João, Igor Custodio; Lucas, André; Schaumburg, Julia; Schwaab, Bernd

Working Paper

Dynamic clustering of multivariate panel data

ECB Working Paper, No. 2577

Provided in Cooperation with:
European Central Bank (ECB)

Suggested Citation: João, Igor Custodio; Lucas, André; Schaumburg, Julia; Schwaab, Bernd (2021) : Dynamic clustering of multivariate panel data, ECB Working Paper, No. 2577, ISBN 978-92-899-4763-3, European Central Bank (ECB), Frankfurt a. M., https://doi.org/10.2866/657440

This Version is available at:
https://hdl.handle.net/10419/237716

EUROPEAN CENTRAL BANK
EUROSYSTEM

Working Paper Series

Igor Custodio Joao, André Lucas,
Julia Schaumburg, Bernd Schwaab

Dynamic clustering of multivariate panel data

No 2577 / July 2021

Abstract

We propose a dynamic clustering model for uncovering latent time-varying group structures in multivariate panel data. The model is dynamic in three ways. First, the cluster location and scale matrices are time-varying to track gradual changes in cluster characteristics over time. Second, all units can transition between clusters based on a Hidden Markov model (HMM). Finally, the HMM's transition matrix can depend on lagged time-varying cluster distances as well as economic covariates. Monte Carlo experiments suggest that the units can be classified reliably in a variety of challenging settings. Incorporating dynamics in the cluster composition proves empirically important in an a study of 299 European banks between 2008Q1 and 2018Q2. We find that approximately 3% of banks transition per quarter on average. Transition probabilities are in part explained by differences in bank profitability, suggesting that low interest rates can lead to long-lasting changes in financial industry structure.

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

JEL classification: G21, C33.

Non-technical summary

This paper proposes a novel statistical model for studying time-varying group (cluster) structures in high-dimensional multivariate panel data. The statistical model is dynamic in multiple ways. First, the cluster means and cluster scale matrices can be time-varying to track gradual changes in cluster characteristics over time. Second, the units of interest can transition between clusters, and can do so with time-varying transition probabilities. We demonstrate that the new 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.

The statistical model is applied to study accounting data for 299 large European banks between 2008Q1 and 2018Q2, tracking their key accounting ratios through the 2008–2009 global financial crisis, the 2010–2012 euro area sovereign debt crisis, as well as the relatively calmer period between 2013 and 2018. Our sample is also 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. Arguably, all these developments have put pressure on banks' business models and their profitability.

We stress three main empirical results. First, we interpret each cluster as identifying a certain bank business model. We distinguish six business model groups (clusters) in the data: A) market-oriented universal banks, B) international diversified lenders, C) fee-focused retail lenders, D) international corporate lenders, E) domestic diversified lenders, and F) domestic retail lenders. The similarities and differences between these groups are discussed in detail in the main text.

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 more reliant on non-market (i.e., central bank and deposit) 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, this finding suggests that monetary policy can have long-lasting effects on banking sector structure via business model transitions.

1 Introduction

Clustering is one of the most frequently-used unsupervised statistical learning techniques, with applications in marketing, psychology, sociology, medical sciences, and other fields; see e.g. McLachlan and Peel (2000) and Aggarwal and Reddy (2014) for textbook treatments. Not until recently, however, have these techniques been considered in economic contexts, where the models often need to account for dynamics in the data and potential model instabilities. For example, banking supervisors routinely need to define peer groups to benchmark supervised banks' risk provisions, profitability, and capital buffers. Bank characteristics, however, can vary significantly over time, for example owing to changes in financial regulation, information technology, and equilibrium interest rates. Some business models can then become less attractive, leading banks to adapt their strategies and thus change peer groups (cluster membership). Clustering each cross section of banks at each time in isolation is possible, but fraught with practical problems and, at best, statistically inefficient. Despite their obvious need, dynamic clustering models for multivariate panel data in economics and finance are rare.

This paper proposes a new dynamic clustering model for studying time-varying group structures in multivariate and potentially high-dimensional panel data. The model is dynamic in multiple ways. First, the cluster means are time-varying to track gradual changes in group (cluster) characteristics over time. The static parameters governing the evolution of the time-varying parameters can be made variable- and cluster-specific if this is preferred by the data at hand. Second, all units can transition between clusters based on a Hidden Markov model (HMM). Third, the HMM's transition probabilities are time-varying and can depend on lagged (time-varying) cluster distances and, potentially, additional conditioning variables. Our modeling framework proves useful for allocating a potentially large number of cross-sectional units with vector-valued measurements into a much smaller number of approximately homogeneous groups, even in fairly complicated dynamic settings, while keeping track of overall trends, cluster memberships, and transitions probabilities. Intuitive filtering recursions are available for all time-varying parameters and each unit's cluster membership probabilities. Finally, the baseline model can also be extended to accommodate inac-

tive states and non-Markovian transition behavior.

Our paper relates to the recent literature on panel models with group-specific heterogeneity patterns, but it also differs from it in terms of scope and model formulation. A widely-regarded contribution in this area is Bonhomme and Manresa (2015) who propose a group fixed effects model which relies on variants of k-means clustering of observations in a first stage. Their main interest, however, lies in the estimation of a set of structural coefficients. Similarly, Lin and Ng (2012) use both threshold regressions and conditional k-means to estimate panel models with grouped heterogeneous slope coefficients. Ando and Bai (2016) extend these models by allowing for a factor error structure and suggest an estimation procedure employing a penalty function proposed by Fan and Li (2001). Su et al. (2016), on the other hand, study generalized linear models with grouped coefficients, using a penalized likelihood approach, and Lu and Su (2017) provide a method for determining the number of groups in that framework.

In contrast to the above models, our model does not include regressors for a single dependent variable, but rather considers a vector of dependent variables. Instead of estimating the structural coefficients in a regression set-up and predicting the outcome variable (which is generally the goal of supervised learning), our main objective is to track and interpret group-specific parameter dynamics in a multivariate location-scale setting. Therefore, it may be seen as a dynamic approach to unsupervised learning for multivariate, potentially high-dimensional panel data. In contrast to the papers above, we allow for switches in group membership over time.

In addition, our paper contributes to the literature on clustering methods with dynamic group compositions, and to the literature on parameter heterogeneity in score-driven models with time-varying parameters.[1] Creal et al. (2014) study the dynamics of market-based credit ratings of U.S. public companies, using a clustering approach with fixed categories and a Bayesian estimation method. Catania (2019) considers a mixture model with dynamic group membership and

[1] We also build on an earlier literature on HMM models (see, for instance, Hamilton, 1989; Frühwirth-Schnatter and Kaufmann, 2008; Hamilton and Owyang, 2012), but also depart from it by allowing all cluster transition probabilities to depend on lagged cluster distances as well as additional conditioning variables. This is relevant for our empirical application, but also non-trivial, given that all cluster location and scale matrix parameters vary over time, and a parsimonious specification is needed to link the possibly high-dimensional matrix of state transition probabilities to the lagged cluster distances and additional conditioning variables.

time-varying parameters, and applies it to a large set of financial asset returns. In contrast to the application used by 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. Finally, we explicitly consider heterogeneity in the parameters that govern the (score-driven) dynamics of the cluster means. By contrast, most multivariate models with such dynamics use a highly restrictive specification for the speed with which the time-varying parameters adjust, typically requiring all adjustment speeds to be the same; see for instance Opschoor et al. (2018) and Lucas et al. (2019). This is increasingly untenable in higher-dimensional settings such as ours. In contrast to these earlier models, we find that is empirically relevant to allow for different speeds of adjustment in both the variable (D) and cluster (J) dimension.

Extensive Monte Carlo experiments suggest that our model is able to accurately classify units into their respective true clusters at each time. It also recovers all relevant time-varying and static parameters, despite the presence of cluster transitions. In our simulations, the cluster classification is almost perfect for sufficiently large distances between the time-varying cluster means. As the unconditional cluster distances decrease and transitions become more frequent, the share of correct classifications decreases, but generally remains high. Allowing for heterogeneity in the cluster mean updating parameters improves the performance of our model. The gains are particularly large if the variables in the model differ with respect to their signal-to-noise ratios. Finally, we compare our method with the hierarchical clustering method of Ward (1963), which is frequently used in the empirical literature on bank business model transitions (Roengpitya et al., 2017; Ayadi et al., 2020). We find that the new model achieves higher classification and filtering accuracy across all settings considered.

