TI 2020-009/III
Tinbergen Institute Discussion Paper

Dynamic clustering of multivariate panel data

André Lucas[1]
Julia Schaumburg[1]
Bernd Schwaab[2]

[1] Vrije Universiteit Amsterdam and Tinbergen Institute
[2] European Central Bank, Financial Research

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

December 2019

Abstract

We propose a dynamic clustering model for studying time-varying group structures in multivariate panel data. The model is dynamic in three ways: First, the cluster means and covariance matrices 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. On average, approximately 3% of banks transition each quarter. Transitions across clusters are related to cluster dissimilarity and differences in bank profitability.

Keywords: dynamic clustering; panel data; Hidden Markov Model; score-driven dynamics; 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 dynamic location-scale mixture model for studying time-varying group structures in multivariate panel data. The model is dynamic in multiple ways: First, the cluster means and covariance matrices are time-varying to track gradual changes in group (cluster) characteristics over time. Second, the units of interest can transition between groups based on a Hidden Markov model (HMM). Finally, the HMM's transition probabilities are time-varying as well. They can depend on lagged cluster distances and, potentially, additional economic covariates. Our modeling framework is useful for allocating a potentially large number of units into a much smaller number of approximately homogeneous groups in fairly complicated dynamic settings, while keeping track of overall trends, group membership probabilities, and group transitions. If appropriate, the baseline model can be extended to accommodate in-active states and non-Markovian transition behavior.

All time-varying location and scale parameters of our dynamic mixture model are driven by the score of the local (time t) objective function; see e.g. Creal et al. (2013) and Harvey (2013). In this approach, the time-varying parameters are perfectly predictable one step ahead. This makes the model observation-driven in the terminology of Cox (1981). In addition, the log-likelihood is known in closed form, facilitating parameter estimation via standard methods. Intuitive filtering recursions are available for all time-varying parameters and cluster membership probabilities.

Extensive Monte Carlo experiments suggest that our model is able to accurately classify the units of interest into their respective true clusters at each time, as well as to simultaneously recover all time-varying location and scale parameters, despite the presence of cluster transitions. In our simulations, the cluster classification is perfect for sufficiently large distances between the time-varying cluster means. As the time-varying cluster means move closer together and cluster transitions become more frequent, the share of correct classifications decreases, but generally remains high. Importantly, if cluster transitions are present but ignored, parameter estimates are biased and classification results can be poor. This calls into question the assumption of time-invariant cluster assignments (see, e.g., Lucas et al., 2019) when studying the group structure in multivariate panel

data sets with a non-negligible time dimension.

We apply our modeling framework to a multivariate panel of accounting data for $N = 299$ European banks between 2008Q1 and 2018Q2, i.e. over $T = 42$ quarters, considering $D = 12$ bank-level variables for $J = 6$ groups of similar banks. We thus track bank 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 and 2018. Our sample, overall, is characterized by a significant increase in post-crisis financial regulation, the introduction of centralized supervision in some countries, increasing competition from FinTech and BigTech firms, as well as declining and ultimately negative monetary policy interest rates. All these developments have put significant pressure on banks' business models.

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 Section 4.3.

We focus on three main empirical results. First, we study whether banks have become less diverse over time. A decrease in financial sector diversity could be problematic from a financial stability perspective. For example, the probability and severity of so-called 'fire sales' could increase if large numbers of banks adopt similar business strategies. We find that our bank business model groups have become less diverse over time in key characteristics such as size, leverage, the share of trading activities, and funding choices.

Second, we study which business model groups have become more or less popular over time. In this regard, our cluster location estimates and bank transitions point in the same direction: since the start of our sample, European banks i) have relied increasingly on fee income to lean against impaired profitability from e.g. low interest rates and increased competition, 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 differences in

cluster-specific point-in-time profitability measures. 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 into high-profitability groups. Vice versa, banks from high-profitability groups are less likely to transition into low-profitability groups. To the extent that low bank profitability is caused by low monetary policy rates for some groups of banks (Brunnermeier and Koby, 2019; Heider et al., 2019), this finding suggests that monetary policy can have long-lasting effects on banking sector structure via 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 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 HMMs.

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 (2019) 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 (2019) to conditional asset return distributions typically satisfies this requirement.

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 (2019), 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 in our sample. We therefore require the cluster components to be characterized by dynamic parameters.

Our paper also contributes to the literature on identifying bank business models. Ayadi and Groen (2015), Roengpitya et al. (2017), and Farne and Vouldis (2017) 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 (Farne and Vouldis, 2017) or bottom-up agglomerative clustering with fixed business model characteristics and a non-chronological time dimension (Ayadi and Groen, 2015; Roengpitya et al., 2017). Instead, we consider a panel framework which allows us to pool information over time while also allowing for a rich set of dynamics.

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 $\boldsymbol{y}_{it} \in \mathbb{R}^{D\times 1}$, 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 location-scale mixture model

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

where $\boldsymbol{\mu}_{c_{it},t}$ is a $D \times 1$ vector of cluster-specific means, and ϵ_{it} is a sequence of independently and identically distributed (i.i.d.) $D \times 1$ vectors 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_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 distributions are also easily accommodated in this framework, but are not considered in this paper.

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.

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

We introduce the short-hand notation $\pi_{jk,t} := \mathbb{P}\{c_{i,t+1} = k | c_t = j\}$, 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 could 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 $d_{jk,t}$. If two or more clusters are temporarily close and overlapping at time $t-1$, then specification (2) may not work well. In such cases (2) can be adapted to, for example, $\mathbf{\Pi}_t = \mathbf{\Pi}_t \left(\bar{\mathcal{D}}_{t-1} \right)$, where $\bar{\mathcal{D}}_{t-1} = H^{-1} \sum_{h=1}^{H} \mathcal{D}_{t-h}$, and where H is a positive integer to be chosen ex-ante. Alternatively, lagged distances can be taken into account as

$$\mathbf{\Pi}_t = \mathbf{\Pi}_t \left(\tilde{\mathcal{D}}_{t-1} \right), \quad \tilde{\mathcal{D}}_{t-1} = \lambda \mathcal{D}_{t-1} + (1-\lambda) \tilde{\mathcal{D}}_{t-2}, \tag{2'}$$

where $0 < \lambda \leq 1$ is a smoothing parameter to be estimated or chosen ex-ante.

To avoid an undue increase in the number of parameters, we parsimoniously model the transition probabilities as

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

where the scalar parameter γ indicates the rate of decay of the transition probabilities in terms of the cluster distances, and $\tilde{d}_{jk,t-1}$ is an element of $\tilde{\mathcal{D}}_{t-1}$. The numerator in (3) is equal to one if

$j = k$, regardless of γ. A higher value for γ leads to lower values of $\exp\left(-\gamma \tilde{d}_{jk,t-1}\right)$ 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 measure cluster proximity we adopt the Mahalanobis distance metric

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

where $\overline{\mathbf{\Sigma}}_t = J^{-1} \sum_{j=1}^{J} \mathbf{\Sigma}_{jt}$ 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 Euclidian distance is a special case of (4) and is obtained by setting $\overline{\mathbf{\Sigma}}_t = \mathrm{I}_D$.

