Dynamic clustering of multivariate panel data*

André Lucas,[(a)] *Julia Schaumburg,*[(a)] *Bernd Schwaab,*[(b)]

[(a)] Vrije Universiteit Amsterdam and Tinbergen Institute

[(b)] European Central Bank, Financial Research

***** preliminary, incomplete, unpolished *****

February 2019

Abstract

We propose a novel observation-driven model for the dynamic clustering of multivariate panel data. The model is dynamic in three 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 matrix depends on lagged cluster distances. Monte Carlo experiments suggest that the units can be classified reliably in a variety of settings. Classification results can be poor, however, if cluster transitions are present but ignored. An empirical study of 312 European banks between 2008Q1–2018Q2 suggests a gradual convergence of some bank business model characteristics as well as a moderate number of transitions between clusters.

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 depend on lagged cluster parameters and are thus time-varying as well. We argue that our modeling framework is useful to robustly allocate a potentially large number of units into approximately homogeneous groups, while keeping track of group memberships over time as well as group transitions. We extend the model to handle non-Markovian transitions as well as the availability of explanatory covariates. Monte Carlo experiments suggest that the units can be classified reliably in a variety of relevant settings. Parameter estimates are biased, however, and classification results can be poor if cluster transitions are present but ignored.

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 cluster mean and covariance matrix parameters as well as for all 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.

We apply our modeling framework to a multivariate panel of $N = 312$ 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-crises period between 2013–2018 characterized by a significant increase in financial regulation as well as ultra-low interest rates.

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

We then 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, assets and derivatives held for trading, risk profile, and retail orientation. The timing of the decline in diversity points towards increased financial sector regulation, as well as increased competition at low interest rates, as possible causes.

Second, we study which business model groups have become more or less popular (populated) over time. The estimated transitions suggest that banks have become less reliant on market funding, rely increasingly on fee income to lean against low interest rates, and have lent increasingly to retail clients. Taken together, banks appear to have moved towards traditional 'bread-and-butter' business strategies. Finally, we study whether bank business model transitions can be explained by profitability (return-on-assets) differences at the group level.

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.

Fruehwirth-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). 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. Instead, we consider a panel framework, which allows us to pool information over time, leading to a more accurate assessment.

We proceed as follows. Section 2 presents a 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 $\boldsymbol{y}_{it} \in \mathbb{R}^{D\times 1}$, where $\boldsymbol{y}_{it}$ 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 $\boldsymbol{y}_{it}$ by the mixture model

$$\boldsymbol{y}_{it} = \boldsymbol{\mu}_{c_{it},t} + \boldsymbol{\epsilon}_{it}, \quad \boldsymbol{\epsilon}_{it} \sim f_{\boldsymbol{\epsilon}}\left(0, \boldsymbol{\Sigma}_{c_{it},t}, \nu_{c_{it}}\right), \tag{1}$$

where $\boldsymbol{\mu}_{c_{it},t}$ is a $D \times 1$ vector of cluster-specific means, and ϵ_{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 $\nu_{c_{it}}$. For example, if f_ϵ 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 $\boldsymbol{\mu}_{jt}$ and disturbance vectors $\boldsymbol{\epsilon}_{it}$ 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.

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

We introduce the short-hand notation $\pi_{jk,t} := \mathbb{P}\left\{c_{i,t+1} = k | c_t = j\right\}$, 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 $\mathbf{\Pi}_t$ contains all transition probabilities $\pi_{jk,t}$ for $j, k = 1, \ldots, J$. We require the rows of $\mathbf{\Pi}_t$ 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

$$\mathbf{\Pi}_t = \mathbf{\Pi}_t\left(\mathcal{D}_{t-1}\right), \tag{2}$$

where $\mathcal{D}_t$ 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 $\mathbf{\Pi}_t$ are then decreasing in $d_{jk,t}$. To avoid an undue increase in the number of parameters, we parsimoniously model the transition probabilities as

$$\pi_{jk,t} = \frac{\exp\left(-\gamma d_{jk,t-1}\right)}{\Sigma_{q=1}^{J} \exp\left(-\gamma d_{jq,t-1}\right)} \quad \text{for } j, k = 1, \ldots, J, \tag{3}$$

where the scalar parameter γ 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 γ. A higher value for γ 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 γ leads to higher transition probabilities. Finally, the multinomial specification in (3) ensures that the rows of $\mathbf{\Pi}_t$ sum to one by construction.

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

$$d_{jk,t} = \sqrt{(\boldsymbol{\mu}_{jt} - \boldsymbol{\mu}_{kt})' \, \overline{\boldsymbol{\Sigma}}_t^{-1} \, (\boldsymbol{\mu}_{jt} - \boldsymbol{\mu}_{kt})}, \tag{4}$$

where $\overline{\boldsymbol{\Sigma}}_t = J^{-1} \sum_{j=1}^{J} \boldsymbol{\Sigma}_{jt}$ 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. The Euclidian distance is obtained as a special case by setting $\overline{\boldsymbol{\Sigma}}_t \equiv \mathrm{I}_D$.

2.2 Time-varying conditional cluster probabilities

This section derives a filtering equation for the conditional probability $\tau_{ij,t|t} := \mathbb{P}[c_{it} = j | \mathcal{F}_t; \boldsymbol{\theta}]$, where $\tau_{ij,t|t}$ denotes the probability that unit i belongs to cluster j at time t given the information set $\mathcal{F}_t$, where $\mathcal{F}_t$ contains observations up to t, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t$. The vector $\boldsymbol{\theta}$ contains the static parameters of the model that need to be estimated.

We start by considering the log-likelihood contribution of observation $\boldsymbol{y}_{it}$,

$$\ell_{it} = \log \left(f \left(\boldsymbol{y}_{it} | \mathcal{F}_{t-1}; \boldsymbol{\theta} \right) \right) = \log \left(\sum_{j=1}^{J} \tau_{ij,t|t-1} f \left(\boldsymbol{y}_{it} | c_{it} = j, \mathcal{F}_{t-1}; \boldsymbol{\theta} \right) \right) \tag{5}$$

where $f \left(\boldsymbol{y}_{it} | c_{it} = j, \mathcal{F}_{t-1}; \boldsymbol{\theta} \right)$ is the density of $\boldsymbol{y}_{it}$ in cluster j, such as the multivariate Student's t or normal density, and $\tau_{ij,t|t-1} := \mathbb{P}[c_{it} = j | \mathcal{F}_{t-1}; \boldsymbol{\theta}]$ is the conditional probability that unit i belongs to cluster j at time t given $\mathcal{F}_{t-1}$. 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 $\boldsymbol{\Pi}_t$. We use this property to update the cluster probabilities as

$$\tau_{ij,t+1|t} = \mathbb{P}[c_{i,t+1} = j | \mathcal{F}_t; \boldsymbol{\theta}] = \sum_{k=1}^{J} \pi_{kj,t} \mathbb{P}[c_{it} = k | \mathcal{F}_t; \boldsymbol{\theta}] = \sum_{k=1}^{J} \tau_{ik,t|t} \pi_{kj,t}. \tag{6}$$

Using a standard Bayes argument, the filtered cluster probabilities are determined by

$$\begin{aligned}\tau_{ij,t|t} = \mathbb{P}[c_{it} = j|\mathcal{F}_t;\boldsymbol{\theta}] &= \frac{\tau_{ij,t|t-1} f(\boldsymbol{y}_{it}|c_{it} = j, \mathcal{F}_{t-1};\boldsymbol{\theta})}{f(\boldsymbol{y}_{it}|\mathcal{F}_{t-1};\boldsymbol{\theta})} \\ &= \frac{\tau_{ij,t|t-1} f(\boldsymbol{y}_{it}|c_{it} = j, \mathcal{F}_{t-1};\boldsymbol{\theta})}{\tau_{i1,t|t-1} f(\boldsymbol{y}_{it}|c_{it} = 1, \mathcal{F}_{t-1};\boldsymbol{\theta}) + \ldots + \tau_{iJ,t|t-1} f(\boldsymbol{y}_{it}|c_{it} = J, \mathcal{F}_{t-1};\boldsymbol{\theta})}.\end{aligned} \tag{7}$$

The filtered cluster probabilities thus update the predicted cluster probabilities $\tau_{ij,t|t-1}$ by using the time t observation $\boldsymbol{y}_{it}$ and its likelihood of coming from the cluster j density $f(\boldsymbol{y}_{it}|c_{it} = j, \mathcal{F}_{t-1};\boldsymbol{\theta})$, normalized by the unconditional data density $f(\boldsymbol{y}_{it}|\mathcal{F}_{t-1};\boldsymbol{\theta})$. This is intuitive: if $\tau_{ij,t|t-1}\, f(\boldsymbol{y}_{it}|c_{it} = j, \mathcal{F}_{t-1};\boldsymbol{\theta})$ is high compared to $\tau_{ik,t|t-1}\, f(\boldsymbol{y}_{it}|c_{it} = k, \mathcal{F}_{t-1};\boldsymbol{\theta})$ for $k \neq j$, then $\boldsymbol{y}_{it}$ 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

For the time-varying means, we use score-driven dynamics as introduced by Creal et al. (2013). We impose further parsimony by using the exponentially weighted score-driven dynamics of Lucas and Zhang (2016), such that

$$\boldsymbol{\mu}_{j,t+1} = \boldsymbol{\mu}_{jt} + \boldsymbol{A}_1 \boldsymbol{S}_{\mu_{jt},t} \cdot \boldsymbol{\nabla}_{\mu_{jt},t}, \tag{8}$$

where the diagonal matrix $\boldsymbol{A}_1 = \boldsymbol{A}_1(\boldsymbol{\theta})$ depends on the vector of unknown static parameters $\boldsymbol{\theta}$, $\boldsymbol{S}_{\mu_{jt},t}$ is a scaling matrix, and the score $\boldsymbol{\nabla}_{\mu_{jt},t}$ is the first derivative of the log-density of $\boldsymbol{y}_{it}$ with respect to $\boldsymbol{\mu}_{jt}$. In our case, the score is given by

$$\begin{aligned}\boldsymbol{\nabla}_{\mu_{jt},t} &= \frac{\partial \ell_t}{\partial \boldsymbol{\mu}_{jt}} = \frac{\partial \left[\Sigma_{i=1}^N \log\left(f\left(\boldsymbol{y}_{it}|\mathcal{F}_{t-1};\boldsymbol{\theta}\right)\right)\right]}{\partial \boldsymbol{\mu}_{jt}} \\ &= \sum_{i=1}^N \frac{\partial}{\partial \boldsymbol{\mu}_{jt}} \log\left(\sum_{j=1}^J \tau_{ij,t|t-1} f\left(\boldsymbol{y}_{it}|c_{it}=j,\mathcal{F}_{t-1};\boldsymbol{\theta}\right)\right) \\ &= \sum_{i=1}^N \tau_{ij,t|t} \cdot \frac{\partial}{\partial \boldsymbol{\mu}_{jt}} \log f\left(\boldsymbol{y}_{it}|c_{it}=j,\mathcal{F}_{t-1};\boldsymbol{\theta}\right) = \sum_{i=1}^N \tau_{ij,t|t} \cdot \boldsymbol{\nabla}^{(j)}_{\mu_{jt},t}, \end{aligned} \tag{9}$$

where $\boldsymbol{\nabla}^{(j)}_{\mu_{jt},t} = \partial \log f\left(\boldsymbol{y}_{it}|c_{it}=j,\mathcal{F}_{t-1};\boldsymbol{\theta}\right)/\partial\boldsymbol{\mu}_{jt}$ is the score of mixture component j. As a closed form expression for the conditional Fisher information matrix of $\boldsymbol{\mu}_{jt}$ is not available, we use an approximation to account for the curvature of the score, namely

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

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 $\boldsymbol{y}_{it}$ coming from regime j.