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. 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 be-

tween 2013 and 2018.[2] Our sample is characterized by a significant increase in post-crisis financial regulation, the introduction of centralized European Central Bank supervision, 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, forcing them to adapt to changes in the external environment. While banks can reasonably be assumed to adhere to a single business model over a more limited time span (Lucas et al., 2019), this assumption becomes increasingly difficult to defend as the sample size T grows to comprise more than a decade of data.

We identify six business model groups (clusters) from banks' accounting data, and highlight two main empirical results. First, business model popularity changed over time. Two business model clusters (international diversified lenders and domestic retail lenders) grew in popularity, one remained approximately stable (fee-focused retail lenders), and the remaining three sank (market-oriented universal banks, international corporate lenders, and domestic diversified lenders). Overall, our transition estimates and cluster location estimates point in the same direction: since the start of our sample, European banks have relied increasingly on fee income to lean against impaired profitability from e.g. low interest rates and increased competition, have become more reliant on non-market (i.e., central bank and deposit) funding, and have lent increasingly to retail clients rather than corporate clients. These industry trends are broadly in line with policy discussions in e.g. ECB (2016) and Ayadi et al. (2020).

Second, we find that bank business model transitions are in part explained by differences in cluster-specific point-in-time profitability measures. Differences in cluster-specific return-on-CET1-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, and 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

[2]Computer code illustrating our dynamic clustering model, as well as the cluster membership probabilities for our sample of European banks, is available at www.gasmodel.com/code.htm.

long-lasting effects on financial industry structure via bank business model transitions.[3]

We proceed as follows. Section 2 presents our observation-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 Dynamic clustering model

2.1 Markov chain cluster transitions

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

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

where $\boldsymbol{\mu}_{c_{it},t}$ is a $D \times 1$ vector of cluster-specific means, and $\boldsymbol{\epsilon}_{it}$ is a $D \times 1$ vector of Student's t-distributed error terms characterized by a zero mean, a (possibly time-varying) and cluster-specific $D \times D$ scale matrix $\boldsymbol{\Sigma}_{c_{it},t}$, and degrees-of-freedom parameter $\nu_{c_{it}}$ for unit i at time t, and where the latent state variable c_{it} is driven by an underlying Markov chain. The multivariate Student's t distribution encompasses the special case of the multivariate normal distribution, for which we can set $\nu_{c_{it}}^{-1} = 0$. Further extensions of (1) can include skewed distributions such as in Lucas et al. (2014, 2017).

We model the transitions from one cluster to the next by a Hidden Markov Model (HMM); see

[3] Financial structure, in turn, has been linked to a variety of issues including economic growth and risk (Popov and Manganelli, 2015), innovation and economic dynamism (Cavalleri et al., 2019), and the ability to fund green initiatives combating climate change (De Haas and Popov, 2021).

e.g. Goldfeld and Quandt (1973), Hamilton (1989), Bhar and Hamori (2004), Fruehwirth-Schnatter (2006), and Bazzi et al. (2017). The latent (hidden) states c_{it} evolve over time as characterized by the HMM's transition dynamics. There are as many Markov Chains as there are units $i = 1, \ldots, N$, all of which are (conditionally on past data) independent. The transition probabilities are restricted to be the same across units. The Markov property implies that the next state depends only on the current state, i.e.

$$\mathbb{P}\left\{c_{i,t+1} = j | c_{it}, c_{i,t-1}, \ldots, c_{i1}\right\} = \mathbb{P}\left\{c_{i,t+1} = j | c_{it}\right\}.$$

We introduce the short-hand notation $\pi_{jkt} := \mathbb{P}\left\{c_{i,t+1} = k | c_{it} = j\right\}$, where π_{jkt} does not depend on i and 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 the transition probabilities $0 \leq \pi_{jkt} \leq 1$ for $j, k = 1, \ldots, J$ at time t. The transition probabilities are common to all units $i = 1, \ldots, N$. We require the rows of $\mathbf{\Pi}_t$ to sum to one, i.e., $\sum_{k=1}^{J} \pi_{jkt} = 1$ for all $j = 1, \ldots, J$. We assume the transition probabilities π_{jkt} vary over time as a function of the time-varying distances between the clusters at time $t-1$. In particular, the transition matrix can be specified 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_{jkt}, where d_{jkt} 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_{jkt}. If two or more clusters are close to each other or even overlapping at time $t-1$, then specification (2) may not work well. In such cases (2) can be adapted to, for example,

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

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

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

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 elements of $\mathbf{\Pi}_t$ are positive and its rows sum to one by construction. As a result, the matrix $\mathbf{\Pi}_t$ need not be symmetric despite the symmetry of the distance measure $d_{jk,t-1}$.

To measure cluster proximity we adopt the distance metric

$$d_{jkt} = \sqrt{(\boldsymbol{\mu}_{jt} - \boldsymbol{\mu}_{kt})' \, \overline{\mathbf{\Sigma}}_t^{-1} \, (\boldsymbol{\mu}_{jt} - \boldsymbol{\mu}_{kt})}, \tag{4}$$

where $\overline{\mathbf{\Sigma}}_t = J^{-1} \sum_{j=1}^{J} \mathbf{\Sigma}_{jt}$.[4] The Euclidian distance between $\boldsymbol{\mu}_{jt}$ and $\boldsymbol{\mu}_{kt}$ is a special case of (4), and is obtained by setting $\overline{\mathbf{\Sigma}}_t = \mathrm{I}_D$. As a result of scaling by $\overline{\mathbf{\Sigma}}_t$, however, cluster distances become invariant to alternative ways of scaling the input variables $\boldsymbol{y_{it}}$. In addition, variables that are less correlated with the others then receive more "weight" in the distance metric. This is often desirable in economic contexts.

2.2 Conditional cluster membership probabilities

In this section we derive a filtering equation for the filtered 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

[4] Alternatively, the distance function $d^*_{jkt} = [(\mathbf{\Sigma}_{jt}^{-1/2}\boldsymbol{\mu}_{jt} - \mathbf{\Sigma}_{kt}^{-1/2}\boldsymbol{\mu}_{kt})'(\mathbf{\Sigma}_{jt}^{-1/2}\boldsymbol{\mu}_{jt} - \mathbf{\Sigma}_{kt}^{-1/2}\boldsymbol{\mu}_{kt})]^{1/2}$ could be used, particularly in settings where the scale matrices differed substantially across clusters. We use (4) in our empirical work since it is numerically more stable and slightly faster to compute.

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; see Section 2.5.

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 predicted conditional probability that unit i belongs to cluster j at time t given $\mathcal{F}_{t-1}$. By the Markov property the predicted conditional state probability $\tau_{ij,t|t-1}$ only depends on the previous state and on elements of the transition matrix $\boldsymbol{\Pi}_t$. We use this property to update the cluster probabilities as

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

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 updating equations (6) – (7) are effectively Hamilton (1989)'s filter applied to our model. The filtered cluster probabilities $\tau_{ij,t|t}$ 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 cluster j's 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 than from cluster k, 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 mean and scale matrix parameters

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). The mean and scale matrix transition equations presented here are identical to those in Lucas et al. (2019), to which we refer for details.

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

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

where $\boldsymbol{\nabla}^{(i)}_{\mu_{jt},t} = \partial \log f\left(\boldsymbol{y}_{it} | c_{it} = j, \mathcal{F}_{t-1}; \boldsymbol{\theta}\right) / \partial \boldsymbol{\mu}_{jt}$ is unit i's contribution to the score of mixture component j. For our case of a mixture of Student's t distributions,

$$\boldsymbol{\nabla}^{(i)}_{\mu_{jt},t} = w_{ijt} \cdot \boldsymbol{\Sigma}_{jt}^{-1} \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt}\right), \tag{10}$$

where the robustness weight $w_{ijt} = \left(1 + \nu_j^{-1} D\right) / \left(1 + \nu_j^{-1} \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt}\right)' \boldsymbol{\Sigma}_{jt}^{-1} \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt}\right)\right) \to 1$ as $\nu_j^{-1} \to 0$. As a closed-form expression for the conditional Fisher information matrix of $\boldsymbol{\mu}_{jt}$ is not

available, we follow Lucas et al. (2019) and use

$$\boldsymbol{S}_{\mu_{jt},t} = \left(\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \mathbb{E} \left[\boldsymbol{\nabla}^{(i)}_{\mu_{jt},t} \left(\boldsymbol{\nabla}^{(i)}_{\mu_{jt},t} \right)' \middle| c_{it} = j \right] \right)^{-1} \tag{11}$$

to correct for the score's curvature. The scaling matrix $\boldsymbol{S}_{\mu_{jt},t}$ is the inverse of the weighted average of each unit's contribution to the conditional Fisher information matrix of regime j, weighted by the filtered probabilities $\tau_{ij,t|t}$. Combining (10) and (11), we have

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

where we refer to Lucas et al. (2019) and Web Appendix A.1 for further details. The transition equation (12) is highly intuitive: the cluster means are updated by the prediction errors of their respective clusters, accounting for the posterior probabilities that an 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 $\boldsymbol{\mu}_{jt}$ is unresponsive to $\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt}$. The weight w_{ijt} provides the parameter paths 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_{ijt}.