2.2 Time-varying conditional cluster probabilities

In this section we derive 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 $\mathcal{F}_t = \{y_t, y_{t-1}, \ldots, y_1\}$ containing the observations up to time 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 $\boldsymbol{y}_{it}$,

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

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, 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 $\mathbf{\Pi}_t$. 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

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

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

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

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

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

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} is unresponsive to $\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt}$.

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

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

where the weight $w_{ij,t} = (1 + \nu_j^{-1} D) \big/ \left(1 + \nu_j^{-1} \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt}\right)' \boldsymbol{\Sigma}^{-1}_{jt} \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt}\right)\right)$ 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

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

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

Following the exponentially weighted score-driven dynamics of Lucas and Zhang (2016), the transition equation for the time-varying covariance matrices $\boldsymbol{\Sigma}_{jt}$ 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 $\boldsymbol{\nabla}^{(j)}_{\Sigma_{jt,t}} = \partial \log f\left(\boldsymbol{y}_{it}|c_{it}=j,\mathcal{F}_{t-1};\boldsymbol{\theta}\right)/\partial \text{vec}(\boldsymbol{\Sigma}_{jt})$. For the scaling matrix, we can take the analogous expression as in (10) and consider

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

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

$$\begin{aligned}\boldsymbol{\nabla}^{(j)}_{\Sigma_{jt},t} &= \tfrac{1}{2}\,\text{vec}\left(\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \boldsymbol{\Sigma}_{jt}^{-1}\,\left((\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})' - \boldsymbol{\Sigma}_{jt}\right)\boldsymbol{\Sigma}_{jt}^{-1}\right) \\ &= \tfrac{1}{2}\,\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \left(\boldsymbol{\Sigma}_{jt}^{-1}\,\otimes \boldsymbol{\Sigma}_{jt}^{-1}\right)\text{vec}\left((\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})' - \boldsymbol{\Sigma}_{jt}\right), \end{aligned} \tag{18}$$

$$\boldsymbol{S}_{\Sigma_{jt},t} = \left(\tfrac{1}{2}\,\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \boldsymbol{\Sigma}_{jt}^{-1} \otimes \boldsymbol{\Sigma}_{jt}^{-1}\right)^{-1}, \tag{19}$$

$$\text{vec}(\boldsymbol{\Sigma}_{j,t+1}) = \text{vec}(\boldsymbol{\Sigma}_{jt}) + \boldsymbol{A}_2\,\frac{\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \text{vec}\left((\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt})' - \boldsymbol{\Sigma}_{jt}\right)}{\sum_{i=1}^{N} \tau_{ij,t|t}}. \tag{20}$$

Unvectorizing (20), we obtain the covariance matrix transition equation

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

Web Appendix A.2 provides a step-by-step derivation of (21). 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

$$\boldsymbol{\nabla}^{(j)}_{\Sigma_{jt},t} = \tfrac{1}{2} \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \left(\boldsymbol{\Sigma}_{jt}^{-1} \otimes \boldsymbol{\Sigma}_{jt}^{-1}\right) \text{vec}\left(w_{ij,t}(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt})(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt})' - \boldsymbol{\Sigma}_{jt}\right), \tag{22}$$

where $w_{ij,t}$ is defined below (13). Pre-multiplying the score by the approximate scaling matrix (19) yields the transition equation

$$\boldsymbol{\Sigma}_{j,t+1} = \boldsymbol{\Sigma}_{jt} + \boldsymbol{A}_2 \frac{\sum_{i=1}^{N} \tau_{ij,t|t} \left[w_{ij,t} \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt}\right) \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt}\right)' - \boldsymbol{\Sigma}_{jt}\right]}{\sum_{i=1}^{N} \tau_{ij,t|t}}, \tag{23}$$

where the Gaussian case (21) is again a special case of (23) for $\nu^{-1} \to 0$.

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 can initialize by any cross-sectional clustering algorithm, such as k-means (Hartigan and Wong (1979)), intelligent k-means (de Amorim and Hennig (2015)), or hierarchical agglomerative clustering (Ward Jr (1963)). For this purpose we use data of $t = 1$ only, $\boldsymbol{y}_{i1}$ for $i = 1, \ldots, N$. Any such algorithm allocates our N observations in D dimensions to J clusters such that e.g. the within-cluster sum of squares is minimized. Alternatively, static clustering with time-varying parameters could be applied to all data $t = 1, \ldots, T$ (e.g. Lucas et al. (2019)).

The initial clustering algorithm provides the cluster probabilities $\tau_{ij,1|1}$. In the case of k-means, or variants thereof, 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 empirical covariance of observations $\boldsymbol{y}_{i1}$ for units i assigned to cluster j. If $\tau_{ij,1|1} \in (0, 1)$ for all i and j then probability-weighted averages over i are appropriate.

The initial $\tau_{ij,1|1}$ 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 $\boldsymbol{\Sigma}_{j1}\left(\tau_{i,1|1}\right)$ to minimize the impact from the initialization procedure.

2.4 Extensions

2.4.1 Non-Markovian transitions

In some settings, economic reasoning suggests 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_{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 extended

transition probability matrix (from row j to column k)

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

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 anew. 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 β a vector of unknown coefficients that need to be estimated. The transition probabilities can then be modeled as

$$\pi_{jk,t} = \frac{\exp\left(-\gamma \tilde{d}_{jk,t-1} + \beta' x_{jk,t}\right)}{\Sigma_{q=1}^{J} \exp\left(-\gamma \tilde{d}_{jq,t-1} + \beta' x_{jq,t}\right)} \quad \text{for } j,k = 1,\ldots,J, \tag{3'}$$

where γ and $\tilde{d}_{jk,t-1}$ 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 $\boldsymbol{\theta} = \{\text{vec}(\boldsymbol{A}_1)', \text{vec}(\boldsymbol{A}_2)', \nu_1,\ldots,\nu_J,\gamma,\beta'\}'$ can be estimated by maximizing the log-likelihood function with respect to $\boldsymbol{\theta}$, that is

$$\mathcal{L}\left(\boldsymbol{\theta}|\mathcal{F}_T\right) = \sum_{t=1}^{T}\sum_{i=1}^{N} \ell_{it}, \tag{24}$$

where the log-likelihood contribution 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 (24) can in principle be carried out by any convenient numerical optimization method. In practice, however, mixture time series models such as ours can imply irregularly shaped log-likelihood surfaces. In such cases standard numerical optimizers are at risk to converge to a local, rather than the global, maximum. More robust optimization 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