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

$$\boldsymbol{\nabla}^{(j)}_{\mu_{jt},t} = \boldsymbol{\Sigma}_{jt}^{-1}(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}), \qquad \boldsymbol{S}_{\mu_{jt},t} = \left(\sum_{i=1}^N \tau_{ij,tt}\boldsymbol{\Sigma}_{jt}^{-1}\right)^{-1}, \tag{11}$$

$$\boldsymbol{\mu}_{j,t+1} = \boldsymbol{\mu}_{jt} + \boldsymbol{A}_1 \frac{\sum_{i=1}^N \tau_{ij,t|t}\cdot(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})}{\sum_{i=1}^N \tau_{ij,t|t}}. \tag{12}$$

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 $\boldsymbol{y}_{it}$ comes from cluster j with negligible probability, then the update of μ_{jt} does not depend on $\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}$.

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

$$\boldsymbol{\nabla}^{(j)}_{\mu_{jt},t} = w_{ij,t} \cdot \boldsymbol{\Sigma}_{jt}^{-1} \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt}\right), \tag{13}$$

where the weight $w_{ij,t} = (1 + \nu_j^{-1}\, D) \,\big/ \big(1 + \nu_j^{-1}\, (\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt})' \, \boldsymbol{\Sigma}_{jt}^{-1} \, (\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt})\big)$ provides the model with a robustness feature: observations $\boldsymbol{y}_{it}$ 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

$$\boldsymbol{\mu}_{j,t+1} = \boldsymbol{\mu}_{jt} + \boldsymbol{A}_1 \frac{\sum_{i=1}^{N} \tau_{ij,t|t} \cdot w_{ij,t} \cdot (\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt})}{\sum_{i=1}^{N} \tau_{ij,t|t}}. \tag{14}$$

The Gaussian transition equation (12) is obtained as as special case of (14) as $\nu^{-1} \to 0$ 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 $\boldsymbol{\Sigma}_{jt}$. Following the exponentially weighted score-driven dynamics of Lucas and Zhang (2016), it is given by

$$\text{vec}(\boldsymbol{\Sigma}_{j,t+1}) = \text{vec}(\boldsymbol{\Sigma}_{jt}) + \boldsymbol{A}_2 \, \boldsymbol{S}_{\Sigma_{jt},t} \cdot \boldsymbol{\nabla}_{\Sigma_{jt},t}, \tag{15}$$

where matrix $\boldsymbol{A}_2 = \boldsymbol{A}_2(\boldsymbol{\theta})$ depends on parameters to be estimated, $\boldsymbol{S}_{\Sigma_{jt},t}$ is a scaling matrix, and $\boldsymbol{\nabla}_{\Sigma_{jt},t}$ is the score. The score dynamics are determined in the same way as for the time-varying cluster means. The score is given by

$$\begin{aligned} \boldsymbol{\nabla}_{\Sigma_{jt},t} &= \tfrac{1}{2} \frac{\partial \ell_t}{\partial \text{vec}(\boldsymbol{\Sigma}_{jt})} = \tfrac{1}{2} \frac{\partial \sum_{i=1}^{N} \log f\left(\boldsymbol{y}_{it} | \mathcal{F}_{t-1}; \boldsymbol{\theta}\right)}{\partial \text{vec}(\boldsymbol{\Sigma}_{jt})} \\ &= \tfrac{1}{2} \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \frac{\partial \log f\left(\boldsymbol{y}_{it} | c_{it} = j, \mathcal{F}_{t-1}; \boldsymbol{\theta}\right)}{\partial \text{vec}(\boldsymbol{\Sigma}_{jt})} = \tfrac{1}{2} \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \boldsymbol{\nabla}^{(j)}_{\Sigma_{jt},t}, \end{aligned} \tag{16}$$

where $\boldsymbol{\nabla}^{(j)}_{\Sigma_{jt},t} = \partial \log f\left(\boldsymbol{y}_{it}|c_{it}=j, \mathcal{F}_{t-1};\boldsymbol{\theta}\right)/\partial \mathrm{vec}(\boldsymbol{\Sigma}_{jt})$. For the scaling matrix, we can take the analogous expression as in (10) and consider

$$\boldsymbol{S}_{\Sigma_{jt},t} = \left(\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \mathbb{E}\left[\boldsymbol{\nabla}^{(j)}_{\Sigma_{jt},t}\,(\boldsymbol{\nabla}^{(j)}_{\Sigma_{jt},t})' \,\middle|\, c_{it}=j, \mathcal{F}_{t-1};\boldsymbol{\theta}\right]\right)^{-1} \tag{17}$$

$$= \left(\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \mathbb{E}\left[-\partial\boldsymbol{\nabla}^{(j)}_{\Sigma_{jt},t}/\partial \mathrm{vec}(\boldsymbol{\Sigma}_{jt})' \,\middle|\, c_{it}=j, \mathcal{F}_{t-1};\boldsymbol{\theta}\right]\right)^{-1}. \tag{18}$$

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

$$\boldsymbol{\nabla}^{(j)}_{\Sigma_{jt},t} = \tfrac{1}{2}\,\mathrm{vec}\left(\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \boldsymbol{\Sigma}_{jt}^{-1}\,\left((\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})' - \boldsymbol{\Sigma}_{jt}\right)\boldsymbol{\Sigma}_{jt}^{-1}\right)$$

$$= \tfrac{1}{2}\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \left(\boldsymbol{\Sigma}_{jt}^{-1} \otimes \boldsymbol{\Sigma}_{jt}^{-1}\right)\mathrm{vec}\left((\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})' - \boldsymbol{\Sigma}_{jt}\right), \tag{19}$$

$$\boldsymbol{S}_{\Sigma_{jt},t} = \left(\tfrac{1}{2}\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \boldsymbol{\Sigma}_{jt}^{-1} \otimes \boldsymbol{\Sigma}_{jt}^{-1}\right)^{-1}, \tag{20}$$

$$\mathrm{vec}(\boldsymbol{\Sigma}_{j,t+1}) = \mathrm{vec}(\boldsymbol{\Sigma}_{jt}) + \boldsymbol{A}_2\,\frac{\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \mathrm{vec}\left((\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})' - \boldsymbol{\Sigma}_{jt}\right)}{\sum_{i=1}^{N} \tau_{ij,t|t}}. \tag{21}$$

Unvectorizing (21), we obtain

$$\boldsymbol{\Sigma}_{j,t+1} = \boldsymbol{\Sigma}_{jt} + \boldsymbol{A}_2\frac{\sum_{i=1}^{N} \tau_{ij,t|t}\left[(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})\,(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})' - \boldsymbol{\Sigma}_{jt}\right]}{\sum_{i=1}^{N} \tau_{ij,t|t}}. \tag{22}$$

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

$$\boldsymbol{\nabla}^{(j)}_{\Sigma_{jt},t} = \tfrac{1}{2}\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \left(\boldsymbol{\Sigma}_{jt}^{-1} \otimes \boldsymbol{\Sigma}_{jt}^{-1}\right)\mathrm{vec}\left(w_{ij,t}(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})' - \boldsymbol{\Sigma}_{jt}\right), \tag{23}$$

where $w_{ij,t}$ is defined below (13). Pre-multiplying the score by the approximate scaling matrix

(20) yields the transition equation

$$\boldsymbol{\Sigma}_{j,t+1} = \boldsymbol{\Sigma}_{jt} + \boldsymbol{A}_2 \frac{\sum_{i=1}^{N} \tau_{ij,t|t} \left[w_{ij,t} \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right) \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right)' - \boldsymbol{\Sigma}_{jt} \right]}{\sum_{i=1}^{N} \tau_{ij,t|t}}, \tag{24}$$

where the Gaussian case (22) is again obtained as a special case of (24) as $\nu^{-1} \to 0$ 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 $\boldsymbol{\mu}_{j1}$, and the cluster covariance matrices $\boldsymbol{\Sigma}_{j1}$ need to be initialized to start the filtering recursions. We initialize by k-means clustering; see e.g. Hartigan and Wong (1979). For this purpose we use data of $t = 1$ only, $\boldsymbol{y}_{i1}$ for $i = 1, \ldots, N$. The k-means algorithm 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 $\tau^{k}_{ij,1|1}$. These probabilities are one for the assigned cluster, and zero for the remaining clusters. Based on these initial cluster assignments, the initial cluster means $\boldsymbol{\mu}_{j1}$ equal the sample average of $\boldsymbol{y}_{i1}$ for units $i = 1, \ldots, N$ for which $\tau^{k}_{ij,1|1}$ equals 1. The initialized covariance matrices $\boldsymbol{\Sigma}_{j1}$ are similarly determined as the covariance of $\boldsymbol{y}_{i1}$ for units i in cluster j.

