Dynamic clustering of multivariate panel data*

André Lucas,\textsuperscript{(a)} Julia Schaumburg,\textsuperscript{(a)} Bernd Schwaab,\textsuperscript{(b)}

\textsuperscript{(a)} Vrije Universiteit Amsterdam and Tinbergen Institute
\textsuperscript{(b)} European Central Bank, Financial Research

*** preliminary, unpolished ***

November 18, 2019

Abstract

We propose a novel observation-driven time series model for studying group structure dynamics in multivariate panel data. The model is dynamic in three ways: First, the cluster means and covariance matrix parameters are time-varying to track gradual changes in cluster characteristics over time. Second, the units of interest can transition between clusters over time based on a Hidden Markov model (HMM). Finally, the HMM’s transition matrix can depend on lagged cluster distances as well as economic covariates. Monte Carlo experiments suggest that the units can be classified reliably in a variety of settings. An empirical study of 299 European banks between 2008Q1 and 2018Q2 suggests that banks have become less diverse over time in key characteristics. Approximately 3% of banks transition each quarter on average. Transitions across clusters are related to differences in bank profitability.

Keywords: dynamic clustering; panel data; Hidden Markov Model; score-driven model; bank business models.

JEL classification: G21, C33.

*Author information: André Lucas, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, Email: a.lucas@vu.nl. Julia Schaumburg, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, Email: j.schaumburg@vu.nl. Bernd Schwaab, European Central Bank, Kaiserstrasse 29, 60311 Frankfurt, Germany, Email: bernd.schwaab@ecb.int. The views expressed in this paper are those of the authors and they do not necessarily reflect the views or policies of the European Central Bank.
1 Introduction

This paper proposes a novel observation-driven model for the dynamic clustering of multivariate panel data. The model is dynamic in several ways: First, the cluster means and covariance matrix parameters are time-varying to track changes in cluster characteristics over time. Second, the units of interest can transition between clusters based on a Hidden Markov model (HMM). Finally, the HMM’s transition probabilities is time-varying and can depend on lagged cluster location and scale parameters. We argue that our modeling framework is useful to robustly allocate a potentially large number of units into approximately homogeneous groups in a fairly complicated setting, while keeping track of all group membership probabilities as well as all group transitions. Finally, the model can be extended to accommodate non-Markovian transitions as well as additional explanatory covariates.

All time-varying parameters of our dynamic clustering model are driven by the score of the local (time \( t \)) objective function using the so-called Generalized Autoregressive Score (GAS) approach developed by Creal et al. (2013); see also Harvey (2013). In this setting, the time-varying parameters are perfectly predictable one step ahead. This makes the model observation-driven in the terminology of Cox (1981). The likelihood is known in closed form through a standard prediction error decomposition, facilitating parameter estimation via standard likelihood-based methods. Straightforward filtering recursions are available for all time-varying parameters as well as for all conditional cluster membership probabilities.

Extensive Monte Carlo experiments suggest that our model is able to reliably classify units of interest into distinct clusters, as well as to simultaneously infer all relevant cluster-specific time-varying parameters in the presence of cluster transitions. In our simulations, the cluster classification is perfect for sufficiently large distances between the time-varying cluster means and sufficiently informative signals relative to the variance of the noise terms. As the time-varying cluster means move closer together, however, and as cluster transitions become more frequent, the share of correct classifications decreases, but generally remains high. Parameter estimates are biased, however, and classification results can be poor if cluster transitions are present but ignored.

We apply our modeling framework to a multivariate panel of \( N = 299 \) European banks between 2008Q1 and 2018Q2, i.e. over \( T = 42 \) quarters, considering \( D = 12 \) bank-level indicator variables for \( J \) groups of similar banks. We thus track banking sector data through the 2008–2009 global
financial crisis, the 2010–2012 euro area sovereign debt crisis, as well as the relatively calmer post-crisis period between 2013–2018 characterized by a significant increase in financial regulation, centralized supervision, increasing competition from FinTech and BigTech firms, as well as ultra-low monetary policy interest rates.

We identify $J = 6$ business model groups (clusters). Specifically, we distinguish A) market-oriented universal banks, B) international diversified lenders, C) fee-focused retail lenders, D) international corporate lenders, E) domestic diversified lenders, and F) domestic retail lenders. The similarities and differences between these groups are discussed in detail in the main text.

We focus on three main empirical results. First, we study whether banks have become more or less similar over time. A decrease in financial sector diversity could be problematic from a financial stability perspective. For example, the probability and severity of fire sales could increase if a large number of banks had adopted similar business strategies. We find that our bank business model groups have become more similar over time in key characteristics such as size, leverage, and the share of trading activities. At the same time, banks have become slightly less similar in terms of their cross-border activities and funding choices.

Second, we study which business model groups have become more or less popular over time. The estimated transitions suggest that, since 2008Q1, banks i) have relied increasingly on fee income to lean against impaired profitability from low interest rates, ii) have become less reliant on market funding, and iii) have lent increasingly to retail clients rather than corporate clients.

Third, we study whether bank business model transitions can be explained by profitability differences. We find this to be the case: Differences in cluster-specific return-on-equity are a significant predictor of business model transitions. Banks are more likely to move away from low-profitability groups and to move on to greener pastures. This is intuitive, but also suggest that ultra-low monetary policy rates may have semi-permanent effects on the industry’s structure via discrete bank business model transitions.

From a methodological point of view, our paper contributes to the literature on clustering of time series data. This literature can be divided into four strands. Static clustering of time series refers to a setting with fixed cluster classification, i.e., each time series is allocated to one cluster over the entire sample period. Dynamic clustering, by contrast, allows for changes in the cluster assignments over time. Each approach can be further split into whether the cluster-specific parameters are constant (static) or time-varying (dynamic).
Wang et al. (2013) is an example of static clustering with static parameters. They cluster time series into different groups of autoregressive processes, where the autoregressive parameters are constant within each cluster and cluster assignments are fixed over time.

Frühwirth-Schnatter and Kaufmann (2008) use static clustering with elements of both static and dynamic parameters. First, they cluster time series into different groups of regression models with static parameters. Later, they generalize this to static clustering into groups of different Hidden Markov Models (HMMs), each switching between two regression models. The HMM can be regarded as a specific form of dynamic parameters for the underlying regression model. Their method is used in Hamilton and Owyang (2012) to differentiate between business cycle dynamics among groups of U.S. states. Also Smyth (1996) clusters time series into groups characterized by different Hidden Markov Models.

Creal et al. (2014) is an example of dynamic clustering with static parameters. They develop a model for credit ratings based on market data. Their main objective is to classify firms into different rating categories over time. They therefore allow for transitions across clusters (dynamic clustering), while the parameters in their underlying mixture model are kept constant.

Finally, Catania (2016) is an example of dynamic clustering with dynamic parameters. He proposes a score-driven dynamic mixture model, which relies on score-driven updates of almost all parameters, allowing for time-varying parameters and changing cluster assignments and time-varying cluster assignment probabilities. Due to the high flexibility of the model, a large number of observations is required over time. The application in Catania (2016) to conditional asset return distributions typically has a sufficiently large number of observations.

Our approach falls in the category of dynamic clustering methods with dynamic parameters. We use dynamic clustering as banks are found to switch their business model infrequently over longer periods of time; see e.g. Ayadi and Groen (2015) and ECB (2016). Also, in contrast to the application used by for instance Catania (2016), our banking data are observed over only a moderate number of time points $T$, while the number of units $N$ and the number of firm characteristics $D$ are high. Given present but infrequent transitions, the properties of bank business models are unlikely to be constant throughout the periods of market turbulence and shifts in bank regulations experienced in our sample. We therefore require the cluster components to be characterized by dynamic parameters using the score-driven framework of Creal et al. (2013).

Our paper also contributes to the literature on identifying bank business models. Roengpitya
et al. (2014), Ayadi et al. (2014), and Ayadi and Groen (2015) also use cluster analysis to identify bank business models. Conditional on the identified clusters, the authors discuss bank profitability trends over time, study banking sector risks and their mitigation, and consider changes in banks’ business models in response to new regulation. Our statistical approach is different in that our clusters are not identified based on single (static) cross-sections of year-end data or bottom-up agglomerative clustering with fixed business model characteristics. Instead, we consider a panel framework which allows us to pool information over time while allowing for a rich set of dynamics over time.

We proceed as follows. Section 2 presents our score-driven dynamic clustering model. Section 3 discusses the outcomes of a variety of Monte Carlo simulation experiments. Section 4 applies the model to European financial institutions. Section 5 concludes. A Web Appendix provides further technical and empirical results.

