Model-based Clustering of Categorical Time Series with Multinomial Logit Classification

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

*Institute for Applied Statistics, Johannes Kepler University Linz, Altenberger Strasse 69, 4040 Linz, Austria
†Department of Economics, Johannes Kepler University Linz, Altenberger Strasse 69, 4040 Linz, Austria
**Chair for Applied Political Economy, University of Mannheim, L7, 3-5, 68131 Mannheim, Germany

Abstract. A common problem in many areas of applied statistics is to identify groups of similar time series in a panel of time series. However, distance-based clustering methods cannot easily be extended to time series data, where an appropriate distance-measure is rather difficult to define, particularly for discrete-valued time series.

Markov chain clustering, proposed by Pamminger and Frühwirth-Schnatter [6], is an approach for clustering discrete-valued time series obtained by observing a categorical variable with several states. This model-based clustering method is based on finite mixtures of first-order time-homogeneous Markov chain models.

In order to further explain group membership we present an extension to the approach of Pamminger and Frühwirth-Schnatter [6] by formulating a probabilistic model for the latent group indicators within the Bayesian classification rule by using a multinomial logit model.

The parameters are estimated for a fixed number of clusters within a Bayesian framework using an Markov chain Monte Carlo (MCMC) sampling scheme representing a (full) Gibbs-type sampler which involves only draws from standard distributions.

Finally, an application to a panel of Austrian wage mobility data is presented which leads to an interesting segmentation of the Austrian labour market.

Keywords: Bayesian Statistics; Transition Matrices; Panel Data; Multinomial Logit Model; Random Utility Model; Markov Chain Monte Carlo; Auxiliary Mixture Sampler; Classification;

MODEL-BASED CLUSTERING

A common problem in many areas of applied statistics is to identify groups of similar time series in a panel of time series. However, distance-based clustering methods cannot easily be extended to time series data, where an appropriate distance-measure is rather difficult to define, particularly for discrete-valued time series, see e.g. the review by Liao [5].

Recently, Frühwirth-Schnatter and Kaufmann [4] demonstrated that model-based clustering based on finite mixture models (Banfield and Raftery [1], Fraley and Raftery [2]) extends to time series data in 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 (sampling) density $p(\mathbf{y}_i|\vartheta_h)$ of the finite mixture model, also called clustering kernel, with unknown model parameter ϑ_h, $h = 1,...,H$, plays a crucial role in the corresponding clustering procedure and has to capture salient features of the observed time series $\mathbf{y}_i$.

Pamminger and Frühwirth-Schnatter [6] focus on clustering discrete-valued time series obtained by observing a categorical variable with several states. For discrete-valued time series it is particularly difficult to define distance measures and model-based clustering appears to be a promising alternative. Pamminger and Frühwirth-Schnatter [6] 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.

The estimation of the classification probabilities to determine the (unknown) group membership is a crucial part of any model-based clustering approach. In order to further explain group membership we present an extension to the Markov chain clustering approach of Pamminger and Frühwirth-Schnatter [6] by formulating a probabilistic (logit-type prior) model for the latent group indicators, like in Frühwirth-Schnatter and Kaufmann [4], within the Bayesian

CP1281, *ICNAAM, Numerical Analysis and Applied Mathematics, International Conference 2010*
edited by T. E. Simos, G. Psihoyios, and Ch. Tsitouras

classification rule by using a multinomial logit model. This allows for the inclusion of individual characteristics in terms of unit-specific time-constant covariates instead of complete prior ignorance.

The model parameters are estimated within a Bayesian framework for a fixed number of clusters using Markov chain Monte Carlo (MCMC) and data augmentation methods. Particularly the parameters of the multinomial logit model are estimated using the very efficient auxiliary mixture sampler of Frühwirth-Schnatter and Frühwirth [3] which is based on the (differenced) random utility model representation of the multinomial logit model. A major advantage of this MCMC sampling scheme is that it results in a (full) Gibbs-type sampler which involves only draws from standard distributions.

Finally, an application of this model-based clustering approach, using a multinomial logit classification step to a large panel of Austrian wage mobility data, as in Pamminger and Frühwirth-Schnatter [6], is presented which leads to an interesting segmentation of the Austrian labour market.

MARKOV CHAIN CLUSTERING

Let $\{y_{it}\}$, $t = 0, \dots, T_i$ be a panel of categorical time series observed for N units $i = 1, \dots, N$ on T_i occasions with y_{it} taking K potential states labeled by $\{1, \dots, K\}$. Let $\mathbf{y}_i = \{y_{i0}, \dots, y_{i,T_i}\}$ denote an individual time series. Model-based clustering assumes that H hidden clusters are present and the clustering kernel $p(\mathbf{y}_i|\vartheta_h)$ with cluster-specific parameter ϑ_h could be used for describing all time series in group h, $h = 1, \dots, H$, i.e. $p(\mathbf{y}_i|S_i, \vartheta_1, \dots, \vartheta_H) = p(\mathbf{y}_i|\vartheta_{S_i})$, where $S_i \in \{1, \dots, H\}$ is a latent group indicator. The (unknown) group indicators $\mathbf{S} = (S_1, \dots, S_N)$ are a priori independent and have to be estimated along with the cluster-specific parameters from the data (see next Section).