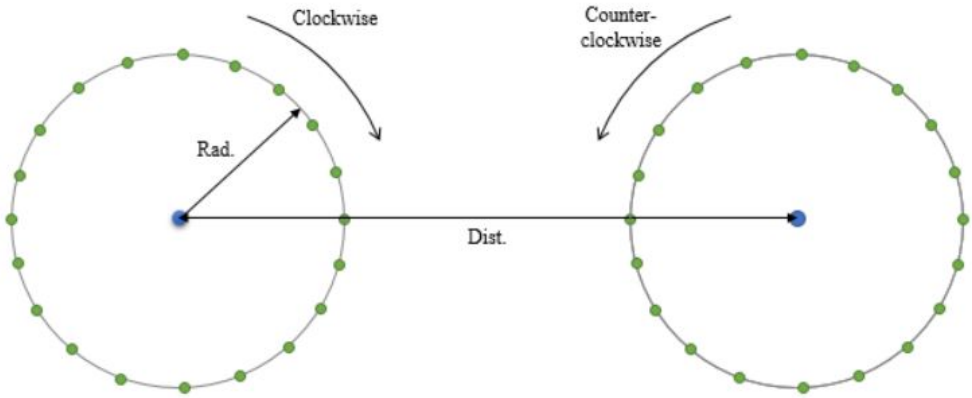
In this section we investigate the ability of our score-driven dynamic clustering model to simultaneously i) correctly classify the units of interest to distinct clusters, and ii) recover the true time-varying transition probabilities that govern cluster transitions. In all cases, we pay particular attention 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. In Section 3.3, we compare the performance of our method to the hierarchical clustering approach that is frequently used in the empirical finance literature on bank business models, see Roengpitya et al. (2014), Roengpitya et al. (2017), Ayadi et al. (2014), and Ayadi and Groen (2015).

We simulate from a mixture of dynamic bivariate normal densities. Specifically, we generate two clusters located around two distinct, time-varying cluster means. These time-varying means move along two non-overlapping circles. Our baseline setting is visualized in Figure 1. At each time t and for each of the two clusters, the units are generated using the mean as given in Figure 1, and a unit covariance matrix. Between time points, units can switch cluster using the HMM structure of the model. Key inputs into our simulations are the transition intensity parameter γ in (3), the distance between the two circle centers, and the sample sizes T and N.

We consider two choices for the transition parameter $\gamma \in \{0.3, 0.5\}$, and two choices of unconditional cluster distance $\in \{10, 12\}$. The circle radius is five in all cases, so that the two time-varying means have a tangency point in the case of the smaller distance, which makes cluster identification harder. 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 $N \in \{100, 300\}$. The number of clusters is fixed at $J = 2$ throughout, which is also imposed during estimation. Finally, in order to prevent too many

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.



switches, especially at the tangency point, we set the distance smoothing parameter λ in (2') to 0.1 for all simulations. In Section 3.3, we investigate the robustness of our method towards different choices of λ during estimation.

The time-varying cluster means evolve either clockwise, or one cluster moves clockwise and the other 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. At the same time, however, it makes it harder for the model to correctly classify each unit. Second, the circle distances become particularly interesting when one circle rotates clockwise and the other one counter-clockwise. The distances then determine how close and how

Table 1: Simulation outcomes I

Mean parameter estimates (av. $\widehat{\gamma}$), average percentage of correct classification (%C), and average mean squared errors (MSEs) for time-varying cluster means. Left panel (const. transition matrix): 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. Right panel (tv. transition matrix): 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 γ determines the frequency of transitions; lower values of γ 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 5 in all cases.

			const. transition matrix						tv. transition matrix					
			$\gamma = 0.3$			$\gamma = 0.5$			$\gamma = 0.3$			$\gamma = 0.5$		
$N/2$	T	dist	av. $\widehat{\gamma}$	%C	MSE	av. $\widehat{\gamma}$	%C	MSE	av. $\widehat{\gamma}$	%C	MSE	av. $\widehat{\gamma}$	%C	MSE
50	20	12	0.295	1.000	0.543	1.861	1.000	0.542	0.266	0.977	0.535	4.023	0.989	0.537
50	40	12	0.295	1.000	0.194	0.495	1.000	0.193	0.271	0.969	0.192	0.445	0.983	0.191
150	20	12	0.298	1.000	0.466	0.502	1.000	0.467	0.271	0.981	0.459	1.294	0.991	0.461
150	40	12	0.300	1.000	0.144	0.503	1.000	0.145	0.280	0.971	0.141	0.463	0.985	0.142
50	20	10	0.294	1.000	0.542	0.493	1.000	0.542	0.282	0.866	2.353	0.589	0.953	0.898
50	40	10	0.295	1.000	0.194	0.492	1.000	0.193	0.435	0.823	3.943	0.431	0.864	2.615
150	20	10	0.298	1.000	0.466	0.498	1.000	0.467	0.292	0.898	0.476	0.476	0.962	0.464
150	40	10	0.300	1.000	0.144	0.502	1.000	0.145	0.331	0.839	1.869	0.458	0.897	0.147

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 model parameters from the simulated data. The static parameters to be estimated include the switching intensity parameter γ, 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. $A_1 = a_1 \mathrm{I}_D$. Initial cluster parameters and allocations are obtained from k-means clustering; see Section 2.3.3 and Web Appendix B for details.

Table 1 reports our simulation results for the 32 settings we consider. The left panel reports the results when both time-varying cluster means move clockwise, and the means are thus equidistant and well-separated at all time points. As a result, the transition probabilities are time-invariant

("const. transition matrix"). In this case, the share of correct classifications is perfect (100%) and the mean tracking performance is not affected by the distance between the circles. Furthermore, the transition intensity parameter γ is estimated accurately if there is a sufficient number of transitions, i.e. if γ is small ($= 0.3$), the sample size is sufficiently large, and the unconditional distance between clusters is not too big. As expected, larger sample sizes improve the model's classification and tracking performance. Increasing the number of time points increases accuracy more than increasing the number of cross-sectional units.

The right-hand panel of Table 1 shows the results for a more challenging setup, where the cluster means start far apart, but they move towards each other in different directions, one clockwise and the other counter-clockwise ("tv. transition matrix"). Since the radii equal five, the two circles with dist$= 10$ have a tangency point after $T/2$ time points. If sample sizes are small, both classification accuracy and the ability of the model to track the time-varying means are affected. This is most severe when $T = 40$ and $N = 100$. However, the average share of correct classifications never drops below 80%, suggesting that the methodology is still useful in such challenging settings.

3.3 Robustness and comparison with benchmark clustering method

Our approach allows for a dynamic allocation of units to clusters over time. To verify whether this leads to an improved cluster assignment compared to a much simpler, static approach, we compare our previous simulation results to the outcome of a hierarchical clustering method of Ward Jr (1963). This method is popular in the empirical literature on bank business models, see Roengpitya et al. (2014), Roengpitya et al. (2017), Ayadi et al. (2014), and Ayadi and Groen (2015), where the method is applied to multivariate panel data. The hierarchical approach then treats each bank-year observation as cross-sectional and groups the entire sample, thereby allowing for cluster switches.