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

2.4 Extensions

2.4.1 Non-Markovian transitions

We expect business model choices to be highly persistent over time. 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 is not explicitly enforced in the current model set-up. Particularly if two clusters are close at a particular moment in time, the probability of swithing from business model A to model

B can be large. Due to the symmetry, the probability of switching back from B to A is then large as well.

In order to accommodate the persistence of business model choices better, 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 mot 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_{itp} 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 $i = (i, 0)$ 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 altered transition probability matrix (from row j to column k)

$$
\begin{array}{cc}
 & \textit{To state } (j,p)\textit{:} \\
\textit{From state } (i,p)\textit{:} & \begin{array}{cccccc} \textit{(1,0)} & \textit{(1,1)} & \textit{(1,2)} & \textit{(2,0)} & \textit{(2,1)} & \textit{(2,2)} \end{array} \\
\begin{array}{c} \textit{(1,0)} \\ \textit{(1,1)} \\ \textit{(1,2)} \\ \textit{(2,0)} \\ \textit{(2,1)} \\ \textit{(2,2)} \end{array} &
\begin{pmatrix}
\pi_{11,t} & 0 & 0 & 0 & 0 & \pi_{12,t} \\
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & \pi_{21,t} & \pi_{22,t} & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0
\end{pmatrix}
\end{array}.
$$

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

The transition probabilities (3) can be extended further to include contemporaneous or lagged variables as additional conditioning variables. For example, banks from low (high) profitability clusters could have a higher (lower) incentive to leave that cluster; see e.g. 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 β a vector of unknown coefficients that need to be estimated. The transition probabilities can then be modeled as

$$\pi_{jk,t} = \frac{\exp\left(-\gamma d_{jk,t-1} + \beta' X_{jk,t}\right)}{\Sigma_{q=1}^{J} \exp\left(-\gamma d_{jq,t-1} + \beta' X_{jq,t}\right)} \quad \text{for } j, k = 1, \ldots, J, \tag{25}$$

where γ 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 in the previous subsections are attractive because the log-likelihood is known in closed form. As a result, parameter estimates can be obtained in a standard way by numerically maximizing the likelihood function. This is a key advantage over the alternative class of parameter-driven models, as considered in for instance Koopman et al. (2011, 2012) and Azizpour et al. (2017). For parameter-driven models, the log-likelihood function is typically not available in closed form and parameter estimation is computationally more cumbersome.

For a given set of observations $y_1, \ldots, y_T$, the vector of unknown parameters

$$\boldsymbol{\theta} = \{\mathrm{vec}(\boldsymbol{A}_1)', \mathrm{vec}(\boldsymbol{A}_2)', \nu_1, \ldots, \nu_J, \gamma\}'$$

can be estimated by maximizing the log-likelihood function with respect to $\boldsymbol{\theta}$, that is

$$\mathcal{L}\left(\boldsymbol{\theta}|\mathcal{F}_T\right) = \sum_{t=1}^{T}\sum_{i=1}^{N}\ell_{it}, \tag{26}$$

where ℓ_{it} is defined in (5). The evaluation of ℓ_{it} is easily incorporated in the filtering process for the latent states. The maximization of (26) can be carried out using a conveniently chosen quasi-Newton optimization method.

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.

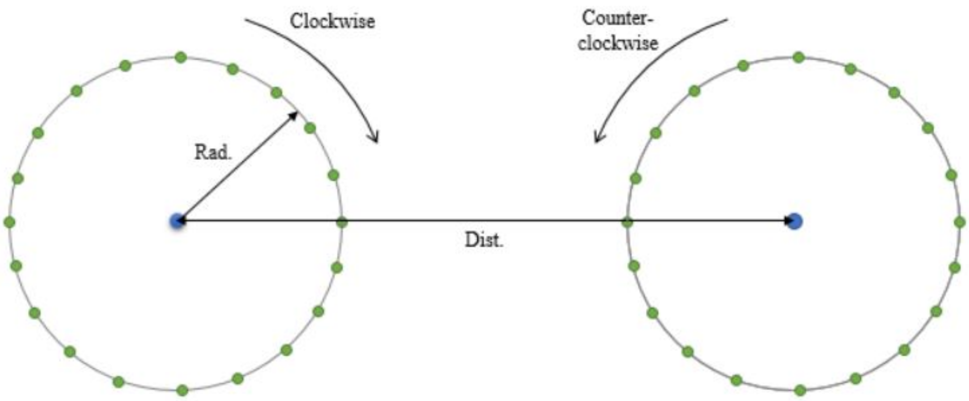
Specifically, we generate data from two clusters located around two different time-varying cluster means. The time-varying means move in two non-overlapping circles over time. Our baseline setting is visualized in Figure 1. Key inputs into our simulations are the transition intensity parameter γ in (3),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

Figure 1: Illustration of DGP: two clusters with time-varying means

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.



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 γ, 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 γ 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 γ 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. $A_1 = a_1 \mathrm{I}_D$.

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 γ can be estimated depends on γ itself. For low values of γ (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 γ (e.g., $\gamma = 0.75$) the model implies fewer transitions, and γ is estimated less precisely. Increasing the cross-sectional dimension N does not help to estimate γ more precisely, as the fraction of switching units to total units remains unchanged. By contrast, increasing the sample size T does help to estimate γ.

We further observe that classification accuracy increases with the distance between the cluster means and with the parameter γ. 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 2 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 γ can be estimated again depends on γ itself. For low values of γ the transition parameter is estimated more precisely. Comparing Tables 1 and 2 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 γ (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

Table 1: Simulation outcomes I: time-invariant transition probabilities

Parameter estimates and average percentage of correct classification across simulation runs. 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. Considered sample sizes are $N = 100, 200$ and $T = 20, 40$. The transition intensity parameter γ determines the frequency of transitions; lower values of γ imply a higher number of transitions in expectation. 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.

DGP			$N = 100, T = 20$		$N = 100, T = 40$	
γ	$rad.$	$dist.$	$\hat{\gamma}$	%cor.	$\hat{\gamma}$	%cor.
0.25	1	4	0.25	0.73	0.25	0.73
0.25	1	8	0.25	0.89	0.25	0.88
0.25	4	4	0.25	0.72	0.25	0.72
0.25	4	8	0.25	0.88	0.25	0.88
0.50	1	4	0.50	0.87	0.50	0.87
0.50	1	8	0.50	0.98	0.50	0.98
0.50	4	4	0.52	0.86	0.50	0.89
0.50	4	8	0.50	0.98	0.50	0.98
0.75	1	4	0.74	0.94	0.75	0.95
0.75	1	8	0.77	1.00	0.75	1.00
0.75	4	4	0.79	0.94	0.75	0.94
0.75	4	8	0.77	1.00	0.75	1.00

DGP			$N = 200, T = 20$		$N = 200, T = 40$	
γ	$rad.$	$dist.$	$\hat{\gamma}$	%cor.	$\hat{\gamma}$	%cor.
0.25	1	4	0.25	0.73	0.25	0.73
0.25	1	8	0.25	0.89	0.25	0.88
0.25	4	4	0.26	0.72	0.25	0.73
0.25	4	8	0.25	0.89	0.25	0.88
0.50	1	4	0.50	0.87	0.50	0.87
0.50	1	8	0.50	0.98	0.50	0.98
0.50	4	4	0.52	0.86	0.50	0.87
0.50	4	8	0.50	0.98	0.50	0.98
0.75	1	4	0.75	0.95	0.75	0.95
0.75	1	8	0.75	1.00	0.75	1.00
0.75	4	4	0.80	0.94	0.75	0.94
0.75	4	8	0.76	1.00	0.75	1.00

Table 2: Simulation outcomes II: time-varying transition probabilities

Parameter estimates and average percentage of correct classification across simulation runs. 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 $N = 100, 200$ and $T = 20, 40$. The transition intensity parameter γ determines the frequency of transitions; lower values of γ imply a higher number of transitions in expectation. 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.

DGP			$N = 100, T = 20$		$N = 100, T = 40$	
γ	$rad.$	$dist.$	$\hat{\gamma}$	%cor.	$\hat{\gamma}$	%cor.
0.25	1	4	0.25	0.72	0.25	0.72
0.25	1	8	0.25	0.88	0.25	0.88
0.25	4	4	0.23	0.65	0.25	0.66
0.25	4	8	0.28	0.79	0.25	0.80
0.50	1	4	0.49	0.85	0.50	0.84
0.50	1	8	0.50	0.98	0.50	0.98
0.50	4	4	0.38	0.69	0.48	0.69
0.50	4	8	0.54	0.85	0.51	0.86
0.75	1	4	0.73	0.91	0.74	0.91
0.75	1	8	0.77	1.00	0.76	1.00
0.75	4	4	0.44	0.70	0.63	0.68
0.75	4	8	0.65	0.87	0.73	0.88

DGP			$N = 200, T = 20$		$N = 200, T = 40$	
γ	$rad.$	$dist.$	$\hat{\gamma}$	%cor.	$\hat{\gamma}$	%cor.
0.25	1	4	0.25	0.72	0.25	0.72
0.25	1	8	0.25	0.88	0.25	0.88
0.25	4	4	0.23	0.65	0.25	0.66
0.25	4	8	0.27	0.80	0.25	0.80
0.50	1	4	0.49	0.85	0.50	0.84
0.50	1	8	0.50	0.98	0.50	0.98
0.50	4	4	0.40	0.69	0.48	0.69
0.50	4	8	0.49	0.85	0.49	0.86
0.75	1	4	0.73	0.91	0.74	0.91
0.75	1	8	0.76	1.00	0.75	1.00
0.75	4	4	0.42	0.69	0.63	0.68
0.75	4	8	0.63	0.87	0.72	0.88

Table 3: 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.

DGP			$N = 100, T = 20$		$N = 100, T = 40$	
γ	$rad.$	$dist.$	%cor.	diff.	%cor.	diff.
0.25	1	4	0.64	-0.09	0.60	-0.13
0.25	1	8	0.73	-0.16	0.67	-0.21
0.25	4	4	0.64	-0.08	0.60	-0.12
0.25	4	8	0.73	-0.15	0.67	-0.21
0.50	1	4	0.73	-0.14	0.66	-0.21
0.50	1	8	0.92	-0.05	0.87	-0.11
0.50	4	4	0.73	-0.13	0.66	-0.23
0.50	4	8	0.93	-0.05	0.87	-0.11
0.75	1	4	0.84	-0.10	0.76	-0.19
0.75	1	8	0.99	-0.01	0.98	-0.02
0.75	4	4	0.83	-0.11	0.76	-0.18
0.75	4	8	0.99	-0.01	0.98	-0.02

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.

