Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2007-30

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

Sylvia Frühwirth-Schnatter and Christoph Pamminger

October 2007

Abstract

Two approaches for model-based clustering of categorical time series based on time-homogeneous first-order Markov chains are discussed. 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 the individual transition matrices deviate from the group means and follow a Dirichlet distributions with unknown group-specific hyperparameters.

Estimation is carried out through Markov chain Monte. 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 labour market.

Keywords: Markov chain Monte Carlo, model-based clustering, panel data, income dynamics, transition matrices, labour market

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 unlabelled a priori. To this aim, clustering techniques are required to determine subsets of similar time series within the panel. While distance-based clustering methods cannot easily be extended to time series data, where an appropriate distance-measure is rather difficult to define, 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, see e.g. the recent review by Liao (2005). 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. For such time series it is particularly difficult to define distance measures and model-based clustering appears quite promising.

The crucial point in model-based clustering is to select appropriate clustering kernels in terms of a sampling density which captures salient features of the observed time series. Various such clustering kernels where suggested for panels with real-valued time series observations in Frühwirth-Schnatter and Kaufmann (2007) and Juárez and Steel (2006). Several papers applied finite mixtures of first-order time-homogeneous Markov chains to cluster discrete-valued time series. Cadez et al. (2000) clustered users according to their behaviour on a web site, while Ramoni et al. (2002) clustered sensor data from mobile robots using this method. 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 labour 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.

Although model-based clustering of discrete-valued time series based on finite mixtures of first-order time-homogeneous Markov chains proved to be useful in these papers, the approach is limited insofar as it implies that within each cluster all individuals follow exactly the same transition behaviour. To be more flexible in this respect we introduce in this paper a more general approach which captures

unobserved heterogeneity within each cluster by allowing individual transition matrices to deviate from the average group-specific transition matrix. This variation is described through a Dirichlet distribution with an unknown group-specific hyperparameter. As the resulting clustering kernel is closely related to the Dirichlet multinomial model, this approach will be referred to as Dirichlet multinomial clustering.

For estimation, we pursue a Bayesian approach extending earlier work by Ridgeway and Altschuler (1998) and Fougère and Kamionka (2003) rather than EM estimation as in Cadez et al. (2000) or Frydman (2005).

The remaining paper is organised as follows. Section 2 deals with Markov chain clustering, while Dirichlet multinomial clustering is discussed in Section 3. In Section 4 we give a short review of some well-known criteria for selecting the number of groups and investigate their behaviour for simulated data. 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 labour market.

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 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, see Frühwirth-Schnatter and Kaufmann (2007).

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 to which group the time series belongs:

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

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 group 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.

In this paper we pursue a Bayesian approach toward estimation. We assume prior independence between $\boldsymbol{\eta}$ and $(\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_H)$ and apply the Dirichlet distribution $\boldsymbol{\eta} \sim$

$\mathcal{D}(\alpha_0, \ldots, \alpha_0)$ which is commonly used in mixture modelling, see e.g. Frühwirth-Schnatter (2006) for more details. For a fixed number of groups MCMC estimation is easily implemented using data augmentation as in Algorithm 1.

Algorithm 1

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

$$\Pr(S_i = h|\mathbf{y}_i, \boldsymbol{\eta}, \boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_H) \propto p(\mathbf{y}_i|\boldsymbol{\vartheta}_h)\eta_h, \quad h = 1, \ldots, H. \tag{2}$$

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 $p(\boldsymbol{\vartheta}_h|\mathbf{S}, \mathbf{y}), \quad h = 1, \ldots, H$.*

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

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

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 of individual i. Note that we condition in (3) on the first observation y_{i0}.

Markov chain clustering is based on the assumption that within each group 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 directly 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 completely unrelated which leads to more flexibility in capturing differences in the transition behaviour between the groups.

2.3 The Bayesian Approach to Markov Chain Clustering

As the likelihood 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 and that each row $\boldsymbol{\xi}_{h,j\cdot}$, $j = 1, \ldots, K$ follows a Dirichlet distribution, $\boldsymbol{\xi}_{h,j\cdot} \sim \mathcal{D}\,(e_{0,j1}, \ldots, e_{0,jK})$, with prior parameters $\mathbf{e}_{0,j\cdot} = \{e_{0,j1}, \ldots, e_{0,jK}\}$. This prior is conjugate to the complete data likelihood and allows straightforward implemention 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:

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

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

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

$N^h_{jk} = \sum_{i:S_i=h} N_{i,jk}$ where $N_{i,jk}$ has been defined in (4) is the total number of transitions from j to k observed in group h and is determined from all individuals that fall into that particular group.

The Bayesian approach offers several advantages in the context of Markov chain clustering compared to EM estimation as in Cadez et al. (2000) 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 might 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 $\boldsymbol{\xi}_h$, $\boldsymbol{\xi}_{h,j\cdot}$:

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

and no estimator for $\boldsymbol{\xi}_{h,j\cdot}$ exists in the M-step. Second, the EM algorithm fails if for a single cell (j,k) no transitions from j to k are observed, i.e. $N_{jk}^h = 0$ for all $h = 1, \ldots, H$. In this case the M-step leads to an estimator of $\xi_{h,jk}$ that lies on the boundary of the parameter space, $\hat{\xi}_{h,jk} = 0$, which 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 **?** 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 likelihood $p(\mathbf{y}|\boldsymbol{\xi})$ with the Dirichlet prior $\mathcal{D}(e_{0,j1}, \ldots, e_{0,jK})$ within a Bayesian approach.

3 Finite Mixtures of Markov Chain Models with Unobserved Heterogeneity

In this section we suggest a generalization of Markov chain clustering which takes unobserved heterogeneity within each cluster into account. This is achieved by allowing the individual transition probabilities to deviate from the average group-specific transition behaviour. This variation is described through a Dirichlet distribution with an unknown group-specific hyperparameter.