Table 2 reports the results. We only consider the case with time-varying transition probabilities, which has proven to be more challenging, see Section 3.2. Again, we vary the transition intensity parameter γ as well as the number of time points and cross-sectional units, and the distance be-

Table 2: Simulation outcomes II

Average percentage of correct classification and average mean squared errors for time-varying cluster means using three methodologies: (1) HMM with correctly specified distance smoothing parameter λ, (2) HMM with misspecified distance smoothing parameter $\lambda = 0.25$, while the true value is 0.1, and (3) the hierarchical clustering method of Ward Jr (1963).

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 γ determines the frequency of transitions; lower values of γ 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 5 in all cases.

			high transition intensity ($\gamma = 0.3$)						low transition intensity ($\gamma = 0.5$)					
			HMM, $\lambda = 0.1$		HMM, $\lambda = 0.25$		hierarch.		HMM, $\lambda = 0.1$		HMM, $\lambda = 0.25$		hierarch.	
$N/2$	T	dist.	%C	MSE	%C	MSE	%C	MSE	%C	MSE	%C	MSE	%C	MSE
50	20	12	0.977	0.535	0.976	0.535	0.885	0.771	0.989	0.537	0.989	0.537	0.886	0.779
50	40	12	0.969	0.192	0.962	0.192	0.885	0.818	0.983	0.191	0.976	0.191	0.885	0.721
150	20	12	0.981	0.459	0.979	0.459	0.886	0.618	0.991	0.461	0.992	0.461	0.886	0.654
150	40	12	0.971	0.141	0.964	0.142	0.883	0.699	0.985	0.142	0.978	0.141	0.885	0.628
50	20	10	0.866	2.353	0.867	2.228	0.828	1.811	0.953	0.898	0.950	1.024	0.833	1.642
50	40	10	0.823	3.943	0.836	1.578	0.829	1.729	0.864	2.615	0.840	3.161	0.835	1.542
150	20	10	0.898	0.476	0.884	0.834	0.828	1.588	0.962	0.464	0.959	0.583	0.830	1.559
150	40	10	0.839	1.869	0.845	0.310	0.830	1.487	0.897	0.147	0.855	0.702	0.829	1.559

tween unconditional means. We report two sets of results from our method: $\lambda = 0.1$ refers to the case with correctly specified distance smoothing parameter (see equation ((2'))), whereas the other value, $\lambda = 0.25$ is imposed during estimation, while the true DGP has $\lambda = 0.1$.

We find that in all settings considered, the HMM method clearly outperforms the hierarchical clustering method in terms of classification accuracy. Also the time-varying means are also recovered more precisely (smaller MSE), except in one setting with small cluster distance, small T, and small N (fifth row in Table 2). The outperformance is robust to a misspecification of the smoothing parameter λ.

4 Empirical application to bank business models

4.1 Data

Our sample consists of $N = 299$ European banks. We observe 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. Missing values occur routinely because some banks report at a quarterly frequency, while others report semi-annually. 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 (55%) 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 values. In practice, experts consider between four and up to more than ten different bank business models; see, for example, Ayadi et al. (2014), SSM (2016), 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 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 ν across clusters and variables following a preliminary data 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.

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	$\frac{\text{Assets held for trading}}{\text{Total assets}}$
	5. Derivatives held for trading	$\frac{\text{Derivatives held for trading}}{\text{Total assets}}$
Risk profile	6. Market vs. credit risks	$\frac{\text{Market risk}}{\text{Credit risk}}$
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	$\frac{\text{Domestic loans}}{\text{Total loans}}$
Funding	12. Deposits to assets ratio	$\frac{\text{Total deposits}}{\text{Total assets}}$

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 Π_t in addition to lagged cluster distances; see (3'). 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).

	M1 No transitions	M2 Markovian transitions	M3 non-Markovian transitions	M4 non-Markovian transitions II
A_1	0.894	0.850	0.813	0.967
	(0.02)	(0.02)	(0.03)	(0.02)
A_2	0.998	0.998	0.993	0.998
	(0.01)	(0.01)	(0.01)	(0.01)
ν	6.595	19.518	8.088	14.723
	(0.07)	(0.06)	(0.06)	(0.05)
γ	-	1.369	1.503	1.313
		(0.01)	(0.02)	(0.02)
β	-	-	-	-17.757
				(0.17)
P	-	0	4	4
DBI	3.14	2.94	2.93	2.92
CHI	11.89	21.08	21.12	21.09
loglik	144,253.2	150,197.1	150,003.1	150,506.9

As a result, we end up with a parsimonious yet highly flexible model with static parameter vector $\theta = \left(A_1, A_2, \nu, \gamma, \beta\right)' \in \mathbb{R}^5$. For the maximization of the likelihood, we used a simulated annealing method. Figure C.3 in the Web Appendix shows plots of directional slices of the log likelihood evaluated at the global optimum.

Table 4 reports parameter estimates and the log-likelihood fit for four different specifications of our dynamic clustering model. Model specifications M1 – M4 use the same initial cluster allocations, initial cluster mean and covariance matrix parameters, and distance smoothing parameter λ.[1]

Model M1 allows for time-varying means and covariance matrices, but rules out transitions

[1]Initial cluster allocations $\tau_{ij,1|1}$ are obtained using the static clustering approach with time-varying parameters of Lucas et al. (2019). Replacing $\tau_{ij,1|1}$ with filtered estimates from a first run, and subsequently re-estimating θ, led to negligible improvements in log-likelihood fit. Specifications M1 – M4 use the same distance smoothing parameter $\lambda = 0.25$ for quarterly data; see (2'). The log-likelihood surface is fairly flat in λ; we treat it as a tuning parameter for this reason.

across groups ($\gamma = 500$). 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 γ increases somewhat, indicating fewer (short-lived) transitions. The degrees-of-freedom parameter ν 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 Π_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 $x_{jt} \equiv \sum_i^N \hat{\tau}_{ij,t|t} \cdot \text{ROE}_{it} / \sum_i^N \hat{\tau}_{ij,t|t}$ 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. The transition matrix $\Pi_t\left(\tilde{\mathcal{D}}_{t-1}, \gamma, \beta\right)$ becomes more asymmetric (viz-a-viz Model M3) as a result. The time-varying parameter paths implied by Models M2 – M4 are visibly different from those implied by Model M1.