## 2 Score-driven dynamic clustering

### 2.1 Hidden Markov Model

We study the dynamic clustering of multivariate panel data $y_{it} \in \mathbb{R}^{D \times 1}$, where $y_{it}$ is a vector containing characteristics $d = 1, \ldots, D$ for unit $i = 1, \ldots, N$ at time $t = 1, \ldots, T$. Each unit belongs to one cluster $j$ at each time point $t$, for $j = 1, \ldots, J$ clusters. Unit $i$’s cluster membership at time $t$ is described by the latent process $c_{it}$, where $c_{it} = j$ if unit $i$ belongs to cluster $j$ at time $t$. We model the multivariate data $y_{it}$ by the location-scale mixture model

$$y_{it} = \mu_{c_{it},t} + \epsilon_{it}, \quad \epsilon_{it} \sim f_\epsilon (0, \Sigma_{c_{it},t}, \nu_{c_{it}}),$$

where $\mu_{c_{it},t}$ is a $D \times 1$ vector of cluster-specific means, and $\epsilon_{it}$ is a $D \times 1$ vector of disturbance terms characterized by a zero mean, a time-varying and cluster-specific $D \times D$ covariance (or scale) matrix $\Sigma_{c_{it},t}$, and possibly additional parameters $\nu_{c_{it}}$. If $f_\epsilon$ is a multivariate Student’s $t$ density, then $\nu_{c_{it}}$ is the degrees of freedom parameter for unit $i$ at time $t$. This encompasses the special case of the normal distribution, for which we can set $\nu_{c_{it}}^{-1} = 0$. Skewed distribution are also easily accommodated in this framework, but are not considered in this paper. We assume that
cluster means $\mu_{jt}$ and disturbance vectors $\epsilon_{it}$ are mutually uncorrelated for all clusters $j$ and at all leads and lags.

We model the transitions from one cluster to the next by a Hidden Markov Model (HMM); see e.g. Goldfeld and Quandt (1973) and Bhar and Hamori (2004). The dynamics of the HMM are characterized by the latent (hidden) states $c_{it}$ that are driven by an underlying Markov chain. The Markov property implies that the next state depends only on the current state, i.e.

$$
\mathbb{P} \left\{ c_{i,t+1} = j \mid c_{i0}, \ldots, c_{it} \right\} = \mathbb{P} \left\{ c_{i,t+1} = j \mid c_{it} \right\}.
$$

We introduce the short-hand notation $\pi_{jk,t} := \mathbb{P} \left\{ c_{i,t+1} = k \mid c_{it} = j \right\}$, where $\pi_{jk,t}$ denotes the possibly time-varying probability of transiting from state $j$ to state $k$ at time $t$.

The $J \times J$ HMM transition matrix $\Pi_t$ contains all transition probabilities $\pi_{jk,t}$ for $j, k = 1, \ldots, J$. We require the rows of $\Pi_t$ to sum to one, i.e., $\sum_{k=1}^{J} \pi_{jk,t} = 1$ for all $j = 1, \ldots, J$. We assume the transition probabilities $\pi_{jk,t}$ vary over time as a function of the time-varying distance between the clusters at time $t - 1$. In particular, we specify the transition matrix as

$$
\Pi_t = \Pi_t (D_{t-1}),
$$

where $D_t$ is a $J \times J$ matrix with elements $d_{jk,t}$, where $d_{jk,t}$ denotes the distance between cluster $j$ and cluster $k$ at time $t$. For example, it is often natural to assume that a unit’s transition from one cluster to another is less likely when the clusters are further apart. Conversely, transitions between nearby (neighboring) clusters may be more likely. The off-diagonal elements of $\Pi_t$ are then decreasing in $d_{jk,t}$. To avoid an undue increase in the number of parameters, we parsimoniously model the transition probabilities as

$$
\pi_{jk,t} = \frac{\exp \left( -\gamma d_{jk,t-1} \right)}{\sum_{q=1}^{J} \exp \left( -\gamma d_{jq,t-1} \right)} \quad \text{for } j, k = 1, \ldots, J,
$$

where the scalar parameter $\gamma$ indicates the rate of decay of the transition probabilities in terms of the cluster distances. The numerator in (3) is equal to one if $j = k$, regardless of $\gamma$. A higher value for $\gamma$ leads to lower values of $\exp \left( -\gamma d_{jk,t-1} \right)$ for $j \neq k$, and therefore to lower transition probabilities and to fewer implied transitions. Vice versa, a lower value for $\gamma$ leads to higher
transition probabilities. Finally, the multinomial specification in (3) ensures that the rows of $\Pi_t$ sum to one by construction. If lagged distances matter for transitions beyond the first lag, then (2) can easily be adapted to, for example, $\Pi_t = \Pi_t \left( \tilde{D}_{t-1} \right)$, where $\tilde{D}_t = \lambda D_t + (1 - \lambda) \tilde{D}_{t-1}$, and $0 < \lambda < 1$ is a parameter to be estimated or chosen ex-ante.

To close the specification of our model, we need to specify the cluster distances $d_{jk,t}$. To measure cluster proximity we adopt the Mahalanobis distance metric

$$d_{jk,t} = \sqrt{\left( \mu_{jt} - \mu_{kt} \right)^\prime \Sigma_t^{-1} \left( \mu_{jt} - \mu_{kt} \right)}, \quad (4)$$

where $\Sigma_t = J^{-1} \sum_{j=1}^{J} \Sigma_{jt}$ is the average scaling matrix across the different clusters. As a result, cluster distances are invariant to adopting a different scaling of input variables. Variables that are less correlated with the others receive more “weight” in the distance metric. The Euclidean distance is obtained as a special case by setting $\Sigma_t \equiv I_D$.

### 2.2 Time-varying conditional cluster probabilities

This section derives a filtering equation for the conditional probability $\tau_{ij,t|t} := \mathbb{P}[c_{it} = j|F_t; \theta]$, where $\tau_{ij,t|t}$ denotes the probability that unit $i$ belongs to cluster $j$ at time $t$ given the information set $F_t$, where $F_t$ contains observations up to $t$, $y_1, \ldots, y_t$. The vector $\theta$ contains the static parameters of the model that need to be estimated.

We start by considering the log-likelihood contribution of observation $y_{it}$,

$$\ell_{it} = \log \left( f(y_{it}|F_{t-1}; \theta) \right) = \log \left( \sum_{j=1}^{J} \tau_{ij,t|t-1} f(y_{it}|c_{it} = j, F_{t-1}; \theta) \right) \quad (5)$$

where $f(y_{it}|c_{it} = j, F_{t-1}; \theta)$ is the density of $y_{it}$ in cluster $j$, and $\tau_{ij,t|t-1} := \mathbb{P}[c_{it} = j|F_{t-1}; \theta]$ is the conditional probability that unit $i$ belongs to cluster $j$ at time $t$ given $F_{t-1}$. By the Markov property the predicted conditional state probability $\tau_{ij,t|t-1}$ only depends on the previous state and on elements of the transition matrix $\Pi_t$. We use this property to update the cluster probabilities as

$$\tau_{ij,t+1|t} = \mathbb{P}[c_{i,t+1} = j|F_t; \theta] = \sum_{k=1}^{J} \pi_{kj,t} \mathbb{P}[c_{it} = k|F_t; \theta] = \sum_{k=1}^{J} \tau_{ik,t} \pi_{kj,t}, \quad (6)$$
Using a standard Bayes argument, the filtered cluster probabilities are determined by

\[
\tau_{ij,t|t} = \mathbb{P}[c_{it} = j | \mathcal{F}_t; \theta] = \frac{\tau_{ij,t-1|t-1} f(y_{it}|c_{it} = j, \mathcal{F}_{t-1}; \theta)}{f(y_{it}|\mathcal{F}_{t-1}; \theta)}
\]

\[
= \frac{\tau_{ij,t|t-1} f(y_{it}|c_{it} = j, \mathcal{F}_{t-1}; \theta)}{\tau_{1,t|t-1} f(y_{it}|c_{it} = 1, \mathcal{F}_{t-1}; \theta) + \ldots + \tau_{iJ,t|t-1} f(y_{it}|c_{it} = J, \mathcal{F}_{t-1}; \theta)}.
\]

The filtered cluster probabilities thus update the predicted cluster probabilities \(\tau_{ij,t|t-1}\) by using the time \(t\) observation \(y_{it}\) and its likelihood of coming from the cluster \(j\) density \(f(y_{it}|c_{it} = j, \mathcal{F}_{t-1}; \theta)\), normalized by the unconditional data density \(f(y_{it}|\mathcal{F}_{t-1}; \theta)\). This is intuitive: if \(\tau_{ij,t|t-1} f(y_{it}|c_{it} = j, \mathcal{F}_{t-1}; \theta)\) is high compared to \(\tau_{ik,t|t-1} f(y_{it}|c_{it} = k, \mathcal{F}_{t-1}; \theta)\) for \(k \neq j\), then \(y_{it}\) is more likely to come from cluster \(j\), and the filtered cluster probability \(\tau_{ij,t|t}\) increases accordingly. Otherwise the filtered cluster probability is adjusted downward. We can use the filtered cluster probabilities \(\tau_{ij,t|t}\) or their predicted counterparts \(\tau_{ij,t|t-1}\) to assign each observation \(i\) at time \(t\) to a specific cluster \(j\). For example, we may assign unit \(i\) to the cluster \(j^*\) for which the filtered cluster probability is maximal, i.e., \(j^* = \arg \max_j \tau_{ij,t|t}\).

