Bayesian Analysis (2010) **5**, Number 2, pp. 345–368

Model-based Clustering of Categorical Time Series

Christoph Pamminger* and Sylvia Frühwirth-Schnatter†

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

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 through the use of clustering techniques. 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. In such an approach, each time series $\mathbf{y}_i$, $i = 1, \ldots, N$, in a panel of N time series is considered to be a single entity and a finite mixture model with H components is assumed as data generating process for $\mathbf{y}_i$. Clustering is achieved as for a traditional finite mixture model by assigning each time $\mathbf{y}_i$ to one of the H groups. The component specific density $p(\mathbf{y}_i|\boldsymbol{\vartheta}_h)$ of the finite mixture model, also called clustering kernel, plays a crucial role in the corresponding clustering procedure and has to capture salient features of the observed time series $\mathbf{y}_i$. Various such clustering kernels were suggested for panels with real-valued time series observations by Frühwirth-Schnatter and Kaufmann (2008). Recently, Juárez and Steel (2010) suggested to use skew-t distributions to capture skewness in the cluster-specific sampling density.

*Department of Applied Statistics, Johannes Kepler University Linz, Austria, mailto:christoph.pamminger@jku.at

†Department of Applied Statistics, Johannes Kepler University Linz, Austria, mailto:sylvia.fruehwirth-schnatter@jku.at

 DOI:10.1214/10-BA606

The present paper focuses on clustering discrete-valued time series obtained by observing a categorical variable with several states. Our application in Section 5 deals with a panel reporting the wage category in successive years for young men entering the Austrian labor market between 1975 and 1980, see Figure 1. The panel contains almost ten thousand of such wage careers and we are searching for clusters of individuals with similar wage mobility behavior.

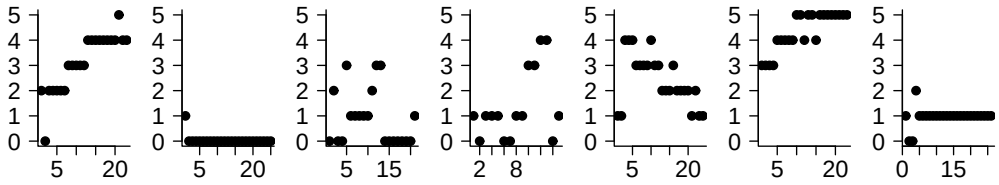

Figure 1: Individual wage mobility time series of seven randomly selected employees; x-axis: time t (in years); y-axis: income class k (k ranging from 0 to 5).

For discrete-valued time series it is particularly difficult to define distance measures and model-based clustering appears to be a promising alternative. We consider clustering kernels which are based on first-order time-homogeneous Markov chain models. One approach, called Markov chain clustering, assumes that all time series within a cluster could be sufficiently described by the same cluster-specific transition matrix. Earlier applications of this approach include Cadez et al. (2003) who clustered users according to their behavior on a web site, Ramoni et al. (2002) who clustered sensor data from mobile robots and Frydman (2005) who considered an application to bond ratings migration. 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. Our second clustering approach, called Dirichlet multinomial clustering, could be viewed as a finite mixture of random-effects models designed specifically to capture unobserved heterogeneity in the transition behavior across time series within the same cluster. Such a model may be regarded as a finite 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.

The remaining paper is organized as follows. Section 2 discusses Markov chain clustering as well as Dirichlet multinomial clustering. Bayesian estimation using a two-block Markov chain Monte Carlo sampler as in Frühwirth-Schnatter and Kaufmann (2008) is considered in Section 3. In Section 4 we give a short review of some well-known criteria for selecting the number of clusters. 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

Let $\{y_{it}\}, t = 0, \ldots, T_i$ be a panel of categorical time series observed for N units $i = 1, \ldots, N$ on T_i occasions with y_{it} taking K potential states labeled by $\{1, \ldots, K\}$. Let $\mathbf{y}_i = \{y_{i0}, \ldots, y_{i,T_i}\}$ denote an individual time series. Model-based clustering assumes that H hidden clusters are present and the clustering kernel $p(\mathbf{y}_i|\boldsymbol{\vartheta}_h)$ with cluster-specific parameter $\boldsymbol{\vartheta}_h$ could be 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. The group indicators $\mathbf{S} = (S_1, \ldots, S_N)$ are a priori independent and $\Pr(S_i = h) = \eta_h$, where $\sum_{h=1}^{H} \eta_h = 1$.

2.1 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 $\xi_{jk} = \Pr(y_{it} = k|y_{i,t-1} = j)$, $j, k = 1, \ldots, K$. Each row of $\boldsymbol{\xi}$ represents a probability distribution over the discrete set $\{1, \ldots, K\}$, i.e. $\sum_{k=1}^{K} \xi_{jk} = 1$.

Markov chain clustering is based on choosing such a Markov chain model with cluster-specific transition matrix $\boldsymbol{\xi}_h$ as clustering kernel. Hence, the group-specific parameter $\boldsymbol{\vartheta}_h$ is equal to $\boldsymbol{\xi}_h$ and the clustering kernel $p(\mathbf{y}_i|\boldsymbol{\xi}_h)$ reads:

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

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 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 $\boldsymbol{\Lambda}_h = \mathrm{Diag}\,(\lambda_{h,1}, \ldots, \lambda_{h,K})$ 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.

2.2 Dirichlet Multinomial Clustering

Model-based clustering using a finite mixture of Markov chain models implies that each time series $\mathbf{y}_i$ is generated by a Markov chain model with individual transition

matrix $\boldsymbol{\xi}_i^s$. While under Markov chain clustering $\boldsymbol{\xi}_i^s$ is equal to the cluster-specific transition matrix $\boldsymbol{\xi}_h$ for all time series in cluster h, we suggest a generalization which takes unobserved heterogeneity within each cluster into account, i.e. for each time series in cluster h the individual transition matrix $\boldsymbol{\xi}_i^s$ is allowed to deviate from the cluster-specific transition matrix $\boldsymbol{\xi}_h$. To describe this deviation, the Dirichlet multinomial model is applied to each row of $\boldsymbol{\xi}_i^s$. We assume that the rows of $\boldsymbol{\xi}_i^s$ are independent and that each row $\boldsymbol{\xi}_{i,j\cdot}^s$, $j = 1, \ldots, K$, follows a Dirichlet distribution with cluster-specific parameter $\mathbf{e}_{h,j\cdot} = (e_{h,j1}, \ldots, e_{h,jK})$:

$$\boldsymbol{\xi}_{i,j\cdot}^s | (S_i = h) \sim \mathcal{D}\left(e_{h,j1}, \ldots, e_{h,jK}\right), \qquad j = 1, \ldots, K. \tag{2}$$

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 (2).

