Model-based Clustering of Time Series - A Review from a Bayesian Perspective

Sylvia Frühwirth-Schnatter*

April 6, 2011

Abstract

Clustering is a widely used statistical tool to determine subsets in a given data set. Frequently used clustering methods are mostly based on distance measures and cannot easily be extended to cluster time series within a panel or a longitudinal data set. The paper reviews recently suggested approaches to model-based clustering of panel or longitudinal data based on finite mixture models. Several approaches are considered that are suitable both for continuous as well as categorical time series observations. Bayesian estimation through Markov chain Monte Carlo methods is described in detail and various criteria to select the number of clusters are reviewed. An application to a panel of marijuana use among teenagers serves as an illustration.

Keywords: Markov chain Monte Carlo, longitudinal data, panel data, model selection, discrete-valued time series, regularization, skewed distributions

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

In many areas of applied statistics, like economics, finance, or public health, data are available in the form of panel or longitudinal data where repeated measurements are taken for a group of subjects at several points in time. Hence, the observations for a single subject form a time series. Standard methods for panel or longitudinal data analysis assume homogeneity across the subjects in the sense that all time series can be described by the same data generating mechanism (Diggle et al., 2002; Hsiao, 2003). Very often, however, it is of substantive interest to learn if heterogeneity is present in the sense that the data generating mechanism is different across the individual time series.

Consider, for example, data on the marijuana use of 237 teenagers taken from five annual waves (1976-80) of the National Youth Survey (Lang et al., 1999). The respondents whose gender is also known were 13 year old in 1976 and reported for five consecutive years their marijuana use in the past year as a categorical variable with the categories “never”, “not more than once a month” and “more than once a month”. The individually observed response profiles, presented in Figure 1 for ten teenagers, indicate that various types of marijuana use are present among these teenagers. It is of interest to identify groups of teenagers with similar marijuana use behavior, to describe the data generating mechanism within each group, and to investigate, if individual characteristics like the gender can be associated with a certain type of marijuana use.

To capture heterogeneity across such time series it is common to assume that groups (or clusters) are present, in the sense that each group can be characterized by a different data generating mechanism. Each time series is considered to be a single entity belonging to one of these groups and all time series within the same group are characterized by the same data generating mechanism. The main issue with such an approach is how the time series are assigned to the various groups.

A common procedure is to define the groups apriori in a deterministic manner, by stratifying the time series according to some observed covariates, and to estimate the data-generating
mechanism independently for each group. It is then examined aposteriori, if the underlying
data generating mechanism differs across the various strata. However, deterministic prior strat-
ification may be incorrect in the sense that other, possibly unobserved, variables define the
true grouping and time series are assigned to the same group, although their underlying data
generating mechanism is actually different. This might introduce a bias, when estimating the
group-specific parameters. Evidence of that is provided in Section 5 where the panel of marijuana
use of teenagers is analyzed in details.

In the recent years, a new approach toward grouping time series became popular which is
based on letting the data themselves define an appropriate grouping. In such an approach, the
mechanism assigning each time series to a particular group is random rather than deterministic
and is captured by a latent (endogenous) variable which is estimated along with the group-
specific characteristics. Each time series has an apriori unknown probability to belong to a
certain cluster. The corresponding prior assignment distribution may be either independent of
the specific time series or may depend on characteristics of the time series such as the covariates
that are typically used for deterministic prior stratification.

Given observed panel data, the prior assignment distribution is updated for each individual
time series using Bayes’ rule to obtain the posterior assignment distribution which forms the
basis for assigning each time to a certain group. Hence, the final grouping of the time series is to
a large extend determined in a data driven manner. Furthermore, the group-specific parameters
in the data generating mechanism are determined simultaneously with the grouping.

This approach is closely related to what has been known in statistics for many decades as
(unsupervised) clustering, see e.g. Everitt et al. (2001). Unsupervised or endogenous clustering
of time series is quite a recent issue, because the extension of common clustering techniques
to time series data is far from trivial. Distance-based clustering methods, for instance, cannot
be easily extended to time series data, where an appropriate distance-measure is rather diffi-
cult to define, see e.g. the review by Liao (2005). An alternative popular approach to cluster
(non-time series) data is model-based clustering based on finite mixture models (Banfield and
Raftery, 1993; Fraley and Raftery, 2002). As opposed to distance-based methods, model-based
clustering extends to time series data in a quite a natural way as shown by Frühwirth-Schnatter
and Kaufmann (2008).1 Heard et al. (2006) provide further empirical evidence in the context of
clustering gene expression time series that model-based clustering leads to meaningful results in
cases where Euclidian-distance-based clustering methods fail.

In model-based clustering, each time series is considered to be a single entity belonging to
one of \( H \) each endogenous groups, where each group is described by a different data generating
mechanism. In terms of mixture modeling this leads to a multivariate finite mixture model
with \( H \) components where the data generating mechanism of each group defines the component
specific distribution (Frühwirth-Schnatter, 2006). Clustering is achieved as for a traditional
finite mixture model by assigning each time series to one of the \( H \) groups.

The group-specific data generating mechanism, also called clustering kernel in the context of
model-based clustering, plays a crucial role in this respect. Clustering kernels for panels of real-
valued time series are typically based on dynamic regression models (Frühwirth-Schnatter and
Kaufmann, 2008; Juárez and Steel, 2010), multivariate normal distributions with parsimonious
covariances matrices specifically designed for longitudinal data (McNicholas and Murphy, 2010),
or GARCH models (Bauwens and Rambouts, 2007). Clustering kernels for panels with discrete-
valued time series observations are based on the probit model (Aßmann and Boysen-Hogrefe,

---

1Research report versions of this paper have been circulated since 2002 and were first presented at the Valencia 7
Meeting on Teneriffe (Spain) in June 2002 and at seminar presentations at the CORE, Louvain-la-Neuve (Belgium)
and the Econometrics Institute, Rotterdam (NL) in October 2002.
2011), the multinomial logit model (Greene and Hensher, 2003) or on Markov chain models (Pamminger and Frühwirth-Schnatter, 2010; Frühwirth-Schnatter et al., 2011). These, and several more, clustering kernels are reviewed in Section 2.

Estimation for model-based clustering may be carried out using maximum likelihood (ML) estimation based on the expectation-maximization (EM) algorithm, see e.g. McLachlan and Peel (2000). The Bayesian approach suggested by Frühwirth-Schnatter and Kaufmann (2008) offers several advantages compared to ML estimation, as will be discussed in Section 3. The Bayesian approach requires the choice of a proper prior distributions for the parameters appearing in the group-specific clustering kernel. While this might appear as a restriction compared to ML estimation, it actually turns out that the prior acts as a convenient tool to regularize the likelihood function which might be highly irregular for mixture models. Practical Bayesian inference is typically carried out using Markov chain Monte Carlo methods based on data augmentation.

A crucial issue in model-based clustering is, of course, how to select the number of clusters present in the panel. Common model selection criteria such as AIC, BIC, DIC, ICL-BIC, AWE, and marginal likelihoods are reviewed in Section 4. Finally, the panel describing marijuana use among teenager introduced above will serve in Section 5 as an illustrative application of model-based clustering of panel data.

2 Model-Based Clustering of Longitudinal or Panel Data Using Finite Mixtures

Longitudinal or panel data arise when repeated measurements \( y_{it} \) are taken on \( N \) subjects, indexed by \( i = 1, \ldots, N \), at a number of points in time, typically indexed by \( t = 0, \ldots, T_i \). Focus is here on relatively short time series where \( \max_i T_i \) is small compared to the number \( N \) of subjects. Subsequently, \( \mathbf{y}_i = \{y_{i0}, \ldots, y_{iT_i}\} \) denotes each individual time series, while \( \mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\} \) refers to the whole panel. The measurement \( y_{it} \) for subject \( i \) on the \( t \)th occasion may be a continuous variable taking values in \( \mathbb{R} \) or \( \mathbb{R}^+ \), or a categorical variable with \( K \) potential states labelled by \( k \in \{1, \ldots, K\} \). In addition, one may observe exogenous covariates \( \mathbf{x}_{it} \) of potential influence on the distribution of the outcome variable \( y_{it} \).

2.1 Clustering time series using finite mixture models

Model-based clustering as introduced by Frühwirth-Schnatter and Kaufmann (2008) is based on formulating a clustering kernel for each individual time series in terms of a sampling density \( p(y_i|\theta) \), where \( \theta \in \Theta \) is an unknown model parameter.

The central assumption is that the \( N \) time series arise from \( H \) hidden groups. Within each group, say \( h \), all time series can be characterized by the same data generating mechanism defined by the density \( p(y_i|\theta_h) \), where \( \theta_h \) is an unknown group-specific parameter. A latent group indicator \( S_i \) taking a value in the set \( \{1, \ldots, H\} \) is introduced for each time series \( y_i \) to indicate which cluster the time series belongs to:

\[
p(y_i|S_i, \theta_1, \ldots, \theta_H) = p(y_i|\theta_{S_i}) = \begin{cases} 
p(y_i|\theta_1), & \text{if } S_i = 1, \\
\vdots & \\
p(y_i|\theta_H), & \text{if } S_i = H.
\end{cases}
\]

In model-based clustering it is common to assume that the group assignment indicators \( S_1, \ldots, S_N \) are independent a priori given a set of parameters \( \eta \), i.e. \( p(S_1, \ldots, S_N|\eta) = p(S_1|\eta) \cdots p(S_N|\eta) \).
This lead to a representation of the marginal distribution \( p(y_i|\theta_H) \), where \( \theta_H = (\vartheta_1, \ldots, \vartheta_H, \eta) \), in terms of a finite mixture distribution with \( H \) components:

\[
p(y_i|\theta_H) = \sum_{h=1}^{H} \Pr(S_i = h|\eta)p(y_i|\vartheta_h).
\]

It is assumed that the subjects in the panel are independent given \( \theta_H \). Hence the data generating mechanism for the entire panel \( y \) is formulated by defining a data generating mechanism independently for each \( y_i \). Hence, to set up model-based clustering for a specific panel, two modelling assumptions have to be made, namely choosing the clustering kernel \( p(y_i|\vartheta_h) \) and choosing a model for the prior assignment distribution \( p(S_i = h|\eta) \), \( h = 1, \ldots, H \).

### 2.2 Choosing the clustering kernel

The crucial point in model-based clustering of time series is to select an appropriate clustering kernel in terms of a sampling density \( p(y_i|\vartheta_h) \) which captures salient features of the observed time series. To address serial correlation among the observations for each subject, model-based clustering of time series data is often based on dynamic clustering kernels, i.e.:

\[
p(y_i|\vartheta_h) = \prod_{t=1}^{T_i} p(y_{it}|y_{i,t-1}, [x_{it}, ]\vartheta_h).
\]

For such a dynamic clustering kernel it is necessary to deal with the initial condition problem, see Subsection 2.4. Other approaches dealing with model-based clustering of panel data ignore serial correlation and assume that the \( T_i \) observations are conditionally independent, given cluster assignment, i.e.:

\[
p(y_i|\vartheta_h) = \prod_{t=1}^{T_i} p(y_{it}|[x_{it}, ]\vartheta_h).
\]

(1) may be called a latent class clustering kernel, because assuming conditional independence, given cluster assignment, is one the corner stones of latent class analysis (Clogg, 1995).