3.1 The Dirichlet Multinomial Model

Dirichlet multinomial clustering is based on the assumption that each individual time series $\mathbf{y}_i$ is generated by a Markov chain model with individual transition matrix $\boldsymbol{\xi}_i^s$. The sampling distribution of $\mathbf{y}_i$ given $\boldsymbol{\xi}_i^s$ is obtained from (3):

$$p(\mathbf{y}_i|\boldsymbol{\xi}_i^s) = \prod_{j=1}^{K} \prod_{k=1}^{K} (\xi_{i,jk}^s)^{N_{i,jk}}. \tag{6}$$

It is possible to estimate each row $\boldsymbol{\xi}_{i,j\cdot}^s$ of $\boldsymbol{\xi}_i^s$ individually under the prior

$$\boldsymbol{\xi}_{i,j\cdot}^s \sim \mathcal{D}(e_{0,j1}, \ldots, e_{0,jK}), \qquad j = 1, \ldots, K, \tag{7}$$

using a Bayesian approach. This hierarchical 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.

This model, however, has certain drawbacks in a clustering context. First of all, the estimated transition matrices are highly dependent on the prior parameters $e_{0,j1}, \ldots, e_{0,jK}$, in particular for short individual time series. By increasing the sum $e_{0,j1} + \ldots + e_{0,jK}$ shrinkage of the row $\boldsymbol{\xi}_{i,j\cdot}^s$ toward the prior average increases.

This disadvantage can be avoided within a Bayesian framework by considering the parameters of the Dirichlet prior as unknown hyperparameters that are estimated from the data. But even so, the standard Dirichlet multinomial model is not really useful in a clustering context, as it is not easy to decide on the basis of the estimated individual transition matrices $\boldsymbol{\xi}_i^s$ which individuals have a similar transition behaviour and which are different in this respect.

3.2 Dirichlet Multinomial Clustering

To adjust the Dirichlet multinomial model to a clustering context, we assume that the time series form H (hidden) groups and that the Dirichlet prior distribution of the individual transition matrix $\boldsymbol{\xi}_i^s$ is different across groups. Within each group $h, h = 1, \ldots, H$ the K rows $\boldsymbol{\xi}_{i,j\cdot}^s$ of $\boldsymbol{\xi}_i^s$ are assumed to be independent, each following a Dirichlet distribution with group-specific prior parameter $\mathbf{e}_{h,j\cdot} = (e_{h,j1}, \ldots, e_{h,jK})$:

$$\boldsymbol{\xi}_{i,j\cdot}^s | (S_i = h) \sim \mathcal{D}\left(e_{h,j1}, \ldots, e_{h,jK}\right), \qquad j = 1, \ldots, K. \tag{8}$$

S_i is the latent group indicator introduced in Subsection 2.1. The parameters $\mathbf{e}_h = \{\mathbf{e}_{h,j\cdot}, j = 1, \ldots, K\}$ of all Dirichlet priors appearing in (8) are treated for each group as unknown hyperparameters that are estimated from the data. In the context of the model-based clustering approach discussed in Subsection 2.1 the resulting model corresponds to a finite mixture of Dirichlet multinomial models and for this reason is called Dirichlet multinomial clustering.

A distinctive advantage of modelling the distribution of heterogeneity in this way is that each group may be entirely described by the group-specific parameter $\mathbf{e}_h$ and 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 for any time series belonging to group h is available in closed form. By combining (6) and (8), we are able to derive the sampling density $p(\mathbf{y}_i | \mathbf{e}_h)$ in the following way:

$$\begin{aligned} p(\mathbf{y}_i | \mathbf{e}_h) &= \int p(\mathbf{y}_i | \boldsymbol{\xi}_i^s) p(\boldsymbol{\xi}_i^s | \mathbf{e}_h) d\boldsymbol{\xi}_i^s = \\ &= \frac{\prod_{j=1}^K \Gamma(\sum_{k=1}^K e_{h,jk})}{\prod_{j=1}^K \prod_{k=1}^K \Gamma(e_{h,jk})} \int \prod_{k=1}^K \prod_{j=1}^K (\xi_{i,jk}^s)^{N_{i,jk} + e_{h,jk} - 1} d\xi_{i,jk}^s = \\ &= \frac{\prod_{j=1}^K \Gamma(\sum_{k=1}^K e_{h,jk})}{\prod_{j=1}^K \prod_{k=1}^K \Gamma(e_{h,jk})} \frac{\prod_{j=1}^K \prod_{k=1}^K \Gamma(N_{i,jk} + e_{h,jk})}{\prod_{j=1}^K \Gamma(\sum_{k=1}^K (N_{i,jk} + e_{h,jk}))}. \end{aligned} \tag{9}$$

It is evident from (9) that this clustering kernel no longer is a first-order Markov process but allows for higher order dependence.

It is illuminating to study the group-specific transition behaviour implied by the parameter $\mathbf{e}_h$ in more detail. Each group may be characterised 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:

$$\xi_{h,jk} = \mathrm{E}(\xi_{i,jk}^s | S_i = h, \mathbf{e}_h) = \frac{e_{h,jk}}{\sum_{k=1}^K e_{h,jk}}. \tag{10}$$

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 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 $\xi_{i,jk}^s$:

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

It can easily be shown that

$$\frac{\mathrm{Var}(\xi_{i,jk}^s | S_i = h, \mathbf{e}_h)}{\mathrm{E}(\xi_{i,jk}^s | S_i = h, \mathbf{e}_h)\ (1 - \mathrm{E}(\xi_{i,jk}^s | S_i = h, \mathbf{e}_h))} = \frac{1}{1 + \sum_{k=1}^K e_{h,jk}}. \tag{11}$$

As the right hand side of (11) is the same for all elements of row $\boldsymbol{\xi}_{i,j\cdot}^s$, a single parameter depending only on the row sum $\sum_{k=1}^K e_{h,jk}$ controls variability for all elements in the j-th 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 $\sum_{k=1}^K e_{h,jk}$, the more variable are the individual transition probabilities in row j 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 $\sum_{k=1}^K e_{h,jk}$ 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 $\boldsymbol{\xi}_i^s$.