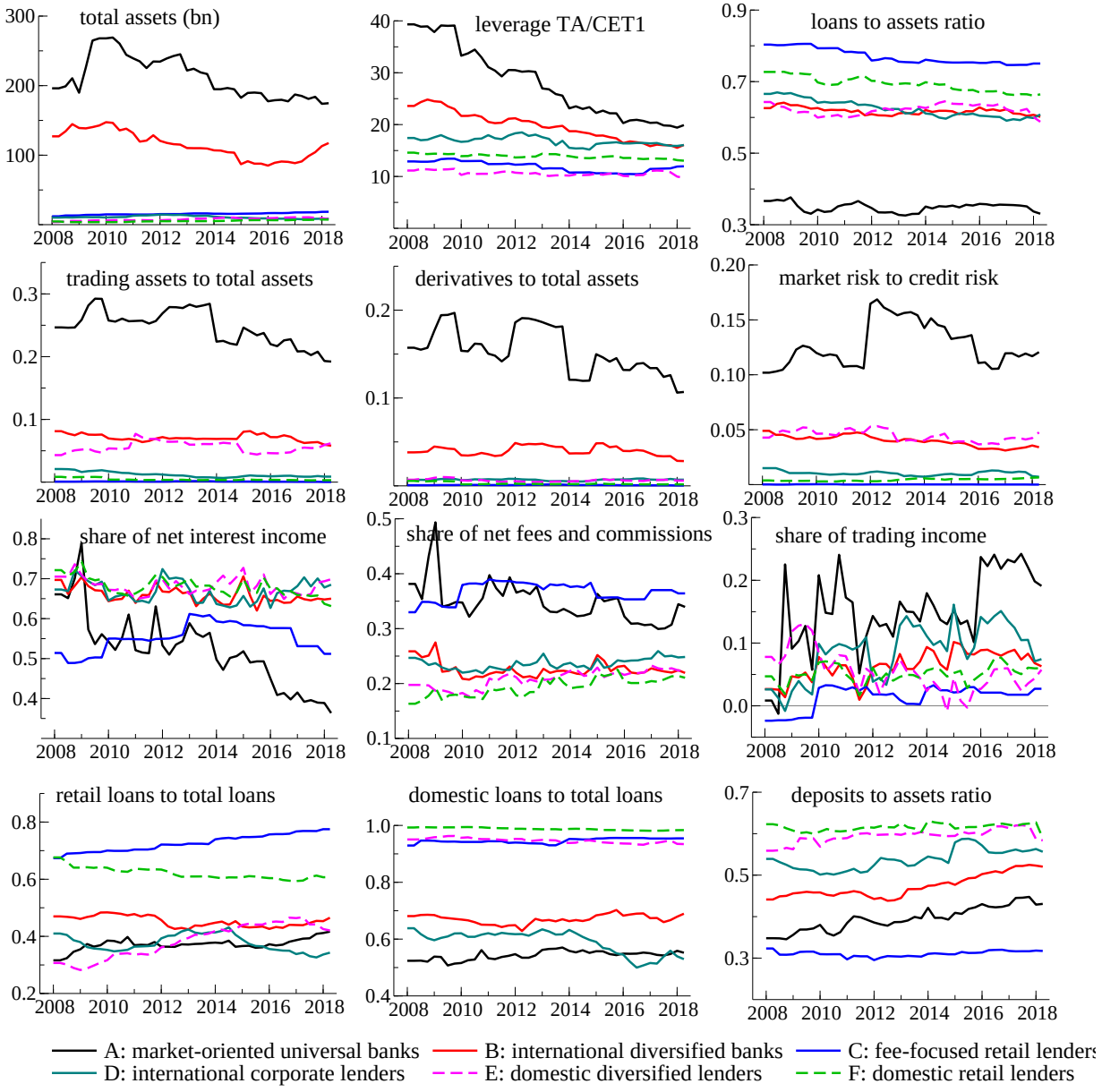
Model specification M4 is strongly preferred in terms of log-likelihood fit, and also does well in terms of non-parametric cluster validation indices (DBI). We therefore select M4 for the remainder of our empirical analysis. Using this specification, we combine model parsimony with the ability to explore 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 clusters. Specifically, we assign labels to the identified clusters to guide intuition and for ease

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 3. 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).



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 $\hat{\sigma}_{j,t|t}(d) = \left(\hat{\Sigma}_{j,t|t}(d,d)\right)^{\frac{1}{2}}$ for variables $d = 1, \ldots, D$. Business model groups A to F are ordered in terms of decreasing median bank size (total assets). Specifically, we distinguish

(A) **Market-oriented universal banks** (8.9% of bank-quarter observations; comprising firms such as Barclays, Credit Suisse, Deutsche Bank, HSBC Holdings, and Royal Bank of Scotland almost all of the time.)

(B) **International diversified banks** (15.0% of obs.; e.g. Banco Santander, Bank of Ireland, BBVA, Cooperative Rabobank, Danske Bank, ING Groep, UniCredit.)

(C) **Fee-focused retail lenders** (7.4% of obs.; e.g. Argenta Bank- en Verzekeringsgroup, all subsidiaries of Caisse Regionale de Credit Agricole, Credit Lyonnais.)

(D) **International corporate lenders** (16.6 % of obs.; e.g. Citadele Banka, Hellenic Bank, Landesbank Saar, Millennium Bank, Sberbank Europe.)

(E) **Domestic diversified lenders** (19.2% of obs.; e.g. ABH Financial, Gazprombank, Spare Bank 1, Swedbank.)

(F) **Domestic retail lenders** (32.9% of obs; e.g. Helgeland Sparebank, Newcastle Building Society, Sparebanken Sør, St. Gallener Kantonalbank.)

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 less diverse 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. Based on between-cluster variation, European banks have become less diverse in terms of size, leverage, loans-to-assets ratio, share of assets held for trading, share of derivatives held for trading, and deposits-to-assets ratio. Arguably, the convergence takes place in such a way (e.g. towards lower size, lower leverage, reduced complexity, and less flighty market funding) that does not signal an immediate financial stability concern.