### 2.3 Time-varying cluster-specific parameters

#### 2.3.1 Time-varying means

Time-variation in location and scale parameters is modeled following the score-driven approach as introduced by Creal et al. (2013) and Harvey (2013). We impose further parsimony by using the exponentially weighted score-driven dynamics of Lucas and Zhang (2016). For the time-varying means, we specify

\[
\mu_{jt,t+1} = \mu_{jt} + A_1 S_{\mu_{jt}, t} \cdot \nabla_{\mu_{jt}, t},
\]

where the diagonal matrix \(A_1 = A_1(\theta)\) depends on the vector of unknown static parameters \(\theta\), \(S_{\mu_{jt}, t}\) is a scaling matrix, and the score \(\nabla_{\mu_{jt}, t}\) is the first derivative of the log-density of \(y_{it}\) with
respect to $\mu_{jt}$. In our case, the score is given by

$$\nabla_{\mu_{jt}, t} = \frac{\partial \ell_t}{\partial \mu_{jt}} = \frac{\partial}{\partial \mu_{jt}} \left[ \sum_{i=1}^{N} \log \left( f(y_{it} | F_{t-1}; \theta) \right) \right]$$

$$= \sum_{i=1}^{N} \frac{\partial}{\partial \mu_{jt}} \log \left( \sum_{j=1}^{J} \tau_{ij,t|t-1} f(y_{it} | c_{it} = j, F_{t-1}; \theta) \right)$$

$$= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \frac{\partial}{\partial \mu_{jt}} \log f(y_{it} | c_{it} = j, F_{t-1}; \theta) = \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \nabla_{\mu_{jt}, t}^{(j)} \tag{9}$$

where $\nabla_{\mu_{jt}, t}^{(j)} = \frac{\partial \log f(y_{it} | c_{it} = j, F_{t-1}; \theta)}{\partial \mu_{jt}}$ is the score of mixture component $j$. As a closed form expression for the conditional Fisher information matrix of $\mu_{jt}$ is not available, we use an approximation to account for the curvature of the score, namely

$$S_{\mu_{jt}, t} = \left( \sum_{i=1}^{N} \tau_{ij,t|t} \cdot E \left[ \nabla_{\mu_{jt}, t}^{(j)} \left( \nabla_{\mu_{jt}, t}^{(j)} \right)' | c_{it} = j \right] \right)^{-1} \tag{10}$$

Our scaling matrix thus takes the weighted average of the conditional Fisher information matrices of each of the regimes $j$, weighted by their filtered posterior probability $\tau_{ij,t|t}$ of observation $y_{it}$ coming from regime $j$.

As a concrete example, consider the case of a mixture of normal distributions. In that case we have

$$\nabla_{\mu_{jt}, t}^{(j)} = \Sigma_{jt}^{-1}(y_{it} - \mu_{jt}), \quad S_{\mu_{jt}, t} = \left( \sum_{i=1}^{N} \tau_{ij,t|t} \Sigma_{jt}^{-1} \right)^{-1}, \tag{11}$$

$$\mu_{jt+1} = \mu_{jt} + A_1 \frac{\sum_{i=1}^{N} \tau_{ij,t|t} \cdot (y_{it} - \mu_{jt})}{\sum_{i=1}^{N} \tau_{ij,t|t}} \tag{12}$$

A detailed derivation of (12) is provided in Web Appendix A.1. The transition equation (12) is highly intuitive: the cluster means are updated by the prediction errors for that cluster, accounting for the posterior probabilities that the observation was drawn from that same cluster. For example, if the posterior probability $\tau_{ij,t|t}$ indicates that observation $y_{it}$ comes from cluster $j$ with negligible probability, then the update of $\mu_{jt}$ does not depend on $y_{it} - \mu_{jt}$.

As a second example, consider a mixture of Student’s $t$ distributions. In that case (9) remains
unchanged, while

\[ \nabla_{\mu_{jt},t}^{(j)} = w_{ij,t} \cdot \Sigma_{jt}^{-1} (y_{it} - \mu_{jt}), \tag{13} \]

where the weight \( w_{ij,t} = \frac{(1 + \nu^{-1} D) \cdot (1 + \nu^{-1} (y_{it} - \mu_{jt})' \Sigma_{jt}^{-1} (y_{it} - \mu_{jt}))}{1 + \nu^{-1} (y_{it} - \mu_{jt})' \Sigma_{jt}^{-1} (y_{it} - \mu_{jt})} \) provides the model with a robustness feature: observations \( y_{it} \) that are outlying given the fat-tailed nature of the Student’s t density receive a reduced impact on the location and volatility dynamics by means of a lower value for \( w_{ij,t} \).

Combining (13) with the approximate scaling function in (11) yields the transition equation

\[ \mu_{j,t+1} = \mu_{jt} + A_1 \sum_{i=1}^{N} \tau_{ij,t} |t| \cdot w_{ij,t} \cdot (y_{it} - \mu_{jt}) \frac{1}{\sum_{i=1}^{N} \tau_{ij,t} |t|}. \tag{14} \]

The Gaussian transition equation (12) is a special case of (14) for \( \nu^{-1} \to 0 \) and \( w_{ij,t} \to 1 \).

### 2.3.2 Time-varying covariance matrices

This section presents the transition equation for the time-varying covariance matrices \( \Sigma_{jt} \). Following the exponentially weighted score-driven dynamics of Lucas and Zhang (2016), it is given by

\[ \text{vec}(\Sigma_{j,t+1}) = \text{vec}(\Sigma_{jt}) + A_2 \cdot S_{\Sigma_{jt},t} \cdot \nabla_{\Sigma_{jt},t}, \tag{15} \]

where matrix \( A_2 = A_2(\theta) \) depends on parameters to be estimated, \( S_{\Sigma_{jt},t} \) is a scaling matrix, and \( \nabla_{\Sigma_{jt},t} \) is the score. The score dynamics are determined in the same way as for the time-varying cluster means. The score is given by

\[ \nabla_{\Sigma_{jt},t} = \frac{1}{2} \frac{\partial \ell_t}{\partial \text{vec}(\Sigma_{jt})} = \frac{1}{2} \frac{\partial}{\partial \text{vec}(\Sigma_{jt})} \sum_{i=1}^{N} \log f(y_{it} | F_{t-1}; \theta) \]

\[ = \frac{1}{2} \sum_{i=1}^{N} \tau_{ij,t} |t| \cdot \frac{\partial \log f(y_{it} | c_{it} = j, F_{t-1}; \theta)}{\partial \text{vec}(\Sigma_{jt})} = \frac{1}{2} \sum_{i=1}^{N} \tau_{ij,t} |t| \cdot \nabla_{\Sigma_{jt},t}^{(j)}, \tag{16} \]

\[ = \frac{1}{2} \sum_{i=1}^{N} \tau_{ij,t} |t| \cdot \nabla_{\Sigma_{jt},t}^{(j)}, \tag{16} \]
where \( \nabla_{\Sigma_{jt}, t} = \partial \log f(y_{it} | c_{it} = j, \mathcal{F}_{t-1}; \theta) / \partial \text{vec}(\Sigma_{jt}) \). For the scaling matrix, we can take the analogous expression as in (10) and consider

\[
S_{\Sigma_{jt}, t} = \left( \sum_{i=1}^{N} \tau_{ij,t} | t \cdot \mathbb{E} \left[ \nabla_{\Sigma_{jt}, t} (\nabla_{\Sigma_{jt}, t})' | c_{it} = j, \mathcal{F}_{t-1}; \theta \right] \right)^{-1}
\]

\[
= \left( \sum_{i=1}^{N} \tau_{ij,t} | t \cdot \mathbb{E} \left[ -\partial \nabla_{\Sigma_{jt}, t} / \partial \text{vec}(\Sigma_{jt})' | c_{it} = j, \mathcal{F}_{t-1}; \theta \right] \right)^{-1}.
\]