3.3 Bayesian Estimation

3.3.1 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 2.3 no conjugate prior allowing straightforward MCMC estimation is available, but the structure of the complete data likelihood to be discussed in Subsection 3.3.2 still suggests to assume that all rows $\mathbf{e}_{h,j\cdot}$ are independent within each group and between groups.

To avoid all problems with empty transitions that have been discussed in Subsection 2.3 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

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

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:

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

The negative multinomial distribution arises as a mixture distribution when the K elements of $\mathbf{e}_{h,j\cdot}$ are assumed to be independent, with $e_{h,jk} - 1 \sim \mathcal{P}\,(\gamma \lambda_{jk})$, where $\gamma \sim \mathcal{G}\,(\alpha, \beta)$. The resulting mixture distribution is equal to $\text{NegMulNom}\,(p_{jk}, \ldots, p_{jk}, \beta)$ with $p_{jk} = \lambda_{jk} / (\alpha + \sum_{k=1}^{K} \lambda_{jk})$.

This suggest to choose in our application $\lambda_{jk} = N_0 \hat{\xi}_{jk}$, 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.

3.3.2 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 (9).

The third step of Algorithm 1 is the only step which is essentially different from the corresponding steps 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:

$$\begin{aligned} p(\mathbf{e}_1, \ldots, \mathbf{e}_H | \mathbf{S}, \mathbf{y}) &\propto \prod_{i=1}^{N} p(\mathbf{y}_i | \mathbf{e}_{S_i}) \prod_{h=1}^{H} p(\mathbf{e}_h) = \prod_{h=1}^{H} p(\mathbf{e}_h) \prod_{i:S_i=h} p(\mathbf{y}_i | \mathbf{e}_{h,j\cdot}) \\ &\propto \prod_{h=1}^{H} \prod_{j=1}^{K} p(\mathbf{e}_{h,j\cdot}) \frac{\Gamma(\sum_{k=1}^{K} e_{h,jk})^{N_h}}{\left(\prod_{k=1}^{K} \Gamma(e_{h,jk}) \right)^{N_h}} \left(\prod_{i:S_i=h} \frac{\prod_{k=1}^{K} \Gamma(N_{i,jk} + e_{h,jk})}{\Gamma(\sum_{k=1}^{K} (N_{i,jk} + e_{h,jk}))} \right), \end{aligned} \tag{12}$$

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

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

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 the 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 iteration. This is possible only because of the special structure of our model which yields a closed form for the density $p(\mathbf{y}_i|\mathbf{e}_{S_i})$.

If needed, it is possible to obtain draws for $\boldsymbol{\xi}_i^s$ for each $i = 1, \ldots, N$ by drawing the jth row 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).

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 groups. Various criteria have been developed in the context of model-based clustering based on multivariate normal distributions some of which are shortly reviewed in Subsection 4.1. To evaluate the performance of these criteria in the somewhat different context of clustering categorical time series based on mixtures of Markov chains a simulation study is carried out in Subsection 4.2.

4.1 A Short Review of Some Criteria for Selecting the Number of Clusters

Most clustering criteria are based on measuring model fit through some kind of likelihood function which is then penalised in an appropriate way to avoid overfitting. For any of these criteria the optimal number H of groups is defined as that value of H which minimises the criterion. Subsequently, $\hat{\boldsymbol{\theta}}^H$ indicates an estimator for the parameter $\boldsymbol{\theta}^H = (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_H, \eta_1, \ldots, \eta_H)$ in a model with H groups.

AIC (Akaike, 1974) and BIC (Schwarz, 1978) penalise the mixture log-likelihood $L(H, \boldsymbol{\theta}^H)$ defined by

$$L(H, \boldsymbol{\theta}^H) = \log p(\mathbf{y}|\boldsymbol{\theta}^H) = \sum_{i=1}^{N} \log \left(\sum_{h=1}^{H} \eta_h \, p(\mathbf{y}_i|\boldsymbol{\vartheta}_h) \right) \tag{13}$$

by model complexity defined as the total number d_H of independent parameters to be estimated in a mixture model with H components:

$$AIC(H) = -2\, L(H, \hat{\boldsymbol{\theta}}^H) + 2\, d_H, \tag{14}$$

$$BIC(H) = -2\, L(H, \hat{\boldsymbol{\theta}}^H) + d_H \log N. \tag{15}$$

BIC is consistent under correct specification of the family of the component densities (Keribin, 2000), while it tends to selected too many components under misspecification. AIC tends to select too many components even for a correctly specified mixture.

Like BIC, approximate weight of evidence (AWE) is derived in Banfield and Raftery (1993) as an approximation to minus twice the log Bayes factor and penalises the complete data log-likelihood defined by

$$L_C(H, \boldsymbol{\theta}^H) = \log\, p(\mathbf{y}, \mathbf{S}|\boldsymbol{\theta}^H) = \sum_{h=1}^{H} \sum_{i=1}^{N} \ln \left(\eta_{S_i} p(\mathbf{y}_i|\boldsymbol{\vartheta}_{S_i}) \right) \tag{16}$$

with model complexity:

$$AWE(H) = -2\, L_C(H, \hat{\boldsymbol{\theta}}^H) + 2\, d_H (\frac{3}{2} + \log N). \tag{17}$$

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 $EN(H, \boldsymbol{\theta}^H)$ given by

$$EN(H, \boldsymbol{\theta}^H) = -\sum_{h=1}^{H} \sum_{i=1}^{N} t_{ih} \log t_{ih} \geq 0, \tag{18}$$

where $t_{ih} = \Pr(S_i = h|\mathbf{y}_i, \boldsymbol{\theta}^H)$ is the posterior classification probability defined in (2). The entropy is a measure of how well the data are classified given the 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) penalises the mixture log-likelihood $L(H, \boldsymbol{\theta}^H)$ by the entropy $EN(H, \hat{\boldsymbol{\theta}}^H)$ rather than by model complexity as in AIC or BIC:

$$CLC(H) = -2\, L(H, \hat{\boldsymbol{\theta}}^H) + 2\, EN(H, \hat{\boldsymbol{\theta}}^H). \tag{19}$$

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 is equal to

$$ICL\text{-}BIC(H) = BIC(H) + 2\, EN(H) \tag{20}$$

and penalises not only model complexity, but also the failure of the mixture model to provide a classification of the data in well-separated clusters. 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.