Hence, a dynamic clustering kernel is defined through the density \( p(y_{it}|y_{i,t-1}, [x_{it}, ]\vartheta_h) \), while a latent class clustering kernel is defined through the density \( p(y_{it}|[x_{it}, ]\vartheta_h) \). The specific choice of these densities depends, of course, on the nature of \( y_{it} \).

#### 2.2.1 Clustering kernels for real-valued time series observations

Clustering kernels for panels with real-valued time series observations \( y_{it} \) are typically based on dynamic regression models (Frühwirth-Schnatter and Kaufmann, 2008; Juárez and Steel, 2010):

\[
y_{it} = \zeta_h + \delta_h y_{i,t-1} + x_{it}\vartheta_h + \sigma_h \varepsilon_{it},
\]

where \( \varepsilon_{it} \sim f(\varepsilon_{it}) \) and all parameters of the clustering kernel are cluster-specific, including the variance of the error term. It makes sense for specific applications to assume that certain parameters are the same across the clusters, for instance to assume that the long-run mean is cluster-specific, while all other coefficients are pooled across clusters (Juárez and Steel, 2010). A latent class regression kernel results, if \( \zeta_h = 0 \) in all clusters. Dynamic regression clustering kernels have been applied, e.g., in monetary policy (Frühwirth-Schnatter and Kaufmann, 2006).
and in economic growth analysis (Frühwirth-Schnatter and Kaufmann, 2008; Paap et al., 2005; Owen et al., 2009; Başturk et al., 2010), while latent class regression kernel turned out to be useful in marketing research (Lenk and DeSarbo, 2000; Frühwirth-Schnatter et al., 2004; Rossi et al., 2005).

Various error models may be assumed in (2), the most common choice being i.i.d. standard normal errors $\varepsilon_{it}$. To achieve robustness against outliers, usually the Student-$t$ distribution is employed. Choosing $\varepsilon_{it}|\omega_i \sim N(0,1/\omega_i)$, $\omega_i \sim G(\nu/2,\nu/2)$, marginally leads to a multivariate Student-$t$ distribution for $\varepsilon_{i0}, \ldots, \varepsilon_{iT_i}$ and introduces robustness against atypical time series as a whole. Choosing $\varepsilon_{it}|\omega_i \sim N(0,1/\omega_i)$, $\omega_{it} \sim G(\nu/2,\nu/2)$, leads to i.i.d. $t_\nu$ errors and introduces robustness against atypical individual observations. To capture conditional skewness within each cluster, Juárez and Steel (2010) assume that $f(\varepsilon_{it})$ arises from a skew Student-$t$ distribution (Fernández and Steel, 1998) and show for panels from economic growth analysis that this kernel is superior to kernels based on the normal distribution.

The simple dynamic structure of (2) has been modified to capture general autocorrelation patterns, eg. by included higher order AR terms (Frühwirth-Schnatter and Kaufmann, 2008). Pure autoregressive clustering kernels of higher order were applied in Ramoni et al. (2002b) to cluster gene expression time series. Bauwens and Rambouts (2007) considered GARCH models as clustering kernels to capture conditional heteroscedasticity for financial time series.

McNicholas and Murphy (2010) pursue a slightly different approach for balanced panels with $T_i \equiv T$ by using a multivariate normal distribution as clustering kernel for the whole time series $y_{it}$, i.e. $y_{it}|\theta_h \sim N_{T+1}(\mu_h, \Sigma_h)$. Based on the Cholesky decomposition $\Sigma_h^{-1} = T_h' D_h^{-1} T_h$, where $D_h = \text{Diag}(\sigma_{h0}^2, \ldots, \sigma_{hT}^2)$ and $T_h$ is lower triangular matrix with unit diagonal, parsimonious covariance structures are achieved by constraining $D_h$ or $T_h$ to be equal across clusters or by imposing the isotropic constraint $D_h = \sigma_h^2 I$, where $I$ is the identity matrix. Since $(y_{iT}, \mu_h)' \Sigma_h^{-1} (y_{i} - \mu_h) = (T_h (y_i - \mu_h))' D_h^{-1} (T_h (y_i - \mu_h))$, it follows that $T_h y_i \sim N_{T+1}(T_h \mu_h, D_h)$. This kernel has a representation involving the conditional density $p(y_{it}|y_{i0}, \ldots, y_{it-1})$, for $t = 1, \ldots, T$:

$$y_{it} = \mu_{ht} + \sum_{j=1}^{t-1} (-\delta_{h,t,t-j}) (y_{i,t-j} - \mu_{ht}) + \sigma_{ht} \varepsilon_{it},$$

(3)

where $\delta_{h,t,t-j}$ are the sub-diagonal elements of $T_h$, $\mu_{ht}$ are the elements of $\mu_h$, and $\varepsilon_{it}$ is i.i.d. standard normal. If $T_h$ is constrained to be zero below the $p$th sub-diagonal, this corresponds to including time-varying AR($p$) coefficients in model (2). The formulation as a dynamic regression model (3) allows to extend this kernel to unbalanced panels.

Several authors (Luan and Li, 2003; Heard et al., 2006) use non-linear regression splines in model-based clustering of gene expression time series. This is another special case of the regression kernel (2) where $x_{it}$ is a vector of non-linear basis functions of the time ordinate $t$.

This clustering kernel also allows to handle longitudinal data where the time points between the individual observations are unequally spaced.

### 2.2.2 Clustering kernels for panels of binary and count data

Clustering kernels for dichotomous or binary panels are often based on the probit model:

$$\Pr(y_{it} = 1|x_{it}, S_i = h) = \Phi(x_{it} \beta_h),$$

(4)

where $\Phi(\cdot)$ is the cdf of the standard normal distribution. Aßmann and Boysen-Hogrefe (2011), for instance, discuss an application of this clustering kernel in industrial organization to model
latent heterogeneity of firm activity in product and process innovation. Alternatively, the logit model may be used as clustering kernel, see e.g. Zhu and Zhang (2004), among many others.

Clustering kernels for panels of count data are typically based on the Poisson distribution:

\[ y_{it}|S_i = h \sim \mathcal{P}(\lambda_{ith}), \quad \log \lambda_{ith} = \mathbf{x}_{it}^\top \vartheta_h, \] (5)

see, again, Zhu and Zhang (2004). Further applications are found e.g. in marketing research (Wedel and DeSarbo, 1995). Roeder et al. (1999) use a zero-inflated version of this model and discuss an application to criminology.

### 2.2.3 Clustering kernels for general discrete-valued time series observations

For more general panels of discrete-valued time series, \( y_{it} \) is a categorical variable with \( K \) potential states labelled by \( k \in \{1, \ldots, K\} \). Any clustering kernel is based on modelling \( \Pr(y_{it} = k|\vartheta_h) \), \( k = 1, \ldots, K \), in terms of group-specific parameters \( \vartheta_h \). In dynamic discrete-valued time series panels, persistence is present and is captured by assuming that \( \Pr(y_{it} = k|y_{i,t-1} = j, S_i = h) \) depends on the previous state \( y_{i,t-1} \). An important building block for clustering dynamic discrete-valued panels is the first-order time-homogeneous Markov chain model characterized by a cluster-specific transition matrix \( \xi_h \), where

\[ \xi_{h,jk} = \Pr(y_{it} = k|y_{i,t-1} = j, S_i = h), \quad j, k = 1, \ldots, K, \quad \sum_{k=1}^K \xi_{h,jk} = 1. \]

Each row \( \xi_{h,j} = (\xi_{h,j1}, \ldots, \xi_{h,jK}) \) of the matrix \( \xi_h \) represents a probability distribution over the discrete set \( \{1, \ldots, K\} \). Markov chain clustering is based on the assumption that within each cluster such a Markov chain model with group-specific transition matrix \( \xi_h \) could be used as clustering kernel, i.e.

\[ p(y_i|\xi_h) = \prod_{j=1}^K \prod_{k=1}^K \xi_{N_{i,jk}}, \] (6)

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 (6) is formulated conditional on the first observation \( y_{i0} \).

Applications of this approach include Cadez et al. (2003) who clustered users according to their behavior on a web site, Ramoni et al. (2002a) who clustered sensor data from mobile robots, and Pamminger and Frühwirth-Schnatter (2010) and Frühwirth-Schnatter et al. (2011) who cluster individual wage careers in the Austrian labor market.

While these approaches assume that the transition matrices \( \xi_1, \ldots, \xi_H \) are entirely unconstrained, other authors use more parsimonious versions of this clustering method. To incorporate a simple form of heterogeneity across individual labor market transition data, Fougère and Kamionka (2003) considered a mover-stayer model in continuous time where \( H = 2, \xi_1 = \mathbf{1} \) and \( \xi_2 \) is unconstrained. Frydman (2005) considered a constrained version where \( \xi_{h,j} \), \( h \geq 2 \), is related to \( \xi_1 \) through \( \xi_h = \mathbf{1} - \Lambda_h (\mathbf{1} - \xi_1) \) where \( \Lambda_h = \text{Diag}(\lambda_{h,1}, \ldots, \lambda_{h,K}) \) with \( 0 \leq \lambda_{h,j} \leq 1/(1 - \xi_{1,jj}) \) for \( j = 1, \ldots, K \). This method is applied to bond ratings migration.

For other panels of discrete-valued time series more general clustering kernels turned out to be useful. The clustering kernel may involve the use of \( k \)th order Markov chains in order to extend the memory of the clustering kernel to the past \( k \) observations, see e.g. Saul and Jordan (1999). If additional covariate information \( \mathbf{x}_{it} \) is available, inhomogeneous Markov chains could
be used as clustering kernels, by modeling the rows of the transition matrix in group \( h \) through a dynamic multinomial logit (MNL) model:

\[
\Pr(y_{it} = k | y_{i,t-1} = j, x_{it}, S_i = h) = \frac{\exp(\lambda_{ith,jk})}{\sum_{l=1}^{K} \exp(\lambda_{ith,kl})}, \quad \lambda_{ith,jk} = \gamma_{h,jk} + x_{it} \beta_{h,k},
\]

where \( k = 1, \ldots, K \) and \( \beta_{h,k} \) is a group and category specific regression parameters capturing the effect of the covariates \( x_{it} \). To achieve identifiability, \( \gamma_{h,j0} = 0 \) for each \( j = 1, \ldots, K \). If no covariates are present, then (7) reduces to Markov chain clustering with the transition matrix \( \xi_h \) being parameterized in terms of \( \gamma_{h,jk} \). A special version of (7) where the potential values of \( x_{it} \) are limited is discussed in Subsection 5.1.

If \( \gamma_{h,jk} = 0, h = 1, \ldots, H, j, k = 1, \ldots, K \), then no persistence is present and (7) reduces to a latent class MNL clustering kernel for repeated categorical observations. A latent class MNL clustering kernel where the covariates rather than the regression coefficient is category specific, i.e. \( \lambda_{ith,k} = x_{it} / \beta_{h,k} \), found numerous application in clustering panels of repeated discrete choices, e.g. in marketing research (Wedel and DeSarbo, 1993) and in transportation research (Greene and Hensher, 2003).

### 2.2.4 Capturing unobserved heterogeneity within a cluster

All clustering kernels discussed so far allow the data generation mechanism to be different across clusters, but assume it to be identical for all subjects within a cluster. An important extension allows for unobserved heterogeneity within a cluster by incorporating random effects into the formulation of the clustering kernel. This has been suggested both for real- and discrete-valued panels. From the view point of mixture modeling, these clustering kernels lead to finite mixtures of random-effects models as marginal distribution (Frühwirth-Schnatter, 2006, Section 8.5).