Similarly, the transition equation for the time-varying cluster scale matrices $\boldsymbol{\Sigma}_{jt}$ is given by

$$\boldsymbol{\Sigma}_{j,t+1} = \boldsymbol{\Sigma}_{jt} + \boldsymbol{A}_2 \frac{\sum_{i=1}^{N} \tau_{ij,t|t} \left[w_{ijt} \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right) \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right)' - \boldsymbol{\Sigma}_{jt} \right]}{\sum_{i=1}^{N} \tau_{ij,t|t}}, \tag{13}$$

where we refer to Web Appendix A.2 for the derivation. Again, the transition equation is highly intuitive: the components of the scale matrix are updated by the difference between the outer product of the prediction errors and the current scale matrix for that cluster, weighted by the filtered probabilities that the observation was drawn from that same cluster. The weights w_{ijt} ensure that outlying observations $\boldsymbol{y}_{it}$ have a reduced impact on the estimated dispersion of data points associated to the respective cluster.

To start the filtering recursions (12)–(13), we require intial values for $\boldsymbol{\mu}_{j1}$ and $\boldsymbol{\Sigma}_{j1}$. In our empirical application, we use the static clustering approach of Lucas et al. (2019) to first obtain time-invariant cluster membership probabilities $\tau_{ij,1:T}$, and then obtain initial cluster means $\boldsymbol{\mu}_{j1}$ and cluster scale matrices $\boldsymbol{\Sigma}_{j1}$ conditional on these $\tau_{ij,1|1} = \tau_{ij,1:T}$; see Web Appendix A.3 for further details. Alternatively, an initialization can be obtained via the k-means algorithm using only data from the first cross-section.

2.4 Extensions

This section considers three extensions to the baseline dynamic clustering model as presented thus far. Our empirical study in Section 4 combines all these extensions. We first discuss parameter heterogeneity across variables $d = 1, \ldots, D$ and/or clusters $j = 1, \ldots, J$. We then allow for non-Markovian transition dynamics. Finally, we incorporate additional explanatory variables to better explain the Markov chain transition dynamics.

2.4.1 Parameter heterogeneity across variables and clusters

The transition equations (12) and (13) specify how the cluster means and cluster scale matrices evolve over time. In the most parsimonious case, two scalar parameters determine by how much the respective scaled scores are adjusted: $\boldsymbol{A}_1 = a_1 \cdot I_D$ and $\boldsymbol{A}_2 = a_2 \cdot I_D$ for scalars a_1 and a_2. Such a parameterization may, however, be too restrictive for the data at hand. For example, the means μ_{djt} of some of the variables $d = 1, \ldots, D$ may vary more strongly over time than others. Similarly, the mean vector may move more quickly for some clusters j than others, for instance if some banks were subjected to stricter supervision than others. These plausible alternatives require the introduction of additional heterogeneity in the score updates.

Focusing on the vectors of means $\boldsymbol{\mu}_{jt}$, we therefore use the updating equation

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

where

$$\boldsymbol{A}_{1,j} = a_1 \cdot I_D + \text{diag}\left(\bar{a}_1^D, \ldots, \bar{a}_D^D\right) + \bar{a}_j^J \cdot I_D, \qquad j = 1, \ldots, J. \tag{15}$$

The additive structure in (15) is both flexible and parsimonious. The diagonal matrix containing $\bar{a}_d^D$ for $d = 1, \ldots, D$ allows for different adjustment speeds across variables y_{dit}. Similarly, $\bar{a}_j^J$ for $j = 1, \ldots, J$ allows for different adjustment speeds across clusters. For identification, we need to restrict two parameters in (15). Without loss of generality, we choose to set $\bar{a}_1^D = \bar{a}_1^J = 0$. In our application to European banks in Section 4, we find that allowing for heterogeneous adjustment speeds as in (15) is empirically relevant.

2.4.2 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 very 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. A period later, the probability of switching back from 2 to 1 may then be 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 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 of transiting from state j to state k. We now model such a transition as a change from state

$j = (j, 0)$ to state (k, P).[5] For $p > 0$, only transitions occur from state (k, p) to state $(k, p-1)$. For instance, if $P = 2$, and $J = 2$, we would get the extended transition probability matrix (from row j to column k)

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

It is clear that the number of parameters is the same as in the benchmark model. The intuition for the above transition matrix is as follows. If a bank starts with business model 1, it can migrate to state $(1, p = 0)$ with probability $\pi_{11,t}$, and to state $(2, p = 2)$ with probability $\pi_{12,t}$. If it migrates to state $(2, p = 2)$, the next period it migrates to state $(2, p = 1)$ with probability 1, and the period after that to state $(2, p = 0)$. Only in state $(2, p = 0)$, the bank is at risk of a business model migration again, namely with probability $\pi_{21,t}$. If such a change hits, the bank would migrate to state $(1, p = 2)$ takes place. Then it again takes 2 periods to land via state $(1, p = 1)$ into state $(1, p = 0)$, where the whole process can start anew. As P is 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.

[5] A similar approach of extending the discrete state-space is found in the credit rating momentum literature (Lando and Skødeberg, 2002; Christensen et al., 2004; Koopman et al., 2009), where 'active' states are introduced to model a cascade of state switches. Here, by contrast, we introduce the opposite of 'inactive' states to prevent economically unmeaningful switching behavior.

2.4.3 Explanatory covariates

The cluster transition matrix $\mathbf{\Pi}_t$ in (2′) could 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 in a low profitability cluster could have an 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 De Groen (2015) and Roengpitya et al. (2017). Using additional conditioning variables allows us to incorporate and test for such effects. Let $\boldsymbol{x}_{jkt}$ denote a vector of observed covariates that may have an impact on transition probability π_{jkt}, and $\boldsymbol{\beta}$ a vector of unknown coefficients that need to be estimated. The transition probabilities can then be modeled as

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

where γ and $\tilde{d}_{jk,t-1}$ are defined below equation (3). The inclusion of covariates results in increased asymmetry in the transition probability matrix. For instance, if the profitability difference between clusters j and k is large and positive, the transition from j into k becomes less likely, whereas that from k into j becomes more likely.

2.5 Parameter estimation

As the model is observation-driven, the log-likelihood is known in closed form as

$$\mathcal{L}\left(\boldsymbol{\theta}|\mathcal{F}_T\right) = \sum_{t=1}^{T}\sum_{i=1}^{N} \ell_{it}, \tag{16}$$

with $\boldsymbol{\theta} = \{a_1, \bar{a}_2^D, \ldots, \bar{a}_D^D, \bar{a}_2^J, \ldots, \bar{a}_J^J, a_2, \nu_1, \ldots, \nu_J, \gamma, \boldsymbol{\beta}'\}'$, and 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 (16) can be carried out by any convenient numerical optimization method.

In the context of our empirical study in Section 4, we observe that, when not allowing for parameter heterogeneity as introduced in Section 2.4.1, the log-likelihood surface can be quite irregular. This is due to one parameter being then used to summarize a range of quite different adjustment speeds. The log-likelihood surface becomes considerably more regular after allowing for parameter heterogeneity.