Despite unobserved heterogeneity, each cluster is characterized by a "typical" cluster-specific transition matrix $\boldsymbol{\xi}_h$ given by the expected value of $\boldsymbol{\xi}_i^s$ in group h. The elements of $\boldsymbol{\xi}_h$ read:

$$\xi_{h,jk} = \mathrm{E}(\xi_{i,jk}^s | S_i = h, \mathbf{e}_h) = \frac{e_{h,jk}}{\sum_{l=1}^K e_{h,jl}}. \tag{3}$$

It follows immediately that each row of $\mathbf{e}_h$ determines the corresponding row in the cluster-specific transition matrix $\boldsymbol{\xi}_h$. The matrices $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_H$ may be compared with the corresponding matrices in the Markov chain clustering approach studied in Subsection 2.1.

The variability of $\boldsymbol{\xi}_i^s$ within each cluster is given by the variance of the individual transition probabilities $\xi_{i,jk}^s$:

$$\mathrm{Var}(\xi_{i,jk}^s | S_i = h, \mathbf{e}_h) = \frac{e_{h,jk} \sum_{l \neq k} e_{h,jl}}{\left(\sum_{l=1}^K e_{h,jl}\right)^2 \left(1 + \sum_{l=1}^K e_{h,jl}\right)}. \tag{4}$$

It can easily be shown that

$$\frac{\mathrm{Var}(\xi_{i,jk}^s | S_i = h, \mathbf{e}_h)}{\mathrm{E}(\xi_{i,jk}^s | S_i = h, \mathbf{e}_h)\ (1 - \mathrm{E}(\xi_{i,jk}^s | S_i = h, \mathbf{e}_h))} = \frac{1}{1 + \sum_{l=1}^K e_{h,jl}}. \tag{5}$$

Thus the row sums $\Sigma_{hj} = \sum_{l=1}^K e_{h,jl}$ 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.

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. This is easily verified by combining the conditional distribution $p(\mathbf{y}_i|\boldsymbol{\xi}_i^s)$ with (2):

$$\begin{aligned} p(\mathbf{y}_i|\mathbf{e}_h) &= \int p(\mathbf{y}_i|\boldsymbol{\xi}_i^s)p(\boldsymbol{\xi}_i^s|\mathbf{e}_h)d\boldsymbol{\xi}_i^s = \\ &= \frac{\prod_{j=1}^K \Gamma(\sum_{k=1}^K e_{h,jk})}{\prod_{j=1}^K \prod_{k=1}^K \Gamma(e_{h,jk})} \frac{\prod_{j=1}^K \prod_{k=1}^K \Gamma(N_{i,jk} + e_{h,jk})}{\prod_{j=1}^K \Gamma(\sum_{k=1}^K (N_{i,jk} + e_{h,jk}))}. \end{aligned} \tag{6}$$

Hence, the clustering kernel may be entirely characterized by the group-specific parameter $\mathbf{e}_h$. It is evident from (6) that this clustering kernel no longer is a first-order Markov process but allows for higher order dependence.

Finally, 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 $\boldsymbol{\xi}_i^s$.

2.3 Clustering Using an Inhomogeneous Markov Chain

If additional covariate information is available, an interesting extension is to use an inhomogeneous Markov chain as clustering kernel. Clustering could be based on modeling the rows of the transition matrix in group h through a dynamic multinomial logit model:

$$\xi_{h,jk} = \Pr(y_{it} = k|y_{i,t-1} = j, S_i = h) = \frac{\exp(\gamma_{h,jk} + \mathbf{x}_{it}\boldsymbol{\beta}_{h,k})}{\sum_{l=1}^K \exp(\gamma_{h,jl} + \mathbf{x}_{it}\boldsymbol{\beta}_{h,l})}, \tag{7}$$

where $\boldsymbol{\beta}_{h,k}$ is a group and category specific regression parameter capturing the effect of the covariates $\mathbf{x}_{it}$. To achieve identifiability, it has to be assumed for each $j = 1, \ldots, K$ that $\gamma_{h,jk_0} = 0$ for some baseline category k_0. If no covariates are present, then (7) reduces to Markov chain clustering with the transition matrix $\boldsymbol{\xi}_h$ being parameterized in terms of $\gamma_{h,jk}$.

3 Bayesian Inference for a Fixed Number of Clusters

The latent group indicators $\mathbf{S}$ are estimated along with the group-specific parameters $(\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_H)$ and the group sizes $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_H)$. We assume prior independence between $\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_H$ and $\boldsymbol{\eta} \sim \mathcal{D}(\alpha_0, \ldots, \alpha_0)$. As in Frühwirth-Schnatter and Kaufmann (2008), we use an MCMC sampler described in **Algorithm 1**, see also the short note by Ridgeway and Altschuler (1998):

1. Bayes' classification for each individual i: draw $S_i, i = 1, \ldots, N$, from the discrete probability distribution $\Pr(S_i = h|\mathbf{y}_i, \boldsymbol{\eta}, \boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_H) \propto p(\mathbf{y}_i|\boldsymbol{\vartheta}_h)\eta_h,\ h = 1, \ldots, H$.
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 $p(\boldsymbol{\vartheta}_h|\mathbf{S}, \mathbf{y}), h = 1, \ldots, H$.

Fougère and Kamionka (2003) applied a similar MCMC sampler to the mover-stayer model. Alternatively, Ramoni et al. (2002) applied a heuristic Bayesian 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

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 Algorithm 1 with $\boldsymbol{\vartheta}_h = \boldsymbol{\xi}_h, h = 1, \ldots, H$. Classification in Step 1 is based on the clustering kernel $p(\mathbf{y}_i|\boldsymbol{\xi}_h)$ defined in (1). The complete data posterior distribution $p(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_H|\mathbf{S}, \mathbf{y})$ appearing in Step 3 is of closed form:

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

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

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

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, see also Section 5.

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 deal with zero transitions, while the EM algorithm breaks down, if no transitions starting from j are observed in group h, i.e. $\sum_{k=1}^{K} N^h_{jk} = 0$ for some

j. In this case the observed-data likelihood function $p(\mathbf{y}|\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_H,\mathbf{S})$ is independent of the jth row $\boldsymbol{\xi}_{h,j\cdot}$ of $\boldsymbol{\xi}_h$ 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, Agresti (1990) suggests to add a small constant, 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 observed-data likelihood $p(\mathbf{y}|\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_H,\mathbf{S})$ with the Dirichlet prior $\mathcal{D}(e_{0,j1},\ldots,e_{0,jK})$ for each row $\boldsymbol{\xi}_{h,j\cdot}$ within a Bayesian approach.

3.2 Bayesian Inference for Dirichlet Multinomial Clustering

In contrast to Subsection 3.1, no conjugate prior allowing straightforward MCMC estimation is available for the group-specific parameters $\mathbf{e}_h, h = 1,\ldots,H$. 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. The structure of the complete-data likelihood suggests to assume that all rows of $\mathbf{e}_{h,j\cdot}$ are independent within and across each group.