For example, for a Gaussian mixture of normals, we obtain

\[
\nabla_{\Sigma_{jt}, t} = \frac{1}{2} \text{vec} \left( \sum_{i=1}^{N} \tau_{ij,t} | t \cdot \Sigma_{jt}^{-1} \left( (y_{it} - \mu_{jt}) (y_{it} - \mu_{jt})' - \Sigma_{jt} \right) \right)
\]

\[
= \frac{1}{2} \sum_{i=1}^{N} \tau_{ij,t} | t \cdot (\Sigma_{jt}^{-1} \otimes \Sigma_{jt}^{-1}) \text{vec} \left( (y_{it} - \mu_{jt}) (y_{it} - \mu_{jt})' - \Sigma_{jt} \right),
\]

\[
S_{\Sigma_{jt}, t} = \left( \frac{1}{2} \sum_{i=1}^{N} \tau_{ij,t} | t \cdot (\Sigma_{jt}^{-1} \otimes \Sigma_{jt}^{-1}) \right)^{-1},
\]

\[
\text{vec}(\Sigma_{j,t+1}) = \text{vec}(\Sigma_{jt}) + A_2 \sum_{i=1}^{N} \tau_{ij,t} | t \cdot \text{vec} \left( (y_{it} - \mu_{jt}) (y_{it} - \mu_{jt})' - \Sigma_{jt} \right) / \sum_{i=1}^{N} \tau_{ij,t} | t
\]

Unvectorizing (21), we obtain the covariance matrix transition equation

\[
\Sigma_{j,t+1} = \Sigma_{jt} + A_2 \sum_{i=1}^{N} \tau_{ij,t} | t \left[ (y_{it} - \mu_{jt}) (y_{it} - \mu_{jt})' - \Sigma_{jt} \right] / \sum_{i=1}^{N} \tau_{ij,t} | t.
\]

Web Appendix A.2 provides a step-by-step derivation of (22). Again, the transition equation is highly intuitive: the components of the covariance matrix are updated by the difference between the outer product of the prediction errors and the current covariance matrix for that cluster, weighted by the filtered probabilities that the observation was drawn from that same cluster.

For a mixture of Student’s t distributions, (16) remains unchanged, while the cluster-specific score is now given by

\[
\nabla_{\Sigma_{jt}, t} = \frac{1}{2} \sum_{i=1}^{N} \tau_{ij,t} | t \cdot (\Sigma_{jt}^{-1} \otimes \Sigma_{jt}^{-1}) \text{vec} (w_{ij,t} (y_{it} - \mu_{jt}) (y_{it} - \mu_{jt})' - \Sigma_{jt}'),
\]

where \( w_{ij,t} \) is defined below (13). Pre-multiplying the score by the approximate scaling matrix...
(20) yields the transition equation

\[ \Sigma_{j,t+1} = \Sigma_{jt} + A_2 \sum_{i=1}^{N} \tau_{ij,t|t} \left[ w_{ij,t} (y_{it} - \mu_{jt}) (y_{it} - \mu_{jt})' \right] \sum_{i=1}^{N} \tau_{ij,t|t} , \]  

where the Gaussian case (22) is again a special case of (24) for \( \nu^{-1} \to 0 \) and \( w_{ij,t} \to 1 \).

### 2.3.3 Initialization of the time-varying parameters

The cluster probabilities \( \tau_{ij,1|1} \), the cluster means \( \mu_{j1} \), and the cluster covariance matrices \( \Sigma_{j1} \) need to be initialized to start the filtering recursions. We can initialize by any static clustering algorithm, including \( k \)-means (Hartigan and Wong (1979)), intelligent \( k \)-means (de Amorim and Hennig (2015)), or hierarchical agglomerative clustering (McLachlan and Peel (2000)). For this purpose we use data of \( t = 1 \) only, \( y_{i1} \) for \( i = 1, \ldots, N \). We choose the \( k \)-means algorithm for simplicity. It allocates our \( N \) observations in \( D \) dimensions to \( J \) clusters such that the within-cluster sum of squares is minimized. Web Appendix B provides the details for this algorithm.

The \( k \)-means clustering algorithm provides the cluster probabilities \( \tau_{ij,1|1}^k \). These probabilities are one for the assigned cluster, and zero for the remaining clusters. Based on these initial cluster assignments, the initial cluster means \( \mu_{j1} \) equal the sample average of \( y_{i1} \) for units \( i = 1, \ldots, N \) for which \( \tau_{ij,1|1}^k \) equals 1. The initialized covariance matrices \( \Sigma_{j1} \) are similarly determined as the empirical covariance of observations \( y_{i1} \) for units \( i \) assigned to cluster \( j \).

The initial \( \tau_{ij,1|1}^k \) can be replaced by the filtered \( \tau_{ij,1|1} \) from (7) once a first estimate of parameters \( \theta \) is available. Alternatively, \( \tau_{ij,4|4} \) could be used for quarterly data. Parameters \( \theta \) can subsequently be re-estimated conditional on \( \tau_{ij,1|1} \), \( \mu_{j1} \left( \tau_{ij,1|1} \right) \), and \( \Sigma_{j1} \left( \tau_{i,1|1} \right) \) to minimize any impact from the initialization procedure.

### 2.4 Extensions

#### 2.4.1 Non-Markovian transitions

In some settings, economic reasoning will suggest that cluster membership is persistent over time. For example, we may expect banks’ business model choices to be highly persistent. Once a bank opts for a different business model, it is extremely unlikely to revert back to the old business model.
the next period. This economic reasoning, however, is not explicitly enforced in the current model set-up. Particularly if two clusters are close at any particular moment in time, the probability of switching from business model (cluster) 1 to 2 can be large. Due to the symmetry, the probability of switching back from 2 to 1 is then large as well.

In order to better accommodate the persistence of business model choices, we can introduce asymmetry in the model: once a bank has changed business model, it becomes ‘inactive’ for a number of periods, meaning that it is not at risk of leaving its current state. Such behavior results in non-Markovian transitions, as the probability of transiting from one business model to the next no longer only depends on the current business model, but also on the fact whether or not there was a business model change over the most recent periods.

The advantage of this new set-up is that it can be accommodated without increasing the number of parameters. Let \( P \) denote the number of periods that a firm is not at risk of changing business model after a business model change. We introduce new states \( c_{it,p} \) for \( p = 1, \ldots, P \), where \( c_{it,0} \) is our old state \( c_{it} \) in which the bank is at risk for transiting from state \( i \) to state \( j \). We now model such a transition as a change from state \( (i,0) \) to state \( (j,p) \). For \( p > 0 \), only transitions occur from state \( (j,p) \) to state \( (j,p - 1) \). For instance, if \( P = 2 \), and \( J = 2 \), we would get the extended transition probability matrix (from row \( j \) to column \( k \))

\[
\begin{pmatrix}
\pi_{11,t} & 0 & 0 & 0 & 0 & \pi_{12,t} \\
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & \pi_{21,t} & \pi_{22,t} & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0
\end{pmatrix}
\]

It is clear that the number of parameters is the same as in the benchmark model. The intuition for the above transition matrix is as follows. If a bank starts with business model 1, it can migrate to state \((1, p = 0)\) with probability \( \pi_{11,t} \), and to state \((2, p = 2)\) with probability \( \pi_{12,t} \). If it migrates to state \((2, p = 2)\), the next period it migrates to state \((2, p = 1)\) with probability 1, and the period
after that to state \((2, p = 0)\). Only in state \((2, p = 0)\), the bank is at risk of a business model migration again, namely with probability \(\pi_{21,t}\). With the remaining probability \(\pi_{22,t}\), its business model remains unchanged. If a change hits with probability \(\pi_{21,t}\), a migration to state \((1, p = 2)\) takes place. Then it takes 2 periods to land via state \((1, p = 1)\) into state \((1, p = 0)\) again, where the whole process can start again. As \(J\) and \(P\) can be chosen by the modeler, this set-up can flexibly accommodate transition-free periods after an initial business model change and prevent erratic, short-lived business model changes.

### 2.4.2 Explanatory covariates

Cluster transition dynamics can be related to explanatory covariates above and beyond what is implied by lagged cluster distances. Fortunately, the transition probabilities (3) can be extended to include contemporaneous or lagged variables as additional conditioning variables. For example, banks from low profitability clusters could have a higher incentive to leave that cluster. Vice versa, banks from high profitability clusters could try to remain there, and not migrate to a lower-profitability cluster; see e.g. Ayadi and Groen (2015) and Roengpitya et al. (2017). Using additional conditioning variables allows us to incorporate and test for such effects. Let \(x_{jk,t}\) be a vector of observed covariates, and \(\beta\) a vector of unknown coefficients that need to be estimated. The transition probabilities can then be modeled as

\[
\pi_{jk,t} = \frac{\exp(-\gamma d_{jk,t-1} + \beta' x_{jk,t})}{\sum_{q=1}^{J} \exp(-\gamma d_{jq,t-1} + \beta' x_{jq,t})} \quad \text{for } j, k = 1, \ldots, J, \quad (25)
\]

where \(\gamma\) and \(d_{jk,t-1}\) are defined below (3) and rows continue to add up to one.