An important building block for clustering discrete-valued time series is the first-order time-homogeneous *Markov chain model* characterised by the transition matrix ξ, where $\xi_{jk} = \Pr(y_{it} = k|y_{i,t-1} = j)$, $j, k = 1, \dots, K$. Each row of ξ represents a probability distribution over the discrete set $\{1, \dots, K\}$, i.e. $\sum_{k=1}^{K} \xi_{jk} = 1$.

The Markov chain clustering approach of Pamminger and Frühwirth-Schnatter [6] is based on choosing such a Markov chain model with cluster-specific transition matrix ξ_h as clustering kernel. Hence, the group-specific parameter ϑ_h is equal to ξ_h and the clustering kernel $p(\mathbf{y}_i|\xi_h)$ reads:

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

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.

Pamminger and Frühwirth-Schnatter [6] assume that the transition matrices $\xi_1, \dots, \xi_H$ are entirely unconstrained which allows a certain amount of flexibility in capturing differences in the transition behavior between the groups.

PROBABILISTIC MODEL FOR THE GROUP INDICATORS

A simple common probability model for the group indicators $\mathbf{S} = (S_1, \dots, S_N)$ under complete prior ignorance is: $\Pr(S_i = h|\eta_1, \dots, \eta_H) = \eta_h$, where η_h is the relative size of group h, i.e. $\sum_{h=1}^{H} \eta_h = 1$.

In order to incorporate unit-specific information $\mathbf{x}_i$ one can also formulate a more general *multinomial logit model* (MNL) for $\mathbf{S}$:

$$\Pr(S_i = h|\beta_2, \dots, \beta_H) = \frac{\exp(\mathbf{x}_i \beta_h)}{1 + \sum_{l=2}^{H} \exp(\mathbf{x}_i \beta_l)}, \tag{2}$$

where $\mathbf{x}_i$ is a row vector of regressors, including 1 for the intercept and $\beta_2, \dots, \beta_H$ are group-specific, unknown parameters. For identifiability reasons we set $\beta_1 = \mathbf{0}$, which means that $h = 1$ is the baseline group and β_h is the effect on the log-odds ratio relative to the baseline.

It can easily be shown that the complete prior ignorance model for $\mathbf{S}$ corresponds to the special case of the MNL without any covariates but with a different parameterisation (see Frühwirth-Schnatter and Kaufmann [4]).

An important aspect of model-based clustering is that we do not know a priori which time series belong to which group. Therefore the group indicators $\mathbf{S}$ have to be estimated along with the group-specific parameters $\xi_1, \dots, \xi_H$ and the regression parameters $\beta_2, \dots, \beta_H$ from the data.

BAYESIAN INFERENCE

The latent group indicators **S** are estimated along with the group-specific parameters $\xi_1,\dots,\xi_H$ as well as $\beta_2,\dots,\beta_H$ from the data for a fixed number of clusters. We further assume prior independence between $\xi_1,\dots,\xi_H$ and $\beta_2,\dots,\beta_H$. Conditional on knowing $\beta_1,\dots,\beta_H$ the observations are mutually independent. $\beta_2,\dots,\beta_H$ follow a priori a (standard) normal distribution with known hyperparameters.

The MCMC sampling algorithm of Pamminger and Frühwirth-Schnatter [6] is now adapted with respect to the multinomial logit classification and described in **Algorithm 1** (for details see below). The result is a (full) Gibbs-type sampler that needs only draws from standard distributions.

Primarily, choose appropriate initial values for the model parameters and then repeat the following steps:

Algorithm 1.

1. *Bayes' classification for each individual i given* $\beta_2,\dots,\beta_H$ *and* $\xi_1,\dots,\xi_H$*:* draw $S_i,\ i=1,\dots,N$ from the discrete probability distribution which combines the likelihood $p(\mathbf{y}_i|\xi_h)$ (which corresponds to the clustering kernel defined in (1)) with the prior (2):

$$\Pr(S_i = h|\mathbf{y}_i,\mathbf{x}_i,\beta_2,\dots,\beta_H,\xi_1,\dots,\xi_H) \propto p(\mathbf{y}_i|\xi_h)\,\frac{\exp(\mathbf{x}_i\beta_h)}{1+\sum_{l=2}^{H}\exp(\mathbf{x}_i\beta_l)}, \quad h=1,\dots,H. \tag{3}$$

2. *Sample transition matrices* $\xi_1,\dots,\xi_H$ *given* **S***:* draw ξ_h from $p(\xi_h|\mathbf{S},\mathbf{y}),\ h=1,\dots,H$. Note that only Dirichlet distributions are involved (for details see Pamminger and Frühwirth-Schnatter [6]).
3. *Sample regression parameters* $\beta_2,\dots,\beta_H$ *given* **S***:* draw $\beta_h,\ h=2,\dots,H$ from the multinomial logit model (2) using the *auxiliary mixture sampler* of Frühwirth-Schnatter and Frühwirth [3] who use data augmentation steps to eliminate non-linearity and non-normality in such a way that finally only draws from standard distributions are necessary.