4.2 Application to Dirichlet Multinomial Clustering

To investigate the performance of the various model selection criteria discussed in the previous subsection in the context of Dirichlet multinomial clustering we consider synthetic data simulated from a Dirichlet multinomial mixture with the true number of groups being equal to three, four or five, respectively. N, T_i and y_{i0}, $i = 1, \ldots, N$ are chosen as in the case study for the Austrian labour market to be studied in Section 5 and the estimators obtained there for $\boldsymbol{\eta}, \mathbf{e}_1, \ldots, \mathbf{e}_H$ and $\mathbf{S}$ are used to simulate the data in this subsection.

For each of the three panels both Markov chain clustering as well as Dirichlet multinomial clustering is carried out for an increasing number of groups ranging from $H = 1, \ldots, 6$. While Dirichlet multinomial clustering corresponds to fitting the correct mixture, the component densities are misspecified for Markov chain clustering. For each panel $AIC(H)$, $BIC(H)$, $AWE(H)$, $CLC(H)$ and $ICL\text{-}BIC(H)$ are computed as in Subsection 4.1 for both clustering methods and are plotted as a function of H in Figure 1. To compute these criteria, $\hat{\boldsymbol{\theta}}^H$ is selected as that posterior draws which maximises the nonnormalised mixture posterior density $p^\star(\boldsymbol{\theta}^H|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta}^H)p(\boldsymbol{\theta}^H)$, because the ML estimator is not available within the framework of MCMC estimation.

From Figure 1 it can be seen that most criteria have a critical point at $H = H_{true}$. BIC and AWE perform best in the sense that they exhibit an obvious minimum and always detect the real number of groups. This holds not only for clustering based on the true component density but also for a model where the component densities are assumed to be a homogeneous Markov chain and therefore are misspecified. Somewhat surprisingly, $CLC(H)$ and $ICL\text{-}BIC(H)$ work fine for Markov chain clustering which is a misspecified model but lead to underfitting for several panels when fitting the correct component density.

5 Application to Austrian Wage Mobility Data

In our application we consider wage mobility in the Austrian labour market. Wage mobility describes chances but also risks of an individual to move between wage categories over time. 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 a sensible assumption 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 behaviour.

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 (Raferzeder and Winter-Ebmer, 2007) and consists of time series for $N = 9\,809$ men entering the labour 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 labelled with 0 up to 5. Category zero corresponds to zero-income or non-employment which is not equivalent to be out of labour 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 in our sample. 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 Model-based Clustering

To give a more detailed picture of this panel several individual time series showing wage mobility for a few employees are presented in Figure 2. 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 groups of individuals with similar wage mobility behaviour. To this aim we apply both Markov chain clustering as well as Dirichlet multinomial clustering for 2 up to 10 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 $\alpha = \beta = 1$, $N_0 = 10$ 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.

Initial values for Markov chain clustering are $\eta_h^{(0)} = 1/H$ and $\boldsymbol{\xi}_h^{(0)} = \hat{\boldsymbol{\xi}}$, where $\hat{\xi}_{jj} = 0.7$ and $\hat{\xi}_{jk} = 0.06$, if $j \neq k$. The estimators obtained by Markov chain clustering are used to define initial values for Dirichlet multinomial clustering with

the same number of groups: $\boldsymbol{\eta}^{(0)} = \hat{\boldsymbol{\eta}}$ and $\mathbf{e}_h^{(0)} = N_0 \hat{\boldsymbol{\xi}}_h$. 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 3.3.2, leading to an average acceptance rate of 0.24 percent.

For each number H of groups we conducted $M = 10\,000$ MCMC iterations for Dirichlet multinomial clustering and $M = 20\,000$ MCMC iterations for Markov chain clustering. For Markov chain clustering we discarded the first 2 000 draws as burn-in. For Dirichlet multinomial clustering we discarded the first 5 000 draws and kept only each 5th draw to reduce the autocorrelation in the MCMC draws.

5.2.1 Markov Chain Versus Dirichlet Multinomial Clustering

For each $H = 2, \ldots, 10$ the group-specific parameters $\boldsymbol{\xi}_h$ (Markov chain clustering) and $\mathbf{e}_h$ (Dirichlet multinomial clustering) are estimated for $h = 1, \ldots, H$ as the posterior means of the stationary MCMC draws. This is a valid estimator as visual inspection of the MCMC draws revealed no label switching, see e.g. Frühwirth-Schnatter (2006, Section 3.5) for an exhaustive review of the label switching problem.

For Dirichlet multinomial clustering, the average group-specific matrix $\boldsymbol{\xi}_h$ is estimated from the posterior draws of $\mathbf{e}_h$ as in (10). For up to $H = 5$ groups the resulting transition matrices are rather similar to the group-specific transition matrices $\boldsymbol{\xi}_h$ obtained under Markov chain clustering for the same number of clusters. Only for a large number of groups many small groups appear and the results are more different.

Table 1 shows for $H = 1, \ldots, 4$ the group-specific unobserved heterogeneity estimated according to expression (11). Unobserved heterogeneity varies considerably between the groups and in some groups also between the rows. In absolute terms, the amount of unobserved heterogeneity is pretty moderate explaining why the average group-specific transition matrices obtained for Dirichlet multinomial clustering are rather similar to the matrices obtained by Markov chain clustering.

5.2.2 Selecting the Number of Groups

The model selection criteria described in Subsection 4.1 are applied to select the number of groups both under Dirichlet multinomial as well as under Markov chain clustering, see Figure 3.