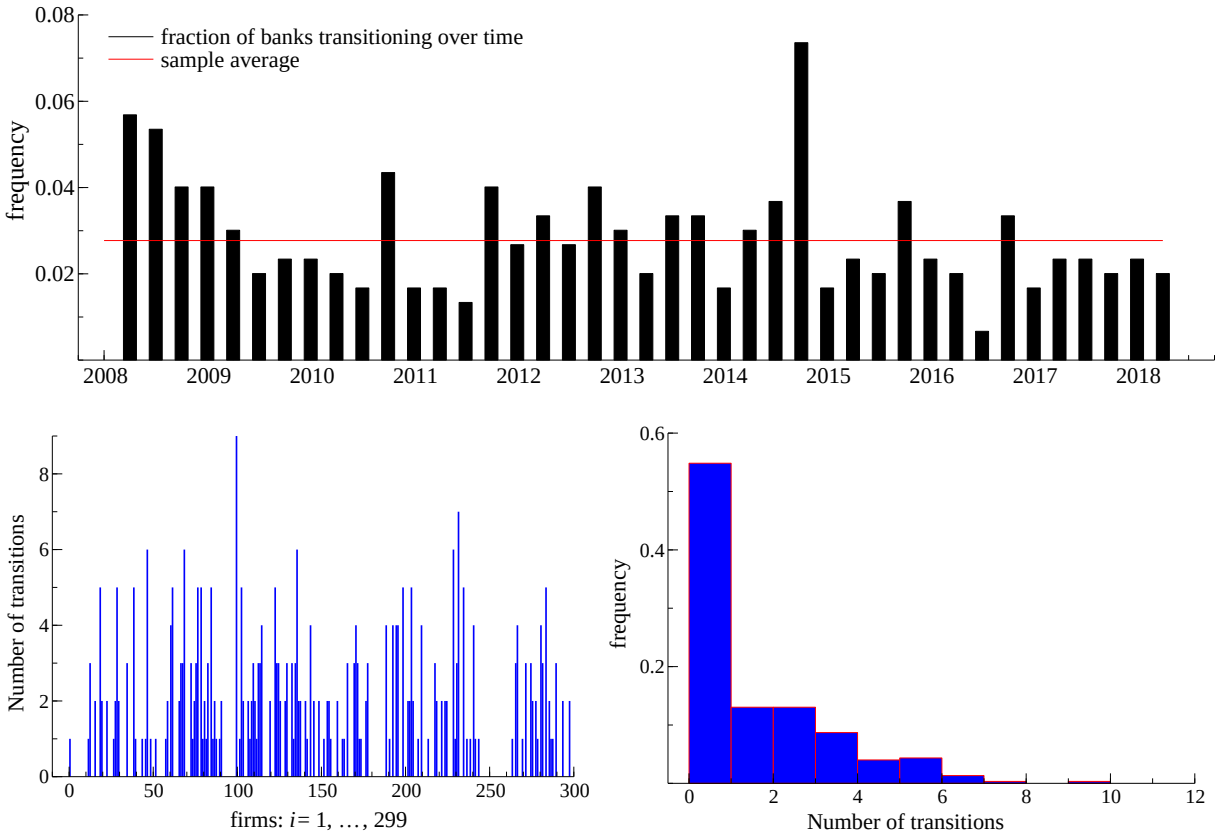
4.5 Group transitions and popularity

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 in transition intensity. Instead, the transition intensity is above-average during the Great Financial Crisis (2008), the peak of the euro area sovereign debt crisis (2012), and in anticipation of centralized SSM banking supervision in the euro area (2014). 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.

Figure 3: Timing and histogram of cluster transitions

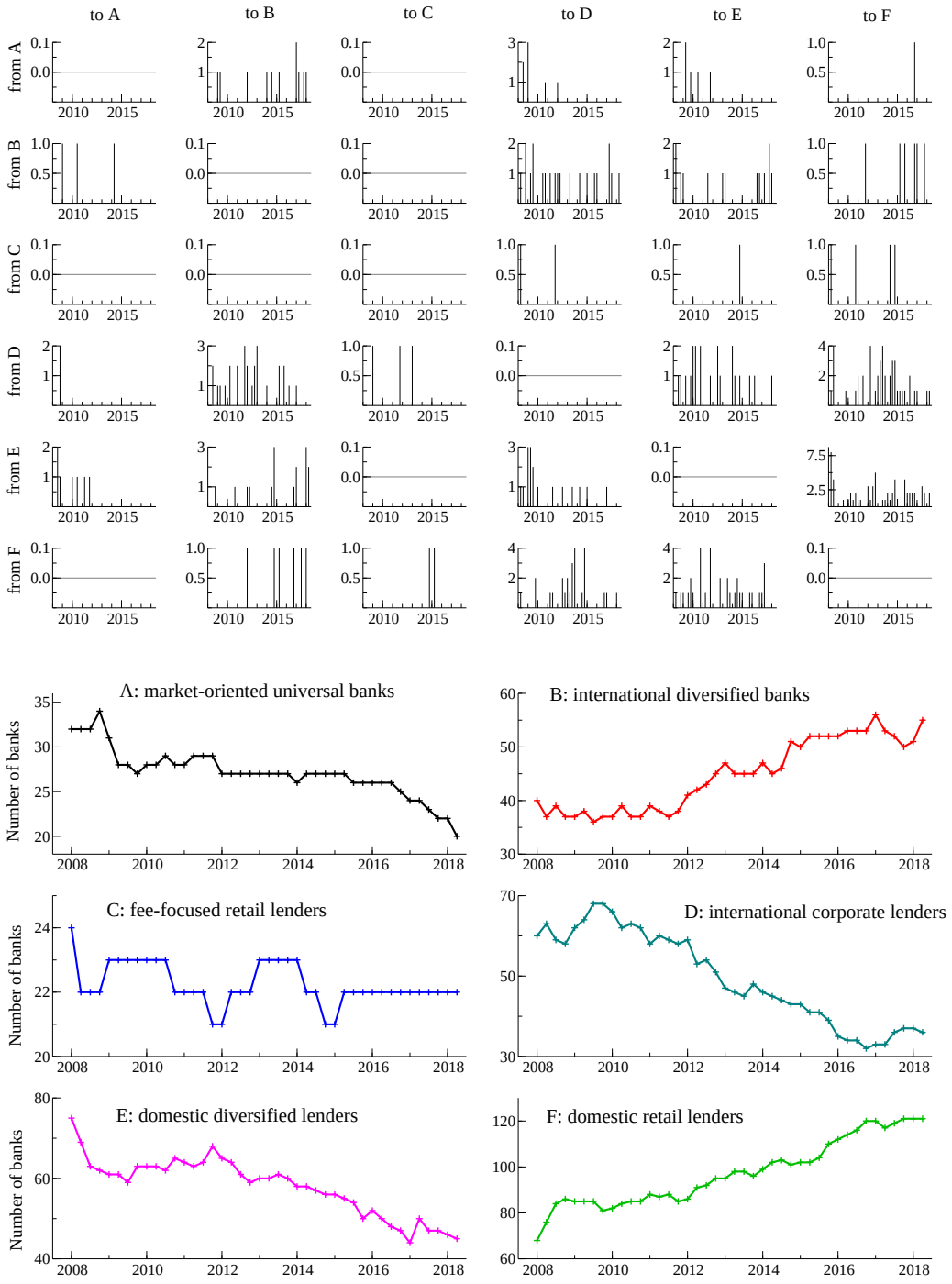
Top panel: black bars indicate the fraction of firms that are estimated to transition at each time t between 2008Q2 and 2018Q2. The red horizontal line indicates the average transition frequency. Bottom left panel: Number of transitions per firm $i = 1, \ldots, 299$. Bottom right panel: histogram of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



The top panel of Figure 4 plots the number of estimated transitions from cluster j (rows) to k (columns) at any time. Most transitions take place between 'nearby' clusters, e.g. between A and B, B and D, D and E, and E and F. 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 A, C, D, and E shrink. The observed industry trends are in line with large banks becoming less reliant 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 $\rightarrow$ F), and banks relying progressively more on more on fee income, possibly to lean against a lower profitability from increasingly low interest rates (D $\rightarrow$ C; D $\rightarrow$ F). These industry trends are in line with Figure 2 and with the discussion in e.g. ECB (2016).

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.



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.2 discusses the evolution of return

on equity (ROE) per bank cluster over time, where bank-specific ROE_{it}s 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 12% over time. Banks in cluster D (international corporate lenders) 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, adding to the move out of D to other business models, as indicated above.