For real-valued time series observations, random coefficients are introduced into the dynamic regression model (2):

\[
y_{it} = \zeta_h + \delta_h y_{i,t-1} + x_{it} \beta^*_{i} + \sigma_h \epsilon_{it}.
\]

Hence, each time series is assumed to be generated by an individual regression coefficient \( \beta^*_i \) which is allowed to deviate from the cluster-specific regression coefficient \( \beta_h \). The deviation with each group is described through a group-specific probability distribution for \( p(\beta^*_i | S_i = h) \) such that the expected value of \( \beta^*_i \) in cluster \( h \) is equal to the average cluster-specific regression coefficient \( \beta_h \), typically \( \beta^*_i | S_i = h \sim N_d(\beta_h, Q(\phi_h)) \), where the variance-covariance matrix depends on unknown parameters \( \phi_h \) controlling the amount of heterogeneity within each group.

It is worth noting that the clustering kernel \( p(y_{i} | \beta_h, \phi_h) \) where \( \beta^*_i \) is integrated out is equal to a normal distribution with a general autocorrelation structure. Hence, despite within-cluster unobserved heterogeneity, each cluster may be characterized by the “typical” cluster-specific regression coefficient \( \beta_h \) and the parameter \( \phi_h \) measuring the amount of heterogeneity in \( \beta^*_i \) within each cluster. Such clustering kernels turned out to be useful e.g. in economic growth analysis (Canova, 2004; Frühwirth-Schnatter and Kaufmann, 2008; Juárez and Steel, 2010) and in marketing research (Lenk and DeSarbo, 2000; Frühwirth-Schnatter et al., 2004).

Clustering based on random-effects kernels has also been developed for discrete-valued time series. Pamminger and Frühwirth-Schnatter (2010) suggested a generalization of Markov chain clustering, called Dirichlet multinomial clustering, which takes unobserved heterogeneity within each cluster into account, by assuming that each time series is generated by an individual transition matrix \( \xi^*_i \) which is allowed to deviate from the cluster-specific transition matrix \( \xi_h \).
The deviation within each group is described through a group-specific probability distribution \( p(\xi_i^s|h_i = h) \) such that the expected value of \( \xi_i^s \) in cluster \( h \) is equal to the average cluster-specific transition matrix \( \xi_h \), i.e. each row \( \xi_{i,j}^s \), \( j = 1, \ldots, K \), of \( \xi_i^s \) follows a Dirichlet distribution, \( \xi_{i,j}^s|h_i = h \sim \mathcal{D}(\phi_{hj_1}, \ldots, \phi_{hj_K}) \). The unknown parameter \( \phi_h = (\phi_{h1}, \ldots, \phi_{hK}) \) controlling the amount of heterogeneity within each group.

A distinctive advantage of modeling within-group unobserved heterogeneity in this way is that the clustering kernel \( p(y_i|\xi_h, \phi_h) \) where \( \xi_i^s \) is integrated out is available in closed form. Hence, despite unobserved heterogeneity, each cluster may be characterized by the “typical” cluster-specific transition matrix \( \xi_h \) and the parameter \( \phi_h \) measuring the amount of heterogeneity in the individual transition matrices \( \xi_i^s \) within each cluster. The clustering kernel \( p(y_i|\xi_h, \phi_h) \) underlying Dirichlet multinomial clustering may be seen as a “robustified” version of the clustering kernel \( p(y_i|\xi_h) \) of first-order Markov chain clustering, because this clustering kernel no longer describes a first-order Markov process, but allows for higher order dependence.

Random-effects have also been introduced into clustering kernels based on the probit model (Aßmann and Boysen-Hogrefe, 2011) as well as the logit model and the MNL model (Rossi et al., 2005). However, no explicit expression for the marginal distribution \( p(y_i|\beta_h, \phi_h) \) is available for these kernels.

### 2.3 Choosing the prior assignment distribution

An important feature of model-based clustering is relaxing the strict grouping of the subjects according to deterministic prior stratification. Assume that deterministic prior stratification of the individual time series is based on the \( H \) distinct values \( z_{i1}^h, \ldots, z_{ih}^h \) of a collection \( z_i \) of subject-specific factors. Then following deterministic assignment rule into the \( H \) groups holds for all subjects \( i \) for \( h = 1, \ldots, H \):

\[
Pr(S_i = h) = I\{z_i = z_i^0\},
\]

where \( I\{\cdot\} \) is the indicator function. Hence the \( i \)th subject belongs to the \( h \)th group, iff \( z_i = z_i^0 \).

Instead of strict assignment according to (8), model-based clustering uses a softer prior assignment rule, based on the prior assignment distribution \( Pr(S_i = h|\eta) \), \( h = 1, \ldots, H \), which depends on unknown parameters \( \eta \) that are estimated from the data. Random prior assignment has the advantage that assignment is adjusted for each subject \( i \), given the observed individual time series \( y_i \), by means of Bayes’ rule:

\[
Pr(S_i = h|y_i, \theta_H) \propto p(y_i|\theta_h)Pr(S_i = h|\eta), \quad h = 1, \ldots, H,
\]

where \( p(y_i|\theta_h) \) is the clustering kernel discussed above. In this sense, the final grouping of the time series is to a large degree data driven. The standard assumption concerning the prior assignment distribution is homogeneity across the subjects, i.e.:

\[
Pr(S_i = h|\eta) = \eta_h,
\]

where \( \sum_{h=1}^H \eta_h = 1 \). Hence, \( \eta_h \) is equal to the relative size of cluster \( h \) in the underlying population of subjects. Homogeneity of random prior assignment implies that each subject has the same probability to belong to a certain cluster, regardless of its specific characteristics.

To obtain a more meaningful model for data where this assumption seems to be unrealistic, observed subject-specific factors \( z_i \) are allowed to impact the prior assignment distribution \( Pr(S_i = h|\eta) \) which is modelled as a multinomial logit (MNL) model:

\[
Pr(S_i = h|\eta) = \frac{\exp(z_i\gamma_h)}{1 + \sum_{l=2}^H \exp(z_i\gamma_l)},
\]
where the row vector $z_i$ includes a constant for the intercept, in addition to the subject specific factors. For identifiability reasons $\gamma_1 = 0$, which means that $h = 1$ is the baseline group and $\gamma_h$ is the effect on the log-odds ratio relative to the baseline. The remaining parameters $\eta = (\gamma_2, \ldots, \gamma_H)$ are group-specific unknown regression coefficients.

As the subject-specific factors $z_i$ typically are the same variables that would be used for a deterministic prior stratification, this approach has been called stratified model-based clustering (Aßmann and Boysen-Hogrefe, 2011). Evidently, stratified random prior assignment through the MNL model (11) relaxes deterministic prior stratification. Stratified random prior assignment coincides with homogeneous prior assignment as in (10), whenever $z_i$ includes only a constant. By testing the regression coefficient corresponding to a certain subject specific factor in $z_i$ in the more general MNL model, important insight is gained into learning which factors determine membership of a certain subject to a specific group. Furthermore, assignment according to (11) allows to deal with the initial conditions problem potentially present for dynamic clustering kernels, see Subsection 2.4 for further details.

Finite mixture models with prior assignment according to (11) have been introduced in the machine learning literature as mixture-of-experts models (Peng et al., 1996) and are also known as smoothly mixing regression model in the econometrics literature (Geweke and Keane, 2007). Mixtures-of-experts models have been applied to stratified model-based clustering of time series in combinations with zero-inflated Poisson kernels (Roeder et al., 1999), dynamic regression clustering kernels (Frühwirth-Schnatter and Kaufmann, 2008), Markov chain clustering (Frühwirth-Schnatter et al., 2011), and latent class MNL clustering kernels (Greene and Hensher, 2003; Aßmann and Boysen-Hogrefe, 2011).

### 2.4 Dealing with the initial condition problem for dynamic clustering kernels

For a dynamic clustering kernels the distribution $p(y_{it}|y_{i,t-1}, [x_{it}],[\vartheta])$ depends on the past observation. With the exception of Frühwirth-Schnatter et al. (2011), model-based clustering using dynamic kernels is carried out conditional on the initial observation $y_{i0}$, by treating this variable as exogenous. This assumption implies that the initial observation $y_{i0}$ is independent of group membership $S_i$. However, it is well-known in the econometrics literature that ignoring a possible dependence between the initial condition and a latent variable capturing unobserved heterogeneity may lead to biased estimators for a dynamic model, see e.g. Wooldridge (2005).

As pointed out by Frühwirth-Schnatter et al. (2011), the initial conditions problem is also relevant for model-based clustering of dynamic time series, where $S_i$ follows a discrete distribution. They consider clustering of discrete-valued panels using simple Markov chain models, however, the solution presented in this paper to handle the initial condition problem is relevant for any dynamic clustering kernel $p(y_{it}|y_{i,t-1}, [x_{it}],[\vartheta])$ with first order dependence and could be applied to clustering real-valued time series without any modification.

Frühwirth-Schnatter et al. (2011) formulate the joint distribution of $y_{i0}, \ldots, y_{i,T}$ and $S_i$ in a way that separates the choice of the clustering kernel density $p(y_{i1}, \ldots, y_{iT}|y_{i0}, [\vartheta])$ which is formulated conditional on $y_{i0}$ and $S_i$ from the choice of a joint model for $y_{i0}$ and $S_i$:

$$p(y_{i0}, \ldots, y_{i,T}, S_i|\vartheta_H) = p(y_{i1}, \ldots, y_{iT}|y_{i0}, [\vartheta])p(y_{i0}, S_i|\vartheta_H).$$

The key issue is to allow for dependence between the initial state $y_{i0}$ and the latent variable $S_i$. Among the two ways of modelling $p(y_{i0}, S_i|\vartheta_H)$, Frühwirth-Schnatter et al. (2011) use the factorization $p(y_{i0}, S_i|\vartheta_H) = p(S_i|y_{i0}, [\vartheta_H])p(y_{i0}|[\vartheta_H])$, which specifies a model for $S_i$ conditional on the initial observation $y_{i0}$ and a marginal model for $y_{i0}$ and extends the “simple solution to the initial conditions problem” (Wooldridge, 2005) to the discrete case. Under the assumption that
\( p(S_i|y_{i0}, \theta_{H,1}) \) and \( p(y_{i0}|\theta_{H,2}) \) have no common parameters, the marginal distribution \( p(y_{i0}|\theta_{H,2}) \) need not be specified explicitly, because it cancels from all posterior distributions.

In terms of the clustering procedure this means that for modeling the random prior assignment distribution \( \Pr(S_i = h|\cdot) \) “simply” the MNL model (11) is used with covariates depending on the initial observation \( y_{i0} \). If the observed time series are continuous, then the initial state \( y_{i0} \) is added to the set of covariates, while for discrete time series indicator variables corresponding to the different states of \( y_{i0} \) are added. This method is applicable regardless of whether additional subject-specific factors are included in \( z_i \), in order to allow for stratified random prior assignment.

3 Bayesian Inference

Another 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 \( S = (S_1, \ldots, S_N) \) are estimated along with the group-specific parameters \( \theta_H = (\varphi_1, \ldots, \varphi_H, \eta) \) from the data.