APPLICATION

We continue the application in Pamminger and Frühwirth-Schnatter [6] considering wage mobility in the Austrian labour market, now in particular with respect to entry conditions. Wage mobility describes chances but also risks of an individual to move between wage categories over time. The intention is to search for groups of individuals with similar wage mobility behavior and to find out whether factors like unemployment rate, skills, entry year or color of the collar have an effect on the probabilities to belong to a certain group.

Thus we apply the suggested model to this panel reporting the wage category for young men entering the Austrian labour market in successive years. The data were taken from the Austrian Social Security Data Base (Zweimüller et al. [7]). The panel consists of time series observations for almost fifty thousand native men entering the labour market between 1975 and 1985 at an age of at most 25 years (for examples see Figure 1). The time series represent gross monthly wages in May of successive years and exhibit individual lengths up to 32 years (the median length is equal to 22). The wage is divided into six categories labelled with 0, corresponding to zero-wage, and 1 to 5, corresponding to the quintiles of the respective income distribution of all men.

Various model selection criteria suggest a four-group solution which also allows sensible interpretations from an economic point of view. We choose the following labelling for each cluster which can be seen from Figure 1 (where some typical group members are depicted) in order to give a short and concise description.

The posterior transition probabilities allow the interpretation that in the *downward* cluster the risk for an individual in any wage category to drop into the no-wage category is much higher than for the other clusters. Also the probability to remain in the no-wage category is quite high. The chance to move out of the no-wage category is quite small.

In the *upward* cluster the chance to move out of the no-wage category into any wage category is quite high compared to the other groups. The chance to move forward into the next higher wage category is in general higher in this group. In the *static* cluster the probabilities to stay in a certain wage category are pretty high. Members of the *mobile* cluster move quickly between the various wage categories.

Using the estimated regression coefficients we can investigate the effect of a certain covariate (representing some entry conditions) on the log odds ratio of belonging to a certain group instead of belonging to the *upward* group (which is chosen to be the baseline).

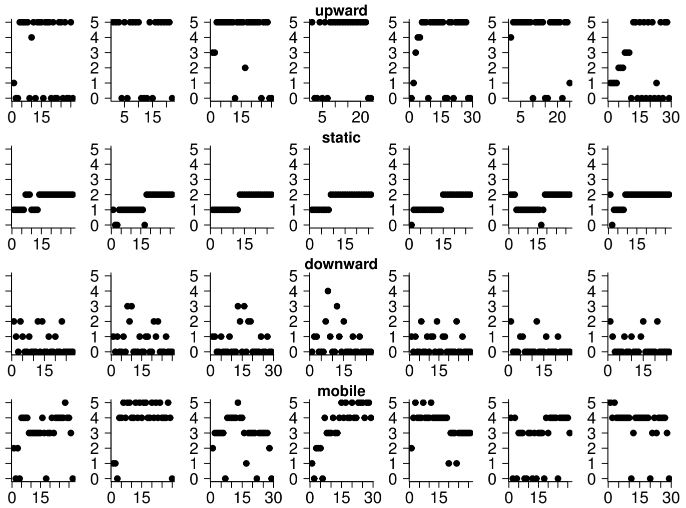

FIGURE 1. Typical group members (the seven individuals with the highest posterior classification probabilities).

The 'risk' to belong to the *static*, *downward* or *mobile* cluster is higher compared to belonging to the *upward* cluster the higher the unemployment rate in the corresponding district at the time of the job entry. With higher education or when an employee started as white collar worker the chance to belong to the *upward* group is considerably higher compared to belonging to any other group.

ACKNOWLEDGMENTS

The second author would like to thank Milan Stehlík for the invitation to give a talk in his session *Statistics and Modelling* at ICNAAM 2010.

The second author's research is supported by the Austrian Science Foundation (FWF) under the grant S 10309 (NRN "The Austrian Center for Labor Economics and the Analysis of the Welfare State", Subproject "Bayesian Econometrics").

REFERENCES

1. Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and Non-Gaussian Clustering. *Biometrics*, 49: 803–821.
2. Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97: 611–631.
3. Frühwirth-Schnatter, S. and Frühwirth, R. (2010). Data augmentation and MCMC for binary and multinomial logit models. In Kneib, T. and Tutz, G. (Eds.): *Statistical Modelling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*, 111–132, Heidelberg: Physica-Verlag.
4. Frühwirth-Schnatter, S. and Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*, 26: 78–89.
5. Liao, T. W. (2005). Clustering of time series data – a survey. *Pattern Recognition*, 38: 1857–1874.
6. Pamminger, C. and Frühwirth-Schnatter, S. (2010). Model-based Clustering of Categorical Time Series. *Bayesian Analysis*, 5(2): 345–368.
7. 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.