3 Simulation study

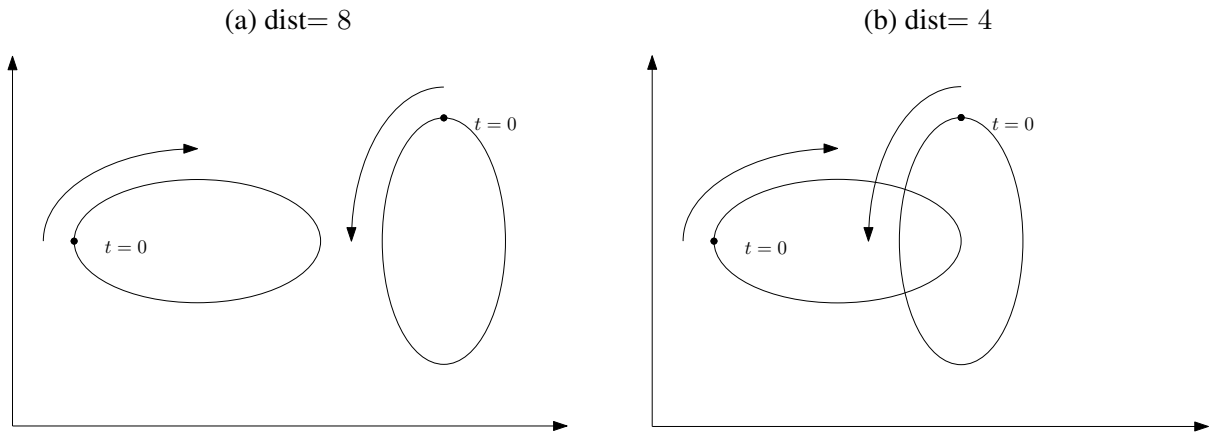
3.1 Simulation design

This section investigates the ability of our dynamic clustering model to simultaneously (i) correctly classify the units of interest into distinct clusters at each point in time, and (ii) recover the true time-varying cluster mean trajectories. 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 impact of allowing for heterogeneous dynamics. Finally, we compare the performance of our method to the hierarchical clustering approach of Ward (1963), which is frequently applied to panel data in the empirical finance literature on bank business models.

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 (possibly overlapping) ellipses, starting from different positions, and in different directions (clockwise and counter-clockwise), such that they move towards each other initially. Consequently, the Euclidean distance between the means is not constant, which implies time-variation in the transition probabilities. At each time t and for each of the two clusters, the units are generated using their respective cluster's mean and a common time-invariant diagonal covariance matrix. From one time point to the next, units can switch cluster according to the HMM structure of the model. Key inputs into our simulations are the transition intensity parameter γ in (3), the distance between the two ellipses' centers, the entries of the scale matrix, and the sample sizes T

Figure 1: Illustration of DGP: two clusters with time-varying means, at two distance settings

We simulate bivariate data $D = 2$ from two clusters $J = 2$. The two true time-varying means move in ellipses that are generated by sinusoid functions. The time-varying cluster means evolve clockwise for one cluster, and counter-clockwise for the other, implying time-variation in cluster distance and transition probabilities. Each cluster starts on different parts of their ellipse, such that their means do not overlap, even when the trajectories cross, as in panel (b).



and N. An illustration of the DGP can be found in Figure 1, for two distance settings.

Heterogeneity across variables and clusters is introduced in two ways. First, using ellipses stretched in different directions, as opposed to circles, implies that the cluster means evolve in different step sizes for each dimension and therefore have different adjustment speeds. Second, by stretching the ellipses along different directions, such that one is "lying down" while the other is "standing up", the two clusters differ from each other. Third, we introduce different signal-to-noise ratios by defining the scale matrices as $\mathbf{\Sigma} = \text{diag}\,(1, \sigma_2^2)$ when generating the data. We examine the tracking and classification performance of our method under a homogeneous version of the mean transition equation (8), in which $\boldsymbol{A}_1 = a_1 \cdot I_2$, and a heterogeneous one, where $\boldsymbol{A}_1 = \text{diag}\,(a_1^1, a_1^2)$; see Section 2.4.1.

We consider two choices for the transition parameter $\gamma \in \{0.25, 0.5\}$, and two choices for the distance between the centers of the two ellipses, dist $\in \{4, 8\}$. The variance of the second variable is chosen as $\sigma_2^2 \in \{1, 8\}$, so that we either have an identity covariance matrix, or a low signal-to-noise ratio for the second variable. The ellipses have radii 2 and 4, so that the trajectories of the two time-varying means intersect twice in the case of the smaller distance. This makes

cluster identification more challenging. The sample sizes are chosen approximately in line with the empirical application with $T = 40$ and $N = 200$, with 100 units in each cluster at time $t = 1$. The number of clusters is fixed at $J = 2$. Finally, to prevent too many switches, especially when the cluster means are close to each other, we set the distance smoothing parameter λ in (2′) to 0.1 for all simulations. The transition probabilities π_{jkt} are time-varying as they depend on the past distances between the two clusters; see equation (3).

We are particularly interested in two issues. First, we examine the impact of allowing for parameter heterogeneity on the performance of the method. If the signal-to-noise ratio is low for one variable (which is the case when its variance is high), the mean update is likely to be slower, which may imply a smaller estimate of a_1^2 in the version with heterogeneous parameters $\boldsymbol{A}_1 = \text{diag}\,(a_1^1, a_1^2)$. Imposing $\boldsymbol{A}_1 = a_1 \cdot I_2$ may then lead to a decrease in accuracy. Second, the lower γ, and the smaller 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. On the other hand, however, many transitions may make it harder for the model to correctly classify each unit. Initial cluster parameters and allocations are obtained from k-means clustering; see e.g. Hartigan and Wong (1979) for details.

3.2 Simulation results

Table 1 presents the results. The left panel ("Homogeneous $\boldsymbol{A}_1$") refers to the model specification with $\boldsymbol{A}_1 = a_1 \cdot I_2$, while the center panel ("Heterogeneous $\boldsymbol{A}_1$") refers to the more flexible specification with $\boldsymbol{A}_1 = \text{diag}\,(a_1^1, a_1^2)$.

When the distance is large (dist $= 8$), both model specifications perform very well. We observe a slightly better tracking error for the mean when using the heterogeneous specification and $\sigma_2^2 = 8$, as indicated by the lower average mean squared error (MSE), but approximately the same classification performance. When the distance is reduced to 4, the classification problem becomes harder and the heterogeneous $\boldsymbol{A}_1$ achieves a considerably better mean tracking performance. A

Table 1: Simulation outcomes

We report average parameter estimates ($\widehat{\gamma}$), average percentage of correct classification (%C), average mean squared errors (MSE) for time-varying cluster means, and log-likelihood (LL). Results are presented for the homogeneous ($\boldsymbol{A}_1 = a_1 \cdot I_2$) and heterogeneous ($\boldsymbol{A}_1 = \text{diag}\left(a_1^1, a_1^2\right)$) cases of equation (8). The rightmost panel displays the results of the hierarchical clustering method of Ward (1963) as a benchmark. The sample size is $N = 200$ and $T = 40$. The transition intensity parameter γ determines the frequency of transitions; lower values of γ imply a higher number of transitions in expectation. Higher σ_2^2 translates to stronger parameter heterogeneity and thus a better comparative performance of the heterogeneous $\boldsymbol{A}_1$. The distance between ellipse centers measures the distinctiveness of clusters. Each setting is simulated 250 times.

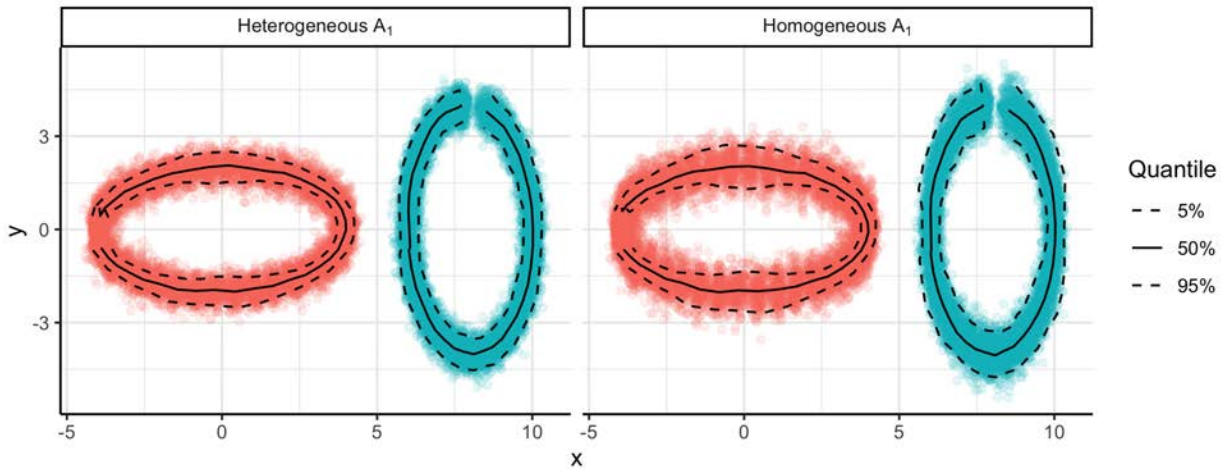
			Homogeneous $\boldsymbol{A}_1$				Heterogeneous $\boldsymbol{A}_1$				Hierarchical	
dist.	σ_2^2	γ	$\widehat{\gamma}$	%C	MSE	LL	$\widehat{\gamma}$	%C	MSE	LL	%C	MSE
4	1	0.25	0.246	0.947	0.141	-900.9	0.234	0.949	0.096	-898.8	0.729	2.161
4	1	0.50	0.470	0.963	0.091	-853.7	0.470	0.963	0.090	-853.4	0.727	2.211
4	8	0.25	0.326	0.776	0.757	-1072.5	0.314	0.781	0.616	-1069.7	0.665	4.474
4	8	0.50	0.582	0.826	0.447	-1036.2	0.561	0.830	0.307	-1034.0	0.674	4.251
8	1	0.25	0.246	0.998	0.083	-875.7	0.246	0.998	0.083	-875.5	0.883	0.796
8	1	0.50	0.493	0.999	0.083	-827.9	0.493	0.999	0.083	-827.7	0.882	0.802
8	8	0.25	0.263	0.990	0.169	-1071.6	0.263	0.991	0.144	-1069.8	0.894	0.871
8	8	0.50	0.537	0.995	0.165	-1024.9	0.536	0.995	0.141	-1023.1	0.898	0.834