5 Conclusion

We proposed a novel observation-driven model for the dynamic clustering of multivariate panel data. The cluster means and covariance matrices are time-varying to track gradual changes in cluster characteristics over time. The model has further flexibility by allowing the units of interest to transition between clusters. This is accomplished 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.

Our empirical study shows that the model, though complex, is computationally tractable as well as sufficiently flexible to answer a range of new empirical questions in multivariate panel data settings. Our results for a sample of 299 European banks between 2008Q1 and 2018Q2 suggest that European banks have become less diverse over time in some key characteristics. In addition, we find a moderate transition intensity between clusters that is related to differences in bank profitability, in line with the notion that currently low profitability entices banks to move out of their current business model and into more profitable, 'nearby' business models.

References

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.

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.

Brunnermeier, M. K. and Y. Koby (2019). The reversal interest rate. *Princeton University working paper*.

Catania, L. (2019). Dynamic adaptive mixture models with an application to volatility and risk. *Journal of Financial Econometrics, forthcoming*.

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.

de Amorim, R. C. and C. Hennig (2015). Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences 324*, 126–145.

ECB (2016). ECB Financial Stability Review, Special Feature C: Adapting bank business models – Financial stability implications. *www.ect.int, 24. November 2016*.

Farne, M. and A. Vouldis (2017). Business models of the banks in the euro area. *ECB working paper 2070*.

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

Goffe, W. L., G. D. Ferrier, and J. Rogers (1994). Global optimization of statistical functions with simulated annealing. *Journal of Econometrics 60*(1-2), 65–99.

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.

Heider, F., F. Saidi, and G. Schepens (2019). Life below zero: Bank lending under negative policy rates. *Review of Financial Studies 32*, 3728–3761.

Lucas, A., J. Schaumburg, and B. Schwaab (2019). Bank business models at zero interest rates. *Journal of Business & Economic Statistics 37*(3), 542–555.

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.

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.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association 58*(301), 236–244.

Web Appendix to

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

*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, Sonnemannstrasse 22, 60314 Frankfurt, Germany, Email: bernd.schwaab@ecb.europa.eu. 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 or the Eurosystem.

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

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

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

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

where

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

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. (2019). We obtain

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

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

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

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

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

where the robustness weight $w_{ij,t}$ 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 vec $(ABC) = (C' \otimes A)$ 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

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

Now transition equation (21) follows directly.

B Sketch of $k-$means algorithm

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 as derived in Section 2 and Web Appendix A. In principle, we can initialize by any cross-sectional clustering algorithm using data of $t = 1$ only, $\boldsymbol{y}_{i1}$, for $i = 1, \ldots, N$. We initialize by k-means in our simulation study for simplicity; see Hartigan and Wong (1979). The k-means algorithm allocates N observations in D dimensions to $k = J$ clusters such that the within-cluster sum of squares is minimized. We sketch the steps below for completeness and ease of reference.

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 10 times for different initial random centers. Chose the one with minimal within-cluster sum of squared errors.

6. **Order** the clusters, e.g. in terms of declining averages for the first variable.

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

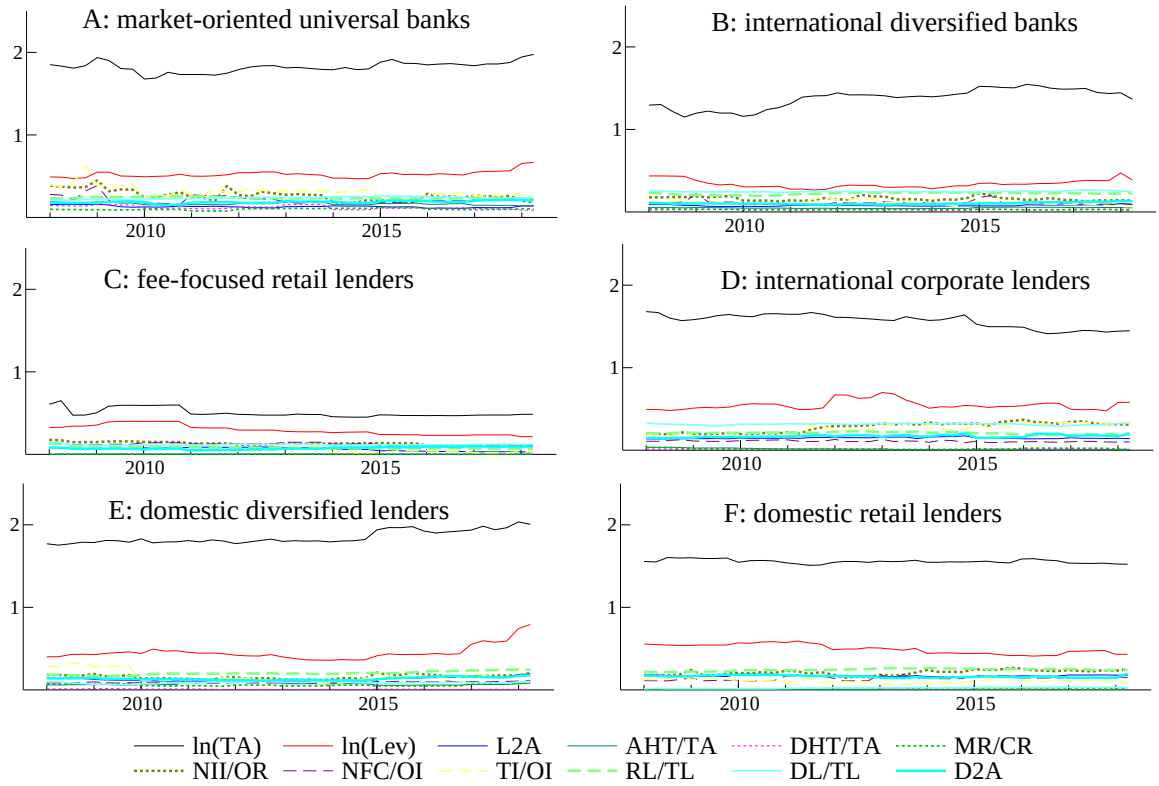
C Additional results

C.1 Estimated cluster standard deviations

Figure C.1 plots the filtered component-specific time-varying standard deviations $\hat{\sigma}_{j,t|t}(d) = \left(\hat{\Sigma}_{j,t|t}(d,d)\right)^{\frac{1}{2}}$ for variables $d = 1, \ldots, 12$. The first two variables, log total assets and log leverage, are the most dispersed across banks within each group A to F. Other variables, such as the share of assets held for trading, and the share of derivatives held for trading, are the least dispersed, particularly for banks in groups C to F.

Figure C.1: Time-varying standard deviations

Filtered time-varying standard deviations $\hat{\sigma}_{j,t|t}(d) = \left(\hat{\Sigma}_{j,t|t}(d,d)\right)^{\frac{1}{2}}$ for variables $d = 1, \ldots, 12$. Each panel contains 12 standard deviation estimates over time, corresponding to the variables listed in Table 3. The standard deviation estimates refer to model specification M4 in Table 4.