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 3 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

The sample under study consists of $N = 312$ European banks, for which we consider quarterly bank-level accounting data from SNL Financial between 2008Q1 and 2018Q2, implying $T = 42$. We assume that differences in banks' business models can be characterized along six dimensions: size, complexity, risk profile, activities, geographical reach, and funding strategy. We select a parsimonious set of $D = 12$ indicators from these six categories. We consider banks' total assets, leverage with respect to CET1 capital [size], net loans to assets ratio, assets held for trading, derivatives held for trading [complexity], ratio of self-reported market risk to credit risk [risk profile], share of net interest income, share of net fees & commissions income, share of trading income, and ratio of retail loans to total loans [activities], ratio of domestic loans to total loans [geography], and the loans to deposits ratio [funding].

We refer to Web Appendix C for a detailed description of our data, including data transformations and SNL Financial field keys. Web Appendix C also discusses our treatment of missing observations and banks' country location.

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, and the average silhouette index (see e.g. Peel and McLachlan (2000)) suggest $J = 6$. Each of these take a local maxium/minimum at this 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.

We make two additional empirical choices. First, we proceed with a model based on a Student's t mixture distribution. Robust models based on fat-tailed mixtures are most appropriate for the fat-tailed bank accounting ratios in our empirical sample. Thin-tailed mixtures have a tendency to treat outlying observations as short-lived cluster transitions which are hard to interpret as salient changes in banks' business models. We treat ν as a robustness parameter and fix it at $\nu = 5$. This allows us to focus on fewer but more economically meaningful cluster transitions. Second, we pool parameters A_1, A_2, and ν across clusters and variables. We checked that such pooling leads to only a minor loss in log-likelihood fit. As a result, $\theta = \left(A_1, A_2, \gamma\right)' \in \mathbb{R}^3$ is of a manageable dimension. Using this parameter specification, we combine model parsimony with the ability to study a high-dimensional array of data.

4.3 Bank business model groups

This section studies the different business models implied by the $J = 6$ different cluster densities. Specifically, we assign labels to the identified cluster to guide intuition and for ease of 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. We distinguish

(A) **International diversified lenders** (16.0% of bank-quarter observations; comprising firms such as BBVA, HSBC Holdings, ING Groep, Santander almost all of the time.)

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

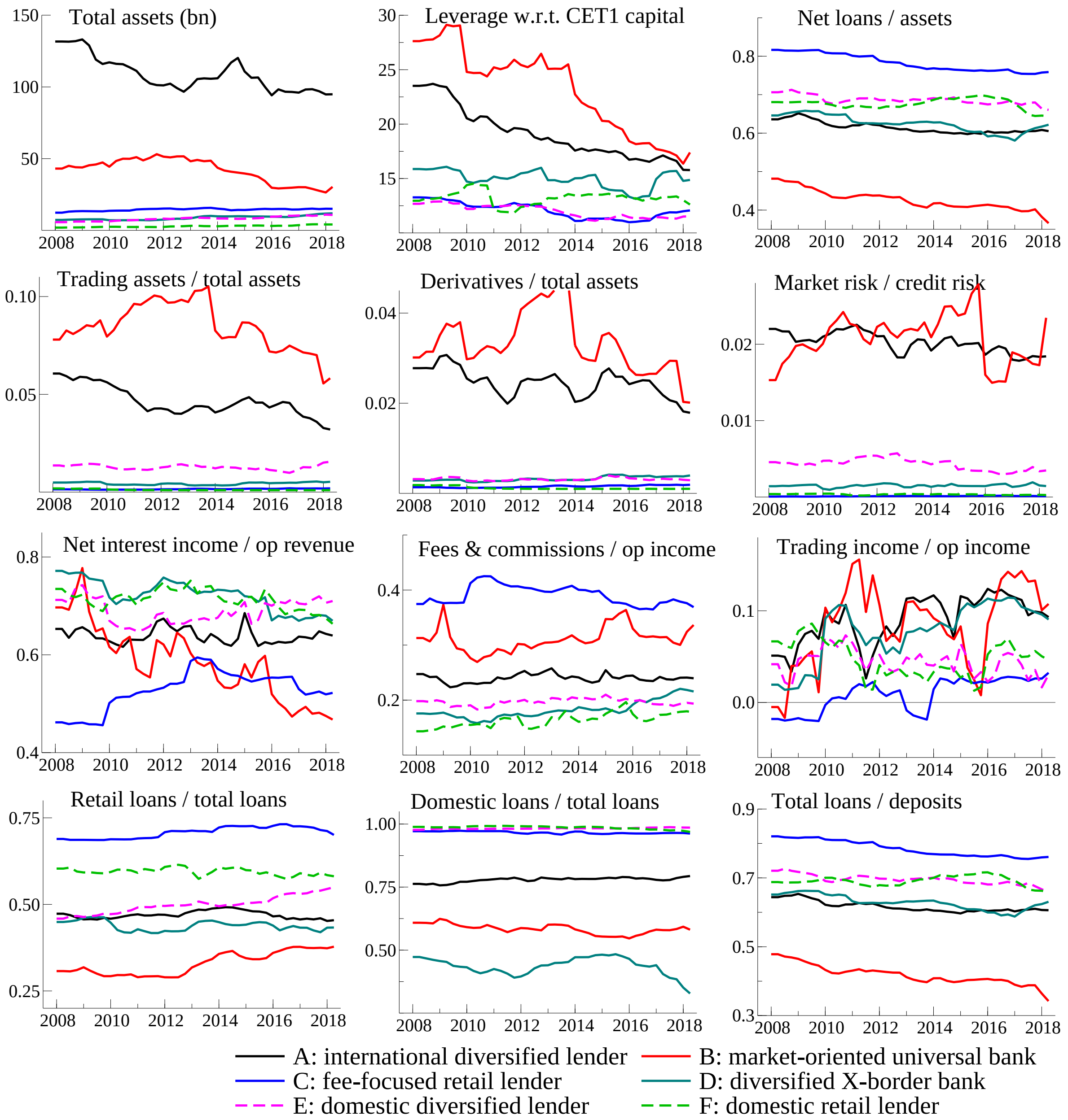
(C) **Fee-focused retail lenders** (8.7% of obs.; e.g. all subsidiaries of Caisse Regionale de Credit Agricole, Credit Lyonnais.)

(D) **Diversified cross-border banks** (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.)

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 $\boldsymbol{y}_{jt}$ and covariance matrices $\boldsymbol{\Sigma}_{jt}$. 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).



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

International diversified lenders (black line) stand out as large institutions, with total assets typically ranging between approximately €100 – 800 bn per firm. As the label suggests, such banks lend significantly across borders and to both retail and corporate clients. The share of non-domestic loans to total loans is approximately 25%, and the share of retail loans ranges between approximately 40–60%. International diversified lenders also serve their corporate customers by trading securities and derivatives on their behalf, resulting in sizeable trading and derivatives books. Funding is obtained from both capital markets as well as customer deposits, as indicated by a moderate loans-to-deposits ratio.

Market-oriented universal banks (red line) comprise large and well-known institutions. Approximately 60% 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 firms at any time between 2008Q1–2018Q2, even though leverage, i.e., total assets to CET1 capital, decrease by more than a third from pre-crisis levels, from approximately 30 to below 20; see 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, including lending. Approximately 40% of loans are cross-border loans.

Fee-focused retail lenders (blue line) achieve most of their income from net fees and commissions (approximately 40%) despite lending mostly to retail customers. Such lenders exhibit a high loans-to-assets ratio of approximately 80%, focus on domestic loans, and receive significant non-deposit funding (e.g. from a parent company). Median total assets are typically below 100 bn per firm.

Diversified cross-border banks (green line) are medium-sized banks with significant cross-border activities. Strikingly, more than half of all loans are cross-border loans. Net interest income accounts for approximately 70% of operating revenue, leaving fee and trading income as relatively less significant sources. Diversified cross-border banks tend to be well-capitalized and lend about equally to both corporate and retail clients.

Domestic diversified lenders (pink dashed line) are relatively numerous, comprising approximately 32% of firms, and are of a small to moderate size. Total assets are typically below €50 bn per firm. Domestic diversified lenders tend to be well capitalized, as implied by relatively low leverage ratios (of typically less than 20). Trading and derivatives books are small. Lending is split approximately evenly between corporate and retail clients. Non-domestic loans are typically below 10%.

Finally, **domestic retail lenders** are the smallest firms, with typically less than €25 bn in total assets. Domestic retail lenders and domestic diversified lenders have much in common. Both types of banks display low leverage, suggesting they are well capitalized. The relatively largest part of their risk is credit risk. 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 retail lenders differ from domestic diversified lenders by their somewhat higher retail orientation, and by a virtual absence of any trading assets, derivatives, and market risk.

4.4 Cluster dissimilarity

This section addresses the question whether banks 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 sale panics could increase if banks had adopted similar business strategies. In addition, contagion across banks could become more likely.

We define a dissimilarity measure as the average distance between cluster means as

$$\begin{aligned} \bar{d}_t &= 2J^{-1}(J-1)^{-1} \sum_{j,k;j<k}^{J} d_{jk,t} \\ &= 2J^{-1}(J-1)^{-1} \sum_{j,k;j<k}^{J} \left((\boldsymbol{\mu}_{jt} - \boldsymbol{\mu}_{kt})' \, \overline{\boldsymbol{\Sigma}}_t^{-1} \, (\boldsymbol{\mu}_{jt} - \boldsymbol{\mu}_{kt}) \right)^{\frac{1}{2}}, \end{aligned} \tag{27}$$

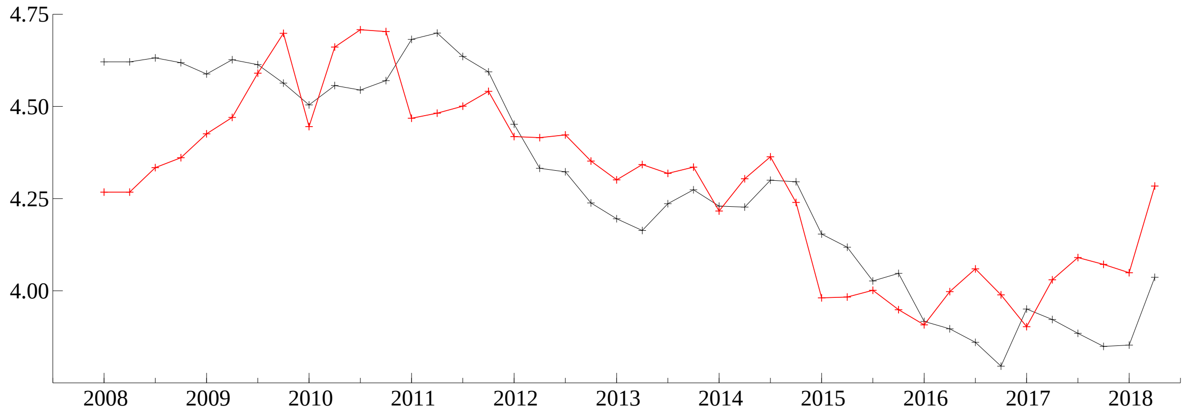
where $\overline{\boldsymbol{\Sigma}}_t^{-1}$ is defined below (4).