lower unconditional distance between clusters implies that there is some overlap of the ellipses; see Figure 1b. In the high-variance case, the heterogeneous model specification achieves a 30% reduction in MSE. This is in line with expectations: The higher σ_2^2, the stronger is the difference in signal-to-noise ratios for the two dimensions, and the more the model benefits from accommodating the different adjustment speeds across variables. The estimates of $\boldsymbol{A}_1$ differ visibly between the heterogeneous and the homogeneous case if $\sigma_2^2 = 8$; we refer to Figure B.1 in Web Appendix B for more details.

Figure 2 shows plots of the filtered mean trajectories from all simulations for the case dist $= 8$, $\sigma_2^2 = 8$, and $\gamma = 0.5$ (corresponding to the last row of Table 1). The associated median and quantiles (5% and 95%), calculated from the distances to the center of the ellipses, are reported as well. The true paths (ellipses) are tracked accurately in either case. Even though Figure 2 does not correspond to one of the cases where the largest MSE reduction are achieved by using a heterogeneous $\boldsymbol{A}_1$, the improvement is still visible: the point cloud in the left-hand panel forms a tighter ring than the one in the right-hand panel. The quantiles are also closer together, showing that the improvement is not driven only by outliers. Appendix B also illustrates other settings, showing that the HMM method consistently achieves more precise mean estimates when $\boldsymbol{A}_1$ is

Figure 2: Simulation results for $dist. = 8$, $\sigma_2^2 = 8$, and $\gamma = 0.5$

The estimated time-varying means from all 250 simulations of this setting are plotted in different colors for each cluster. The 5% and 95% quantiles, calculated from the distances to the center of the ellipses, are plotted as well. Both versions of $\boldsymbol{A}_1$ achieve good tracking of the means, but the improvement afforded by the heterogeneous version is clearly visible. The quantiles also sit more tightly together, showing that the effect is not driven by outliers.



heterogeneous.

Both classification and cluster mean tracking work better when γ is higher. This is intuitive, as a higher value of $\gamma = 0.5$ leads to fewer cluster transitions, simplifying the classification problem. The increased number of transitions associated with a low $\gamma = 0.25$ is only helpful for estimating this particular parameter. We then obtain parameter estimates $\widehat{\gamma}$ that are closer to its true value.

The rightmost panel of Table 1 presents the outcomes for the hierarchical clustering method of Ward (1963). This method is popular in the empirical literature on bank business models, see, for instance, Ayadi and De Groen (2015), Roengpitya et al. (2017), and Ayadi et al. (2020), where it is also applied to multivariate panel data. The hierarchical approach treats each bank-year observation as cross-sectional, ignoring the ordering in time, and groups the entire sample, thereby allowing for (implied) cluster transitions. We find that in all settings considered, the HMM method clearly outperforms the hierarchical clustering method in terms of classification accuracy. It also dramatically lowers the MSE, meaning that the time-varying means are recovered more precisely.

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 and 2018Q2, implying $T = 42$. We assume that differences in banks' business models can be characterized along six categories: size, complexity, risk profile, activities, geographical reach, and funding. These categories are approximately in line with supervisory practices; see e.g. SSM (2016) and Farne and Vouldis (2017). We select a parsimonious set of $D = 12$ indicators to cover these six categories. Specifically, we consider banks' total assets, leverage with respect to CET1 capital [size], net loans to assets ratio, assets held for trading, derivatives held for trading [complexity], the ratio of market risk to credit risk [risk profile], share of net interest income, share of net fees & commissions income, share of trading income, ratio of retail loans to total loans [activities], ratio of domestic loans to total loans [geography], and the deposits to assets ratio [funding].

Web Appendix C provides a detailed discussion of our data, including banks' geographical locations, data transformations, and SNL Financial field keys. We also discuss our handling of missing observations.

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 Davies-Bouldin index (DBI), average silhouette coefficient (ASC), and the Hartigan rule (see e.g. Peel and McLachlan, 2000) point to $J = 5$ or $J = 6$. 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.

Table 2: Parameter estimates

Parameter estimates and cluster validation indices for different model specifications. Model M1 allows for time-varying means and scale matrices but rules out transitions across groups ($\gamma = 500$). The initial clustering is obtained as in Lucas et al. (2019). Model M2 allows for Markovian transitions across groups; see Section 2.1. Model M3 restricts M2 by ruling out transitory transitions that last less than five quarters by imposing $P = 4$ inactive states; see Section 2.4.2. 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 Section 2.4.3. Finally, Model M5 allows the A_1 matrix parameters to differ across variables $d = 1, \ldots, 12$ and clusters $j = 1, \ldots, 6$; see Section 2.4.1. Standard errors in parentheses are constructed from the numerical second derivatives of the log-likelihood function. Heterogeneity parameters $\bar{a}_d^D$ and $\bar{a}_j^J$ are reported in italics if significant at a 5% significance level. We set $\bar{a}_{size}^D = \bar{a}_A^J \equiv 0$ for identification. We also report two cluster validation indices: the Davis-Bouldin index (DBI; the smaller the better), and the average Silhouette Coefficient (ASC; the larger the better).

	M1 No transitions	M2 Markovian transitions	M3 non-Markovian transitions	M4 non-Markovian transitions & covariate	M5 non-Markovian transitions & covariate & heterogeneity
a_1	0.894	0.850	0.813	0.967	1.439
	(0.02)	(0.02)	(0.03)	(0.02)	(0.11)
a_2	0.998	0.998	0.993	0.998	0.999
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)
ν	6.595	19.518	8.088	14.723	13.401
	(0.07)	(0.06)	(0.06)	(0.05)	(0.08)
γ	-	1.369	1.503	1.313	1.283
		(0.01)	(0.02)	(0.02)	(0.01)
β	-	-	-	-17.757	-13.908
				(0.17)	(0.26)
$\bar{a}_{lev}^D$	-	-	-	-	*-0.250*
$\bar{a}_{comp}^D$	-	-	-	-	*-0.392*
$\bar{a}_{risk}^D$	-	-	-	-	*-0.192*
$\bar{a}_{act}^D$	-	-	-	-	*-0.595*
$\bar{a}_{geo}^D$	-	-	-	-	*-1.018*
$\bar{a}_{fund}^D$	-	-	-	-	*-0.424*
$\bar{a}_B^J$	-	-	-	-	*0.315*
$\bar{a}_C^J$	-	-	-	-	*0.169*
$\bar{a}_D^J$	-	-	-	-	0.120
$\bar{a}_E^J$	-	-	-	-	0.104
$\bar{a}_F^J$	-	-	-	-	*-0.213*
P	-	0	4	4	4
DBI	3.14	2.94	2.93	2.92	2.74
ASC $\times 100$	1.65	3.39	3.34	3.35	5.43
loglik	144,253.2	150,197.1	150,003.1	150,506.9	150,680.7

Table 2 reports parameter estimates and several goodness-of-fit measures including the log-likelihood value as well as DBI and ASC, for five different versions of our dynamic clustering model. All specifications M1–M5 use the same initial cluster allocations, and thus also the same initial values for all time-varying parameters. Initial cluster allocations $\tau_{ij,1|1}$ are obtained using the static clustering approach with time-varying parameters of Lucas et al. (2019).[6] We choose a distance smoothing parameter as $\lambda = 0.25$; see (2′).[7] In addition, all specifications are based on mixtures of Student's t distributions. This allows us to be robust to one-off windfall effects and joint outliers in bank accounting ratios. We pool parameters A_2 and ν across clusters and variables following a preliminary data analysis.