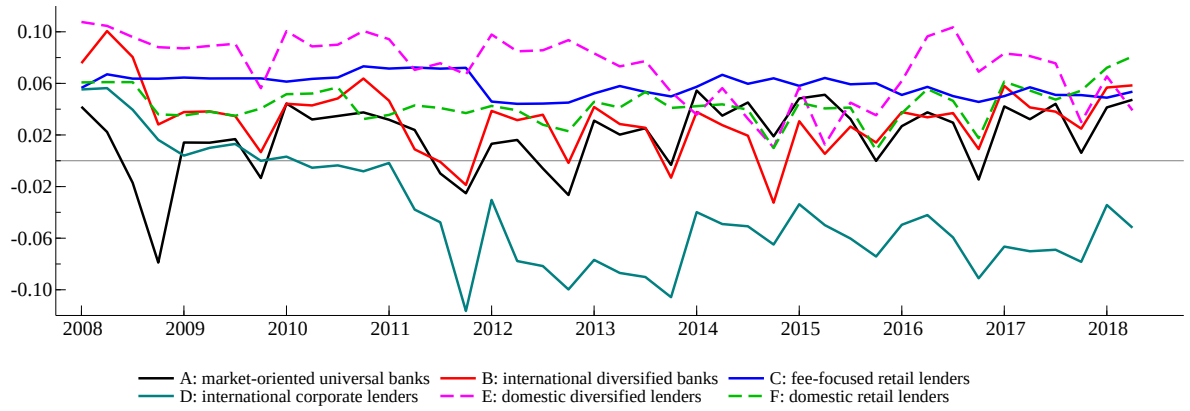


C.2 Bank profitability

The cluster transitions underlying Figures 2 – 4 are in part explained by differences in bank profitability. Figure C.2 below plots the return on equity (ROE) per cluster over time. Bank-specific observations ROE_{it} are weighted by the conditional probability $\tau_{ij,t|t}$ that bank i belongs to cluster j; see Section 4.2. ROE is not used as an input variable for the clustering; see Table 3. European banks' ROE tend to vary between approximately -2% and 12% over time. Banks assigned to cluster D (the "international corporate lenders") are an exception. 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, adding to the move out of D to other business models.

Figure C.2: Bank profitability

Return on equity (ROE) per cluster. Bank-specific observations ROE_{it}s are weighted by the conditional probability $\tau_{ij,t|t}$ that bank i belongs to cluster j.

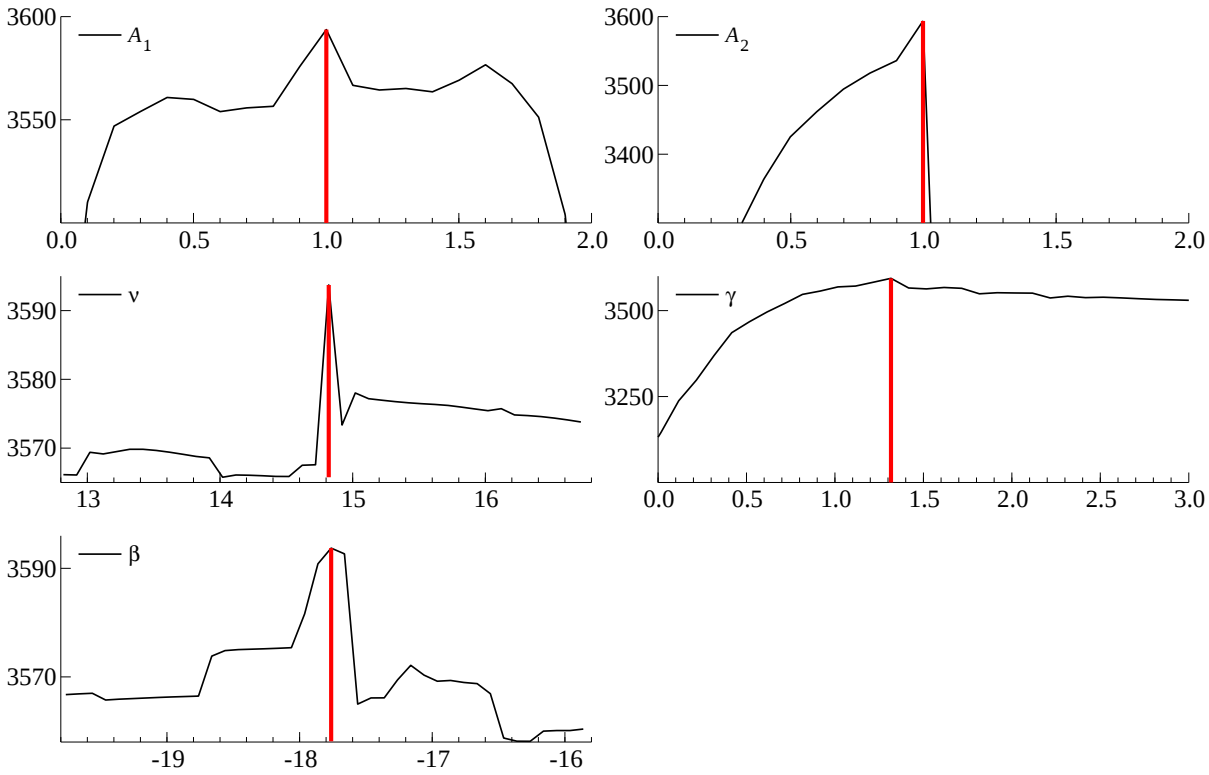


C.3 Likelihood slices

Mixture time series models such as ours can imply incalcitrant log-likelihood surfaces. Robust optimization methods such as simulated annealing (see, e.g., Goffe et al., 1994) can have an advantage over repeatedly re-running standard gradient-based optimization methods such as MaxBFGS in such cases. Figure C.3 plots directional slices through the log-likelihood function evaluated at the global optimum. Several local maxima are visible in which standard gradient-based methods are at risk of getting stuck.

Figure C.3: Log-likelihood slices

We report directional slices through the log-likelihood function (24) evaluated at the global optimum for $\theta = \left(A_1, A_2, \nu, \gamma, \beta\right)' \in \mathbb{R}^5$.



C.4 Filtered cluster probabilities and computer code

Most banks $i = 1, \ldots, 299$ are allocated fairly unequivocally to one cluster j at any time t. A file containing banks' filtered membership probabilities $\hat{\tau}_{ij,t|t}$ is available from the authors. Computer code will be made available at https://www.gasmodel.com/code.htm.

References

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

Goffe, W. L., G. D. Ferrier, and J. Rogers (1994). Global optimization of statistical functions with simulated annealing. *Journal of Econometrics 60*(1-2), 65–99.

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

Lucas, A., J. Schaumburg, and B. Schwaab (2019). Bank business models at zero interest rates. *Journal of Business & Economic Statistics 37*(3), 542–555.

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.