Figure 3 plots our cluster dissimilarity measure (27) over time. Bank business model groups on average become more similar (less dissimilar) during most of our sample. A pronounced drop in the dissimilarity metric occurs between 2011Q1 and 2016Q4, coinciding with a post-crisis increase in financial regulation and supervision, as well as rock-bottom central bank interest rates.

Figure D.2 in Web Appendix D.2 plots the dissimilarity metric for each $d = 1, \ldots, D$ variable

Figure 3: Cluster dissimilarity metric over time

Average cluster distance between 2008Q1 and 2018Q2. We plot (27) based on Euclidian (in black) and Mahalanobis (in red) distances. The red line is matched in mean to the black one for visibility.



separately. Banks become more similar in key characteristics such as size, leverage, assets and derivatives held for trading, risk profile, and retail orientation. At the same time, banks become less similar in terms of loans-to-assets ratios, domestic loans to total loans ratios, and loans to deposits ratios.

4.5 Cluster popularity and transitions

Do certain bank business model groups become more or less popular over time? Figure 4 plots the total number of banks allocated to each cluster at each time. Clusters A, C, and F become more populated over time, while clusters B, D, and E become less populated. This is in line with large banks becoming less reliant on market funding (B $\to$ A), banks relying more on more on fee income to lean against a lower profitability from low interest rates (D $\to$ C), and smaller, domestically-active banks lending relatively more to retail clients than to corporate clients (E $\to$ F).

Figure 5 reports the fraction of firms that are estimated to have transitioned from one cluster to another at each t. The (slight) increase in transition frequency can be explained by the decreasing cluster dissimilarity; see Section 27. Transitions have a higher tendency to occur at year-end, as some banks report only annually. Approximately 3% of the N banks transition each quarter. We do not observe any obvious cyclical variation in cluster transitions between crisis and non-crisis times.

Figure 4: Cluster popularity

The number of banks i allocated to cluster $j = 1, \ldots, 6$ at each time t between 2008Q1 and 2018Q2.

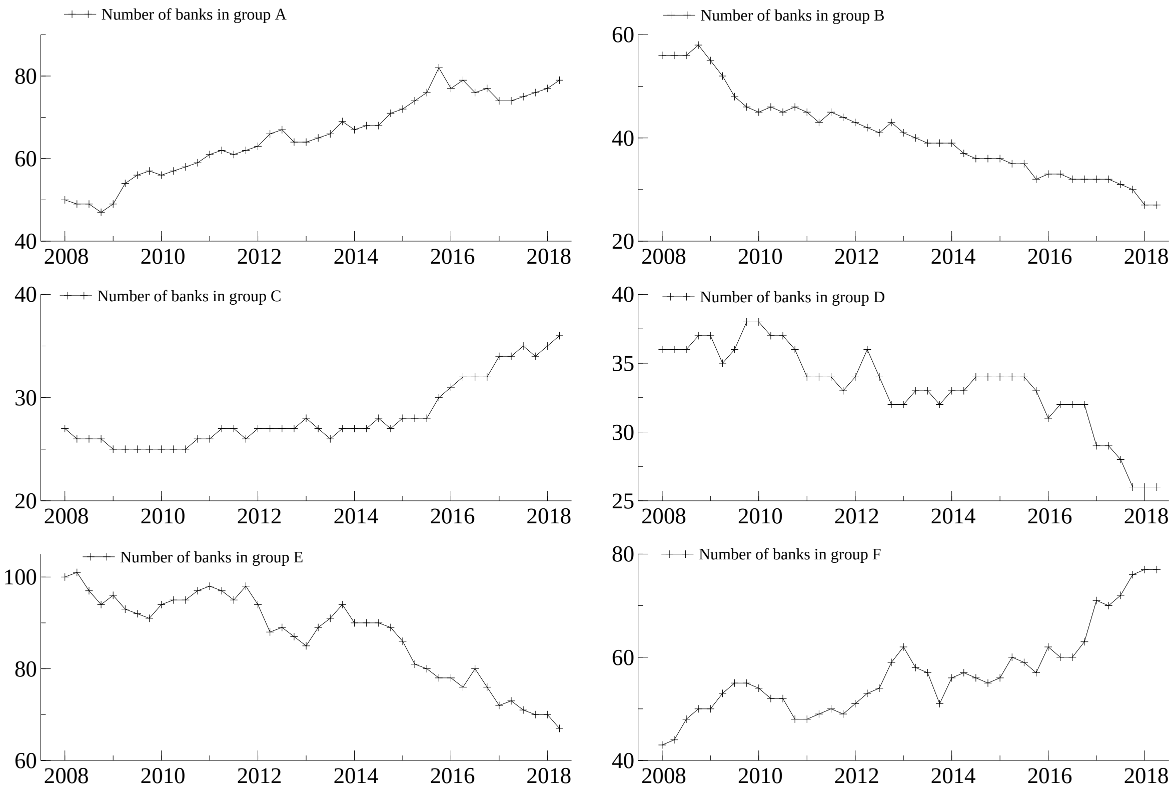
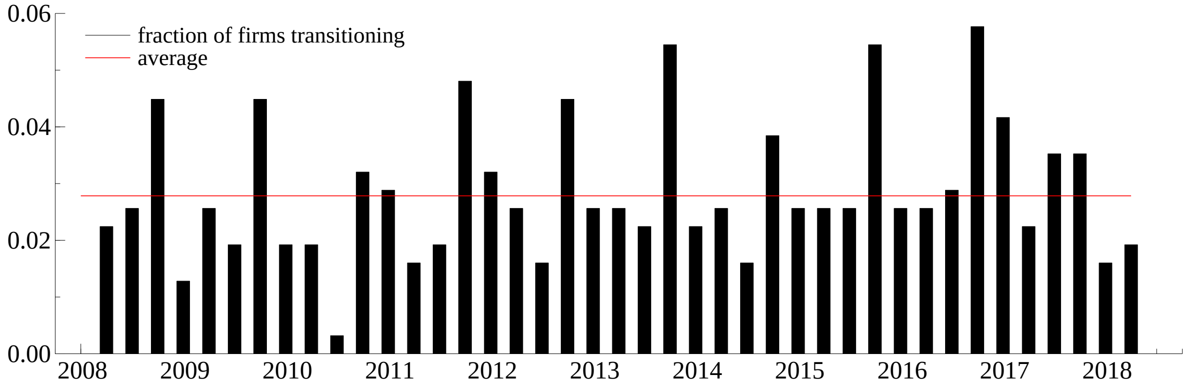


Figure 5: Timing of cluster transitions

The fraction of firms that transition over time, and average estimated transition frequency.



By construction, model-implied transitions to nearby clusters are more likely than to farther apart clusters; see (3). As a result, estimated transition frequencies differ across business model clusters. Web Appendix D.3 reports the average estimated cluster transition matrix, as well as the model-implied transitions. We observe most transitions between clusters E $\rightarrow$ F (75), F $\rightarrow$ E (67), E $\rightarrow$ A (37), B $\rightarrow$ A (27), D $\rightarrow$ F (22), and A $\rightarrow$ E (20).

4.6 Can cluster transitions be explained by bank profitability?

[To be added.]

5 Conclusion

We proposed a novel observation-driven model for the dynamic clustering of multivariate panel data, and subsequently applied it to a sample of 312 European banks between 2008Q1 and 2018Q2. Our empirical results suggest a gradual convergence of a subset of bank business model characteristics over time as well as a moderate number of transitions between clusters.

References

Abadir, K. and J. Magnus (2005). *Matrix Algebra*. Cambridge University Press.

Ayadi, R., E. Arbak, and W. P. de Groen (2014). Business models in European banking: A pre- and post-crisis screening. *CEPS discussion paper*, 1–104.

Ayadi, R. and W. P. D. Groen (2015). Bank business models monitor Europe. *CEPS working paper*, 0–122.

Azizpour, S., K. Giesecke, and G. Schwenkler (2017). Exploring the sources of default clustering. *Journal of Financial Economics*.

Bankscope (2014). Bankscope user guide. Bureau van Dijk, Amsterdam, January 2014. Available to subscribers.

Bhar, R. and S. Hamori (2004). *Hidden Markov models: Applications to financial economics*. Boston: Kluwer Academic Publishers.

Catania, L. (2016). Dynamic adaptive mixture models. *University of Rome Tor Vergata, unpublished working paper*.

Cox, D. R. (1981). Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics 8*, 93–115.

Creal, D., S. Koopman, and A. Lucas (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics 28*(5), 777–795.

Creal, D. D., R. B. Gramacy, and R. S. Tsay (2014). Market-based credit ratings. *Journal of Business & Economic Statistics 32*, 430–444.

Fruehwirth-Schnatter, S. and S. Kaufmann (2008). Model-based clustering of multiple time series. *Journal of Business and Economic Statistics 26*, 78–89.

Goldfeld, S. M. and R. E. Quandt (1973). A Markov model for switching regressions. *Journal of Econometrics 1*(1), 3–15.

Hamilton, J. D. and M. T. Owyang (2012). The propagation of regional recessions. *The Review of Economics and Statistics 94*, 935–947.

Hartigan, J. A. and M. A. Wong (1979). A k-means clustering algorithm. *Applied Statistics 28*(1), 100–108.