To take dependencies between the elements of row $\mathbf{e}_{h,j\cdot}$ 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 $p_{jk} = N_0\hat{\xi}_{jk}/(\alpha+N_0)$. The prior density 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:

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

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 $\gamma \sim \mathcal{G}(\alpha,\beta)$. Marginally, after integrating over γ, $\mathbf{e}_{h,j\cdot} - 1 \sim \text{NegMulNom}(p_{j1},\ldots,p_{jK},\beta)$ with $p_{jk} = \lambda_{jk}/(\alpha + \sum_{l=1}^K \lambda_{jl})$. This representation suggests the 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. $\alpha = \beta = 1$.

The parameters $\mathbf{e}_1,\ldots,\mathbf{e}_H$, $\boldsymbol{\eta}$ and the hidden indicators $\mathbf{S}$ are jointly estimated using Algorithm 1 where $\boldsymbol{\vartheta}_h = \mathbf{e}_h$. Classification in Step 1 is based on the clustering kernel

$p(\mathbf{y}_i|\boldsymbol{\vartheta}_h) = p(\mathbf{y}_i|\mathbf{e}_h)$ defined in (6). To implement Step 3 the complete data posterior distribution $p(\mathbf{e}_1, \ldots, \mathbf{e}_H|\mathbf{S}, \mathbf{y})$ has to be derived:

$$p(\mathbf{e}_1, \ldots, \mathbf{e}_H|\mathbf{S}, \mathbf{y}) \propto \prod_{i=1}^{N} p(\mathbf{y}_i|\mathbf{e}_{S_i}) \prod_{h=1}^{H} p(\mathbf{e}_h) \propto \prod_{h=1}^{H} \prod_{j=1}^{K} p(\mathbf{e}_{h,j\cdot}|\mathbf{y}, \mathbf{S})$$

$$p(\mathbf{e}_{h,j\cdot}|\mathbf{y}, \mathbf{S}) \propto p(\mathbf{e}_{h,j\cdot}) \frac{\Gamma(\sum_{k=1}^{K} e_{h,jk})^{N_h}}{\left(\prod_{k=1}^{K} \Gamma(e_{h,jk})\right)^{N_h}} \left(\prod_{i:S_i=h} \frac{\prod_{k=1}^{K} \Gamma(N_{i,jk} + e_{h,jk})}{\Gamma(\sum_{k=1}^{K} (N_{i,jk} + e_{h,jk}))} \right), \quad (8)$$

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 $p(\mathbf{e}_{h,j\cdot}|\mathbf{y}, \mathbf{S})$ is no longer of closed form. 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. We propose each element $e_{h,jk}$ to be updated independently from a discrete random walk proposal density $q(e_{h,jk}|e_{h,jk}^{(m-1)})$. 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 sampling of the individual transition matrices $\boldsymbol{\xi}_1^s, \ldots, \boldsymbol{\xi}_N^s$ because the special structure of the distribution of heterogeneity underlying Dirichlet multinomial clustering leads to a closed form density $p(\mathbf{y}_i|\mathbf{e}_{S_i})$. Such a step would be extremely time consuming as it requires drawing the K rows $\boldsymbol{\xi}_{i,j\cdot}^s$ of $\boldsymbol{\xi}_i^s$ for each $i = 1, \ldots, N$ from $\boldsymbol{\xi}_{i,j\cdot}^s|(S_i = h, \mathbf{e}_h, \mathbf{y}) \sim \mathcal{D}(e_{h,j1} + N_{i,j1}, \ldots, e_{h,jK} + N_{i,jK})$. For the labor market application in Section 5, for instance, dealing with nearly 10 000 time series and 6 categories, this would require sampling from about 60 000 Dirichlet distributions for each MCMC sweep.

3.3 Label Switching and Post-Processing MCMC

Like for any finite mixture model, label switching may occur during MCMC sampling both for Markov chain clustering as well as for Dirichlet multinomial clustering, see Jasra et al. (2005) or Frühwirth-Schnatter (2006, Section 3.5) for a recent review.

In some applications of our clustering methods it may be sufficient to constrain the parameter space appropriately to prevent label switching. For instance, when clustering binary time series where the cluster-specific transition matrix $\boldsymbol{\xi}_h$ is characterized by the two persistence probabilities $\xi_{h,11}$ and $\xi_{h,22}$, it might be possible to identify simple constraints such as $\xi_{1,11} < \ldots < \xi_{H,11}$ or $\xi_{1,22} < \ldots < \xi_{H,22}$. However, it may be difficult

or even impossible to extend this method to clustering time series with more than two categories K.

Subsequently, we follow Frühwirth-Schnatter (2006, p. 96f) to identify the finite mixture model both for Markov chain clustering as well as Dirichlet multinomial clustering. We apply k-means clustering to all MH posterior draws of the vector $\mathbf{x}_{m,h} = (\xi_{h,11}^{(m)}, \ldots, \xi_{h,KK}^{(m)})^T$ containing the posterior draws of the group-specific persistence probabilities. The whole method is based on the idea that MCMC draws belonging to the same group will cluster around the same point in the point process representation. If label switching occurred between subsequent draws, then the classification sequence resulting from k-means clustering indicates how to rearrange the group-specific parameters. Provided that the simulation clusters are well-separated, the classification sequence $(d_1^{(m)}, \ldots, d_H^{(m)})$ corresponding to $(\mathbf{x}_{m,1}, \ldots, \mathbf{x}_{m,H})$ 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 $(\boldsymbol{\vartheta}_1, \eta_1)^{(m)}, \ldots, (\boldsymbol{\vartheta}_H, \eta_H)^{(m)}$. Finally, 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.

4 Selecting the Number of Clusters

Let $\boldsymbol{\theta}_H = (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_H, \eta_1, \ldots, \eta_H)$ denote the parameter in a finite mixture model with H components and let d_H be the number of parameters. Let $p(\mathbf{y}|\boldsymbol{\theta}_H)$ denote the likelihood function for fixed H, while $p(\mathbf{y}, \mathbf{S}|\boldsymbol{\theta}_H)$ denotes the complete-data likelihood function.

The selection of H may be based on the posterior distribution $p(H|\mathbf{y}) \propto p(\mathbf{y}|H)p(H)$ which is determined either by computing the marginal likelihood $p(\mathbf{y}|H)$ for various values of H or by running some model space methods, see e.g. Frühwirth-Schnatter (2006, Chapter 4 and 5). However, selecting H in this way does not necessarily lead to H distinct clusters.

This potential discrepancy is 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, and n is the sample size. In the present context of panel data it is not obvious how to choose n (Kass and Raftery 1995). As each time series is modeled independently by a mixture model, 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 correct BIC penalty should be based on some measure of information in the data, as derived e.g. by Kim (1998) in the context of non-stationary time series models.

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 mixtures. $BIC(H)$ 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 (1995), it is not surprising that selecting H as to maximize the marginal likelihood $p(\mathbf{y}|H)$ or the posterior probability $p(H|\mathbf{y})$ may not be adequate either, as demonstrated in various applications of model-based clustering, see e.g. Frühwirth-Schnatter and Pyne (2010).