For Dirichlet multinomial clustering AWE and CLC take a minimum at $H = 4$, while $ICL\text{-}BIC$ leads to an underfitting model as for the simulated data considered in Subsection 4.2. For Markov chain clustering AWE, CLC and $ICL\text{-}BIC$ suggest a mixture with 5 clusters.

Both AIC and BIC lead to strongly overfitting models. For Markov chain clustering BIC suggests more than 10 groups while this number is equal to 7 for Dirichlet multinomial clustering. Whereas BIC has been a reliable criterion for the simulated data considered in Subsection 4.2, it appears to be sensitive to misspecification of the clustering kernel which is likely to be present for this real-world data set even under Dirichlet multinomial clustering.

5.2.3 Analysing Wage Mobility

Based on the criteria discussed in Subsection 5.2.2 we proceed with discussing the four-group solution under Dirichlet multinomial clustering which also allows very sensible interpretations from an economic point of view.

The most interesting features are the estimated average group-specific transition matrices which are visualised in Figure 4 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.

A remarkable difference in the transition behaviour of individuals belonging to different groups is evident from these figures. 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 group 1 and group 3 than for the other groups.

The probability to remain in the no-income category is located in the top left cell and is much higher in group 3 then in the other groups. 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 group 1 and 3 than for the other groups. In group 4 chances are high to move into any wage category while in group 2 only the chance to move in wage category one is comparatively high.

The main diagonals of these matrices refer to the probabilities to remain in a certain wage category. Persistence is pretty high except for group 1. In group 1 all rows are pretty close to a uniform distribution meaning that the members of this group move quickly between the wage categories. The upper secondary diagonal represents the chance to move forward into the next higher wage category, which is much higher in group 4 than in the other groups.

These obvious differences in the one-step ahead transition matrices between the groups have a strong impact on the wage mobility of the group members. The first column of Tables 2 shows the initial wage distribution $\boldsymbol{\pi}_{h,0}$ for each group which is estimated from the observed category y_{i0} for all individuals i being classified to group h. The subsequent columns show the estimated wage distributions $\boldsymbol{\pi}_{h,t} = \boldsymbol{\pi}_{h,0}\boldsymbol{\xi}_h^t$ after a period of t years. There are little differences between the groups 1, 2, and 4 in the beginning, but in the long run considerable differences in the wage distribution become evident due to the observed differences in wage mobility. Members of group 1 have a much higher risk to end up in the no-income category than members of group 2 and 4. In the long-run, however, members of group 2 are disadvantaged and end up in lower wage categories while members in group 4 move into the highest wage categories. Finally, individuals in group 3 have a much higher probability to start in the no-income category and about 60 % of the members of this group have no income in the long-run.

The last column where $t = \infty$ is in 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 group 1 and 3 the equilibrium distribution is reached after only a few years whereas in group 2 and 4 this distribution is reached after about two decades.

Posterior Classifications

Next we study how individuals are assigned to the four wage mobility groups using the estimated posterior classification probabilities $t_{ih} = \Pr(S_i = h|\mathbf{y}_i, \boldsymbol{\theta}^4)$. Any employee is allocated to that group which exhibits the maximum posterior probability. The posterior classification probabilities shown in Table 3 for the first 6 individuals indicate that some individuals are allocated with high probability to a particular group (e.g. employee no. 5 to group 3) whereas others are not (e.g. employee no. 4). To assess the uncertainty of the classifications we estimate the entropy of the posterior classification matrix as in (18), giving $EN(4) = 4\,005.21$. While this entropy is far from a perfect classification with zero entropy, it is much smaller than $N \log(4) = 13\,598.16$ which is the entropy of a classification rule, where all individuals are assigned randomly according to the relative group sizes.

To assess differences in the uncertainty of the classifications between the groups, the contribution $\sum_i t_{ih} \ln t_{ih}$ of group h to the total entropy $EN(4)$, divided by the number of individuals assigned to that group for the sake of comparability, is shown in Table 4. The smaller that value the better the individuals are allocated to the corresponding group. Individuals in group 4 have on average the lowest misclassification risk, while it is highest for individuals in group 1.

Finally, to get an even better understanding of the various wage mobility groups, five members that have a very high classification probability to belong to a particular group were selected and their individual time series are plotted in Figure 5 for all four groups. This figure further emphasises the interpretation of the wage groups obtained above. Group 1 obviously represents the more flexible and fluctuating employees. Typical members of group 2 stay mainly in the lowest wage category. Group 3 contains the employees who fall into the no-income category more often and remain there much longer than members of the other groups. Finally, group 4 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 labour market has never been documented before.

6 Concluding Remarks

In this paper we presented approaches for model-based clustering of categorical time series based on time-homogeneous first-order Markov chains with unknown transition matrices. While in the Markov chain clustering approach the individual transition probabilities are fixed to a group-specific transition matrix, we suggested a new approach called Dirichlet multinomial clustering where it is assumed that within each group unobserved heterogeneity is still existent. We allow the individual transition matrices to deviate from the group means and described this variation for each row through a Dirichlet distribution with unknown hyperparameters.

An application of this approach to a panel of Austrian wage mobility data lead to a segmentation of the employees into four groups. The group-specific transition behaviour described through the transition matrices turned out to be very different between the groups and leads to meaningful interpretations from an economic point of view.

For other panels of discrete-valued time series other clustering kernels might be

sensible. One important alternative is to model each row of the transition matrix with a multinomial logit model with random intercept and to assume that the dynamic regression parameters and the variance of the random intercept are different between the groups. MCMC estimation, however, is more involved because no explicit form for the marginal model is available. On the other hand, this clustering kernel allows the inclusion of additional covariates for each individual time series and allows to capture higher order dependence by including not only the immediate past, but also a longer history of the time series as predictor.