Estimation of the unknown parameters \( \theta_H \) for a fixed number of clusters \( H \) may be carried by maximizing the observed-data likelihood function \( p(y|\theta_H) \):

\[
p(y|\theta_H) = \prod_{i=1}^{N} \left( \sum_{h=1}^{H} \Pr(S_i = h|\eta) p(y_i|\varphi_h) \right),
\]

using the expectation-maximization (EM) algorithm (McLachlan and Peel, 2000). The EM algorithm has been applied to clustering time series using multivariate normal distributions (McNicholas and Murphy, 2010) and Markov chains (Cadez et al., 2003; Frydman, 2005).

Alternatively, a Bayesian approach may be pursued toward estimating \( \theta_H \). The Bayesian approach requires the choice of a prior distribution \( p(\theta_H) \) which is combined with the observed-data likelihood function \( p(y|\theta_H) \) to derive the posterior distribution of \( \theta_H \) given \( y \) as \( p(\theta_H|y) \propto p(y|\theta_H)p(\theta_H) \). Since this posterior distribution does not have any closed form, practical Bayesian inference is typically carried out by sampling \( M \) draws from the posterior distribution \( p(\theta_H|y) \), using Markov chain Monte Carlo (MCMC) methods based on data augmentation, a method that has been introduced for finite mixture models by Diebolt and Robert (1994), see also Gamerman and Lopes (2006) for a review of MCMC inference in general. Alternative Bayesian approaches use a deterministic heuristic search in the partition space, exploiting the posterior distribution \( p(S|y) \) of the allocations \( S \).

3.1 Choosing the prior distribution

In mixture modeling, the prior distribution has to be proper in order to guarantee that the posterior distribution is well-defined (Frühwirth-Schnatter, 2006, Section 3). It is usual to assume prior independence between the parameters \( \varphi_1, \ldots, \varphi_H \) defining the group-specific clustering kernels and the parameters \( \eta \) defining the prior allocation distribution.

For a homogenous prior allocation distribution, \( \eta = (\eta_1, \ldots, \eta_H) \) is assumed to follow a Dirichlet prior, \( \eta \sim D(\alpha_0, \ldots, \alpha_0) \). The choice of \( \alpha_0 \) may be based on a recent result by Rousseau and Mengersen (2010) which states that choosing \( \alpha_0 > \dim(\varphi_h)/2 \) asymptotically leads to non-empty clusters, whereas choosing \( \alpha_0 < \dim(\varphi_h)/2 \) asymptotically leads to empty clusters, if the model is overfitting. For the mixtures-of-experts model, each parameter \( \gamma_h \) in \( \eta = (\gamma_2, \ldots, \gamma_H) \) independently follows a normal prior distribution.
Priors for $\theta_h$ are chosen to be conditionally conjugate to the clustering kernel $p(y_i|\theta_h)$, if possible. For the dynamic regression clustering kernel (2), for instance, this suggests a normal distribution for the regression parameters $\theta_h$ and an inverted Gamma distribution for the error variance $\sigma^2_h$. For the Markov chain clustering kernel (6), this suggests a Dirichlet distribution for the rows of group specific transition matrix $\xi_h$.

While the need to choose a proper prior might appear restrictive compared to ML estimation, it actually turns out that a proper prior acts as a convenient tool to regularize the likelihood function which often is irregular for finite mixtures, as the following two examples illustrate.

Consider first clustering $N$ continuous time series observations using the simple normal distribution $N(\mu_h, \sigma^2_h)$ which is a very special case of (2) as clustering kernel. Assume that $H = 2$ and that $\mu_1, \sigma^2_1$ and $\eta_1$ are known, whereas only $\mu_2$ and $\sigma^2_2$ are unknown. Assume that for time series $i$ only a single observation $y_i$ is available. The contribution of this particular time series to the likelihood function reads:

$$p(y_i|\mu_2, \sigma^2_2) = \eta_1 f_N(y_i; \mu_1, \sigma^2_1) + (1 - \eta_1) f_N(y_i; \mu_2, \sigma^2_2),$$

where $f_N(y; \mu, \sigma^2)$ is the density of the $N(\mu, \sigma^2)$-distribution evaluated at $y$. Like for a univariate finite mixture distribution (Kiefer and Wolfowitz, 1956), the likelihood function is unbounded at $\mu_2 = y_i$ and $\sigma^2_2 = 0$. With $c_1 = \eta_1 f_N(y_i; \mu_1, \sigma^2_1)$,

$$p(y_i|\mu_2 = y_i, \sigma^2_2) = c_1 + \frac{1 - \eta_1}{\sqrt{2\pi \sigma^2_2}} \Rightarrow \lim_{\sigma^2_2 \to 0} p(y_1, \ldots, y_N|\mu_2 = y_i, \sigma^2_2) = \infty.$$

Hence, ML estimation based on maximizing the likelihood function $p(y_1, \ldots, y_N|\mu_2, \sigma^2_2)$ is likely to fail for poor starting values.

The Bayesian approach combines the likelihood with the “regularization” prior $1/\sigma^2_2 \sim G(c_0, C_0)$ with $C_0 > 0$ which yields:

$$p(y_i|\mu_2 = y_i, \sigma^2_2) p(\sigma^2_2) \propto \left( c_1 + \frac{1 - \eta_1}{\sqrt{2\pi \sigma^2_2}} \right) \left( \frac{1}{\sigma^2_2} \right)^{c_0+1} \exp \left( -\frac{C_0}{\sigma^2_2} \right).$$

Evidently, the prior penalizes the likelihood as $\sigma^2_2 \to 0$, therefore the pole at $\sigma^2_2 = 0$ disappears:

$$\lim_{\sigma^2_2 \to 0} p(y_1, \ldots, y_N|\mu_2 = y_i, \sigma^2_2) = 0.$$

As a result, the posterior distribution is much more well-behaved than the likelihood function close to $\sigma^2_2 = 0$.

Next, consider clustering discrete-valued time series which are expected to be persistent using the simple Markov chain clustering kernel (6). For persistent time series, it is likely that certain transitions are never observed, in particular, if the observed time series are rather short. However, the EM algorithm used for maximizing the likelihood function breaks down, if no transitions starting from $j$ are observed in the entire cluster $h$, because the complete data likelihood function $p(y|\xi_1, \ldots, \xi_H, S)$ is independent of the $j$th row $\xi_{h,j}$ of $\xi_h$ and no estimator for $\xi_{h,j}$ 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, 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 $Pr(S_t = h|y_1, \eta, \hat{\xi}_1, \ldots, \hat{\xi}_H)$ for all observations in all groups in the subsequent E-step. To avoid these problems, it has been suggested to add a small constant, e.g. $c_{0,jk} = 0.5$, to the number of observed transitions (Agresti, 1990). However, this is equivalent to using the Dirichlet prior $D(c_{0,j1}, \ldots, c_{0,jK})$ for each row $\xi_{h,j}$ within a Bayesian approach.
3.2 Markov chain Monte Carlo

Practical Bayesian inference is typically carried out by sampling, for a fixed value \(H\) of cluster, \(M\) draws from the posterior distribution \(p(\theta_H | y)\) using MCMC methods. For model based clustering, simple Markov chain Monte Carlo sampling in combination with data augmentation may be applied. Based on some initial classification \(S\), the following steps are repeated during a burn-in period to achieve convergence and additional \(M\) iterations are performed to produce the desired number of draws:

(a) Sample the component parameters \(\vartheta_1, \ldots, \vartheta_H\): draw \(\vartheta_h\) from \(p(\vartheta_h | S, y) \propto p(y_i | \vartheta_h) p(\vartheta_h)\) for each \(h = 1, \ldots, H\).

(b) Sample the unknown parameters \(\eta\) in the prior allocation distribution from \(p(\eta | S) \propto \prod_{i=1}^{N} p(S_i | \eta)p(\eta)\).

(c) Bayes’ classification for each subject \(i\): draw \(S_i, i = 1, \ldots, N\), from the discrete probability distribution \(Pr(S_i = h | y_i, \theta_H)\) given in (9).

Step (c) is easily implemented for arbitrary clustering kernels, and is typically the most time-consuming step of the sampler. For random-effects clustering kernels, the marginal clustering kernel where the random effects are integrated out is used, if available.

Closed form Gibbs sampling of \(\vartheta_h\) in Step (a) is possible, whenever the prior \(p(\vartheta_h)\) is conditionally conjugate to the clustering kernel \(p(y_i | \vartheta_h)\). For the dynamic regression clustering kernel (2), for instance, the posterior is a normal distribution for \(\beta_h\) and an inverted Gamma distribution for \(\sigma^2_h\) (Fruhwirth-Schnatter and Kaufmann, 2008; Juarez and Steel, 2010). For the Markov chain clustering kernel (6), the posterior for \(\xi_h\) is a Dirichlet distribution (Pamminger and Fruhwirth-Schnatter, 2010). Under lack of a closed form posterior, a Metropolis-Hasting step has to be implemented to sample \(\vartheta_h\) as, for instance, for Dirichlet multinomial clustering (Pamminger and Fruhwirth-Schnatter, 2010) and for the probit kernel (4) (Assmann and Boysen-Hogrefe, 2011). For further details, the reader is referred to the relevant papers.

Under a homogenous prior allocation distribution, \(\eta = (\eta_1, \ldots, \eta_H)\) is sampled in Step (b) from the Dirichlet posterior distribution \(D(\alpha_1, \ldots, \alpha_H)\) where \(\alpha_h = \# \{S_i = h\} + \alpha_0\). For a mixture-of-expert models, the regression coefficients \(\eta = (\gamma_2, \ldots, \gamma_H)\) are sampled conditional on \(S\) from the posterior distribution \(p(\gamma_2, \ldots, \gamma_H | S)\), where the likelihood \(p(S_i | \gamma_2, \ldots, \gamma_H)\) is obtained from the MNL model (11). To sample \(\gamma_2, \ldots, \gamma_H\), Fruhwirth-Schnatter et al. (2011) recommend to apply auxiliary mixture sampling in the differenced random utility model representation of the MNL model (Fruhwirth-Schnatter and Fruhwirth, 2010), because this method is superior to other MCMC methods in terms of the effective sampling rate.

3.3 Label Switching and Post-Processing MCMC

As for any finite mixture model, label switching may occur during MCMC sampling (Fruhwirth-Schnatter, 2006, Section 3.5). Many useful methods have been developed to force a unique labeling on the posterior draws from the posterior distribution, see Sperrin et al. (2010) for a recent review.

Based on Fruhwirth-Schnatter (2006, p.97), Fruhwirth-Schnatter (2011) suggested recently a simple relabeling technique which is based on k-means clustering in the point process representation of the MCMC draws and operates in the parameter space \(\Theta\) or an even smaller subspace \(\tilde{\Theta} \subset \Theta\) rather than in the high-dimensional space \(\Theta^K\), as most papers do. To identify a mixture model with \(H\) components, k-means clustering with \(H\) clusters is applied to all \(MH\) posterior...
draws of a sub-vector $\tilde{\vartheta}_h \in \tilde{\Theta}$ of the group-specific parameter $\vartheta_h$. Provided that the mixture model is not overfitting, $H$ clusters are present in these $MH$ posterior draws and MCMC draws belonging to the same group will be assigned to the same cluster during k-means clustering. Hence the classification sequences resulting from k-means clustering show how to rearrange the group specific parameters, even if label switching occurred during sampling. Details are provided in Appendix A. This method has been successfully applied in several papers on model-based clustering (Pamminger and Frühwirth-Schnatter, 2010; Frühwirth-Schnatter et al., 2011) and is illustrated further in Subsection 5.2.