Several alternative criteria such as approximate weight of evidence $AWE(H)$ (Banfield and Raftery 1993) are able to identify the correct number of clusters even when the component densities are misspecified. Biernacki and Govaert (1997) expressed $AWE(H)$ as a criterion which penalizes the complete data log-likelihood function 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 quality of the resulting partition measured through 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 Algorithm 1. The entropy 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 likelihood function by the entropy rather than by model complexity, i.e. $CLC(H) = -2\ \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_H) + 2\, EN(H, \hat{\boldsymbol{\theta}}_H)$, while 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 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 (Raferzeder and Winter-Ebmer 2007). Transition between the wage categories is described by a transition matrix which determines the income career and career progressions for an individual. Since from an economical point of view it is expected that the income career and career progression is different between employees we apply both Markov chain clustering as well as Dirichlet multinomial clustering to find groups of employees with similar wage mobility behavior.

5.1 Data Description

The data were 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 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 1 up to 6 groups. Concerning prior distributions, we choose $\alpha_0 = 4$ for the Dirichlet prior of the weight distribution $\boldsymbol{\eta}$ (Frühwirth-Schnatter 2006) and take the prior information concerning the wage categories into account. First of all, wages show considerable persistence and staying in the same wage category is more likely than moving to another wage category. Second, wage categories are ordered, hence transitions into adjacent categories are more likely than into any other category. To incorporate these aspects into the prior, the following matrix $\boldsymbol{\xi}^*$,

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

is chosen as prior mean. For Markov chain clustering the hyperparameter of the Dirichlet prior is selected as $\mathbf{e}_{0,j\cdot} = N_0 \times \boldsymbol{\xi}^*_{j\cdot}$ with $N_0 = 10$, while for Dirichlet multinomial clustering the hyperparameters of the negative multinomial distribution are chosen as $\alpha = \beta = 1$, $N_0 = 10$ and $\hat{\boldsymbol{\xi}}_h = \boldsymbol{\xi}^*$. Alternative hyperparameters were considered but showed negligible differences in the results.

We start MCMC estimation by choosing initial values for the group-indicators S_i, $i = 1, \ldots, N$. For $H > 1$ we find an initial clustering by use of a k-means routine implemented in R running on the transition frequencies. For each number H of groups we simulated 10 000 MCMC draws after a burn-in of 15 000 draws. 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, leading to an average acceptance rate of 0.255.

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 clustering as well as under Markov chain clustering, see Table 1. For all criteria depending on the sample size n we consider $n = \sum_{i=1}^{N} T_i$ and $n = N$. The estimators $\hat{\boldsymbol{\theta}}_H$ and $(\hat{\boldsymbol{\theta}}_H^C, \hat{\mathbf{S}})$ are approximated by the MCMC posterior draw maximizing, respectively, the log-likelihood function and the complete data log-likelihood function. Since this introduces a certain random element into computing these criteria, several independent MCMC runs were performed in order to study the effect on model selection. We found that the variance of a particular model selection criterion across independent MCMC runs did not effect the final choice of H. The criteria reported in Table 1 are based on determining for each H the optimal estimators across all independent MCMC runs.

		BIC		*AWE*			*ICL-BIC*	
H	*AIC*	n_1	n_2	n_1	n_2	*CLC*	n_1	n_2
Markov chain clustering								
1	406098.7	406314.4	406405.8	406680.2	406862.9	406038.7	406314.4	406405.8
2	394117.6	394556.2	394742.0	396551.3	396922.8	396710.2	397270.8	397456.6
3	390431.6	391093.2	391373.4	394870.1	395430.5	395965.8	396811.3	397091.5
4	387757.7	388642.2	389016.8	393730.0	394479.2	394996.6	396127.1	396501.7
5	386243.2	387350.6	387819.6	393313.7	394251.7	394528.3	395943.8	396412.7
6	385221.0	386551.3	387114.7	393606.3	394733.1	394651.5	396351.8	396915.3
Dirichlet multinomial clustering								
1	401183.0	401441.9	401551.5	401880.8	402100.1	401111.0	401441.9	401551.5
2	392033.4	392558.4	392780.7	394842.1	395286.7	395002.9	395673.9	395896.2
3	389080.9	389871.9	390206.9	394206.1	394876.0	395176.2	396187.2	396522.2
4	386851.8	387908.8	388356.5	393567.7	394463.1	394683.3	396034.4	396482.1
5	385852.4	387175.5	387735.9	393894.4	395015.0	394947.0	396638.1	397198.5
6	385004.7	386594.0	387267.0	394466.5	395812.6	395308.4	397339.6	398012.6

Table 1: Model selection criteria for various numbers H of clusters for Markov chain clustering as well Dirichlet multinomial clustering; criteria depending on sample size are computed with sample size $n_1 = N$ and sample size $n_2 = \sum_{i=1}^{N} T_i$.

For both clustering kernels, *AIC* and *BIC* decrease with increasing H and suggest at least 6 components. For *BIC* this holds irrespective of sample size n. However, as outlined in Section 4, *BIC* is likely to be overfitting, since 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 all time series in this panel. Thus it is very likely that two or even more components in the Markov mixture model correspond to groups with rather similar transition behavior, rather than to distinct clusters. 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 $H = 4$, again, irrespective of sample size n. Somewhat surprisingly, *CLC* and *ICL-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.

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 is in general preferred to Markov chain clustering. First, this indicates that unobserved heterogeneity is present in the clusters even after accounting for differences in the typical cluster-specific transition behavior. Second, Dirichlet multinomial clustering is expected to exhibit a higher robustness to untypical group members. It should be noted that the difference in BIC 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.

5.4 Empirical Results

To provide additional insight, we decided to discuss the four-cluster solution for both clustering methods in more detail. The MCMC draws are identified as described in Subsection 3.3 by applying k-means clustering to the MCMC draws $\mathbf{x}_{m,h} = (\xi_{h,00}^{(m)}, \ldots, \xi_{h,55}^{(m)})$, $h = 1, \ldots, 4$, $m = 1, \ldots, M$. For Dirichlet multinomial clustering, posterior draws for $\boldsymbol{\xi}_h$ are obtained by applying the nonlinear transformation (3) to each MCMC draw of $\mathbf{e}_h$.

Pairwise scatter plots of the persistence probabilities $\xi_{h,00}^{(m)}$, $\xi_{h,11}^{(m)}$ and $\xi_{h,22}^{(m)}$ are provided for illustration in Figure 2. Evidently, the MCMC draws form four well-separated simulation clusters. Thus it is not surprising that 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 finite mixture model.

Analyzing Wage Mobility

To analyze wage mobility in the different clusters we investigate the posterior distribution of the group-specific transition matrix $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_4$. Posterior inference is summarized for Dirichlet multinomial clustering in Table 2, reporting for each cell $\mathrm{E}(\xi_{h,jk}|\mathbf{y})$ and $\mathrm{SD}(\xi_{h,jk}|\mathbf{y})$, while Table 3 reports the inefficiency factors for two clusters. 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 these results, we assign a labeling to the various clusters, namely "low wage", "flexible", "unemployed", and "climbers" which will be further corroborated by the long-run wage distribution as well as by the wage careers of typical group members to be discussed later in this subsection.