Both for Dirichlet multinomial clustering as well as for clustering based on a multinomial logit model with random intercept a single parameter controls both the amount of heterogeneity as well as correlation among the transition probabilities within each row. For Dirichlet multinomial clustering, for instance, exactly the same expression which controls unobserved heterogeneity in (11) also determines dependence between two arbitrarily chosen individual transition probabilities in the jth row $\boldsymbol{\xi}^s_{i,j\cdot}$, just with the opposite sign:

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

To obtain more flexibility in the distribution of heterogeneity, a hierarchical multinomial logit model, following e.g. Rossi et al. (2005), may be considered where all dynamic regression parameters are random effects. However, in its most general form this clustering kernel involves the estimation of a high-dimensional covariance matrix of the random effects distribution for each group and for this reason might be intractable for the purposes of clustering short individual time series, unless sensible constraints on the covariance matrix of the random effects are introduced.

Acknowledgements

We thank Rudolf Winter-Ebmer for making available the data set and 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 grant P 17 959 ("Gibbs sampling for discrete data")

References

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

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 (2000). Visualization of navigation patterns on a web site using model-based clustering. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 280–284.

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 S. Kaufmann (2007). Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*. in print.

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

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

Juárez, M. A. and M. F. J. Steel (2006). Model-based clustering of non-gaussian panel data. Technical Report CRiSM Working Paper 06-14, University of Warwick.

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

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

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.

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 Sudies.

Table 1: Amount of unobserved heterogeneity under Dirichlet multinomial clustering for $1, \ldots, 4$ groups according to equation (11) (multiplied by factor 10^2).

$\frac{10^2}{1+\sum_{k=1}^{K} e_{h,jk}}$	$H=1$	$H=2$		$H=3$			$H=4$			
row j	$h=1$	$h=1$	$h=2$	$h=1$	$h=2$	$h=3$	$h=1$	$h=2$	$h=3$	$h=4$
0	3.486	6.250	1.451	3.325	0.842	6.790	2.661	2.916	0.789	5.556
1	2.222	1.218	2.005	2.862	1.440	0.767	2.579	0.535	1.148	1.184
2	2.856	1.850	2.820	2.601	1.484	0.769	2.991	0.667	1.120	0.888
3	2.547	1.685	3.930	3.747	3.496	1.326	3.661	1.922	3.319	1.587
4	2.214	1.431	4.551	4.460	2.659	1.043	6.775	1.802	2.401	1.292
5	0.783	0.540	2.576	3.223	0.884	0.499	3.270	2.307	1.094	0.476

Table 2: Wage distributions $\boldsymbol{\pi}_{h,t}$ over the wage categories 0 to 5 for different years t in the different groups

Group 1	$t=0$	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=10$	$t=20$	$t=\infty$
0	0.0993	0.2682	0.3129	0.3225	0.3231	0.3218	0.3184	0.3180	0.3180
1	0.5989	0.3952	0.3260	0.3003	0.2896	0.2847	0.2791	0.2786	0.2786
2	0.1891	0.1435	0.1274	0.1212	0.1186	0.1174	0.1160	0.1159	0.1159
3	0.0711	0.0992	0.1062	0.1081	0.1087	0.1090	0.1093	0.1094	0.1094
4	0.0322	0.0569	0.0726	0.0815	0.0865	0.0894	0.0932	0.0936	0.0936
5	0.0094	0.0370	0.0549	0.0663	0.0734	0.0778	0.0840	0.0845	0.0845
Group 2	$t=0$	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=10$	$t=20$	$t=\infty$
0	0.0733	0.0711	0.0683	0.0657	0.0635	0.0616	0.0555	0.0510	0.0492
1	0.7345	0.6561	0.5953	0.5471	0.5084	0.4771	0.3855	0.3268	0.3054
2	0.1278	0.1829	0.2193	0.2430	0.2580	0.2671	0.2740	0.2568	0.2447
3	0.0410	0.0499	0.0614	0.0733	0.0843	0.0942	0.1235	0.1357	0.1363
4	0.0199	0.0275	0.0373	0.0478	0.0586	0.0692	0.1150	0.1653	0.1906
5	0.0035	0.0125	0.0184	0.0231	0.0271	0.0308	0.0465	0.0644	0.0738
Group 3	$t=0$	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=10$	$t=20$	$t=\infty$
0	0.3339	0.4374	0.5061	0.5518	0.5819	0.6016	0.6312	0.6264	0.6215
1	0.4427	0.2974	0.2171	0.1716	0.1451	0.1294	0.1061	0.1019	0.1011
2	0.1193	0.1401	0.1333	0.1200	0.1072	0.0968	0.0743	0.0694	0.0690
3	0.0631	0.0602	0.0600	0.0583	0.0554	0.0523	0.0424	0.0397	0.0396
4	0.0281	0.0376	0.0425	0.0451	0.0461	0.0462	0.0428	0.0412	0.0415
5	0.0129	0.0274	0.0410	0.0533	0.0642	0.0737	0.1033	0.1214	0.1274
Group 4	$t=0$	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=10$	$t=20$	$t=\infty$
0	0.0538	0.1151	0.0948	0.0787	0.0685	0.0620	0.0499	0.0457	0.0447
1	0.5427	0.3006	0.2004	0.1473	0.1166	0.0978	0.0649	0.0542	0.0515
2	0.2403	0.2971	0.2928	0.2673	0.2393	0.2147	0.1489	0.1187	0.1104
3	0.1076	0.1781	0.2368	0.2681	0.2800	0.2805	0.2420	0.1993	0.1839
4	0.0405	0.0780	0.1236	0.1654	0.2001	0.2269	0.2748	0.2599	0.2470
5	0.0151	0.0311	0.0516	0.0732	0.0955	0.1181	0.2196	0.3221	0.3625

Table 3: Posterior classification probabilities: The probabilities of individuals to be allocated to each group.