All classification sequences derived by k-means clustering are expected to be permutations of $\{1, \ldots, H\}$, if the mixture is not overfitting the number $H$ of clusters. If a small fraction of non-permutations is present, then only the subsequence of the $M_0$ identified draws is used for posterior inference known to be sensitive to label switching like parameter estimation and classification. Like any other relabeling method, labeling through k-means clustering is sensitive to overfitting the number of clusters. A high fraction of non-permutations should be taken as a serious warning that the selected number $H$ of clusters has been a poor choice and the mixture is overfitting the number of components, see also Frühwirth-Schnatter (2011).

3.4 Bayesian heuristic search in the partition space

Bayesian approaches which are not based on MCMC methods have been considered by several authors for specific clustering kernels, see e.g. Ramoni et al. (2002a, 2002b), and Heard et al. (2006). These methods operate in the partition space characterized by the allocations $S$ and, as opposed to standard MCMC methods, determine $H$ simultaneously with $S$.

These methods can be extended to arbitrary clustering kernel, provided that a conjugate prior exists for the cluster-specific parameter $\vartheta_h$ and a homogeneous prior allocation distribution is chosen. In this case, it is possible to work out the marginal likelihood function $p(y|S)$ where $\vartheta_1, \ldots, \vartheta_H$ are integrated out as well as the marginal prior distribution $p(S)$ where $\eta$ is integrated out. This allows to express $p(S|y)$ in terms of $p(S)$ times the product of $H$ cluster-specific marginal likelihoods which are easily determined from a few subject-specific sufficient statistics.

A heuristic search method is applied for finding a good partition $S$ that maximized $p(S|y)$. The search procedure starts from some initial partition, e.g. assigning each time series to its own cluster, and applies a bottom-up strategy by iteratively merging clusters according to a deterministic rule, until only a single cluster is left. To each $H$ corresponds an “optimal” partition $S_H$ and the optimal number of clusters is defined by the partition $S_H$ that maximizes $p(S|y)$.

4 Selecting the Number of Clusters

Despite much research effort, it is still an open issue how to select the number $H$ of clusters in an optimal manner, when model-based clustering is applied to real data. This section reviews the most interesting approaches that have been put forward in the past decades.

Subsequently, $d_H$ denotes the number of distinct unknown elements of $\theta_H$. $\hat{\theta}_H$ denotes the ML estimator maximizing the observed-data likelihood function (12), while $(\hat{\theta}_H^C, \hat{S})$ maximizes the complete-data likelihood function $p(y, S|\theta_H)$, given by:

$$\log p(y, S|\theta_H) = \sum_{i=1}^{N} \log (\Pr(S_i = h|\eta)p(y_i|\theta_{S_i})).$$ (13)
These estimators are not directly available from MCMC estimation and are approximated by the MCMC posterior draw maximizing, respectively, the observed-data and the complete data log-likelihood function. Since this introduces a certain random element into computing model choice criteria, several independent MCMC runs should be performed in order to study the effect on model selection. Pamminger and Frühwirth-Schnatter (2010) found that the variance of a particular model selection criterion across independent MCMC runs did not effect the final choice of $H$. Given independent MCMC runs, the optimal estimator across all runs is selected for final model choice.

4.1 AIC and BIC

The difficulties with identifying a sensible number of clusters are particularly well-documented for the BIC criterion (Schwarz, 1978) where $H$ is selected as to minimize $\text{BIC}(H)$ defined by:

$$\text{BIC}(H) = -2 \log p(y|\hat{\theta}_H) + d_H \log n. \quad (14)$$

BIC is known to be consistent for the number of components in a finite mixture, if the component density is correctly specified (Keribin, 2000), however, it tends to choose models with too few components in small data sets (Biernacki et al., 2000). Other simulation studies reported in Biernacki et al. (2000) and McLachlan and Peel (2000, Section 6.11) show that BIC will overrate the number of components in the mixture model under misspecification of the component density. Hence, BIC overfits the number of clusters in particular when clustering large panels (Pamminger and Frühwirth-Schnatter, 2010; Frühwirth-Schnatter et al., 2011), because it is unlikely that the clustering kernel is correctly specified for all time series in the panel.

A further difficulty with BIC is that the appropriate sample size $n$ is not obvious in a time series context. As each time series is modeled independently, the number time series seems to be natural, i.e. $n = N$. On the other hand, since multiple observations are available for each time series, one may use the total number of observations, i.e. $n = \sum_{i=1}^N T_i$ (Owen et al., 2009).

The AIC criterion (Akaike, 1974) defined by

$$\text{AIC}(H) = -2 \log p(y|\hat{\theta}_H) + 2 d_H, \quad (15)$$

is independent of the sample size. However, AIC is known to be inconsistent and overfits the number of clusters $H$ also in a time series context (Pamminger and Frühwirth-Schnatter, 2010; Frühwirth-Schnatter et al., 2011).

4.2 Marginal likelihoods

Several authors follow the traditional Bayesian approach to select $H$ as to maximize the posterior distribution $p(H|y) \propto p(y|H)p(H)$ where the marginal likelihood $p(y|H) = \int p(y|\theta_H)p(\theta_H)d\theta_H$ is obtained by integrating the observed-data likelihood with respect to the prior.

Computing the marginal likelihood for a finite mixture model is quite a numerical challenge (Frühwirth-Schnatter, 2004) as are model space MCMC methods that sample $H$ jointly with $\theta_H$ (Frühwirth-Schnatter, 2006, Chapter 4 and 5). Nevertheless, this approach often led to a sensible choice of $H$ in practical panel analysis (Frühwirth-Schnatter and Kaufmann, 2008; Juárez and Steel, 2010; Aßmann and Boysen-Hogrefe, 2011).

However, like $\text{BIC}(H)$ which is an asymptotic approximation to $-2 \log p(y|H)$ (Kass and Raftery, 1995), it may not be adequate to select $H$ as to maximize $p(y|H)$ or $p(H|y)$. Simulation studies show that, even if the component densities are correctly specified, the marginal
likelihood underrates $H$, if the heterogeneity across the clusters is small compared to within-cluster heterogeneity (Frühwirth-Schnatter and Kaufmann, 2008). On the other hand, if the component densities are misspecified, $H$ tends be overrated (Frühwirth-Schnatter and Pyne, 2010). The marginal likelihood selected too many clusters, for instance, when clustering kernels based on the normal distribution are applied to data with skew component-specific distributions (Juárez and Steel, 2010).

Finally, the marginal likelihood might strongly depend on the particular choice of the prior distribution, see Frühwirth-Schnatter and Pyne (2010) among many others, and some sensitivity analysis has to be carried in order to detect spurious over- and underfitting.

4.3 The AWE Criterion

A criterion often is able to identify the correct number of clusters even when the component densities are misspecified is the approximate weight of evidence AWE criterion (Banfield and Raftery, 1993). Biernacki and Govaert (1997) expressed $\text{AWE}(H)$ as a criterion which penalizes the complete data log-likelihood function with model complexity, i.e.

$$
\text{AWE}(H) = -2 \log p(y, \hat{S} | \hat{\theta}_H) + 2 d_H \left( \frac{3}{2} + \log n \right).
$$

As for BIC, the sample size $n$ is not well-defined. Nevertheless, in various practical applications of model-based clustering, such as clustering genetic data using skew-normal and skew-t distributions (Frühwirth-Schnatter and Pyne, 2010) or clustering wage careers in the Austrian labor market using Markov chain clustering (Pamminger and Frühwirth-Schnatter, 2010; Frühwirth-Schnatter et al., 2011), the number of clusters suggested by the AWE criterion turned out to be sensible from, respectively, a genetic or economic point of view.

4.4 Entropy based criteria

None of the criteria discussed so far 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. To address this issue, various criteria have been suggested that involve the quality of the resulting partition which is measured through the entropy $\text{EN}(\theta_H)$ defined by

$$
\text{EN}(\theta_H) = -\sum_{h=1}^{H} \sum_{i=1}^{N} \text{Pr}(S_i = h | y_i, \theta_H) \log t_{ih}(\theta_H) \geq 0,
$$

where $\text{Pr}(S_i = h | y_i, \theta_H)$ is the posterior classification probability defined in (9). The entropy is close to 0 if the resulting clusters are well-separated and increases with increasing overlap of the mixture components.

The classification likelihood criterion CLC (Biernacki and Govaert, 1997) penalizes the log likelihood function by the entropy rather than by model complexity, however, this criterion seems to work well only for well-separated clusters with a fixed weight distribution.

The ICL-BIC criterion (McLachlan and Peel, 2000) penalizes the log likelihood function both by model complexity and the entropy, i.e.

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

Simulation studies indicate that ICL-BIC may identify the correct number of clusters for (multivariate) continuous data even under a misspecified multivariate normal clustering kernel (McLachlan and Peel, 2000, Section 6.11), but it tends to fail for discrete-valued data, even if the true
model is used as clustering kernel (Biernacki et al., 2010). Biernacki et al. (2010) recommend to use the exact integrated classification likelihood (ICL) for discrete-valued data which is defined as \( \text{ICL}(H) = \int p(y, \hat{S}|\theta_H)p(\theta_H|y)d\theta_H \), where \( \hat{S} \) is a posterior estimator of the unknown allocations \( S \). While this criterion showed good performance for latent class models (Biernacki et al., 2010), it still underrated the number of cluster dramatically compared to AWE for Markov chain clustering (Frühwirth-Schnatter et al., 2011).

4.5 DIC

An alternative popular criterion for Bayesian model selection which is easily computed from the MCMC draws is the deviance information criterion DIC (Spiegelhalter et al., 2002). However, as discussed by Celeux et al. (2006), the application of DIC to finite mixture models is not without problems. A first difficulty is the choice of the appropriate likelihood function which could either be the observed-data likelihood function \( p(y|\theta_H) \), the complete-data likelihood function \( p(y, S|\theta_H) \), or the conditional likelihood \( p(y|S, \theta_H) \). Second, the calculation of DIC requires an estimator of the unknown parameter \( \theta_H \) which may suffer from label switching as discussed above, making DIC unstable. Finally, DIC involving the complete-data or the conditional likelihood function requires some way of handling the problem that \( S \) is unobserved, either by integrating with respect to the posterior \( p(S|y, H) \) or by using an estimator \( \hat{S} \) of \( S \).

In reaction to these difficulties, Celeux et al. (2006) investigate in total eight different DIC criteria. DIC\(_2\), for instance, focuses on the observed-data likelihood and considers the allocations as nuisance parameters:

\[
\text{DIC}_2(H) = 2 \log p(y|\hat{\theta}_H^M) - 4E_{\theta_H}(\log p(y|\theta_H)|y),
\]

where the posterior mode estimator \( \hat{\theta}_H^M \) which is invariant to label switching is obtained from the observed-data posterior \( p(\theta_H|y) \).