Model M1 allows for time-varying means and scale matrices, but rules out transitions across clusters ($\gamma = 500$). If the data generating process featured cluster transitions then model M1 would be misspecified. Any business model transitions would then result in observations that do not match the static cluster assignments, and thus can be interpreted as outlying in this sense. We note the low estimate for the degrees-of-freedom parameter $\nu \approx 6.5$, which may be partly attributable to this effect, resulting in more probability mass in the joint tail of the Student's t densities (1).

Model M2 allows for Markovian transitions across clusters 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 and joint outliers as short-lived cluster transitions. Such short-lived transitions are hard to interpret economically as meaningful changes in 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 2.4.2. The decay parameter γ increases somewhat, indicating fewer (short-lived) transitions. The degrees-of-freedom parameter ν decreases slightly to accommodate more frequent outlying

[6] Replacing $\tau_{ij,1|1}$ with filtered estimates from a first run, and subsequently re-estimating θ, leads to negligible improvements in log-likelihood fit.

[7] The log-likelihood surface is fairly flat in λ, see Figure D.3 in Web Appendix D; we treat it as a tuning parameter for this reason.

observations. The insistence on inactive states is reflected in a drop in log-likelihood fit. The improvement compared to the model without transitions (M1), however, is still strong.

Model M4 extends M3 by allowing for an additional explanatory variable to impact the transition probabilities $\mathbf{\Pi}_t$; see Section 2.4.3. We choose x_{jkt} as the difference in probability-weighted return-on-equity (ROE) of banks allocated to clusters j and k at time t. Specifically, we let $x_{jt} \equiv \sum_i^N \hat{\tau}_{ij,t|t} \cdot \text{ROE}_{it} / \sum_i^N \hat{\tau}_{ij,t|t}$ denote the filtered ROE for banks in cluster j at time t, where $\hat{\tau}_{ij,t|t}$ corresponds to the estimated filtered cluster membership probability for bank i. Then $x_{jkt} := x_{jt} - x_{kt}$ denotes the differences in ROE between clusters j and k. These differences can now be collected in a matrix $\mathbf{X}_t$ with typical element x_{jkt}. The transition matrix $\mathbf{\Pi}_t \left(\tilde{\mathcal{D}}_{t-1}, \mathbf{X}_t \right)$ may now become more asymmetric compared to Model M3. The coefficient β is highly significant and has an intuitive interpretation: if a bank's current cluster j has higher profitability than cluster k, that bank is less likely to shift out of j into k and conversely, more likely to switch from k to j. The log-likelihood increases by about 500 points from M3 to M4, and 300 points from M2 to M4.

Finally, Model M5 extends M4 by allowing for additional parameter heterogeneity in the score updates for the cluster means, both in the variable (D) and cluster (J) dimension; see Section 2.4.1. This results in a further log-likelihood increase of about 175 points, suggesting that our data are subject to significant parameter heterogeneity. The time-varying parameter paths implied by Model M5 are visibly different from those implied by M1–M4 (not shown). The values of the adjustment parameters $\bar{a}_d^D$ and $\bar{a}_j^J$ also are often intuitive. The largest negative coefficients are for geographical concentration of the loan portfolio (geo), followed by share of retail loans (act). More negative $\bar{a}_d^D$ coefficients imply lower rates of adjustment, which makes sense for these bank variables, as banks' physical location and distribution channels cannot easily be changed. The cluster-specific adjustment coefficients $\bar{a}_j^J$ are also subject to significant heterogeneity, with the lowest value taken for banks in group F. This group contains the smallest banks in our sample (domestic retail lenders; see the discussion below), which continued to rely on traditional business strategies and were subject to the least changes in financial supervision; see e.g. Nouy (2016).

Model specification M5 is not only strongly preferred in terms of log-likelihood fit, but also

does well in terms of non-parametric cluster validation indices (DBI, ASC). We therefore select M5 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 Group transitions and popularity

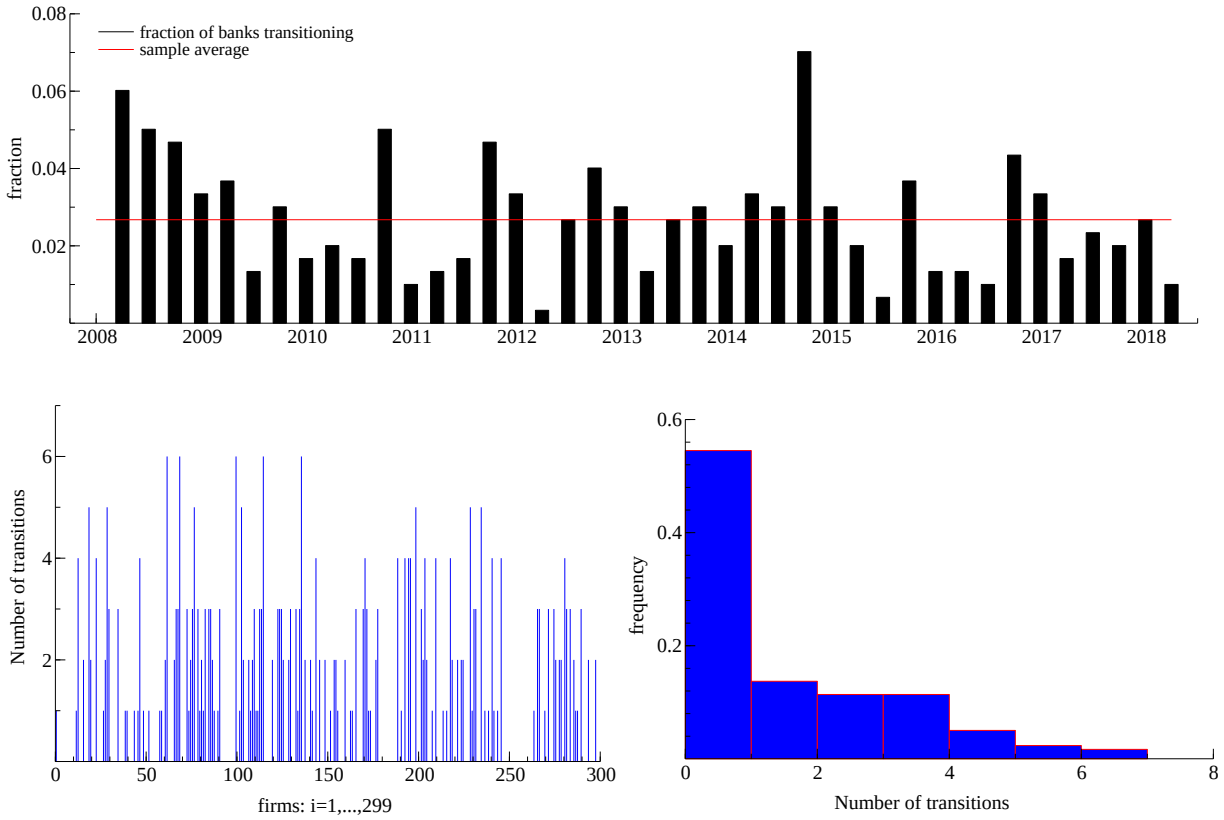
Using the empirical estimates, we distinguish six bank business model groups: (A) market-oriented universal banks, including globally systemically important banks (G-SIBs) such as Barclays, Deutsche Bank, and HSBC; (B) international diversified lenders, including large banking groups such as Banco Santander, BBVA, and ING Group; (C) fee-focused retail lenders, which achieve most of their income from fees and commissions despite lending almost exclusively to domestic retail customers; (D) international corporate lenders; and the larger groups of smaller sized banks under the labels (E) domestic diversified lenders and (F) domestic retail lenders. These labels are chosen in line with the evolution of all time-varying means of the different variables and the identities of the firms in each cluster. Our labeling is approximately in line with the examples given in SSM (2016, p.10).

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. We do not observe an obvious time trend in transition intensity. Instead, the transition intensity is above-average during the global financial crisis (2008), the euro area sovereign debt crisis (2011 – 2012), and, interestingly, is highest at the start of centralized banking supervision within the European Central Bank's Single Supervisory Mechanism (SSM) in the euro area (2014Q4). On average, approximately 3% of the $N = 299$ banks are estimated to transition each quarter (horizontal red line in the top panel of Figure 3). In 2014Q4, by contrast, more than 6% of banks are estimated to have transitioned across groups, suggesting a strong response of European banks to stricter bank supervision (Ampudia et al., 2020).

The bottom left panel of Figure 3 reports the total number of transitions per firm $i = 1, \ldots, 299$.

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).