	Group 1	Group 2	Group 3	Group 4
1	0.0001	0.0001	0.2765	0.7233
2	0.0098	0.0000	0.9901	0.0001
3	0.1205	0.0001	0.1170	0.7623
4	0.0147	0.4768	0.0073	0.5011
5	0.0006	0.0000	0.9994	0.0000
6	0.2792	0.0001	0.7205	0.0002
		⋮		

Table 4: Contribution of each group to the total entropy $EN(4)$ (absolute and relative to group size)

Group h	1	2	3	4
$\sum_i t_{ih} \ln t_{ih}$	930.72	972.28	943.73	1 158.47
$\sum_i t_{ih} \ln t_{ih}/N_h$	0.62	0.57	0.43	0.26

Figure 1: Model selection criteria for various numbers H of clusters for Markov chain clustering (MCC) and Dirichlet multinomial clustering (DMC) for simulated data where the true number of clusters is equal to 3 (top), 4 (medium) and 5 (bottom)

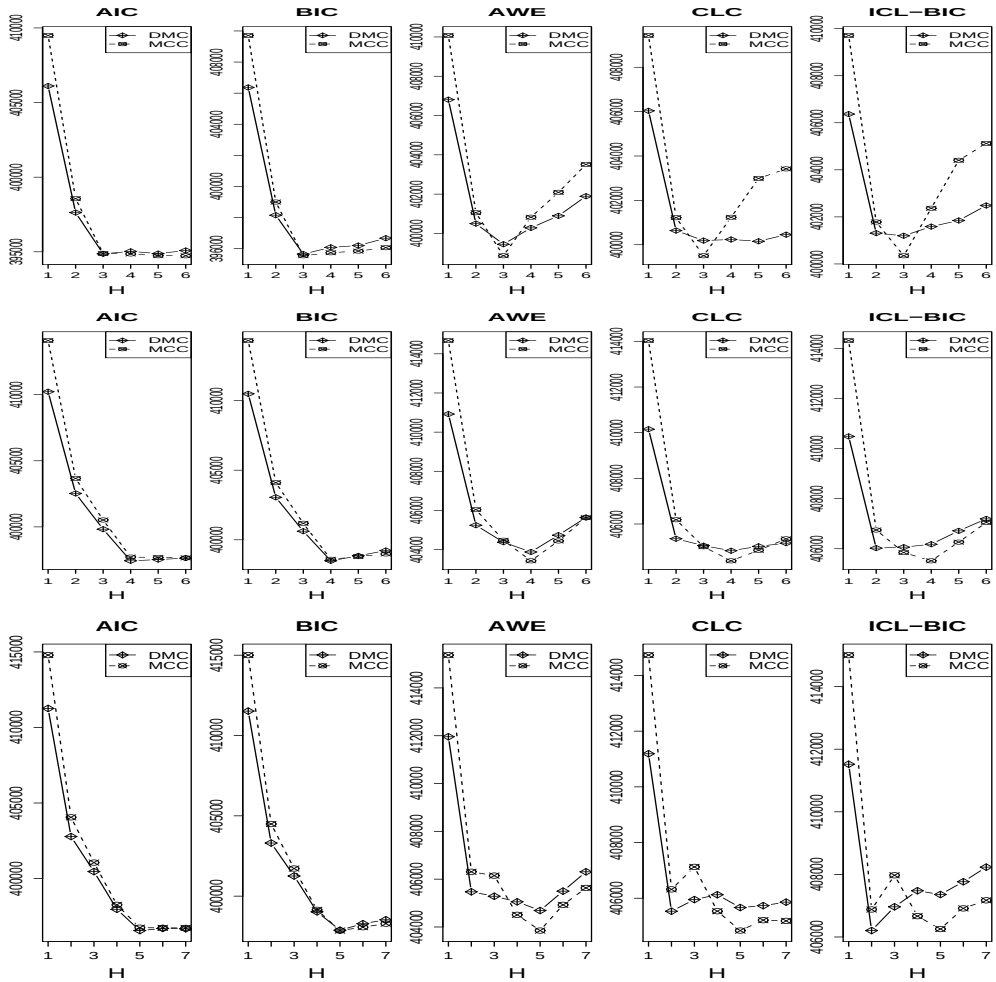

Figure 2: Individual wage mobility time series of nine selected employees (time t on x-axis and income class k on y-axis).

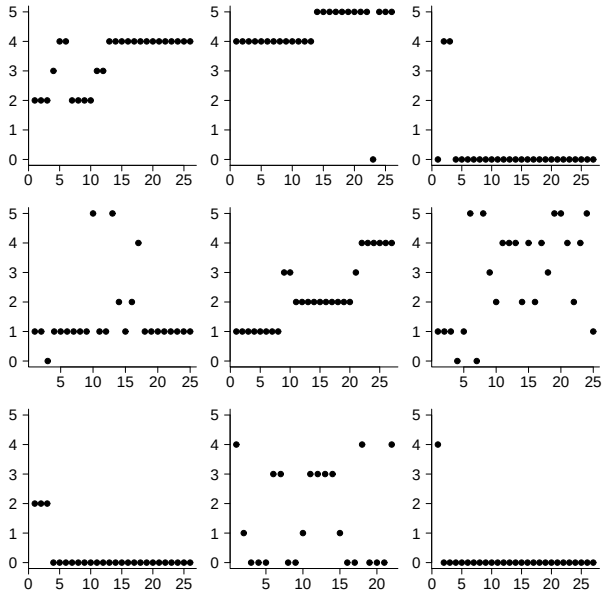

Figure 3: Model selection criteria for various numbers H of clusters for Markov chain clustering (MCC) and Dirichlet multinomial clustering (DMC) for the Austrian labour market data