Based on several simulation studies, Celeux et al. (2006) recommend using DIC\(_4\) which is based on computing first DIC for the complete-data likelihood function and then integrating with respect to the posterior \( p(S|y, H) \). The application of this criterion requires the computation of the complete-data estimator \( E(\theta_H|y, S) \) for each draw from the posterior \( p(S|y, H) \) which is straightforward only for mixture models, where the complete-data posterior \( p(\theta_H|y, S) \) is of closed form.

Celeux et al. (2006) show that by substituting \( E(\theta_H|y, S) \) by the posterior mode estimator \( \hat{\theta}_H^M \), a criterion is obtained which penalizes DIC\(_2\) by the expected entropy:

\[
\text{DIC}_{4a}(H) = \text{DIC}_2(H) + 2E_{\theta_H}(\text{EN}(\theta_H)|y).
\]

Both DIC\(_2\) as well as DIC\(_{4a}\) are easily estimated by substituting all expectations \( E_{\theta_H}(\cdot|y) \) by the average over the MCMC draws from the posterior \( p(\theta_H|y) \). Label switching is not a problem here, because both \( \log p(y|\theta_H) \) as well as \( \text{EN}(\theta_H) \) are invariant to changing the labeling of the groups.

However, despite the easy with which DIC criteria may be computed, they do not appear to be reliable criteria to select the number of clusters \( H \). A recent application of model-based clustering using multivariate skew-normal and skew-t mixtures (Frühwirth-Schnatter and Pyne, 2010) revealed extreme sensitivity of DIC with respect to the prior distribution which was even higher than the well-known sensitivity of the marginal likelihood.
5 An Illustrative Application

5.1 Clustering (dynamic) discrete-valued panels using generalized transition models

In this section, clustering of dynamic discrete-valued time series using the dynamic MNL model (7) is discussed for cases where covariate information \( x_{it} \) is available, but the potential values of \( x_{it} \) are limited.

Assume, for instance, that \( x_{it} \) only contains a single dummy variable like the gender \( g_i \) of subject \( i \), apart from the constant. If all possible combinations of the immediate history \( y_{i,t-1} \) and the gender \( g_i \) are indexed by \( j = 1, \ldots, J \), then the dynamic MNL model (7) reduces to a generalized transition matrix \( \xi_h \), with the \( j \)th row describing the conditional distribution of \( y_{it} \), given a state of the history \( H_{it} = (y_{i,t-1}, g_i) \) equals \( j \):

\[
\xi_{h,j} = \Pr(y_{it} = k | H_{it} = j, S_i = h), \quad j = 1, \ldots, J.
\]

If no covariate information is available, then \( H_{it} = (y_{i,t-1}) \) and (18) reduces to Markov chain clustering as discussed in Subsection 2.2.3.

This clustering kernel is modified in an obvious manner if the covariate \( x_{it} \) is multidimensional and assumes more than two different design points. It may be applied to clustering panels without dynamic by choosing \( H_{it} = x_{it} \), to capture higher order dependence by choosing \( H_{it} = (y_{i,t-2}, y_{i,t-1}) \), or to define time-inhomogeneous Markov chains by choosing \( H_{it} = (y_{i,t-1}, t) \).

Given a specific set of \( J \) histories \( H_{it} \), the clustering kernel for the whole time series reads:

\[
p(y_i | \xi_h) = \prod_{j=1}^{J} \prod_{k=1}^{K} \xi_{N_{i,jk}},
\]

where, for each time series \( i \), \( N_{i,jk} = \# \{ y_{it} = k, H_{it} = j, t = 1, \ldots, T_i \} \) is the number of transitions into state \( k \) given a history of type \( j \). Note that (19) is formulated conditional on the first observation \( y_{i0} \).

To perform Bayesian inference, the following prior choices are made. The rows \( \xi_{h,j} \) of the generalized transition matrix \( \xi_h \) follow the Dirichlet distribution

\[
\xi_{h,j} \sim D (e_{0,j1}, \ldots, e_{0,jK}), \quad e_{0,jk} = \max(N_0 \hat{\xi}_{jk}, 0.5),
\]

where a lower bound of 0.5 is introduced to avoid problems with empty cells. The prior is centered around the empirical estimator \( \hat{\xi} \) of \( \xi \) under the assumption of homogeneity in the transition behavior, i.e.

\[
\hat{\xi}_{jk} = \frac{N_{jk}}{N_j},
\]

where \( N_{jk} = \sum_{i=1}^{N} N_{i,jk} \) is the total number of transitions into state \( k \) given a history of type \( j \) and \( N_j = \sum_{k=1}^{K} N_{jk} \) is the total number of observations with history \( j \). The hyperparameter \( N_0 \) captures the amount of shrinkage of \( \xi_h \) toward \( \hat{\xi} \). Evidently, the larger \( N_0 \) the stronger is shrinkage.

MCMC estimation as described in Subsection 3.2 is easily applied, as the Dirichlet prior (20) is conditionally conjugate to the clustering kernel (19). Therefore, the various rows \( \xi_{h,j} \) of \( \xi_h \) are conditionally independent a posteriori, given \( S \) and \( y \), and are be sampled row-by-row from a total of \( JH \) Dirichlet distributions:

\[
\xi_{h,j} | S, y \sim D \left( e_{0,j1} + N_{j1}^h, \ldots, e_{0,jK} + N_{jK}^h \right), \quad j = 1, \ldots, J, \ h = 1, \ldots, H,
\]

17
where \( N_{jk}^h = \sum_{i: S_i = h} N_{i,jk} \) is the total number of transitions into state \( k \) observed in group \( h \) given a history of type \( j \) and is determined from the transitions \( N_{i,jk} \) for all individuals falling into that particular group.

5.2 Application to the Marijuana Data

Several sets of histories are considered to cluster the marijuana data introduced in Section 1 using the clustering kernels introduced in the previous subsection. First, it is assumed that the distribution of \( y_{it} \) depends on the gender and changes every year, i.e. \( M_1: \mathcal{H}_{it} = (y_{i,t-1}, t) \). This is the only latent class clustering kernel to be considered.

All remaining clustering kernels capture the dependence between subsequent observations by including \( y_{i,t-1} \) into the history \( \mathcal{H}_{it} \), such as Markov chain clustering, i.e. \( M_2: \mathcal{H}_{it} = (y_{i,t-1}) \). More complex clustering kernels assume that the Markov chain used for clustering is inhomogeneous over time, i.e. \( M_3: \mathcal{H}_{it} = (y_{i,t-1}, t) \), inhomogeneous over units, i.e. \( M_4: \mathcal{H}_{it} = (y_{i,t-1}, g_i) \), or both, i.e. \( M_5: \mathcal{H}_{it} = (y_{i,t-1}, t, g_i) \).

These clustering kernels are combined for increasing values of \( H \) with homogeneous random prior assignment and stratified random prior assignment, using gender as a covariate for stratification. However, it is not entirely clear, if all those models are identified, see Grün and Leisch (2008) for identifiability problems with mixtures of multinomial distributions.
Table 1: Marijuana data; various model choice criteria applied to five different clustering kernels in combination with homogeneous random prior assignment; log \( p(y) \), DIC \(_2\), and DIC \(_{4a}\) are based on uniform priors on the weight distribution \((\eta_1, \ldots, \eta_H)\) and on all rows of the generalized transition matrix \(\xi_{h,j}\).

<table>
<thead>
<tr>
<th>(H)</th>
<th>(d_H)</th>
<th>(\log p(y))</th>
<th>AIC</th>
<th>BIC</th>
<th>ICL-BIC</th>
<th>DIC (_2)</th>
<th>DIC (_{4a})</th>
<th>AWE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16</td>
<td>-789.0</td>
<td>1541.0</td>
<td>1596.5</td>
<td>1596.5</td>
<td>1536.1</td>
<td>1536.1</td>
<td>1732.0</td>
</tr>
<tr>
<td>2</td>
<td>33</td>
<td>-696.1</td>
<td>1309.9</td>
<td><strong>1424.3</strong></td>
<td><strong>1460.9</strong></td>
<td>1294.9</td>
<td><strong>1339.7</strong></td>
<td><strong>1728.9</strong></td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>-691.8</td>
<td>1284.1</td>
<td>1457.5</td>
<td>1523.3</td>
<td><strong>1267.6</strong></td>
<td>1342.9</td>
<td>1911.2</td>
</tr>
</tbody>
</table>

Clustering kernel \(\mathcal{M}_1: \mathcal{H}_{it} = (t, g_i)\)

<table>
<thead>
<tr>
<th>(H)</th>
<th>(d_H)</th>
<th>(\log p(y))</th>
<th>AIC</th>
<th>BIC</th>
<th>ICL-BIC</th>
<th>DIC (_2)</th>
<th>DIC (_{4a})</th>
<th>AWE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>-607.4</td>
<td>1196.6</td>
<td><strong>1217.4</strong></td>
<td><strong>1217.4</strong></td>
<td>1196.2</td>
<td><strong>1196.2</strong></td>
<td><strong>1268.2</strong></td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>-611.0</td>
<td><strong>1195.4</strong></td>
<td>1240.5</td>
<td>1440.2</td>
<td>1198.6</td>
<td>1336.3</td>
<td>1366.4</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>-614.7</td>
<td>1204.9</td>
<td>1274.3</td>
<td>1531.7</td>
<td>1198.0</td>
<td>1438.4</td>
<td>1471.6</td>
</tr>
</tbody>
</table>

Clustering kernel \(\mathcal{M}_2: \mathcal{H}_{it} = (y_{i,t-1})\)

<table>
<thead>
<tr>
<th>(H)</th>
<th>(d_H)</th>
<th>(\log p(y))</th>
<th>AIC</th>
<th>BIC</th>
<th>ICL-BIC</th>
<th>DIC (_2)</th>
<th>DIC (_{4a})</th>
<th>AWE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24</td>
<td>-621.3</td>
<td><strong>1209.1</strong></td>
<td><strong>1292.3</strong></td>
<td><strong>1292.3</strong></td>
<td>1196.2</td>
<td><strong>1196.2</strong></td>
<td><strong>1495.6</strong></td>
</tr>
<tr>
<td>2</td>
<td>49</td>
<td>-636.6</td>
<td>1223.7</td>
<td>1393.7</td>
<td>1531.7</td>
<td>1198.0</td>
<td>1438.4</td>
<td>1471.6</td>
</tr>
<tr>
<td>3</td>
<td>74</td>
<td>-646.8</td>
<td>1273.8</td>
<td>1530.4</td>
<td>1616.0</td>
<td><strong>1194.1</strong></td>
<td>1291.2</td>
<td>1220.3</td>
</tr>
</tbody>
</table>

Clustering kernel \(\mathcal{M}_3: \mathcal{H}_{it} = (y_{i,t-1}, t)\)

<table>
<thead>
<tr>
<th>(H)</th>
<th>(d_H)</th>
<th>(\log p(y))</th>
<th>AIC</th>
<th>BIC</th>
<th>ICL-BIC</th>
<th>DIC (_2)</th>
<th>DIC (_{4a})</th>
<th>AWE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>-613.8</td>
<td><strong>1200.4</strong></td>
<td><strong>1242.0</strong></td>
<td><strong>1242.0</strong></td>
<td>1197.4</td>
<td><strong>1197.4</strong></td>
<td><strong>1343.6</strong></td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>-622.3</td>
<td>1207.4</td>
<td>1294.1</td>
<td>1463.0</td>
<td>1200.9</td>
<td>1299.2</td>
<td>1526.2</td>
</tr>
<tr>
<td>3</td>
<td>38</td>
<td>-628.3</td>
<td>1228.6</td>
<td>1360.4</td>
<td>1526.0</td>
<td>1202.3</td>
<td>1361.9</td>
<td>1713.9</td>
</tr>
</tbody>
</table>