### 2.5 Parameter estimation

Observation-driven multivariate time series models such as the score-driven model introduced above are attractive because the log-likelihood is known in closed form. Parameter estimates can therefore be obtained in a standard way by numerically maximizing the likelihood function. For a given set of observations \(y_1, \ldots, y_T\), the vector of unknown parameters \(\theta = \{\text{vec}(A_1)' , \text{vec}(A_2)' , \nu_1, \ldots, \nu_J, \gamma, \beta'\}'\) can be estimated by maximizing the log-likelihood function with respect to \(\theta\),
that is

\[ \mathcal{L}(\theta|\mathcal{F}_T) = \sum_{t=1}^{T} \sum_{i=1}^{N} \ell_{it}, \]  

(26)

where \( \ell_{it} \) is defined in (5). The evaluation of \( \ell_{it} \) is easily incorporated in the filtering process for the latent states.

The maximization of (26) can in principle be carried out by any convenient numerical optimization method. In practice, however, mixture time series models such as ours can imply somewhat unruly log-likelihood surfaces. In such cases standard numerical optimizers are at risk to converge to a local, rather than the global, maximum. Non-gradient based methods such as simulated annealing (see e.g. Goffe et al. (1994)) can then have an advantage over repeatedly re-running standard gradient-based methods.

3 Simulation study

3.1 Simulation design

This section investigates the ability of our score-driven dynamic clustering model to simultaneously

i) correctly classify the units of interest to distinct clusters, and

ii) recover the true time-varying transition probabilities that govern cluster transitions. In all cases, we pay particular attention to the sensitivity of the estimation approach and the filtering algorithm to the (dis)similarity of the clusters, the intensity at which transitions take place, and the number of units per cluster.

We simulate from a mixture of dynamic bivariate densities. These densities are composed of sinusoid mean functions and i.i.d. disturbance terms that are drawn from a bivariate Gaussian distribution. The covariance matrices are chosen to be time-invariant identity matrices.

Specifically, we generate data from two clusters located around two different time-varying cluster means. The time-varying means move in two non-overlapping circles over time. Our baseline setting is visualized in Figure 1. Key inputs into our simulations are the transition intensity parameter \( \gamma \) in (3), the distance between the two circle centers, and the radius of each circle.

We consider three different choices for the transition parameter \( \gamma \in \{0.25, 0.50, 0.75\} \), two choices of radius \( \in \{1, 4\} \), and two choices of unconditional cluster distance \( \in \{4, 8\} \). The sample
We simulate bivariate data $D = 2$ from two clusters $J = 2$. The two time-varying means move in circles that are generated by sinusoid functions. Blue dots indicate the clusters’ unconditional means (circle centers). Green dots indicate the evolution of time-varying cluster means over time. The time-varying cluster means evolve either clockwise, keeping the cluster data equidistant in expectation, or one circle moves clockwise and the other one counter-clockwise, implying time-variation in cluster distance and transition probabilities. Radius (Rad.) refers to the radius of the true mean circles and is a measure of the signal-to-noise ratio of the time-variation in means relative to the variance of the error terms. Distance (Dist.) is the distance between circle centers and measures the distinctiveness of the two clusters in expectation.

sizes are chosen to resemble typical sample sizes in studies of banking data. We thus keep the number of time points small to moderate, considering $T \in \{20, 40\}$, and set the number of cross-sectional units equal to $N \in \{100, 200\}$. The number of clusters is fixed at $J = 2$ throughout. We also assume $J = 2$ is known during estimation. In total, we thus consider 96 different DGPs.

The time-varying cluster means evolve either clockwise, or one circle moves clockwise and the other one counter-clockwise. In the former case, the data drawn from the different clusters are equidistant in expectation. In the latter case, the transition probabilities $\pi_{jk,t}$ are time-varying as they depend on distances between cluster means at $t - 1$; see (3).

We are particularly interested in two issues. First, the lower $\gamma$, and the lower the distance between the two clusters, the more cluster transitions occur and the more informative the data are about such transitions. We expect that more frequent transitions should increase the precision with which $\gamma$ can be estimated, but also make it harder for the model to correctly classify each unit. Second, the radii become particularly interesting when one circle rotates clockwise and the other one counter-clockwise. The radii then determine how close and how far the cluster means
can come together and move apart from each other. Time-varying cluster distance implies time-
variation in the transition probabilities. This time-variation could have an effect on both $\gamma$ and
classification accuracy.

3.2 Simulation results

Using the score-driven model set-up and estimation methodology from Section 2, we classify the
data points and estimate the cluster parameters from the simulated data. The static parameters
to be estimated include the distinct entries of the covariance matrices, and the diagonal elements
of the smoothing matrix $A_1$, which, for simplicity, we assume to be equal across dimensions and
clusters, i.e. $A_1 = a_1 I_D$. 
Table 1: Simulation outcomes

Mean parameter estimates, average percentage of correct classification, and average mean squared errors for time-varying cluster means. Upper panel: The time-varying cluster means evolve clockwise from the same initial position relative to their respective circle center. The simulated cluster data are thus equidistant in expectation, implying time-invariant transition probabilities. Lower panel: One time-varying cluster mean evolves clockwise and the other one counter-clockwise. The cluster distance thus varies over time, also implying time-varying transition probabilities across clusters.

Considered sample sizes are $N = 100, 300$ and $T = 20, 40$. The transition intensity parameter $\gamma$ determines the frequency of transitions; lower values of $\gamma$ imply a higher number of transitions in expectation. Distance (dist.) is the distance between circle centers and measures the distinctiveness of clusters. The circle radius equals 2 in all cases.

<table>
<thead>
<tr>
<th>$N/2$</th>
<th>$T$</th>
<th>dist.</th>
<th>$\hat{\gamma}$</th>
<th>% corr.</th>
<th>MSE</th>
<th>$\hat{\gamma}$</th>
<th>% corr.</th>
<th>MSE</th>
<th>$\hat{\gamma}$</th>
<th>% corr.</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>20</td>
<td>10</td>
<td>0.2506</td>
<td>1.0000</td>
<td>0.1433</td>
<td>0.5039</td>
<td>1.0000</td>
<td>0.1425</td>
<td>2.6626</td>
<td>1.0000</td>
<td>0.1425</td>
</tr>
<tr>
<td>50</td>
<td>40</td>
<td>10</td>
<td>0.2501</td>
<td>1.0000</td>
<td>0.0630</td>
<td>0.5012</td>
<td>1.0000</td>
<td>0.0627</td>
<td>1.9361</td>
<td>1.0000</td>
<td>0.0627</td>
</tr>
<tr>
<td>150</td>
<td>20</td>
<td>10</td>
<td>0.2505</td>
<td>1.0000</td>
<td>0.1027</td>
<td>0.5021</td>
<td>1.0000</td>
<td>0.1028</td>
<td>1.4885</td>
<td>1.0000</td>
<td>0.1029</td>
</tr>
<tr>
<td>150</td>
<td>40</td>
<td>10</td>
<td>0.2499</td>
<td>1.0000</td>
<td>0.0400</td>
<td>0.5027</td>
<td>1.0000</td>
<td>0.0401</td>
<td>0.7708</td>
<td>1.0000</td>
<td>0.0400</td>
</tr>
<tr>
<td>50</td>
<td>20</td>
<td>6</td>
<td>0.2494</td>
<td>0.9837</td>
<td>0.1464</td>
<td>0.4980</td>
<td>0.9851</td>
<td>0.1444</td>
<td>0.7511</td>
<td>0.9860</td>
<td>0.1436</td>
</tr>
<tr>
<td>50</td>
<td>40</td>
<td>6</td>
<td>0.2483</td>
<td>0.9923</td>
<td>0.0642</td>
<td>0.4973</td>
<td>0.9936</td>
<td>0.0638</td>
<td>0.7461</td>
<td>0.9944</td>
<td>0.0631</td>
</tr>
<tr>
<td>150</td>
<td>20</td>
<td>6</td>
<td>0.2488</td>
<td>0.9964</td>
<td>0.1050</td>
<td>0.4986</td>
<td>0.9976</td>
<td>0.1044</td>
<td>0.7509</td>
<td>0.9984</td>
<td>0.1036</td>
</tr>
<tr>
<td>150</td>
<td>40</td>
<td>6</td>
<td>0.2487</td>
<td>0.9930</td>
<td>0.0405</td>
<td>0.4982</td>
<td>0.9940</td>
<td>0.0403</td>
<td>0.7497</td>
<td>0.9947</td>
<td>0.0403</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$N/2$</th>
<th>$T$</th>
<th>dist.</th>
<th>$\hat{\gamma}$</th>
<th>% corr.</th>
<th>MSE</th>
<th>$\hat{\gamma}$</th>
<th>% corr.</th>
<th>MSE</th>
<th>$\hat{\gamma}$</th>
<th>% corr.</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>20</td>
<td>10</td>
<td>0.2511</td>
<td>0.9876</td>
<td>0.1434</td>
<td>0.5055</td>
<td>0.9877</td>
<td>0.1411</td>
<td>0.7792</td>
<td>0.9878</td>
<td>0.1406</td>
</tr>
<tr>
<td>50</td>
<td>40</td>
<td>10</td>
<td>0.2497</td>
<td>0.9955</td>
<td>0.0633</td>
<td>0.5005</td>
<td>0.9957</td>
<td>0.0625</td>
<td>0.7568</td>
<td>0.9958</td>
<td>0.0626</td>
</tr>
<tr>
<td>150</td>
<td>20</td>
<td>10</td>
<td>0.2505</td>
<td>0.9997</td>
<td>0.1030</td>
<td>0.5054</td>
<td>0.9997</td>
<td>0.1032</td>
<td>0.7638</td>
<td>0.9998</td>
<td>0.1031</td>
</tr>
<tr>
<td>150</td>
<td>40</td>
<td>10</td>
<td>0.2499</td>
<td>0.9956</td>
<td>0.0401</td>
<td>0.5018</td>
<td>0.9957</td>
<td>0.0404</td>
<td>0.7559</td>
<td>0.9958</td>
<td>0.0403</td>
</tr>
<tr>
<td>50</td>
<td>20</td>
<td>6</td>
<td>0.2393</td>
<td>0.8013</td>
<td>0.1440</td>
<td>0.4480</td>
<td>0.8040</td>
<td>0.1441</td>
<td>0.5990</td>
<td>0.8064</td>
<td>0.1443</td>
</tr>
<tr>
<td>50</td>
<td>40</td>
<td>6</td>
<td>0.2406</td>
<td>0.8324</td>
<td>0.0657</td>
<td>0.4482</td>
<td>0.8349</td>
<td>0.0652</td>
<td>0.6059</td>
<td>0.8378</td>
<td>0.0656</td>
</tr>
<tr>
<td>150</td>
<td>20</td>
<td>6</td>
<td>0.2400</td>
<td>0.7770</td>
<td>0.1025</td>
<td>0.4527</td>
<td>0.7795</td>
<td>0.1026</td>
<td>0.6130</td>
<td>0.7821</td>
<td>0.1017</td>
</tr>
<tr>
<td>150</td>
<td>40</td>
<td>6</td>
<td>0.2412</td>
<td>0.7923</td>
<td>0.0410</td>
<td>0.4534</td>
<td>0.7950</td>
<td>0.0408</td>
<td>0.6147</td>
<td>0.7978</td>
<td>0.0409</td>
</tr>
</tbody>
</table>
Table 1 reports our simulation results when both time-varying cluster means move clockwise. The cluster means are thus equidistant. As a result, the transition probabilities are time-invariant. We observe that the precision with which $\gamma$ can be estimated depends on $\gamma$ itself. For low values of $\gamma$ (e.g., $\gamma = 0.25$), the model implies relatively frequent transitions between clusters. The transition parameter can therefore be estimated fairly precisely. For higher values of $\gamma$ (e.g., $\gamma = 0.75$) the model implies fewer transitions, and $\gamma$ is estimated less precisely. Increasing the cross-sectional dimension $N$ does not help to estimate $\gamma$ more precisely, as the fraction of switching units to total units remains unchanged. By contrast, increasing the sample size $T$ does help to estimate $\gamma$.

