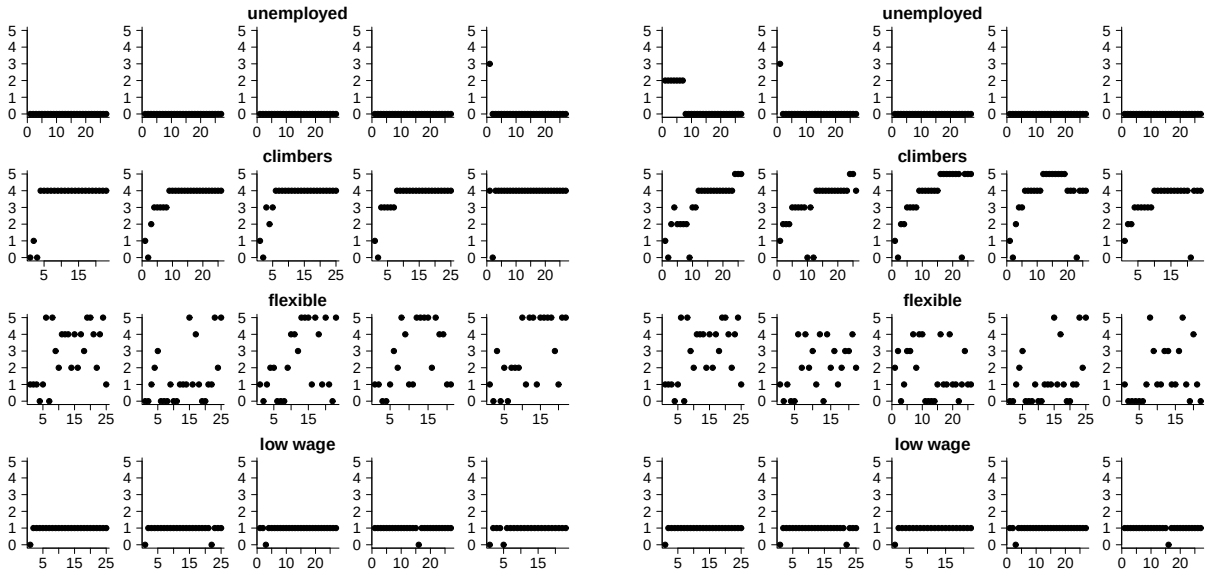


Figure 6: Typical group members within each cluster: wage careers of the five individuals with the highest posterior classification probability; left hand side: Markov chain clustering; right hand side: Dirichlet multinomial clustering

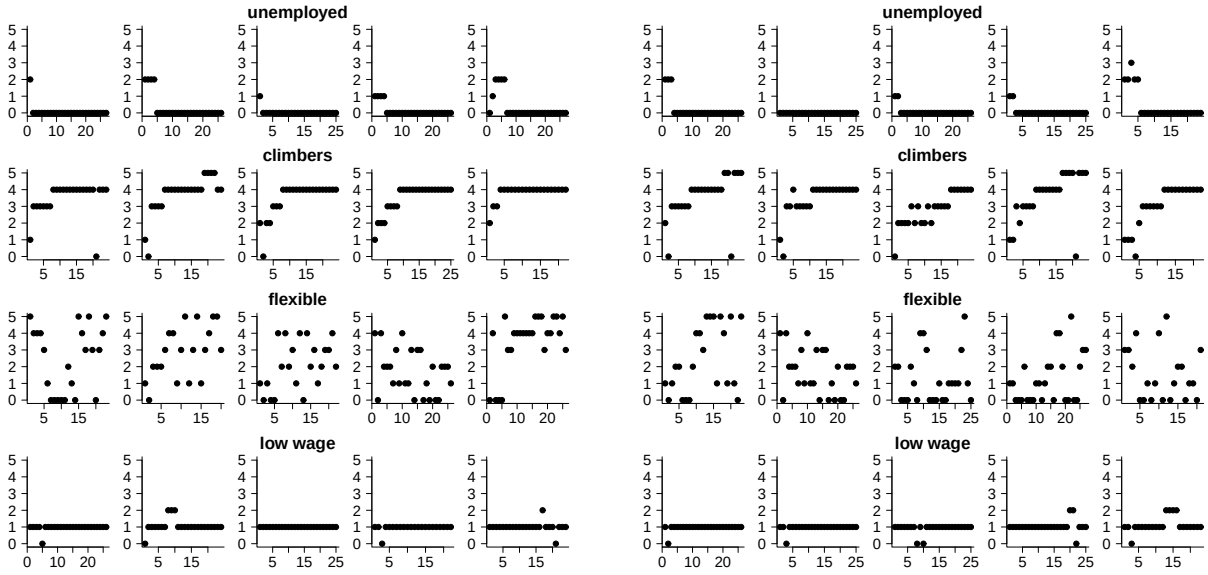


Figure 7: Typical group members within each cluster: wage careers of the individuals no. 10, 25, 50, 100 and 200 in the posterior classification probability ranking; left hand side: Markov chain clustering; right hand side: Dirichlet multinomial clustering