A remarkable difference in the transition behavior of individuals belonging to differ-

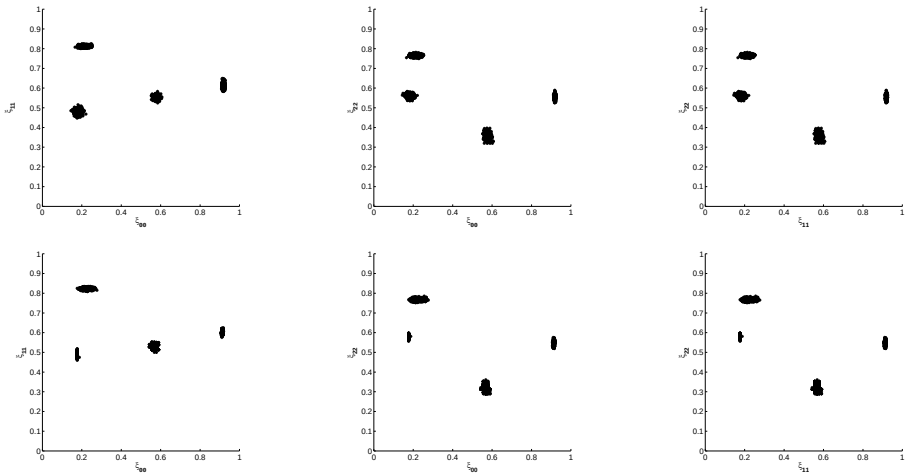

Figure 2: Scatter plots of the MCMC draws of the persistence probabilities $(\xi_{h,00}, \xi_{h,11})$ (left hand side), $(\xi_{h,00}, \xi_{h,22})$ (middle) and $(\xi_{h,11}, \xi_{h,22})$ (right hand side) obtained under Markov chain clustering (top) and Dirichlet multinomial clustering (bottom).

ent clusters is evident from Figure 3. Consider, for instance, the first column containing the risk for an individual 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 risk to remain in the no-income category is located in the top left cell and is much higher in the "unemployed" cluster than in 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. Finally, 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 higher in the "climbers" cluster than in the other clusters.

These differences in the transition matrix have a strong impact on wage mobility and the long-run wage career of the group members. Figure 4 shows the posterior expectation $\mathrm{E}(\boldsymbol{\pi}_{h,t}|\mathbf{y}, \boldsymbol{\pi}_{h,0})$ of the cluster-specific wage distribution $\boldsymbol{\pi}_{h,t} = \boldsymbol{\pi}_{h,0}\boldsymbol{\xi}_h^t$ after a period of t years. The initial wage distribution $\boldsymbol{\pi}_{h,0}$ is estimated from the initial wage category y_{i0} observed for all individuals i being classified to group h. For $t = 100$, the wage distribution is practically equal to the equilibrium distribution of the transition matrix $\boldsymbol{\xi}_h$. 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.

			"unemployed"			
	0	1	2	3	4	5
0	0.913(0.220)	0.047(0.121)	0.016(0.040)	0.008(0.020)	0.008(0.020)	0.008(0.020)
1	0.216(0.666)	0.602(0.778)	0.139(0.491)	0.020(0.294)	0.011(0.184)	0.011(0.184)
2	0.184(0.678)	0.098(0.546)	0.547(0.951)	0.136(0.595)	0.023(0.368)	0.012(0.135)
3	0.171(0.932)	0.033(0.234)	0.130(0.919)	0.456(1.436)	0.177(1.119)	0.033(0.234)
4	0.120(1.043)	0.024(0.266)	0.024(0.266)	0.089(0.940)	0.563(1.586)	0.180(1.467)
5	0.050(0.574)	0.010(0.102)	0.010(0.102)	0.010(0.102)	0.028(0.457)	0.892(0.778)
			"climbers"			
	0	1	2	3	4	5
0	0.176(0.035)	0.235(0.151)	0.235(0.046)	0.176(0.035)	0.118(0.023)	0.059(0.012)
1	0.164(0.518)	0.491(0.916)	0.249(0.600)	0.064(0.323)	0.025(0.231)	0.007(0.077)
2	0.062(0.248)	0.068(0.287)	0.580(0.676)	0.255(0.486)	0.028(0.180)	0.006(0.053)
3	0.038(0.133)	0.013(0.044)	0.093(0.349)	0.647(0.597)	0.196(0.403)	0.013(0.044)
4	0.026(0.061)	0.013(0.030)	0.013(0.030)	0.102(0.252)	0.756(0.412)	0.091(0.311)
5	0.027(0.151)	0.004(0.021)	0.004(0.021)	0.004(0.021)	0.043(0.272)	0.918(0.316)
			"low wage"			
	0	1	2	3	4	5
0	0.225(1.763)	0.502(1.307)	0.172(1.118)	0.051(0.643)	0.025(0.309)	0.025(0.309)
1	0.067(0.245)	0.823(0.419)	0.094(0.309)	0.008(0.144)	0.004(0.028)	0.004(0.028)
2	0.041(0.238)	0.083(0.395)	0.768(0.540)	0.095(0.469)	0.007(0.059)	0.007(0.059)
3	0.024(0.249)	0.024(0.293)	0.111(1.079)	0.736(1.455)	0.093(0.682)	0.012(0.110)
4	0.022(0.482)	0.014(0.209)	0.014(0.209)	0.042(0.653)	0.773(1.323)	0.136(1.174)
5	0.025(0.773)	0.022(0.659)	0.017(0.358)	0.017(0.357)	0.377(2.807)	0.542(3.502)
			"flexible"			
	0	1	2	3	4	5
0	0.568(0.903)	0.244(0.562)	0.081(0.216)	0.054(0.144)	0.027(0.072)	0.027(0.072)
1	0.255(0.733)	0.530(0.866)	0.107(0.382)	0.054(0.191)	0.027(0.096)	0.027(0.096)
2	0.217(0.698)	0.214(0.763)	0.322(1.244)	0.153(0.620)	0.063(0.414)	0.031(0.207)
3	0.167(0.286)	0.112(0.191)	0.112(0.191)	0.386(1.048)	0.167(0.286)	0.056(0.095)
4	0.145(0.412)	0.072(0.206)	0.072(0.206)	0.145(0.412)	0.421(1.647)	0.145(0.412)
5	0.116(0.923)	0.088(1.083)	0.035(0.440)	0.035(0.440)	0.151(1.058)	0.574(2.064)

Table 2: 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.