We further observe that classification accuracy increases with the distance between the cluster means and with the parameter $\gamma$. This is intuitive. In both cases there are less transitions between clusters, and the model needs to reassign units less often, leading to an improved classification accuracy. The classification accuracy remains approximately unchanged for different values of $N$, $T$, and circle radii. This is intuitive as these parameters do not influence the distance between the time-varying cluster means.

Table ?? reports our simulation results when the cluster means move in different directions (one clockwise and the other counter-clockwise). As a result, the cluster means are not equidistant, and most cluster transitions are concentrated around specific times.

We observe that the precision with which $\gamma$ can be estimated again depends on $\gamma$ itself. For low values of $\gamma$ the transition parameter is estimated more precisely. Comparing Tables 1 and ?? reveals that time-variation in the transition probabilities makes parameter estimation more challenging. Increasing the sample size $T$, however, is again beneficial.

We further observe that, for higher values of $\gamma$ (e.g., $\gamma = 0.75$), some problems appear when the unconditional distance between clusters is eight and the radius is four. The minimum (maximum) distance between the two simulated means is the circle distance minus (plus) two times the radius. As a result, the time-varying cluster means meet once. This is less of a problem as the model corrects wrong assignments over time when the distance between the cluster means increases. When circle radius and unconditional distance between clusters are both equal to four, the time-varying cluster means cross twice. The model has difficulties assigning each unit to its corresponding cluster when this occurs, and classification accuracy suffers as a result. Classification quality is lowest in cases when the cluster paths intersect, and best when the clusters are furthest apart.
Table 2: Simulation outcomes III: \( k \)-means clustering, ignoring cluster transitions

Average percentage of correct classification across simulation runs. The time-varying cluster means evolve clockwise implying time-invariant transition probabilities. Considered sample sizes are \( N = 100 \) and \( T = 20, 40 \). The transition intensity parameter is set to \( \gamma = 10^7 \), wrongly implying no transitions across clusters. Radius (rad.) refers to the radius of the true mean circles and is a measure of the signal-to-noise ratio. Distance (dist.) is the distance between circle centers and measures the distinctiveness of clusters. %cor. is the fraction of correct cluster assignments across simulation runs. diff is the difference with respect to the top panel of Table 1. We omit the case of \( N = 200 \) since the results are similar to the \( N = 100 \) case.

<table>
<thead>
<tr>
<th>DGP ( \gamma )</th>
<th>rad.</th>
<th>dist.</th>
<th>( N = 100, T = 20 )</th>
<th>( N = 100, T = 40 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0.25 )</td>
<td>1</td>
<td>4</td>
<td>0.64 -0.09</td>
<td>0.60 -0.13</td>
</tr>
<tr>
<td>( 0.25 )</td>
<td>1</td>
<td>8</td>
<td>0.73 -0.16</td>
<td>0.67 -0.21</td>
</tr>
<tr>
<td>( 0.25 )</td>
<td>4</td>
<td>4</td>
<td>0.64 -0.08</td>
<td>0.60 -0.12</td>
</tr>
<tr>
<td>( 0.25 )</td>
<td>4</td>
<td>8</td>
<td>0.73 -0.15</td>
<td>0.67 -0.21</td>
</tr>
<tr>
<td>( 0.50 )</td>
<td>1</td>
<td>4</td>
<td>0.73 -0.14</td>
<td>0.66 -0.21</td>
</tr>
<tr>
<td>( 0.50 )</td>
<td>1</td>
<td>8</td>
<td>0.92 -0.05</td>
<td>0.87 -0.11</td>
</tr>
<tr>
<td>( 0.50 )</td>
<td>4</td>
<td>4</td>
<td>0.73 -0.13</td>
<td>0.66 -0.23</td>
</tr>
<tr>
<td>( 0.50 )</td>
<td>4</td>
<td>8</td>
<td>0.93 -0.05</td>
<td>0.87 -0.11</td>
</tr>
<tr>
<td>( 0.75 )</td>
<td>1</td>
<td>4</td>
<td>0.84 -0.10</td>
<td>0.76 -0.19</td>
</tr>
<tr>
<td>( 0.75 )</td>
<td>1</td>
<td>8</td>
<td>0.99 -0.01</td>
<td>0.98 -0.02</td>
</tr>
<tr>
<td>( 0.75 )</td>
<td>4</td>
<td>4</td>
<td>0.83 -0.11</td>
<td>0.76 -0.18</td>
</tr>
<tr>
<td>( 0.75 )</td>
<td>4</td>
<td>8</td>
<td>0.99 -0.01</td>
<td>0.98 -0.02</td>
</tr>
</tbody>
</table>

Our approach allows for a dynamic allocation of units to clusters over time. We now verify whether this leads to an improved cluster assignment compared to a much simpler, static approach. For this purpose we compare our previous simulation results to the outcome of a \( k \)-means clustering approach based on averaged (over \( T \)) data. The static approach is misspecified since the true data is characterized by sample transitions.

Table 2 reports the simulation outcomes. The fraction of correctly assigned clusters is lower for the static model. For \( T = 20 \) we see that the percentage of correct cluster assignments is decent as long as cluster transitions are infrequent (\( \gamma = 0.75 \)) and the unconditional distance between the cluster centers is large. As the time horizon increases, and as cluster transitions become more frequent, the static model performs increasingly worse.
4 Empirical application to bank business models

4.1 Data

Our sample under study consists of $N = 299$ European banks, for which we consider quarterly bank-level accounting data from SNL Financial between 2008Q1 – 2018Q2, implying $T = 42$. Banks that underwent distressed mergers, were acquired, or ceased to operate for other reasons during that time, are excluded from the analysis. We assume that differences in the remaining banks’ business models can be characterized along six dimensions: size, complexity, risk profile, activities, geographical reach, and funding. We select a parsimonious set of $D = 12$ indicators to cover these six categories. Table 3 lists the respective indicators.