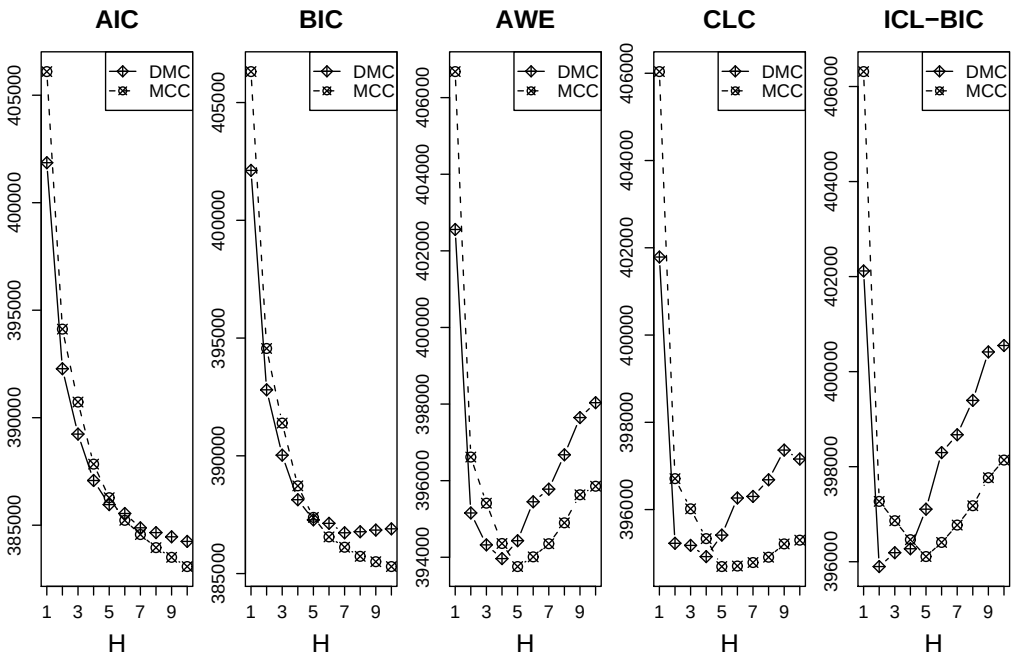

Figure 4: Visualisations of estimated posterior transition probabilities $\hat{\boldsymbol{\xi}}_1, \hat{\boldsymbol{\xi}}_2, \hat{\boldsymbol{\xi}}_3, \hat{\boldsymbol{\xi}}_4$, where the circular areas are proportional to the size of the corresponding entry in the transition matrix. Estimated group sizes $\hat{\boldsymbol{\eta}}$ (mixing proportions) are indicated in brackets.

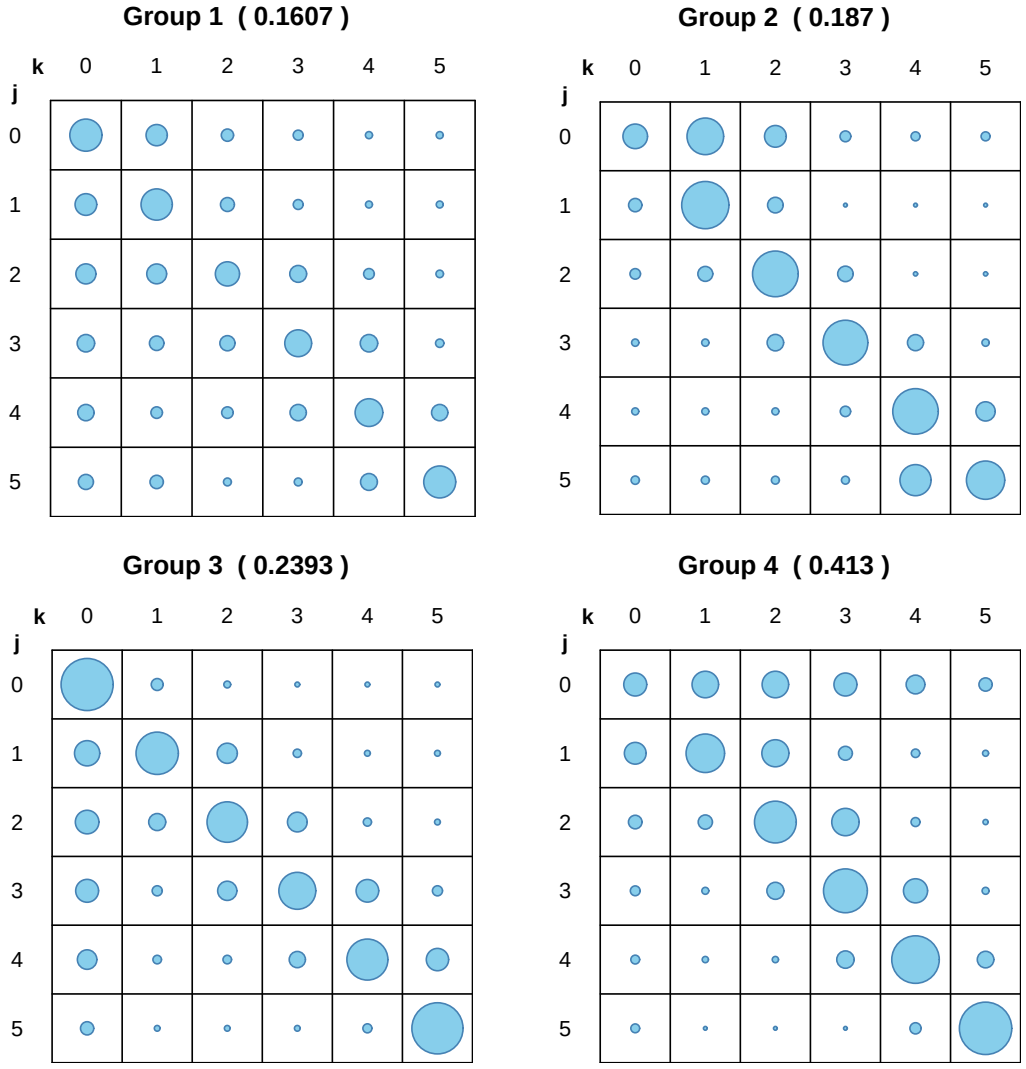

Figure 5: Typical group members: Selected time series of group members, who exhibit a posterior classification probability of virtually one.