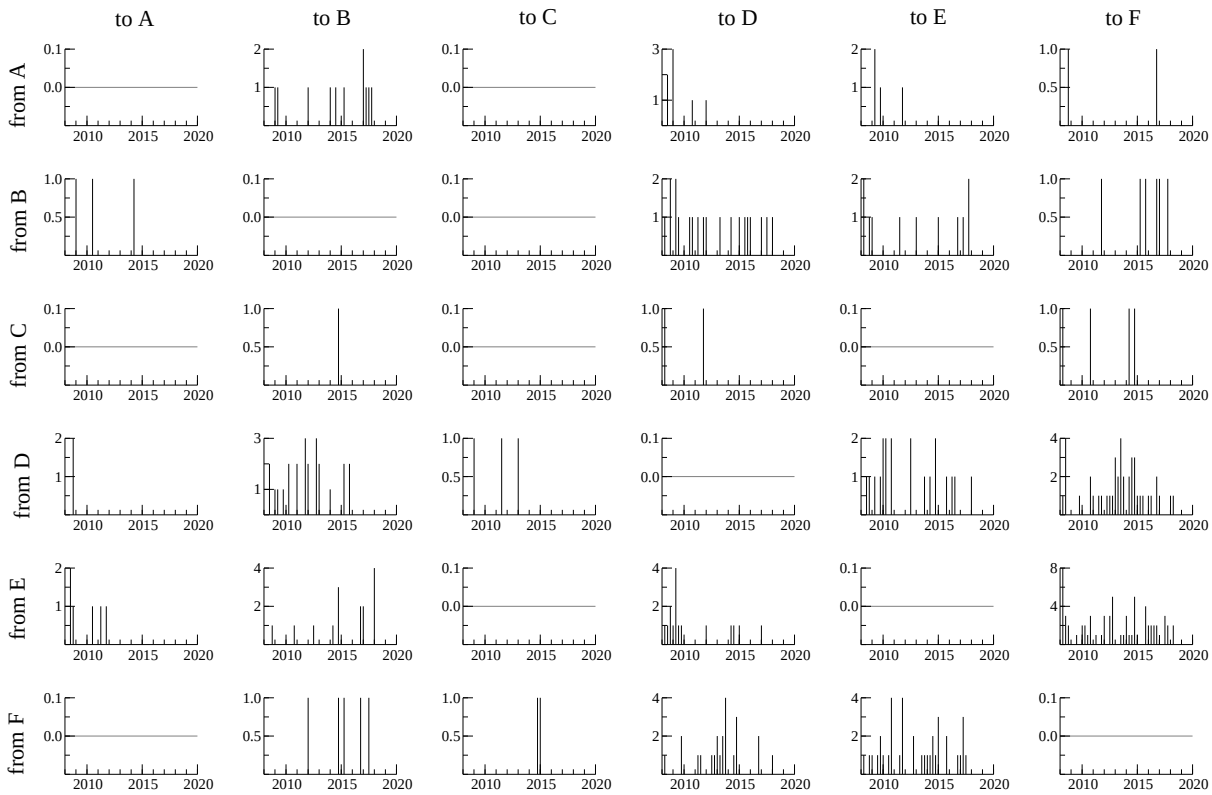
The bottom right panel of Figure 3 provides a histogram of firms' transition counts. The total number of transitions per firm range between 0 and 6. More than half of the banks never experience any transition (54%). If a certain bank transitions more than a few times, then that bank may be located between two or more clusters and may be hard to classify as a result.

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. Cluster C is relatively isolated, in that it receives relatively few in- and outflows.

Figure 5 shows the total number of banks allocated to each cluster over time. Clusters B and F grow in popularity over time, while clusters A, D, and E shrink and cluster C remains approximately stable. The observed trends are in line with large banks becoming less reliant on

Figure 4: Cluster transitions and popularity

Number of transitions from cluster j (rows) to cluster k (columns) over time.

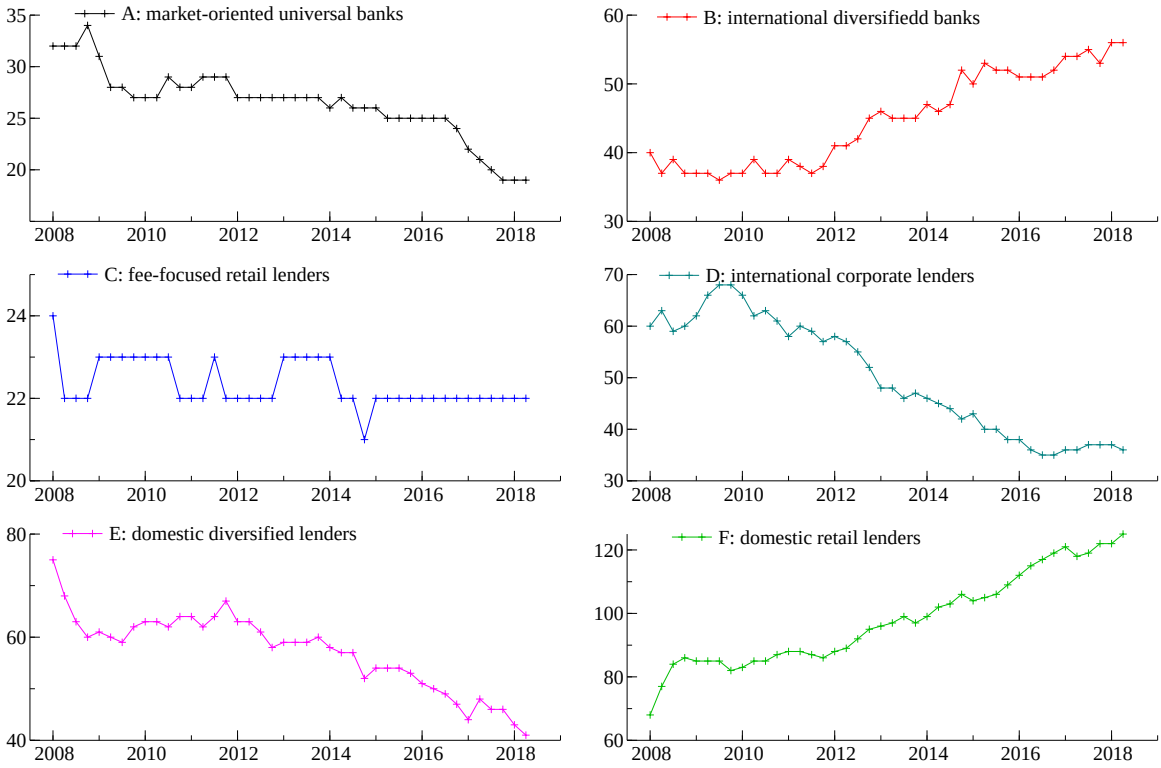


market funding and scaling back trading activities (A $\rightarrow$ B), domestically-active banks lending relatively more to retail clients rather than to corporate clients (D $\rightarrow$ B; D $\rightarrow$ F; E $\rightarrow$ F), and banks relying progressively more on more on fee income, possibly to lean against a lower profitability from increasingly low interest rates (D $\rightarrow$ C; D $\rightarrow$ F). These industry trends are approximately in line with the variable- and cluster-specific filtered means shown in Figure D.1 in Web Appendix D, and with the discussions in e.g. ECB (2016) and Ayadi et al. (2020).

The cluster transitions underlying Figures 3 – 5 are in part explained by differences in bank profitability across clusters; see Section 4.2. Web Appendix D.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

Figure 5: Cluster transitions and popularity

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



sovereign debt crisis in mid-2010, and remains negative until the end of the sample, adding to the move out of D and into other business models, as indicated above.

5 Conclusion

We proposed a novel model for the dynamic clustering of multivariate panel data. The cluster mean and scale matrices are time-varying to track gradual changes in cluster characteristics over time. The dynamics are governed by parameters allowing for heterogeneity in the across variables and clusters. The model incorporates further flexibility by allowing the units of interest to transition between clusters. This is accomplished by 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 give new answers to a range of empirical questions in multivariate panel data settings. Our results for a sample of 299 European banks between 2008Q1 and 2018Q2 suggest that banks' transition intensity between clusters is related to differences in bank profitability, particularly for international diversified lenders. This is 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

Aggarwal, C. C. and C. K. Reddy (2014). *Data Clustering. Algorithms and Applications*. Chapman & Hall/CRC.

Ampudia, M., T. Beck, and A. Popov (2020). Out with the new, in with the old? Bank supervision and the composition of firm investment. *CEPR discussion paper 16225*.

Ando, T. and J. Bai (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics 31*(1), 163–191.