Our multivariate panel data is unbalanced in the time dimension. Missing values occur routinely because some banks are reporting at a quarterly frequency, while others report at an annual or semi-annual frequency. We remove such missing values by substituting the most recently available observation for that variable.

We consider banks at their highest level of consolidation. In addition, however, we also include large subsidiaries of bank holding groups in our analysis provided that a complete set of data is available in the cross-section. Most banks are located in the euro area (54%) and the European Union (E.U., 73%). European non-E.U. banks are located in Norway (12%), Switzerland (4%), and other countries (11%).

4.2 Model selection

We chose the number of clusters $J$ based on the analysis of cluster validation criteria and in line with common choices in the literature. Distance-based cluster validation indices, such as the Calinski-Harabasz index, Davies-Bouldin index, average silhouette index, and the Hardigan rule (see e.g. Peel and McLachlan (2000)) point to $J = 5$ or $J = 6$. Each of these take an extremum at these value. In practice, experts consider between four and up to more than ten different bank business models; see, for example, Ayadi et al. (2014) and Bankscope (2014, p. 299). The larger the number of groups, however, the harder the results are to interpret. With these considerations in mind, in line with related literature, and to be conservative, we choose $J = 6$ clusters for our subsequent empirical analysis.
### Table 3: Indicator variables

Bank-level panel data variables for the empirical analysis. We consider $D = 12$ indicator variables covering six different categories. The third column explains which transformation is applied to each indicator before the statistical analysis.

<table>
<thead>
<tr>
<th>Category</th>
<th>Variable</th>
<th>Transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Size</strong></td>
<td>1. Total assets</td>
<td>$\ln\left(\frac{\text{Total assets}}{\text{CET1 capital}}\right)$</td>
</tr>
<tr>
<td></td>
<td>2. CET1 capital (leverage)</td>
<td>$\ln\left(\frac{\text{Total assets}}{\text{CET1 capital}}\right)$</td>
</tr>
<tr>
<td><strong>Complexity</strong></td>
<td>3. Net loans to assets</td>
<td>$\frac{\text{Total loans - loan loss reserves}}{\text{Total assets}}$</td>
</tr>
<tr>
<td></td>
<td>4. Assets held for trading</td>
<td>$\frac{\text{Assets held for trading}}{\text{Total assets}}$</td>
</tr>
<tr>
<td></td>
<td>5. Derivatives held for trading</td>
<td>$\frac{\text{Derivatives held for trading}}{\text{Total assets}}$</td>
</tr>
<tr>
<td><strong>Risk profile</strong></td>
<td>6. Market vs. credit risks</td>
<td>$\frac{\text{Market risk}}{\text{Credit risk}}$</td>
</tr>
<tr>
<td><strong>Activities</strong></td>
<td>7. Share of net interest income</td>
<td>$\frac{\text{Net interest income}}{\text{Operating revenue}}$</td>
</tr>
<tr>
<td></td>
<td>8. Share of net fees &amp; commission income</td>
<td>$\frac{\text{Net fees and commissions}}{\text{Operating income}}$</td>
</tr>
<tr>
<td></td>
<td>9. Share of trading income</td>
<td>$\frac{\text{Trading income}}{\text{Operating income}}$</td>
</tr>
<tr>
<td></td>
<td>10. Retail orientation</td>
<td>$\frac{\text{Retail loans}}{\text{Retail and corporate loans}}$</td>
</tr>
<tr>
<td><strong>Geography</strong></td>
<td>11. Domestic loans ratio</td>
<td>$\frac{\text{Domestic loans}}{\text{Total loans}}$</td>
</tr>
<tr>
<td><strong>Funding</strong></td>
<td>12. Deposits to assets ratio</td>
<td>$\frac{\text{Total deposits}}{\text{Total assets}}$</td>
</tr>
</tbody>
</table>

**Note:** Total Assets are all assets owned by the company (SNL key field 131929). Net loans to assets are loans and finance leases, net of loan-loss reserves, as a percentage of all assets owned by the bank (226933). Assets held for trading are acquired principally for the purpose of selling in the near term (224997). Derivatives held for trading are derivatives with positive replacement values not identified as hedging or embedded derivatives (224997). Market risk and credit risk (248881, 248880) are reported by the company. P&L variables are expressed as percentages of operating revenue (248959) or operating income (249289). Retail loans are expressed as a percent of retail and corporate loans (226957). Domestic loans are in percent of total loans by geography (226960). The deposits-to-assets ratio is computed from the loans-to-deposits ratio (248919) and loans-to-asset ratio (226933). Total deposits comprise both retail and commercial deposits.
Table 4: Parameter estimates

Parameter estimates and cluster validation indices for different model specifications. Model M1 allows for time-varying means and covariance matrices but rules out transitions across groups \((\gamma^{-1} = 0)\). Model M2 allows for Markovian transitions across groups; see (3). Model M3 restricts M2 by ruling out transitory transitions that last less than five quarters \((P = 4\) inactive states); see Section 2.4.1. Model M4 allows differences in banks’ profitability (return on equity) between clusters to influence the Markov chain transition probabilities \(\Pi_t\) in addition to lagged cluster distances; see (25). M1 – M4 use the same initial means and covariance matrices. Standard errors in parentheses are constructed from the numerical second derivatives of the log-likelihood function. We also report two cluster validation indices: the Davis-Bouldin index (DBI; the smaller the better), and the Calinski-Harabasz index (CHI; the larger the better).

<table>
<thead>
<tr>
<th>M1</th>
<th>M2 (Markovian transitions)</th>
<th>M3 (non-Markovian transitions)</th>
<th>M4 (non-Markovian transitions II)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.894</td>
<td>0.850</td>
<td>0.813</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>A2</td>
<td>0.998</td>
<td>0.998</td>
<td>0.993</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>(\nu)</td>
<td>6.595</td>
<td>19.518</td>
<td>8.088</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.06)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>(\gamma)</td>
<td>-</td>
<td>1.369</td>
<td>1.503</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>(\beta)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>-</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>DBI</td>
<td>3.14</td>
<td>2.94</td>
<td>2.93</td>
</tr>
<tr>
<td>CHI</td>
<td>11.89</td>
<td>21.08</td>
<td>21.12</td>
</tr>
<tr>
<td>loglik</td>
<td>144,253.2</td>
<td>150,197.1</td>
<td>150,003.1</td>
</tr>
</tbody>
</table>

We proceed with a model based on a mixture of Student’s \(t\) distributions. This allows us to be robust to potential one-off effects and outliers in bank accounting ratios. In addition, we pool parameters \(A_1, A_2,\) and \(\nu\) across clusters and variables. As a result, \(\theta = (A_1, A_2, \nu, \gamma, \beta)' \in \mathbb{R}^5\) is of a manageable dimension.

Table 4 reports parameter estimates and the log-likelihood fit for four different specifications of our dynamic clustering model. Model M1 allows for time-varying means and covariance matrices, but rules out transitions across groups \((\gamma^{-1} = 0)\). Cluster transitions are then treated as joint outliers, leading to a low degrees-of-freedom parameter of \(\nu \approx 6.5\). Model M2 allows for Markovian transitions across groups in line with (3). The log-likelihood fit improves considerably as a result. The degrees-of-freedom parameter becomes less extreme as well.

The nonlinear model M2 may have a tendency, however, to treat one-off accounting windfalls as short-lived cluster transitions. Such short-lived transitions are hard to interpret economically as...
meaningful changes in banks’ business models. Model M3 restricts M2 by ruling out transitory transitions that last a year or less by requiring \( P = 4 \) inactive states; see Section 2.4.1. The decay parameter \( \gamma \) increases somewhat as a result, indicating fewer (short-lived) transitions. The degrees-of-freedom parameter \( \nu \) decreases to accommodate more frequent outlying observations. The insistence on inactive states is reflected in a noticeable drop in log-likelihood fit.

Finally, Model M4 extends M3 by allowing an additional explanatory variable to influence the transition probabilities \( \Pi_t \); see Section 2.4.2. We chose \( x_{jk,t} \) as the difference in probability-weighted return on equity (ROE) of banks allocated to clusters \( j \) and \( k \) at time \( t \). Specifically, let \( x_{jt} := \frac{\sum_N \hat{\tau}_{ij,t} \cdot \text{ROE}_{it}}{\sum_N \hat{\tau}_{ij,t}} \) be the filtered ROE for banks in cluster \( j \) at time \( t \). Then \( x_{jk,t} := x_{jt} - x_{kt} \) denotes the differences in ROE between clusters \( j \) and \( k \). As a by-product, this specification allows the transition matrix \( \Pi_t \) to become non-symmetric.