Harvey, A. C. (2013). *Dynamic models for volatility and heavy tails, with applications to financial and economic time series*. Number 52. Cambridge University Press.

Koopman, S. J., A. Lucas, and B. Schwaab (2011). Modeling frailty correlated defaults using many macroeconomic covariates. *Journal of Econometrics 162 (2)*, 312–325.

Koopman, S. J., A. Lucas, and B. Schwaab (2012). Dynamic factor models with macro, frailty, and industry effects for u.s. default counts: the credit crisis of 2008. *Journal of Business and Economic Statistics 30(4)*, 521–532.

Lucas, A., J. Schaumburg, and B. Schwaab (2018). Bank business models at zero interest rates. *Journal of Business & Economic Statistics*, in press.

Lucas, A. and X. Zhang (2016). Score driven exponentially weighted moving average and value-at-risk forecasting. *International Journal of Forecasting 32(2)*, 293–302.

Opschoor, A., A. Lucas, P. Januw, and D. J. van Dijk (2018). New HEAVY models for fat-tailed realized covariances and returns. *Journal of Business and Economic Statistics 36*(4), 643–657.

Peel, D. and G. J. McLachlan (2000). Robust mixture modelling using the t distribution. *Statistics and Computing 10*, 339–348.

Roengpitya, R., N. Tarashev, and K. Tsatsaronis (2014). Bank business models. *BIS Quarterly Review*, 55–65.

Roengpitya, R., N. Tarashev, K. Tsatsaronis, and A. Villegas (2017). Bank business models: popularity and performance. *BIS working paper 682*.

Smyth, P. (1996). Clustering sequences with hidden markov models. *Advances in Neural Information Processing Systems 9*, 1–7.

SSM (2016). SSM SREP methodology booklet. *available at www.bankingsupervision.europa.eu, accessed on 14 April 2016.*, 1–36.

Wang, Y., R. S. Tsay, J. Ledolter, and K. M. Shrestha (2013). Forecasting simultaneously high-dimensional time series: A robust model-based clustering approach. *Journal of Forecasting 32*(8), 673–684.

Web Appendix to

“Dynamic clustering of multivariate panel data”

André Lucas, Julia Schaumburg, Bernd Schwaab

A Derivation of the scaled scores

A.1 Time-varying mean dynamics

The scaled score for updating the time-varying j-th cluster mean $\boldsymbol{\mu}_{jt}$ is given by

$$\boldsymbol{s}_{\mu_{jt},t} = \boldsymbol{S}_{\mu_{jt},t} \cdot \boldsymbol{\nabla}_{\mu_{jt},t}, \tag{A.1}$$

where $\boldsymbol{S}_{\mu_{jt},t}$ is the scaling matrix and $\boldsymbol{\nabla}_{\mu_{jt},t}$ is the score of the predictive likelihood at time t. Starting with the score, and using the fact that $\tau_{ij,t|t-1}$ does not depend on $\boldsymbol{\mu}_{j,t}$ due to the transition probability matrix Π_t depending on the lagged cluster distances only as formulated in equations (2) and (6), we have

$$\begin{aligned}
\boldsymbol{\nabla}_{\mu_{jt},t} &= \frac{\partial \ell_t}{\partial \boldsymbol{\mu}_{jt}} = \frac{\partial \sum_{i=1}^{N} \log f\left(\boldsymbol{y}_{it}|\mathcal{F}_{t-1};\boldsymbol{\theta}\right)}{\partial \boldsymbol{\mu}_{jt}}, \\
&= \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\mu}_{jt}} \log f\left(\boldsymbol{y}_{it}|\mathcal{F}_{t-1};\boldsymbol{\theta}\right) \\
&= \sum_{i=1}^{N} \frac{1}{f\left(\boldsymbol{y}_{it}|\mathcal{F}_{t-1};\boldsymbol{\theta}\right)} \cdot \frac{\partial}{\partial \boldsymbol{\mu}_{jt}} f\left(\boldsymbol{y}_{it}|\mathcal{F}_{t-1};\boldsymbol{\theta}\right) \\
&= \sum_{i=1}^{N} \frac{f(\boldsymbol{y}_{it}|c_{it}=j,\mathcal{F}_{t-1})\tau_{ij,t|t-1}}{f\left(\boldsymbol{y}_{it}|\mathcal{F}_{t-1};\boldsymbol{\theta}\right)} \cdot \frac{\frac{\partial}{\partial \boldsymbol{\mu}_{jt}} f\left(\boldsymbol{y}_{it}|\mathcal{F}_{t-1};\boldsymbol{\theta}\right)}{f(\boldsymbol{y}_{it}|c_{it}=j,\mathcal{F}_{t-1})\tau_{ij,t|t-1}} \\
&= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \frac{\frac{\partial}{\partial \boldsymbol{\mu}_{jt}} f\left(\boldsymbol{y}_{it}|\mathcal{F}_{t-1};\boldsymbol{\theta}\right)}{f(\boldsymbol{y}_{it}|c_{it}=j,\mathcal{F}_{t-1})\tau_{ij,t|t-1}} \\
&= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \frac{\frac{\partial}{\partial \boldsymbol{\mu}_{jt}} \sum_{j=1}^{J} \tau_{ij,t|t-1} f\left(\boldsymbol{y}_{it}|c_{it}=j,\mathcal{F}_{t-1};\boldsymbol{\theta}\right)}{f(\boldsymbol{y}_{it}|c_{it}=j,\mathcal{F}_{t-1})\tau_{ij,t|t-1}} \\
&= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \frac{\frac{\partial}{\partial \boldsymbol{\mu}_{jt}} \left(\tau_{ij,t|t-1} \cdot f\left(\boldsymbol{y}_{it}|c_{it}=j,\mathcal{F}_{t-1};\boldsymbol{\theta}\right)\right)}{f(\boldsymbol{y}_{it}|c_{it}=j,\mathcal{F}_{t-1})\tau_{ij,t|t-1}} \\
&= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \frac{\partial}{\partial \boldsymbol{\mu}_{jt}} \log\left(\tau_{ij,t|t-1} \cdot f\left(\boldsymbol{y}_{it}|c_{it}=j,\mathcal{F}_{t-1};\boldsymbol{\theta}\right)\right) \\
&= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \frac{\partial}{\partial \boldsymbol{\mu}_{jt}} \log f\left(\boldsymbol{y}_{it}|c_{it}=j,\mathcal{F}_{t-1};\boldsymbol{\theta}\right) \\
&= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \boldsymbol{\nabla}^{(j)}_{\mu_{jt},t}.
\end{aligned}$$

In case of a mixture of D-dimensional Student's t distributions, we have

$$f\left(\boldsymbol{y}_{it}|c_{it}=j,\mathcal{F}_{t-1};\boldsymbol{\theta}\right)=\frac{\Gamma\left(\frac{\nu_j+D}{2}\right)}{\Gamma\left(\frac{\nu_j}{2}\right)(\pi\nu_j)^{D/2}\left|\Sigma_j\right|^{1/2}}\left(1+\frac{(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})'\boldsymbol{\Sigma}_{jt}^{-1}(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})}{\nu_j}\right)^{-(\frac{\nu_j+D}{2})}. \tag{A.2}$$

Taking derivatives of the log of (A.2), we obtain

$$\boldsymbol{\nabla}^{(j)}_{\mu_{jt},t}=w_{ij,t}\cdot\boldsymbol{\Sigma}_{jt}^{-1}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right), \tag{A.3}$$

where

$$w_{ij,t}=(1+\nu_j^{-1}D)\Big/\left(1+\nu_j^{-1}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)'\boldsymbol{\Sigma}_{jt}^{-1}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)\right). \tag{A.4}$$

Equation (A.3) contains the unscaled score. We scale the score by the weighted average of the conditional Fisher information matrices for the Gaussian setting $\nu_j^{-1}=0$, using the posterior probabilities $\tau_{ij,t|t}$ as weights; compare Lucas et al. (2018). We obtain

$$\begin{aligned}
\boldsymbol{S}^{-1}_{\mu_{jt},t} &= \sum_{i=1}^{N}\tau_{ij,t|t}\cdot\left(-\mathbb{E}\left[\frac{\partial\boldsymbol{\nabla}^{(j)}_{\mu_{jt},t}}{\partial\boldsymbol{\mu}'_{jt}}\,\middle|\,c_{it}=j,\mathcal{F}_{t-1};\boldsymbol{\theta}\right]\right)\\
&= \sum_{i=1}^{N}\tau_{ij,t|t}\cdot\mathbb{E}\left[\boldsymbol{\nabla}^{(j)}_{\mu_{jt},t}\left(\boldsymbol{\nabla}^{(j)}_{\mu_{jt},t}\right)'\,\middle|\,c_{it}=j,\mathcal{F}_{t-1};\boldsymbol{\theta}\right]\\
&= \sum_{i=1}^{N}\tau_{ij,t|t}\cdot\mathbb{E}\left[\boldsymbol{\Sigma}_{jt}^{-1}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})'\,\boldsymbol{\Sigma}_{jt}^{-1}\,\middle|\,c_{it}=j,\mathcal{F}_{t-1};\boldsymbol{\theta}\right]\\
&= \sum_{i=1}^{N}\tau_{ij,t|t}\cdot\boldsymbol{\Sigma}_{jt}^{-1}\,\mathbb{E}\left[(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})'\,\middle|\,c_{it}=j,\mathcal{F}_{t-1};\boldsymbol{\theta}\right]\,\boldsymbol{\Sigma}_{jt}^{-1}\\
&= \sum_{i=1}^{N}\tau_{ij,t|t}\cdot\boldsymbol{\Sigma}_{jt}^{-1}\,\boldsymbol{\Sigma}_{jt}\,\boldsymbol{\Sigma}_{jt}^{-1}\\
&= \sum_{i=1}^{N}\tau_{ij,t|t}\cdot\boldsymbol{\Sigma}_{jt}^{-1},
\end{aligned} \tag{A.5}$$

where we used the fact that for the Gaussian case $w_{ij,t}=1$.