The long-run 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 than 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 remaining clusters there is little difference between their 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 "flex-

Row	"low wage"						"flexible"					
1	14.7	8.19	7.95	24.6	29.7	29.7	6.86	5.75	7.76	7.76	7.76	7.76
2	5.22	12.7	6.73	34.1	42.9	42.9	10.1	9.23	13.9	13.9	13.9	13.9
3	6.18	7.9	10.3	12.3	41.5	41.5	4.34	6.00	7.44	13.1	28.8	28.3
4	24.1	18.8	20.6	21.4	7.16	38.0	5.72	5.72	5.72	5.72	5.72	5.72
5	35.6	40.7	40.7	12.4	13.3	12.6	9.01	9.01	9.01	9.01	9.01	9.01
6	5.26	4.76	39.7	39.4	23.1	27.2	12.8	12.6	31.3	31.3	9.43	15.7

Table 3: Inefficiency factors of the MCMC draws obtained for each row $j = 1, \ldots, 6$ of the cluster-specific transition matrices $\boldsymbol{\xi}_{h,j\cdot}$ for two clusters.

ible" cluster have a much higher risk to end up in the no-income category, members of the "low wage" cluster end up in lower wage categories, while members of the "climbers" cluster move into the highest wage categories.

Analyzing Unobserved Heterogeneity

To analyze how much unobserved heterogeneity is present in the various clusters, we report in Table 4 the posterior expectation of the variance of the individual transition probabilities $\xi^s_{i,jk}$ defined in (4) as well as posterior expectation and standard deviation of the row-specific unobserved heterogeneity measure defined in (5). These measures vary 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 "low wage" cluster. In general, persistence probabilities have higher variances than the off-diagonal elements.

Apart from a few exceptions, 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 (DMC) are similar to the ones obtained by Markov chain clustering (MCC). 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. For illustration, Figure 5 shows the expected posterior difference $\mathrm{E}(\xi_{h,jk}|\mathbf{y}, \mathrm{DMC}) - \mathrm{E}(\xi_{h,jk}|\mathbf{y}, \mathrm{MCC})$ for the "unemployment" and "low wage" cluster. We observe the biggest differences in the "low wage" cluster, where the expected chance to remain in the highest wage category is 54.2 % under DMC, while for MCC the expected chance is as large as 99.5 %. In general, differences occur mainly for the persistence probabilities with MCC overrating persistence in relation to DMC. 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.

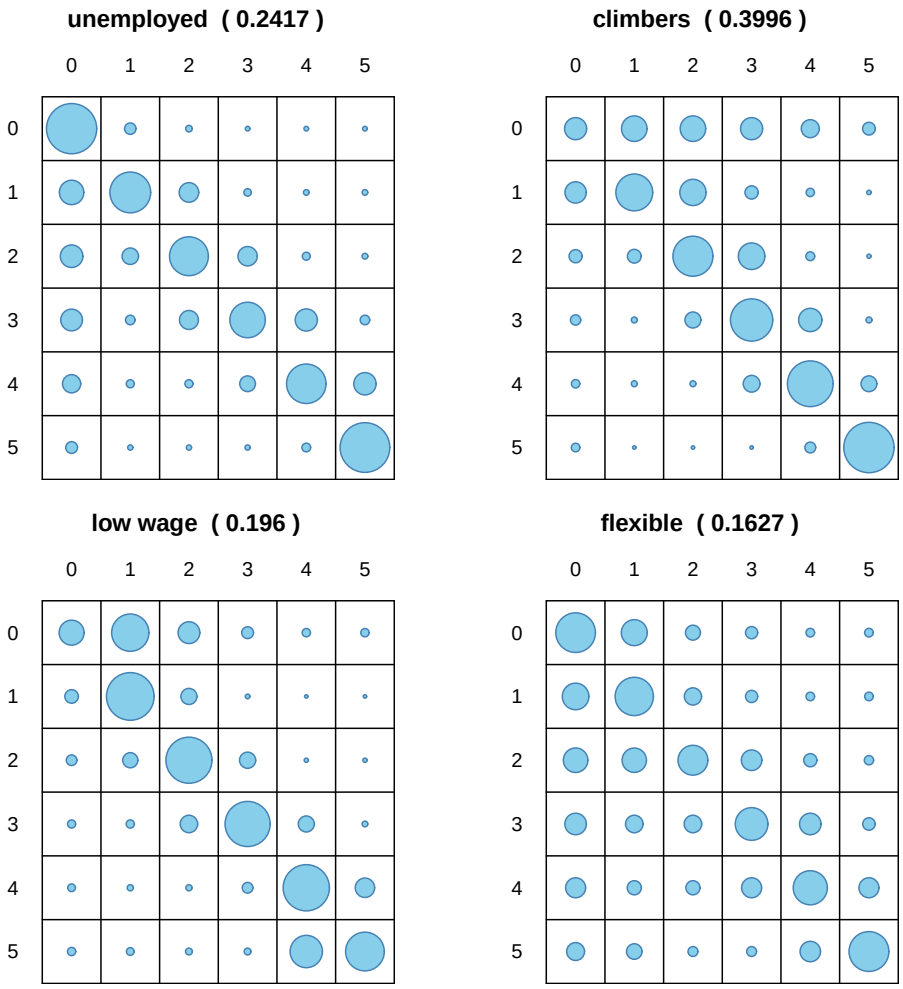

Figure 3: Visualization of posterior expectation of the transition matrices $\boldsymbol{\xi}_1$, $\boldsymbol{\xi}_2$, $\boldsymbol{\xi}_3$, and $\boldsymbol{\xi}_4$ obtained by Dirichlet multinomial clustering. The circular areas are proportional to the size of the corresponding entry in the transition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

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 $\hat{t}_{ih} = \mathrm{E}(t_{ih}(\boldsymbol{\theta}_4)|\mathbf{y})$ is estimated over the last 10 000 MCMC draws for MCC and over

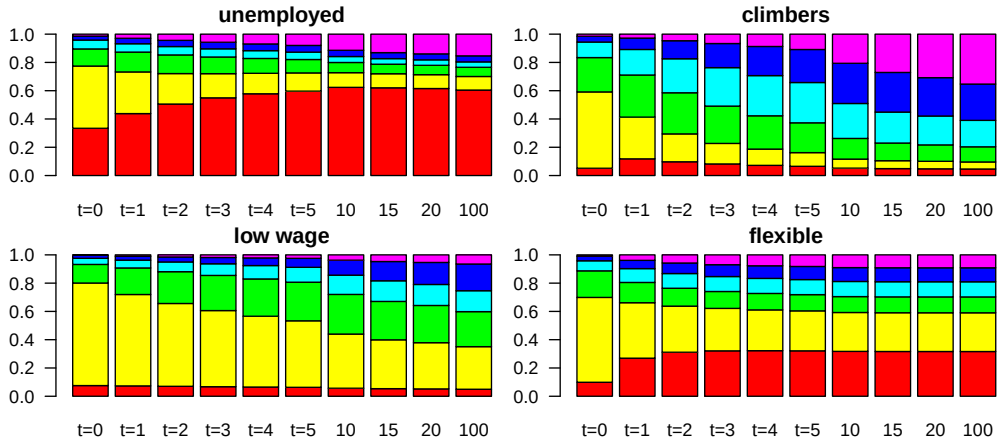

Figure 4: Posterior expectation of the wage distribution $\boldsymbol{\pi}_{h,t}$ over the wage categories 0 to 5 after a period of t years in the various clusters.