Model specification M4 is preferred in terms of log-likelihood fit, and also does well in terms of the non-parametric cluster validation indices (DBI). We therefore select M4 for the remainder of our empirical analysis. Using this specification, we combine model parsimony with the ability to explore a rich set of questions given the data at hand.

4.3 Bank business model groups

This section studies the different bank business models (strategic groups) implied by \( J = 6 \) different cluster densities. Specifically, we assign labels to the identified clusters to guide intuition and for ease of later reference. These labels are chosen in line with Figure 2 and the identities of the firms in each cluster. In addition, our labeling is approximately in line with the examples given in SSM (2016, p.10).

Figure 2 plots the cluster median estimates for each indicator variable and business model cluster. Web Appendix C.1 presents the filtered cluster-specific time-varying standard deviations \( \Sigma_{jt}(d, d) \) for variables \( d = 1, \ldots, D \). Specifically, we distinguish

(A) **Market-oriented universal banks** (17.9% of obs.; e.g. Barclays, Credit Suisse, Deutsche Bank, Royal Bank of Scotland.)

(B) **International diversified banks** (16.0% of bank-quarter observations; comprising firms such as BBVA, HSBC Holdings, ING Groep, Santander almost all of the time.)
Figure 2: Time-varying cluster medians

Filtered cluster medians for twelve indicator variables; see Table C.1 The cluster medians coincide with the cluster means unless the variable is transformed; see the last column of Table C.1 in Web Appendix C. The cluster mean estimates are based on a t-mixture model with $J = 6$ clusters and time-varying cluster means $y_{jt}$ and covariance matrices $\Sigma_{jt}$. We distinguish large diversified lenders (black line), market-funded universal banks (red line), fee-focused retail lenders (blue line), diversified X-border banks (green line), domestic diversified lenders (purple dashed line), and domestic retail lenders (green dashed line).
(C) **Fee-focused retail lenders** (8.7% of obs.; e.g. all subsidiaries of Caisse Regionale de Credit Agricole, Credit Lyonnais.)

(D) **International corporate lenders** (11.5 % of obs.; e.g. Bank of Cyprus, HSBC Bank Malta, Raiffeisen Bank International, ProCredit Holding.)

(E) **Domestic diversified lenders** (32.1% of obs.; e.g. Erste Bank Hungary, Jaeren Sparebank, Nordea Bank Danmark, Swedbank.)

(F) **Domestic retail lenders** (13.8% of obs; e.g. Berner Kantonalbank, Coventry Building Society, Helgeland Sparebank, Newcastle Building Society.)

**Market-oriented universal banks** (A: solid black line) comprise large and well-known institutions. Approximately half of operating revenue tends to come from interest-bearing assets such as loans and securities holdings. This leaves net fees & commissions as well as trading income as significant other sources. Market-oriented universal banks are the most leveraged (highest total-assets-to-CET1-capital ratio) firms at any time between 2008Q1 and 2018Q2. This is the case even though leverage decreases strongly for these firms from pre-crisis levels, from approximately 40 to 20; see panel 2 of Figure 2. Market-oriented universal banks hold the largest trading and derivative books, both in absolute terms and relative to total assets. Naturally, such large banks engage in significant cross-border activities: approximately 50% of loans are cross-border loans; see panel 11 of Figure 2.

**International diversified lenders** (B: solid red line) are large institutions that lend significantly across borders (approximately 30% on average) and approximately equally to retail and corporate clients. International diversified lenders also serve their corporate customers by trading securities and derivatives on their behalf, resulting in non-negligible trading and derivatives books. Funding is obtained from capital markets as well as customer deposits, as indicated by a moderate deposits-to-assets ratio.

**Fee-focused retail lenders** (C: solid blue line) achieve most of their income from fees and commissions despite lending almost exclusively to domestic retail customers. Such fees could e.g. be servicing fees associated with loans that are ultimately moved off these banks’ balance sheets. Banks in this group exhibit a high loans-to-assets ratio of approximately 80%, and receive significant non-deposit funding, e.g. from a parent company. All subsidiaries of Credit Agricole
(Caisse Regionale de Credit Agricole Mutuel) are located in this group.

**International corporate lenders** (D: solid green line) lend internationally and mainly to corporate clients. On average approximately one in two loans are arranged across borders. Net interest income accounts for approximately 70% of operating revenue, leaving fee and trading income as relatively less significant sources.

**Domestic diversified lenders** (E: dashed pink line) and **domestic retail lenders** (F: dashed green line) are relatively numerous and of a small to moderate size. Domestic diversified lenders and domestic retail lenders have much in common: Both types of banks display low leverage, suggesting they are well capitalized. Neither group holds significant amounts of securities or derivatives in trading portfolios. Approximately two-thirds of income comes from interest-bearing assets, making it the dominant source of income. Domestic diversified lenders differ from domestic retail lenders by their lower retail orientation, and their higher trading assets and market risk.

### 4.4 Convergence

Figure 2 suggests that banks may have become more similar over time in important dimensions. A decrease in financial sector diversity could in principle be problematic from a financial stability perspective. For example, the probability and severity of fire sales could increase if more and more banks adopt similar business strategies. Web Appendix D discusses this issue. European banks have become more similar in terms of size, leverage, and share of trading activities. Simultaneously, banks have become less similar in terms of their cross-border lending and funding choices. Arguably, the convergence takes place in variables that do not signal an immediate financial stability concern.

### 4.5 Cluster transitions

The HMM part of our dynamic clustering model allows us to study cluster transitions across business model groups in detail. The top panel of Figure 3 reports the fraction of firms that are estimated to have transitioned to another cluster at each $t$ between 2008Q2 and 2018Q2. A transition here refers to a change in the most-likely cluster (Bayes classifier). Transitions are more likely to take place at year-end. This is intuitive, as some banks report only annually. We do not observe an obvious time trend or a difference in transition intensity between crises (2008–2009, 2010 – 2012)
Figure 3: Timing and histogram of cluster transitions

Top panel: black bars indicate the fraction of firms that are estimated to transition at each time \( t \) between 2008Q2 and 2018Q2. The red horizontal line indicates the average transition frequency. Bottom left panel: Number of transitions per firm \( i = 1, \ldots, 299 \). Bottom right panel: histogram of cluster transitions. A transition refers to a change in the most-likely cluster (Bayes classifier).

and more benign times (2013 – 2018). On average, approximately 3% of the \( N = 299 \) banks transition each quarter.

The bottom left panel of Figure 3 reports the total number of transitions per firm \( i = 1, \ldots, 299 \). The bottom right panel of Figure 3 provides a histogram of firms’ transition counts. The total number of transitions per firm range between 0 and 9. More than half of the banks never transition (55%). If a certain bank transitions more than a few times, then that bank may be located between two or more clusters and is hard to classify as a result.

The top panel of Figure 4 plots the number of estimated transitions from cluster \( j \) (rows) to \( k \) (columns) at any time. The bottom panel of Figure 4 plots the total number of banks allocated to each cluster over time. Clusters B and F grow in popularity over time, while the remaining clusters shrink in size. The observed industry trends are in line with large banks becoming less reliant
Figure 4: Cluster transitions and popularity
Top panel: Number of transitions from cluster $j$ (rows) to cluster $k$ (columns) over time. Bottom panel: The number of banks $i$ allocated to cluster $j = 1, \ldots, 6$ at each time $t$ between 2008Q1 and 2018Q2.

on market funding and scaling back trading and market-making activities ($A \rightarrow B$), domestically-active banks lending relatively more to retail clients rather than to corporate clients ($D \rightarrow B$, $D \rightarrow$
F, E → F), and banks relying progressively more on fee income, possibly to lean against a lower profitability from increasingly low interest rates (D → C, D → F).

The cluster transitions underlying Figures 3 – 4 are in part explained by differences in bank profitability across clusters; see Section 4.2. Web Appendix C reports and discusses the evolution of return on equity (ROE) per bank cluster over time, where bank-specific ROEs are weighted by the filtered probability that bank $i$ belongs to cluster $j$ at time $t$. ROE for European banks is usually positive and varies between approximately -2% and 10% over time. The international corporate lenders in cluster $D$ are an exception in that their ROE turns negative at onset of the euro area sovereign debt crisis in mid-2010, and remains negative until the end of the sample.

5 Conclusion

We proposed a novel observation-driven model for the dynamic clustering of multivariate panel data. The cluster means and covariance matrices can be time-varying to track gradual changes in cluster characteristics over time. In addition, the units of interest can transition between clusters based on a Hidden Markov model (HMM) with time-varying transition probabilities that are in turn related to lagged cluster distances and/or economic variables.

We applied the model to a sample of 299 European banks between 2008Q1 and 2018Q2. Our empirical results suggest that European banks have become more similar over time in some key characteristics. In addition, we find a moderate transition intensity between clusters that is related to differences in bank profitability.

References