Clustering kernel \(\mathcal{M}_4: \mathcal{H}_{it} = (y_{i,t-1}, g_i)\)

<table>
<thead>
<tr>
<th>(H)</th>
<th>(d_H)</th>
<th>(\log p(y))</th>
<th>AIC</th>
<th>BIC</th>
<th>ICL-BIC</th>
<th>DIC (_2)</th>
<th>DIC (_{4a})</th>
<th>AWE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>48</td>
<td>-637.8</td>
<td><strong>1239.5</strong></td>
<td><strong>1405.9</strong></td>
<td><strong>1405.9</strong></td>
<td>1200.6</td>
<td><strong>1200.6</strong></td>
<td><strong>1812.4</strong></td>
</tr>
<tr>
<td>2</td>
<td>97</td>
<td>-670.2</td>
<td>1306.1</td>
<td>1642.5</td>
<td>1687.0</td>
<td>1199.9</td>
<td>1249.0</td>
<td><strong>1266.5</strong></td>
</tr>
<tr>
<td>3</td>
<td>146</td>
<td>-689.7</td>
<td>1403.4</td>
<td>1909.8</td>
<td>1969.7</td>
<td>1195.7</td>
<td>1265.8</td>
<td>1308.0</td>
</tr>
</tbody>
</table>

5.2.1 Deterministic prior stratification

The clustering kernels \(\mathcal{M}_4: \mathcal{H}_{it} = (y_{i,t-1}, g_i)\) and \(\mathcal{M}_5: \mathcal{H}_{it} = (y_{i,t-1}, t, g_i)\) in combination with \(H = 1\) correspond to deterministic prior stratification of the teenagers into two groups according to gender. Model \(\mathcal{M}_4\) assumes a homogeneous Markov chain within each group, while \(\mathcal{M}_5\) assumes a time-inhomogeneous Markov chain. The prior follows the Dirichlet distribution (20) with \(N_0 = 2.5\), in which case the posterior distribution is available in closed form, which each row of the gender specific transition matrices following a Dirichlet distribution.

For illustration, Figures 2 displays, both for male and female teenagers, the posterior distributions of several transition probabilities as a function of time, namely the probability \(\xi_{t,11}\) that a non-user in year \(t - 1\) is still a non-user in year \(t\), the probability \(\xi_{t,12}\) that a non-user in year \(t - 1\) became a light user in year \(t\), and the probability \(\xi_{t,23}\) that a light user in year \(t - 1\) became a heavy user in year \(t\).

From these figures we would conclude that female teenagers have a high probability to remain in the no-user category, while male teenagers have quite a high risk to move out of the no-user category. The subsequent investigations show that teenagers may, indeed, be clustered into these two groups with respect to marijuana use, however, unobserved factors, not the gender, are relevant whether a teenager belongs to one group or the other.
5.2.2 Choosing the clustering kernel and the number of clusters

All model selection criteria reviewed in Section 4 are applied to compare the various clustering kernels and to choose the number $H$ of clusters for a specific clustering kernel. For each clustering kernel and each number of clusters, $M = 10000$ MCMC draws were generated after a burn-in of 5000 draws.

The corresponding criteria are reported in Table 1. While AIC, BIC, ICL-BIC, and AWE are independent of the prior, the marginal likelihood $\log p(y)$, DIC$_2$, and DIC$_{4a}$ are computed under assuming uniform priors for the weight distribution ($\eta_1, \ldots, \eta_H$) and for all rows $\xi_{h,j}$ of the generalized transition matrix in all groups in which case the posterior distribution is nothing but the integrated observed-data likelihood function. For $H = 1$, the marginal likelihood is computed explicitly:

$$p(y|M, H = 1) = \prod_{j=1}^{J} \frac{\Gamma(\sum_{k=1}^{K} e_{0,jk}) \Gamma(e_{0,jk} + N_{jk})}{\Gamma(e_{0,jk})},$$

where $N_{jk}$ and $N_j$ have been defined after (21). For $H > 1$, bridge sampling is used (Frühwirth-Schnatter, 2004).

All criteria lead to a clear decision in favor of dynamic clustering kernels as opposed to the latent class clustering kernel, hence there evidently exists a persistent behavior in the use of marijuana among teenagers. However, quite a discrepancy may be observed across the various criteria concerning the “optimal” number of clusters for each dynamic clustering kernel as well as the best dynamic clustering kernel. AIC prefers simple Markov chain clustering with two clusters. For all other dynamic clustering kernels, only one cluster is selected by AIC. DIC$_2$ prefers the time-inhomogeneous Markov chain as clustering kernel in combination with $H = 3$. Interestingly, DIC$_2$ selects three clusters for any clustering kernels involving $t$, but only one cluster for the remaining clustering kernels. AWE prefers the most complex clustering kernel $M_5$ in combination with $H = 2$.

The remaining four criteria, namely $\log p(y)$, BIC and ICL-BIC and DIC$_{4a}$ decide for homogeneity for all dynamic clustering kernels. $\log p(y)$, BIC and ICL-BIC prefer a single homogeneous Markov chain model to more complex models, while DIC$_{4a}$ is undecided between a homogeneous and time-inhomogeneous Markov chain. In summary, most of these criteria do not indicate any kind of unobserved heterogeneity.

Concerning the marginal likelihood, this underfitting could be caused priors which are too uninformative. Hence, sensitivity of model selection with respect to the prior is considered in Table 2 for all dynamic clustering kernels. The weight distribution ($\eta_1, \ldots, \eta_H$) follows apriori the $D(\alpha_0, \ldots, \alpha_0)$-distribution with $\alpha_0 = (K - 1)/2 + 4$ (Rousséau and Mengersen, 2010). The prior for $\xi_{h,j}$ is selected as in (20) with the shrinkage factor $N_0$ being varied over the values $\{1, 2, 2.5, 3, 4\}$. Increasing $N_0$ increases the marginal likelihood for a homogeneous model and decreases the marginal likelihood for any model with $H > 1$ clusters. Irrespective of the underlying dynamic clustering kernel, three clusters are selected for $N_0 = 1$, whereas a single cluster is selected for $N_0 = 4$. Contrary to the uniform prior, which preferred the most simple clustering kernel $M_2$ with $H = 1$, the shrinkage prior in combination with a small value of $N_0$ prefers the most complex clustering kernel $M_5$ in combination with $H = 3$.

However, it turns out that all models with $H = 3$ are overfitting. For illustration, Table 2 also reports the fraction $M_0/M$ of posteros draws that could be identified using k-means clustering as described in Subsection 3.3, where all persistence probabilities were used for k-means clustering. We observe quite a high fraction $M_0/M$ for $H = 2$, whereas $M_0/M$ is much smaller than one for $H = 3$. Frühwirth-Schnatter (2011) found that a fraction $M_0/M$ of identifiable
Table 2: Marijuana data; log marginal likelihood $\log p(y)$ and fraction $M_0/M$ for four different dynamic clustering kernels in combination with homogeneous random prior assignment; $D(\alpha_0, \ldots, \alpha_0)$-prior on the weight distribution $\eta = (\eta_1, \ldots, \eta_H)$ with $\alpha_0 = (K-1)J/2+4$, priors on the rows $\xi_{h,j}$ of the generalized transition matrix equals (20) with various values of $N_0$.

<table>
<thead>
<tr>
<th>Clustering kernel $M_2$: $\mathcal{H}<em>{it} = (y</em>{i,t-1})$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
</tr>
<tr>
<td>1 -607.4</td>
<td>-606.0</td>
<td>-605.5</td>
<td>-605.1</td>
<td>-604.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2 -596.0</td>
<td>-598.2</td>
<td>-600.0</td>
<td>-601.7</td>
<td>-605.7</td>
<td>0.91</td>
<td>0.93</td>
<td>0.89</td>
<td>0.87</td>
</tr>
<tr>
<td>3 -593.5</td>
<td>-597.2</td>
<td>-600.3</td>
<td>-602.6</td>
<td>-608.3</td>
<td>0.32</td>
<td>0.41</td>
<td>0.43</td>
<td>0.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clustering kernel $M_3$: $\mathcal{H}<em>{it} = (y</em>{i,t-1}, t)$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
</tr>
<tr>
<td>1 -621.2</td>
<td>-615.5</td>
<td>-613.7</td>
<td>-612.0</td>
<td>-609.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2 -581.4</td>
<td>-591.6</td>
<td>-596.5</td>
<td>-606.3</td>
<td>-623.3</td>
<td>0.98</td>
<td>0.93</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>3 -576.5</td>
<td>-590.1</td>
<td>-599.4</td>
<td>-611.7</td>
<td>-637.3</td>
<td>0.47</td>
<td>0.50</td>
<td>0.51</td>
<td>0.51</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clustering kernel $M_4$: $\mathcal{H}<em>{it} = (y</em>{i,t-1}, g_i)$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
</tr>
<tr>
<td>1 -613.8</td>
<td>-611.0</td>
<td>-610.0</td>
<td>-609.2</td>
<td>-607.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2 -593.3</td>
<td>-598.2</td>
<td>-601.3</td>
<td>-604.8</td>
<td>-612.5</td>
<td>0.89</td>
<td>0.92</td>
<td>0.68</td>
<td>0.88</td>
</tr>
<tr>
<td>3 -590.1</td>
<td>-598.8</td>
<td>-603.6</td>
<td>-608.4</td>
<td>-621.0</td>
<td>0.48</td>
<td>0.41</td>
<td>0.24</td>
<td>0.41</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clustering kernel $M_5$: $\mathcal{H}<em>{it} = (y</em>{i,t-1}, t, g_i)$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
<td>$N_0$</td>
</tr>
<tr>
<td>1 -634.8</td>
<td>-623.7</td>
<td>-619.8</td>
<td>-616.6</td>
<td>-611.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2 -564.3</td>
<td>-587.5</td>
<td>-602.7</td>
<td>-618.4</td>
<td>-648.9</td>
<td>0.85</td>
<td>0.78</td>
<td>0.84</td>
<td>0.87</td>
</tr>
<tr>
<td>3 -543.4</td>
<td>-583.9</td>
<td>-601.1</td>
<td>-629.7</td>
<td>-675.9</td>
<td>0.40</td>
<td>0.39</td>
<td>0.44</td>
<td>0.43</td>
</tr>
</tbody>
</table>

Draws considerably smaller than one indicates that the mixture is overfitting the number of clusters $H$. On the other hand, if the selected number of components is smaller than the maximum number of components with a high fraction of identifiable draws, underfitting is present, because the prior is too diffuse and more shrinkage is needed.

Hence, we conclude that two clusters are present. If we rule out all priors that lead either to over- or underfitting for the majority of clustering kernels, we end up with choosing $N_0 = 2.5$ or $N_0 = 3$. Among these priors, the time-inhomogeneous clustering kernel $M_3$ is preferred.

5.2.3 Analyzing the two-cluster time-inhomogeneous Markov chain model

Model selection suggested to choose the two-cluster time-inhomogeneous Markov chain model. To reduce the simulation error, MCMC estimation was repeated with $M = 50,000$ draws after a burn-in of 10,000 draws under prior (20) where $N_0 = 2.5$. The mixture model was identified using the 12-dimensional subvector containing all persistence probabilities at all time points. $M_0 = 48,624$ among the $M = 50,000$ draws could be identified in this way and are used for further inference.