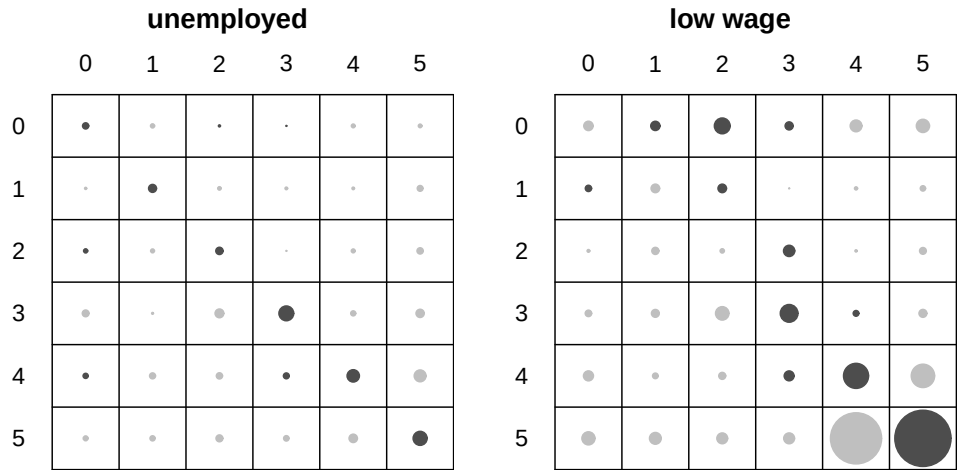

Figure 5: Each cell shows the difference between the posterior expectation of the cluster-specific transition matrices $\boldsymbol{\xi}_h$ obtained by Dirichlet multinomial clustering (DMC) and Markov chain clustering (MCC); dark gray: $\mathrm{E}(\xi_{h,jk}|\mathbf{y}, \mathrm{MCC}) > \mathrm{E}(\xi_{h,jk}|\mathbf{y}, \mathrm{DMC})$, light gray: $\mathrm{E}(\xi_{h,jk}|\mathbf{y}, \mathrm{MCC}) < \mathrm{E}(\xi_{h,jk}|\mathbf{y}, \mathrm{DMC})$ (minimal difference: -0.4525, maximal difference: 0.3744).

the last 1000 MCMC draws for DMC (for CPU running time reasons) with a thinning parameter equal to 10. Each employee is then allocated to that cluster which exhibits the maximum posterior probability, i.e. $\hat{S}_i$ is defined such 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.

"unemployed"							
row j	0	1	2	3	4	5	$100/(1+\Sigma_{hj})$
0	6.247	3.555	1.225	0.617	0.617	0.617	0.785(0.020)
1	21.168	29.916	14.830	2.425	1.567	1.567	1.246(0.115)
2	20.082	11.921	33.173	15.676	3.547	1.799	1.336(0.122)
3	45.030	10.306	35.066	79.321	47.562	10.306	3.200(0.280)
4	27.507	6.781	6.781	20.714	64.105	39.037	2.596(0.216)
5	4.306	0.855	0.855	0.855	2.626	8.775	0.923(0.050)
"climbers"							
row j	0	1	2	3	4	5	$100/(1+\Sigma_{hj})$
0	80.738	99.962	99.962	80.738	57.670	30.757	5.556(0.000)
1	11.723	21.428	15.988	5.124	2.171	0.737	0.858(0.032)
2	3.251	3.545	13.553	10.584	1.533	0.310	0.556(0.019)
3	4.932	1.689	10.876	29.779	20.605	1.689	1.300(0.022)
4	3.242	1.642	1.642	11.570	23.584	10.588	1.281(0.033)
5	1.069	0.169	0.169	0.169	1.711	3.114	0.410(0.011)
"low wage"							
row j	0	1	2	3	4	5	$100/(1+\Sigma_{hj})$
0	39.146	55.264	31.169	9.621	4.926	4.926	2.212(0.185)
1	2.692	6.263	3.642	0.359	0.180	0.180	0.425(0.006)
2	2.142	4.236	9.786	4.743	0.305	0.305	0.553(0.011)
3	2.682	2.695	11.231	22.255	9.561	1.357	1.163(0.071)
4	2.733	2.733	2.733	7.351	29.198	18.956	1.650(0.111)
5	4.765	4.188	3.569	3.569	44.497	46.598	1.872(0.261)
"flexible"							
row j	0	1	2	3	4	5	$100/(1+\Sigma_{hj})$
0	65.173	49.016	19.916	13.672	7.033	7.033	2.652(0.077)
1	50.251	65.833	25.581	13.571	6.981	6.981	2.642(0.090)
2	53.139	52.344	67.911	40.216	18.890	9.773	3.122(0.195)
3	74.856	53.290	53.290	125.494	74.856	28.338	5.330(0.124)
4	82.087	44.478	44.478	82.087	163.497	82.087	6.686(0.094)
5	34.880	26.682	10.826	10.826	42.957	80.337	3.273(0.354)

Table 4: Posterior expectation of the variance of the individual transition probabilities $100\xi^s_{i,jk}$ (in percent) in the various clusters as defined in (4); last column: posterior expectation and, in parenthesis, posterior standard deviation of the amount of unobserved heterogeneity in row j defined in (5) as $1/(1+\Sigma_{hj})$ and multiplied by a factor 100

Table 5 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 70.3 % (MCC) and 71.5 % (DMC) to their respective groups. For 1 out of 4 individuals assignment probability amounts to at least 98.9 % (MCC) and 98.1 % (DMC). Segmentation power is the highest for the "unemployed" and the lowest for the "flexible" cluster. Markov chain clustering has a slightly higher segmentation

power than Dirichlet multinomial clustering in these clusters, while the segmentation power is smaller for the other clusters.

	Markov chain clustering			Dirichlet multinomial clustering		
	1st Qu.	Median	3rd Qu.	1st Qu.	Median	3rd Qu.
"unemployed"	0.8902	0.9930	0.9994	0.8616	0.9844	0.9976
"climbers"	0.6829	0.8803	0.9732	0.7199	0.8852	0.9635
"low wage"	0.6487	0.8671	0.9796	0.6587	0.8682	0.9748
"flexible"	0.6574	0.8719	0.9775	0.6419	0.8436	0.9660
overall	0.7026	0.9113	0.9891	0.7154	0.9064	0.9811

Table 5: 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.

To obtain 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 which shows for both clustering methods the members with the 1st, 5th, 10th, 20th and 50th highest classification probability to belong to a particular cluster. This figure further emphasizes 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.