Inserting (A.5) and (A.3) into (A.1) yields the scaled score

$$
\begin{aligned}
\boldsymbol{s}_{\mu_{jt},t} &= \boldsymbol{S}_{\mu_{jt},t} \cdot \boldsymbol{\nabla}_{\mu_{jt},t} \\
&= \left(\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \boldsymbol{\Sigma}_{jt}^{-1} \right)^{-1} \sum_{i=1}^{N} \tau_{ij,t|t} \cdot w_{ij,t} \cdot \boldsymbol{\Sigma}_{jt}^{-1} \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right) \\
&= \boldsymbol{\Sigma}_{jt} \boldsymbol{\Sigma}_{jt}^{-1} \left(\sum_{i=1}^{N} \tau_{ij,t|t} \right)^{-1} \sum_{i=1}^{N} \tau_{ij,t|t} \cdot w_{ij,t} \cdot \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right) \\
&= \frac{\sum_{i=1}^{N} \tau_{ij,t|t} \cdot w_{ij,t} \cdot \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right)}{\sum_{i=1}^{N} \tau_{ij,t|t}}.
\end{aligned}
$$

Transition equation (12) now follows directly.

A.2 Time-varying covariance matrix dynamics

The scaled score for the time-varying cluster covariance matrix parameters is

$$
\boldsymbol{s}_{\Sigma_{jt},t} = \boldsymbol{S}_{\Sigma_{jt},t} \cdot \boldsymbol{\nabla}_{\Sigma_{jt},t}, \tag{A.6}
$$

where $\boldsymbol{S}_{\Sigma_{jt},t}$ is the scaling matrix and $\boldsymbol{\nabla}_{\Sigma_{jt},t}$ is the score. The score is given by

$$
\boldsymbol{\nabla}_{\Sigma_{jt},t} = \frac{\partial \ell_t}{\partial \text{vec}(\boldsymbol{\Sigma}_{jt})} = \frac{\partial \left[\sum_{i=1}^{N} \ln \left(f \left(\boldsymbol{y}_{it} | \mathcal{F}_{t-1}; \boldsymbol{\theta} \right) \right) \right]}{\partial \text{vec}(\boldsymbol{\Sigma}_{jt})},
$$

where we can take the derivatives with respect to a general matrix $\boldsymbol{\Sigma}_{jt}$ rather than a symmetric matrix. Using the arguments in Proposition 3 of Opschoor et al. (2018), this gives the same steps for the free elements in $\boldsymbol{\Sigma}_{jt}$.

The initial derivations follow the same steps as for the time-varying mean; see Web Appendix A.1. Leaving these steps out, taking the log of (A.2) and omitting the terms that do not depend on Σ_{jt}, we arrive at

$$
\boldsymbol{\nabla}_{\Sigma_{jt},t} = \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \left(- \frac{\partial}{\partial \text{vec}(\boldsymbol{\Sigma}_{jt})} \frac{1}{2} \ln |\boldsymbol{\Sigma}_{jt}| - \frac{\partial}{\partial \text{vec}(\boldsymbol{\Sigma}_{jt})} \left[\left(\frac{\nu_j + D}{2} \right) \ln \left(1 + \frac{(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt})' \boldsymbol{\Sigma}_{jt}^{-1} (\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt})}{\nu_j} \right) \right] \right).
$$

Following Abadir and Magnus (2005) for the derivative of the log of the determinant of the covariance matrix, and for the derivative of a matrix inside a quadratic form, and using $\text{vec}(ABC) = (C' \otimes A)\text{vec}(B)$, we obtain

$$
\begin{aligned}
\boldsymbol{\nabla}_{\Sigma_{jt},t} &= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \text{vec}\left(-\tfrac{1}{2}\left(\boldsymbol{\Sigma}_{jt}^{-1}\right)' + \tfrac{1}{2}\left(\boldsymbol{\Sigma}_{jt}^{-1}\right)' w_{ij,t}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)'\left(\boldsymbol{\Sigma}_{jt}^{-1}\right)'\right) \\
&= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \text{vec}\left(-\tfrac{1}{2}\left(\boldsymbol{\Sigma}_{jt}'\right)^{-1} + \tfrac{1}{2}\left(\boldsymbol{\Sigma}_{jt}'\right)^{-1} w_{ij,t}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)'\left(\boldsymbol{\Sigma}_{jt}'\right)^{-1}\right) \\
&= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \text{vec}\left(-\tfrac{1}{2}\boldsymbol{\Sigma}_{jt}^{-1} + \tfrac{1}{2}\boldsymbol{\Sigma}_{jt}^{-1} w_{ij,t}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)'\boldsymbol{\Sigma}_{jt}^{-1}\right) \\
&= \tfrac{1}{2}\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \text{vec}\left(\boldsymbol{\Sigma}_{jt}^{-1}\left(w_{ij,t}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)' - \boldsymbol{\Sigma}_{jt}\right)\boldsymbol{\Sigma}_{jt}^{-1}\right) \\
&= \tfrac{1}{2}\left(\boldsymbol{\Sigma}_{jt}\otimes\boldsymbol{\Sigma}_{jt}\right) \cdot \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \text{vec}\left(w_{ij,t}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)' - \boldsymbol{\Sigma}_{jt}\right),
\end{aligned}
\tag{A.7}
$$

where $w_{ij,t}$ is defined in (A.4)

Next, we derive the scaling matrix, which we take as the weighted average of Fisher information matrices given $\nu_j^{-1} = 0$ for all j. We have

$$
\begin{aligned}
\boldsymbol{S}_{\Sigma_{jt},t}^{-1} &= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \mathbb{E}\left[\boldsymbol{\nabla}_{\Sigma_{jt},t}\boldsymbol{\nabla}_{\Sigma_{jt},t}' \,\middle|\, c_{it}=j, \mathcal{F}_{t-1};\boldsymbol{\theta}\right] \\
&= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \left(-\mathbb{E}\left[\frac{\partial \boldsymbol{\nabla}_{\Sigma_{jt},t}}{\partial \text{vec}(\boldsymbol{\Sigma}_{jt})'} \,\middle|\, c_{it}=j, \mathcal{F}_{t-1};\boldsymbol{\theta}\right]\right) \\
&= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \left(-\mathbb{E}\left[\frac{\partial}{\partial \text{vec}(\boldsymbol{\Sigma}_{jt})'}\tfrac{1}{2}\text{vec}\left(\boldsymbol{\Sigma}_{jt}^{-1}\left(\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)' - \boldsymbol{\Sigma}_{jt}\right)\boldsymbol{\Sigma}_{jt}^{-1}\right) \,\middle|\, c_{it}=j, \mathcal{F}_{t-1};\boldsymbol{\theta}\right]\right) \\
&= -\tfrac{1}{2}\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \mathbb{E}\left[\frac{\partial}{\partial \text{vec}(\boldsymbol{\Sigma}_{jt})'}\text{vec}\left(\boldsymbol{\Sigma}_{jt}^{-1}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)'\boldsymbol{\Sigma}_{jt}^{-1} - \boldsymbol{\Sigma}_{jt}^{-1}\right) \,\middle|\, c_{it}=j, \mathcal{F}_{t-1};\boldsymbol{\theta}\right] \\
&= -\tfrac{1}{2}\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \left\{\mathbb{E}\left[-\left(\mathrm{I}\otimes\boldsymbol{\Sigma}_{jt}^{-1}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)'\right)\left(\boldsymbol{\Sigma}_{jt}^{-1}\otimes\boldsymbol{\Sigma}_{jt}^{-1}\right) \,\middle|\, c_{it}=j, \mathcal{F}_{t-1};\boldsymbol{\theta}\right] + \right. \\
&\qquad \mathbb{E}\left[-\left(\boldsymbol{\Sigma}_{jt}^{-1}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)'\otimes\mathrm{I}\right)\left(\boldsymbol{\Sigma}_{jt}^{-1}\otimes\boldsymbol{\Sigma}_{jt}^{-1}\right) \,\middle|\, c_{it}=j, \mathcal{F}_{t-1};\boldsymbol{\theta}\right] - \\
&\qquad \left. \mathbb{E}\left[-\left(\boldsymbol{\Sigma}_{jt}^{-1}\otimes\boldsymbol{\Sigma}_{jt}^{-1}\right) \,\middle|\, c_{it}=j, \mathcal{F}_{t-1};\boldsymbol{\theta}\right]\right\} \\
&= \tfrac{1}{2}\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \mathbb{E}\left[\left(\boldsymbol{\Sigma}_{jt}^{-1}\otimes\boldsymbol{\Sigma}_{jt}^{-1}\right) \,\middle|\, c_{it}=j, \mathcal{F}_{t-1};\boldsymbol{\theta}\right] = \tfrac{1}{2}\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \left(\boldsymbol{\Sigma}_{jt}^{-1}\otimes\boldsymbol{\Sigma}_{jt}^{-1}\right),
\end{aligned}
$$

where we used again $\text{vec}\,(ABC) = (C'\otimes A)\,\text{vec}\,(B)$, and $\partial\text{vec}(A^{-1})/\partial\text{vec}(A)' = -((A')^{-1}\otimes A^{-1})$ for a general matrix A. Pre-multiplying the score by the scaling matrix, we obtain the scaled score

$$\begin{aligned}\boldsymbol{s}_{\Sigma_{jt},t} &= \left(\frac{1}{2}\sum_{i=1}^{N}\tau_{ij,t|t}\cdot\left(\boldsymbol{\Sigma}_{jt}^{-1}\otimes\boldsymbol{\Sigma}_{jt}^{-1}\right)\right)^{-1}\times \\ &\qquad\left(\frac{1}{2}\left(\boldsymbol{\Sigma}_{jt}\otimes\boldsymbol{\Sigma}_{jt}\right)\cdot\sum_{i=1}^{N}\tau_{ij,t|t}\cdot\text{vec}\left(w_{ij,t}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)'-\boldsymbol{\Sigma}_{jt}\right)\right) \\ &= \frac{\sum_{i=1}^{N}\tau_{ij,t|t}\cdot\text{vec}\left(w_{ij,t}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)'-\boldsymbol{\Sigma}_{jt}\right)}{\sum_{i=1}^{N}\tau_{ij,t|t}}.\end{aligned} \tag{A.8}$$

Transition equation (22) now follows directly.

B Sketch of $k-$means algorithm

1. **Initialization:** initialize random centers for the J clusters in D dimensions.

2. **Assignment:** assign each observation, for a total of N observations, to the closest cluster according to Euclidean distance. $\tau_{ij,1|0} = \begin{cases} 1 \text{ for } \min_j \sqrt{(\boldsymbol{y}_{i1} - \boldsymbol{\mu}_{j1})' (\boldsymbol{y}_{i1} - \boldsymbol{\mu}_{j1})} \\ 0 \text{ else} \end{cases}$.