Table 3 reports the posterior means $E(\xi_{h,j} | y)$ and $E(\eta_{h} | y)$. The two groups are roughly of equal size, with the first group being slightly larger. A characteristic difference is evident for the two groups of teenagers. In group 1, never-users have a high probability to remain never-users throughout the whole observation period, whereas this probability is much smaller for the second group right from the beginning and drops to only 45% in the last year, see also Figure 3.
Table 3: Marijuana data; clustering kernel $\mathcal{M}_3$: $\mathcal{H}_{it} = (y_{i,t-1}, t)$ with $H = 2$ in combination with homogeneous random prior assignment; posterior mean $E(\xi_{h,\cdot}|y)$ arranged as transition matrices for each $t = 1, \ldots, 4$; the estimated group sizes $\hat{\eta}_h$ are equal to the posterior mean $E(\eta_h|y)$.

<table>
<thead>
<tr>
<th>$t$</th>
<th>Group 1 ($\hat{\eta}_1 = 0.564$)</th>
<th>Group 2 ($\hat{\eta}_2 = 0.436$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.928 0.038 0.034</td>
<td>0.760 0.211 0.029</td>
</tr>
<tr>
<td></td>
<td>0.498 0.167 0.335</td>
<td>0.342 0.146 0.513</td>
</tr>
<tr>
<td></td>
<td>0.224 0.181 0.596</td>
<td>0.175 0.231 0.595</td>
</tr>
<tr>
<td>2</td>
<td>0.894 0.087 0.019</td>
<td>0.695 0.242 0.063</td>
</tr>
<tr>
<td></td>
<td>0.099 0.327 0.574</td>
<td>0.228 0.389 0.384</td>
</tr>
<tr>
<td></td>
<td>0.097 0.170 0.734</td>
<td>0.100 0.223 0.676</td>
</tr>
<tr>
<td>3</td>
<td>0.904 0.040 0.056</td>
<td>0.752 0.181 0.067</td>
</tr>
<tr>
<td></td>
<td>0.167 0.653 0.181</td>
<td>0.312 0.410 0.278</td>
</tr>
<tr>
<td></td>
<td>0.042 0.267 0.692</td>
<td>0.128 0.149 0.723</td>
</tr>
<tr>
<td>4</td>
<td>0.933 0.037 0.030</td>
<td>0.449 0.426 0.124</td>
</tr>
<tr>
<td></td>
<td>0.199 0.637 0.164</td>
<td>0.462 0.428 0.111</td>
</tr>
<tr>
<td></td>
<td>0.146 0.116 0.738</td>
<td>0.050 0.104 0.845</td>
</tr>
</tbody>
</table>

Figure 3: Marijuana data; clustering kernel $\mathcal{M}_3$: $\mathcal{H}_{it} = (y_{i,t-1}, t)$ for $H = 2$ in combination with homogeneous random prior assignment; posterior distributions of the transition probabilities $\xi_{t,11}$, $\xi_{t,12}$, and $\xi_{t,23}$ plotted over $t = 1, \ldots, 4$. 

---

22
Table 4: Marijuana data; clustering kernel $\mathcal{M}_3: \mathcal{H}_t = (y_{i,t-1}, t)$ in combination with stratified prior assignment. Posterior expectation $E(\gamma_2|y)$, where $\gamma_2$ is the coefficient in the logit model modelling the probability to belong to group 2

| $z_i$                          | $E(\gamma_2|y)$ | 95% HPD      | $E(\gamma_2|y)$ | 95% HPD      | $E(\gamma_2|y)$ | 95% HPD      |
|-------------------------------|-----------------|--------------|-----------------|--------------|-----------------|--------------|
| constant                      | -0.69           | (-1.75, 0.35)| -0.75           | (-1.78, 0.3) | -0.67           | (-1.9, 0.19) |
| marijuana use in 1976         | -0.07           | (-1.7, 1.43) | -              | -            | -              | -            |
| male                          | 0.276           | (-0.71, 1.22)| 0.27           | (-0.69, 1.17)| -              | -            |
| log $p(y)$                    | -598.5          | -            | -598.2          | -            | -597.9          | -            |

5.2.4 Stratified random prior assignment

The two types present in Figure 3 show a similar behavior as the types obtained under deterministic prior stratification according to gender in Figure 2. To verify if male teenagers may, indeed, be associated with the second group, while female teenagers may be associated with the first group, stratified random prior assignment through the mixtures-of-experts model (11) is considered, using gender $g_i$ as subject-specific covariate. In addition, the initial observation is included as discussed in Subsection 2.4 through a dummy variable $D_{i0}$ indicating if the teenager used marijuana, light or heavy, in 1976. Since $H = 2$, the MNL model (11) reduces to a logit model for $\Pr(S_i = 2|g_i, D_{i0})$ with a single regression coefficient $\gamma_2$. Prior (20) with $N_0 = 2.5$ is used for the rows $\xi_{h,j}$, while $\gamma_2 \sim \mathcal{N}(0, I)$.

Table 4 shows the estimated parameters in a logit model were, respectively, both covariates, only the gender, or no covariates are included. The various models are compared using the marginal likelihood. The logit model with a constant corresponds to random prior assignment as in Subsection 5.2.3, however, the prior distribution for $\Pr(S_i = 2)$ is slightly different, explaining the differences in the marginal likelihoods between Table 2 ($-596.5$) and Table 4 ($-597.9$).

From Table 4 we find that the initial condition whether a teenager started with using marijuana is not significant which might be explained by the small fraction of teenagers in the panel indicating this behavior. Male teenagers do have a slightly higher probability to belong to the second group, see also Figure 4. However, also this covariate is not significant. The marginal likelihood indicates mild preference of homogeneous over stratified random prior assignment.

It is possible to compare in Table 2 homogeneous random prior assignment into two endogenous groups with deterministic prior assignment into two homogenous groups using gender for stratification. Regardless of the choice of the prior hyperparameter $N_0$, the marginal likelihood is larger for endogenous clustering into two groups, both if a homogeneous first-order Markov chain is used as group-specific model ($\mathcal{M}_2, H = 2$ versus $\mathcal{M}_4, H = 1$) as well as if a time-inhomogeneous first-order Markov chain is used ($\mathcal{M}_3, H = 2$ versus $\mathcal{M}_5, H = 1$).

Hence, identifying female teenagers with the mostly non-user group and male teenagers with the other group, as was done under deterministic prior stratification, is largely misleading. Both male and female teenagers have about the same risk to belong to the second group. Unobserved factors or latent traits other than the gender determine whether a teenager belongs to one group or the other.

As a further robustness check, the panel was split into a female panel with $N = 120$ time series and a male panel with $N = 117$ time series. Model-based clustering using a time-varying first-order Markov chain ($\mathcal{M}_3$) was applied independently to both panels, using the same prior as in Subsection 5.2.3. The corresponding marginal likelihoods are reported in Table 5 and confirm that in both panels additional unobserved heterogeneity is present which may be captured
Figure 4: Marijuana data; clustering kernel $\mathcal{M}_3$: $\mathcal{H}_{lt} = (y_{l,t-1}, t)$ with $H = 2$ in combination with stratified random prior assignment using gender; posterior distribution of the probability to belong to group 2 for female (left hand side) and male (right hand side) teenagers.

Table 5: Marijuana data split into a female and a male panel; clustering kernel $\mathcal{M}_3$: $\mathcal{H}_{lt} = (y_{l,t-1}, t)$ in combination with homogenous random prior assignment applied independently to each panel; marginal likelihoods for numbers $H$ of clusters

<table>
<thead>
<tr>
<th></th>
<th>females</th>
<th></th>
<th>males</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H$</td>
<td>log $p(y)$</td>
<td>$M_0/M$</td>
<td>log $p(y)$</td>
</tr>
<tr>
<td>1</td>
<td>-285.5</td>
<td>-</td>
<td>-333.7</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>-275.3</td>
<td>0.924</td>
<td>-321.3</td>
<td>0.712</td>
</tr>
<tr>
<td>3</td>
<td>-275.7</td>
<td>0.434</td>
<td>-322.3</td>
<td>0.353</td>
</tr>
</tbody>
</table>

by assuming two endogenous groups. Some of the group-specific transition probabilities are displayed for both panels in Figure 5. Evidently, the two types of marijuana user identified before are present both for female and male teenagers. Using deterministic prior stratification according to gender and assuming homogeneity in the use of marijuana within these groups ignores the presence of this unobserved heterogeneity and leads to misleading conclusions.

Acknowledgements

I would like to thank the Editor, Hans-Hermann Bock, for inviting me to contribute this extensive review. Special thanks go to Sylvia Kaufmann and Christoph Pamminger for inspiring and fruitful cooperations in the field of model-based clustering of time series throughout many years. This research has been supported by the Austrian Science Foundation (FWF) under the grant S 10309-G14 (NRN “The Austrian Center for Labor Economics and the Analysis of the Welfare State”, Subproject “Bayesian Econometrics”).

A Details on labelling the MCMC output

The following scheme demonstrates how labeling using k-means clustering works:

(a) Apply k-means clustering with $H$ clusters to all $MH$ posterior draws of $\tilde{\vartheta}_h^{(m)}$, $h = 1, \ldots, H$, $m = 1, \ldots, M$. This delivers a classification index $\tilde{I}_h^{(m)}$ taking a value in $\{1, \ldots, H\}$ for each of the $MH$ posterior draws.
(b) For each \( m = 1, \ldots, M \), construct the classification sequence \( \rho_m = (r_1^{(m)}, \ldots, r_H^{(m)}) \) and check, if \( \rho_m \) is a permutation of \( \{1, \ldots, H\} \). In this case, a unique labelling is achieved by reordering the draws in the following way:

   (c1) Relabel the hidden allocations through \( \rho_m \): substitute \( S^{(m)} = (S_1^{(m)}, \ldots, S_N^{(m)}) \) by \( (\rho_m(S_1^{(m)}), \ldots, \rho_m(S_N^{(m)})) \).

   (c2) Use the inverse \( (r_1, \ldots, r_H) = (\rho_m^{-1}(1), \ldots, \rho_m^{-1}(H)) \) of \( \rho_m \) to relabel the group-specific parameters:

   (c2-a) substitute \( \vartheta_1^{(m)}, \ldots, \vartheta_H^{(m)} \) by \( \vartheta_{r_1}^{(m)}, \ldots, \vartheta_{r_H}^{(m)} \);

   (c2-b) for a homogeneous random prior allocation distribution, substitute \( \eta_1^{(m)}, \ldots, \eta_H^{(m)} \) by \( \eta_{r_1}^{(m)}, \ldots, \eta_{r_H}^{(m)} \);

   (c2-c) under stratified random prior assignment, substitute \( \gamma_2^{(m)}, \ldots, \gamma_H^{(m)} \) by \( \gamma_{r_2}^{(m)} - \gamma^*, \ldots, \gamma_{r_H}^{(m)} - \gamma^* \), where \( \gamma^* = \gamma_{r_1}^{(m)} \). Subtracting \( \gamma^* \) ensures that the regression coefficient of the baseline is equal to 0 in the identified model.

References