6 Concluding Remarks

In this paper we discussed 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 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 criteria indicated that Dirichlet multinomial clustering outperforms Markov chain clustering and that for

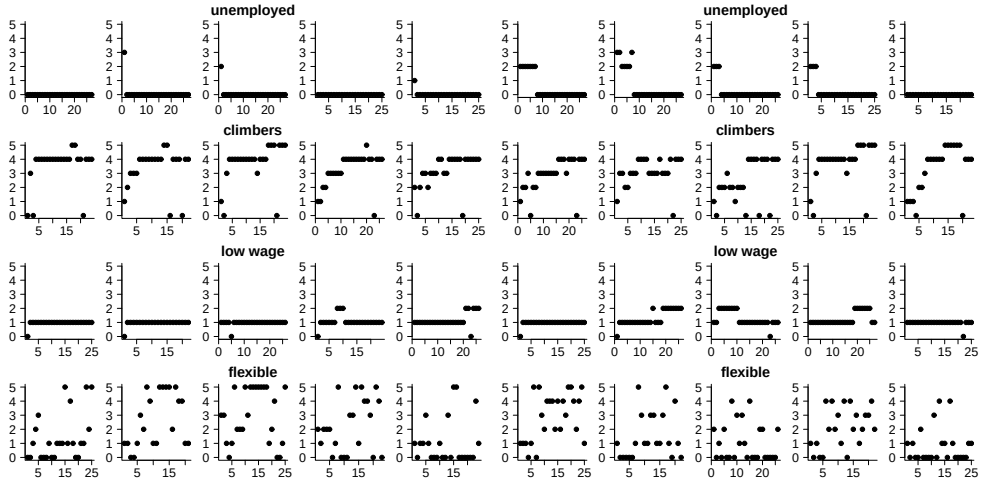

Figure 6: Typical group members within each cluster: wage careers of the individuals no. 1, 5, 10, 20 and 50 in the posterior classification probability ranking; left hand side: Markov chain clustering; right hand side: Dirichlet multinomial clustering.

this cohort study 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". The amount of unobserved heterogeneity within each cluster turned out to be small compared to the differences between the clusters, hence model-based clustering turned out to be robust to the choice of clustering kernel and the meaning of the clusters obtained by Markov chain clustering under a four-group solution was comparable to Dirichlet multinomial clustering. We found for both clustering methods that the segmentation power of the four-group solution is rather high: 3 out of 4 individuals are assigned with at least 70.3 % (MCC) and at least 71.5 % (DMC) probability to their respective cluster.

We conclude that both clustering kernels are a sensible tool for model-based clustering of discrete-valued panel data, in particular, if no background information is available. For our case study we were able to identify various sensible types of wage careers, although important covariates like education or profession were unobserved. Nevertheless, other clustering kernels might be sensible for clustering discrete-valued time series. One interesting extension of our approach is to use a kth order instead of a first-order Markov chain 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.

A very general clustering kernel is obtained through a dynamic multinomial logit model with random effects, see e.g. Rossi et al. (2005). Such a clustering kernel is able to capture rather general dependence patterns in the distribution of unobserved hetero-

geneity, while under Dirichlet multinomial clustering the dependence structure is rather restricted. 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 (5). However, using a dynamic multinomial logit model with random effects as clustering kernel complicates MCMC estimation considerably, because no explicit expression for the marginal distribution where the random effects are integrated out is available. Thus we leave this for future research.

Acknowledgments

We would like to thank the Editor and the referees for their detailed comments which helped us to improve the manuscript considerably. In addition, 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 first 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. 351

Akaike, H. (1974). "A new look at statistical model identification." *IEEE Transactions on Automatic Control*, 19: 716–723. 353

Banfield, J. D. and Raftery, A. E. (1993). "Model-based Gaussian and Non-Gaussian Clustering." *Biometrics*, 49: 803–821. 345, 354

Biernacki, C., Celeux, G., and Govaert, G. (2000). "Assessing a mixture model for clustering with the integrated completed likelihood." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22: 719–725. 353

Biernacki, C. and Govaert, G. (1997). "Using the classification likelihood to choose the number of clusters." *Computing Science and Statistics*, 29: 451–457. 353, 354

Cadez, I., Heckerman, D., Meek, C., Smyth, P., and White, S. (2003). "Model-Based Clustering and Visualization of Navigation Patterns on a Web Site." *Data Mining and Knowledge Discovery*, 7(4): 399–424. 346, 350

Fougère, D. and Kamionka, T. (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. 346, 347, 350

Fraley, C. and Raftery, A. E. (2002). "Model-based clustering, discriminant analysis, and density estimation." *Journal of the American Statistical Association*, 97: 611–631. 345

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer. 352, 353, 355, 361

Frühwirth-Schnatter, S. and Kaufmann, S. (2008). "Model-based clustering of multiple time series." *Journal of Business & Economic Statistics*, 26: 78–89. 345, 346, 349

Frühwirth-Schnatter, S. and 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. 354

Frydman, H. (2005). "Estimation in the mixture of Markov chains moving with different speeds." *Journal of the American Statistical Association*, 100: 1046–1053. 346, 347, 350

Hsiao, C. (2003). *Analysis of Panel Data*. New York: Cambridge University Press, 2 edition. 360

Jain, N. and Warnes, G. R. (2006). "Balloon Plot." *R News*, 6(2): 35–38. URL http://CRAN.R-project.org/doc/Rnews/ 357

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling." *Statistical Science*, 20: 50–67. 352

Juárez, M. A. and Steel, M. F. J. (2010). "Model-based Clustering of non-Gaussian Panel Data based on skew-t distributions." *Journal of Business & Economic Statistics*, 28: 52–66. 345

Kass, R. E. and Raftery, A. E. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90: 773–795. 353, 354

Keribin, C. (2000). "Consistent estimation of the order of mixture models." *Sankhya A*, 62: 49–66. 353

Kim, J.-Y. (1998). "Large sample properties of posterior densities, Bayesian information criterion and the likelihood principle in nonstationary time series models." *Econometrica*, 66: 359–380. 353

Liao, T. W. (2005). "Clustering of time series data – a survey." *Pattern Recognition*, 38: 1857–1874. 345

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics. New York: Wiley. 354

Raferzeder, T. and Winter-Ebmer, R. (2007). "Who is on the rise in Austria: Wage mobility and mobility risk." *Journal of Economic Inequality*, 5(1): 39–51. 354

Ramoni, M., Sebastiani, P., and Cohen, P. (2002). "Bayesian Clustering by Dynamics." *Machine Learning*, 47: 91–121. 346, 350

Ridgeway, G. and Altschuler, S. (1998). "Clustering Finite Discrete Markov Chains." In *Proceedings of the Section on Physical and Engineering Sciences*, 228–229. American Statistical Association. 349

Rossi, P. E., Allenby, G. M., and McCulloch, R. (2005). *Bayesian Statistics and Marketing*. Chichester: Wiley. 365

Saul, L. K. and Jordan, M. I. (1999). "Mixed memory Markov models: Decomposing complex stochastic processes as mixture of simpler ones." *Machine Learning*, 37: 75–87. 365

Schwarz, G. (1978). "Estimating the dimension of a model." *The Annals of Statistics*, 6: 461–464. 353

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. 355

Zweimüller, J., Winter-Ebmer, R., Lalive, R., Kuhn, A., Wuellrich, J.-P., Ruf, O., and Büchi, S. (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. 354