3. **Update:** recalculate the cluster centers as the average of the observations assigned to that cluster $\boldsymbol{\mu}_{j1} = \frac{\sum_{i=1}^{N} \tau_{ij,1|0} \cdot \boldsymbol{y}_{i1}}{\sum_{i=1}^{N} \tau_{ij,1|0}}$.

4. **Convergence 2:** return to step 2, and repeat until convergence of within-cluster sum of squared errors.

5. **Convergence 1:** return to step 1, and repeat 1000 times for different initial random centers. Chose the one with minimal within-cluster sum of squared errors.

6. **Calculate initial covariance matrices:** estimate covariance matrix from the observations that were assigned to each cluster. $\boldsymbol{\Sigma}_{j1} = \frac{\sum_{i=1}^{N} \tau_{ij,1|0} \cdot (\boldsymbol{y}_{i1} - \boldsymbol{\mu}_{j1})(\boldsymbol{y}_{i1} - \boldsymbol{\mu}_{j1})'}{\sum_{i=1}^{N} \tau_{ij,1|0}}$.

C Data details

Our sample under study consists of $N = 312$ European banks, for which we consider quarterly bank-level accounting data from SNL Financial between 2008Q1 – 2018Q2. 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 strategies. We select a parsimonious set of $D = 12$ indicators to cover these six categories. Table C.1 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 (backfilling). If variables are missing in the beginning of the sample, we use the most adjacent future value. In the cross-section, we require at least one entry for each variable and bank.

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%).

Table C.1 also reports the data transformation used in the applied modeling. For example, some ratios lie strictly within the unit interval. We transform such ratios into unbounded continuous variables by mapping them through an inverse Probit transform. We take natural logarithms of large numbers, such as total assets, CET1 capital, assets held for trading, etc.

Table C.1: 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. $\Phi^{-1}(\cdot)$ denotes the inverse Probit transform.

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

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 loans-to-deposits ratio are loans held for investment, before reserves, as a percent of total deposits, the latter comprising both retail and commercial deposits (248919).

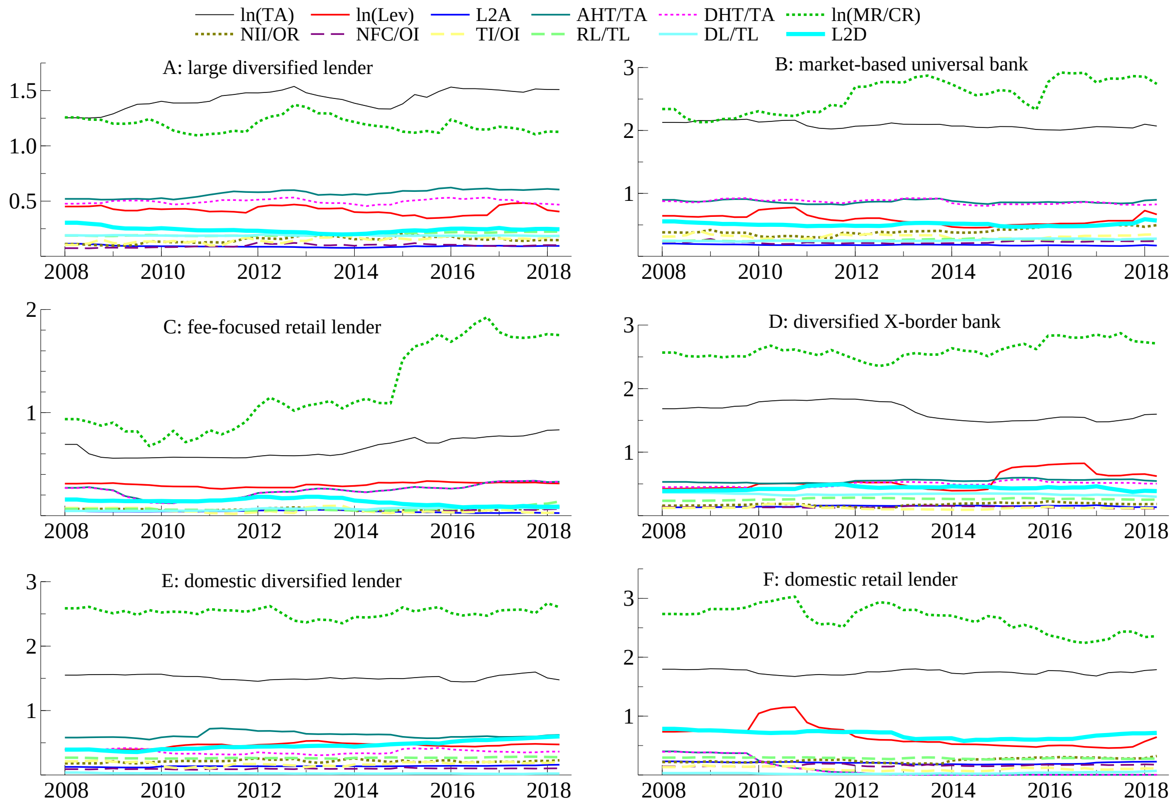
Web Appendix D: Additional results

Web Appendix D.1: Estimated cluster standard deviations

Figure D.1 plots the filtered component-specific time-varying standard deviations $\sigma_{jt}(d) = \left(\Sigma_{jt}(d,d)\right)^{\frac{1}{2}}$ for variables $d = 1, \ldots, D$. The off-diagonal elements of Σ_{jt} are not reported. The variables (log) total assets and (log) market risk to credit risk are particularly dispersed across units within a given component. Overall, the standard deviations remain approximately stable over time even for clusters that contain fewer banks.

Figure D.1: Time-varying standard deviations

Filtered time-varying standard deviations $\sigma_{jt}(d) = \left(\Sigma_{jt}(d,d)\right)^{\frac{1}{2}}$ for variables $d = 1, \ldots, D$. Each panel refers to a business model component A – F, and contains twelve standard deviation estimates over time (one for each variable in Table C.1). Mean and standard deviation estimates are based on a t-mixture model with $J = 6$ components and dynamic covariance matrices Σ_{jt}.

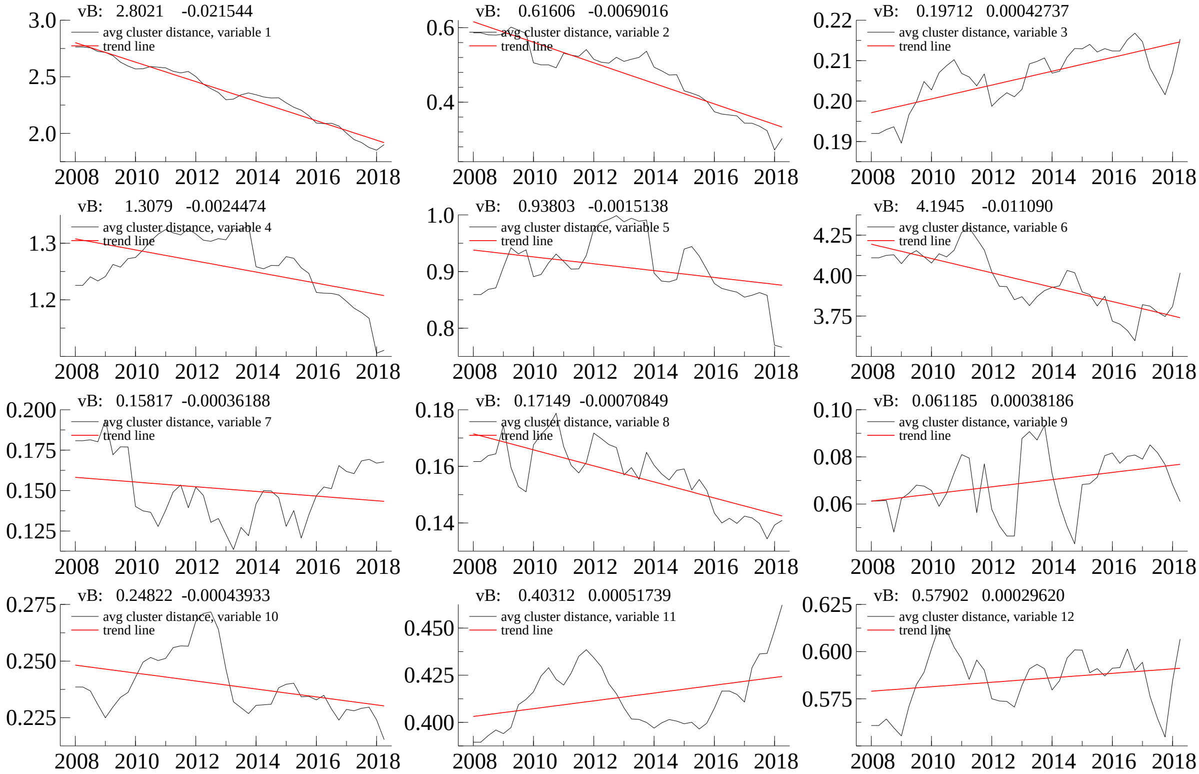


Web Appendix D.2: Variable-specific dissimilarity metrics

Figure D.2 reports variable $d = 1, \ldots, 12$-specific cluster distances across $J \times (J-1)/2$ cluster means.

Figure D.2: Variable-specific dissimilarity measures

Dissimilarity measures for $d = 1, \ldots, 12$ variables.



Web Appendix D.3: Average transition probabilities

[To be added.]

Web Appendix E: Filtered cluster probabilities

Figures X – X plot the filtered component membership probabilities $\tau_{ij,t|t}$ as given in (7). Most firms are fairly unequivocally allocated to one component at a time.

[To be added.]

D Cluster medians without transitions

Figure F.1 plots time-varying component median estimates based on the methodology of Lucas et al. (2018). This methodology proceeds under the assumption that cluster transitions are absent. The component median estimates are visibly different from the estimates reported in Figure 2, suggesting that cluster transitions are economically important when studying the evolution of bank business models over time.

Figure F.1: Time-varying component medians

Filtered component medians for twelve indicator variables; see Table C.1. The component medians coincide with the component means unless the variable is transformed; see the last column of Table C.1 in Web Appendix C. The component mean estimates are based on a t-mixture model with $J = 6$ components and time-varying component means μ_{jt} and covariance matrices Σ_{jt}.