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., P. Bongini, B. Casu, and D. Cucinelli (2020). Bank business model migrations in Europe: Determinants and effects. *British Journal of Management*.

Ayadi, R. and W. P. De 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.

Bazzi, M., F. Blasques, S. J. Koopman, and A. Lucas (2017). Time varying transition probabilities for Markov regime switching models. *Journal of Time Series Analysis 38*, 458–478.

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

Bonhomme, S. and E. Manresa (2015). Grouped patterns of heterogeneity in panel data. *Econometrica 83*(3), 1147–1184.

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*.

Cavalleri, M. C., A. Eliet, P. McAdam, F. Petroulakis, A. Soares, and I. Vansteenkiste (2019). Concentration, market power and dynamism in the euro area. *ECB working paper 2253*, 0–68.

Christensen, J. H., E. Hansen, and D. Lando (2004). Confidence sets for continuous-time rating transition probabilities. *Journal of Banking & Finance 28*(11), 2575–2602.

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 Haas, R. and A. Popov (2021). Finance and green growth. *mimeo*.

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

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*(456), 1348–1360.

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

Fruehwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer.

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

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

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica 57*, 357–384.

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.

Koopman, S. J., R. Kräussl, A. Lucas, and A. B. Monteiro (2009). Credit cycles and macro fundamentals. *Journal of Empirical Finance 16*(1), 42–54.

Lando, D. and T. M. Skødeberg (2002). Analyzing rating transitions and rating drift with continuous observations. *Journal of banking & finance 26*(2-3), 423–444.

Lin, C.-C. and S. Ng (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods 1*(1), 42–55.

Lu, X. and L. Su (2017). Determining the number of groups in latent panel structures with an application to income and democracy. *Quantitative Economics 8*(3), 729–760.

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., B. Schwaab, and X. Zhang (2014). Conditional euro area sovereign default risk. *Journal of Business and Economic Statistics 32 (2)*, 271–284.

Lucas, A., B. Schwaab, and X. Zhang (2017). Modeling financial sector joint tail risk in the euro area. *Journal of Applied Econometrics 32*(1), 171–191.

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.

McLachlan, G. and D. Peel (2000). *Finite Mixture Models*. Wiley.

Nouy, D. (2016). Adjusting to new realities – banking regulation and supervision in Europe. Speech by Daniele Nouy, Chair of the ECB's Supervisory Board, at the European Banking Federation's SSM Forum, Frankfurt, 6 April 2016.

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

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

Popov, A. and S. Manganelli (2015). Financial development, sectoral reallocation, and volatility: international evidence. *Journal of International Economics 96*(2), 323–337.

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

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

Su, L., Z. Shi, and P. C. Phillips (2016). Identifying latent structures in panel data. *Econometrica 84*(6), 2215–2264.

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

A Derivation of the parameter updating equations

A.1 Time-varying mean dynamics

For the time-varying means, we specify

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

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 $\boldsymbol{\mu}_{jt}$. Starting with the score, and using the fact that $\tau_{ij,t|t-1}$ does not depend on $\boldsymbol{\mu}_{jt}$ due to the transition probability matrix $\boldsymbol{\Pi}_t$ 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}^{(i)}_{\mu_{jt},t}.
\end{aligned}$$

where $\boldsymbol{\nabla}^{(i)}_{\mu_{jt},t} = \partial \log f\left(\boldsymbol{y}_{it}|c_{it}=j,\mathcal{F}_{t-1};\boldsymbol{\theta}\right)/\partial \boldsymbol{\mu}_{jt}$ is the score of mixture component j associated with observation $\boldsymbol{y}_{it}$. In case of a mixture of D-dimensional Student's t distributions, we have

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

Taking derivatives of the log of (A.2) with respect to $\boldsymbol{\mu}_{jt}$, we obtain

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

where

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

Turning to the scaling matrix, 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

$$\boldsymbol{S}_{\mu_{jt},t} = \left(\sum_{i=1}^{N}\tau_{ij,t|t}\cdot\mathbb{E}\left[\boldsymbol{\nabla}^{(i)}_{\mu_{jt},t}\left(\boldsymbol{\nabla}^{(i)}_{\mu_{jt},t}\right)'\,\middle|\,c_{it}=j\right]\right)^{-1} \tag{A.5}$$

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. Using the fact that for the Gaussian case $\nu_j^{-1} = 0 \implies w_{ij,t} = 1$ we obtain

$$
\begin{aligned}
\boldsymbol{S}_{\mu_{jt},t}^{-1} &= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \left(-\mathbb{E}\left[\frac{\partial \boldsymbol{\nabla}_{\mu_{jt},t}^{(i)}}{\partial \boldsymbol{\mu}_{jt}'} \,\middle|\, c_{it} = j, \mathcal{F}_{t-1}; \boldsymbol{\theta} \right] \right) \\
&= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \mathbb{E}\left[\boldsymbol{\nabla}_{\mu_{jt},t}^{(i)} \left(\boldsymbol{\nabla}_{\mu_{jt},t}^{(i)} \right)' \,\middle|\, c_{it} = j, \mathcal{F}_{t-1}; \boldsymbol{\theta} \right] \\
&= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \mathbb{E}\left[\boldsymbol{\Sigma}_{jt}^{-1} \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right)\left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right)' \boldsymbol{\Sigma}_{jt}^{-1} \,\middle|\, c_{it} = j, \mathcal{F}_{t-1}; \boldsymbol{\theta} \right] \\
&= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \boldsymbol{\Sigma}_{jt}^{-1} \, \mathbb{E}\left[\left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right)\left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right)' \mid c_{it} = j, \mathcal{F}_{t-1}; \boldsymbol{\theta} \right] \boldsymbol{\Sigma}_{jt}^{-1} \\
&= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \boldsymbol{\Sigma}_{jt}^{-1} \, \boldsymbol{\Sigma}_{jt} \, \boldsymbol{\Sigma}_{jt}^{-1} \\
&= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \boldsymbol{\Sigma}_{jt}^{-1}, \qquad \text{(A.6)}
\end{aligned}
$$

Inserting (A.6) and (A.3) into (A.1) yields the transition equation

$$
\begin{aligned}
\boldsymbol{\mu}_{j,t+1} &= \boldsymbol{\mu}_{jt} + \boldsymbol{A}_1 \boldsymbol{S}_{\mu_{jt},t} \cdot \boldsymbol{\nabla}_{\mu_{jt},t} \\
&= \boldsymbol{\mu}_{jt} + \boldsymbol{A}_1 \left(\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \boldsymbol{\Sigma}_{jt}^{-1} \right)^{-1} \sum_{i=1}^{N} \tau_{ij,t|t} \cdot w_{ij,t} \cdot \boldsymbol{\Sigma}_{jt}^{-1} \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right) \\
&= \boldsymbol{\mu}_{jt} + \boldsymbol{A}_1 \boldsymbol{\Sigma}_{jt} \boldsymbol{\Sigma}_{jt}^{-1} \left(\sum_{i=1}^{N} \tau_{ij,t|t} \right)^{-1} \sum_{i=1}^{N} \tau_{ij,t|t} \cdot w_{ij,t} \cdot \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right) \\
&= \boldsymbol{\mu}_{jt} + \boldsymbol{A}_1 \frac{\sum_{i=1}^{N} \tau_{ij,t|t} \cdot w_{ij,t} \cdot \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right)}{\sum_{i=1}^{N} \tau_{ij,t|t}}, \qquad \text{(A.7)}
\end{aligned}
$$

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

A.2 Time-varying scale matrix dynamics

The transition equation for the time-varying scale matrices $\boldsymbol{\Sigma}_{jt}$ is given by

$$\text{vec}(\boldsymbol{\Sigma}_{j,t+1}) = \text{vec}(\boldsymbol{\Sigma}_{jt}) + \boldsymbol{A}_2 \, \boldsymbol{S}_{\Sigma_{jt},t} \cdot \boldsymbol{\nabla}_{\Sigma_{jt},t}, \tag{A.8}$$

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.

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

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 $\boldsymbol{\Sigma}_{jt}$, we arrive at

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

Following Abadir and Magnus (2005) for the derivative of the log of the determinant of the scale 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

$$\begin{aligned}
\boldsymbol{\nabla}_{\Sigma_{jt},t} &= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \text{vec}\left(-\tfrac{1}{2}\left(\boldsymbol{\Sigma}_{jt}^{-1}\right)' + \tfrac{1}{2}\left(\boldsymbol{\Sigma}_{jt}^{-1}\right)' w_{ij,t}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)'\left(\boldsymbol{\Sigma}_{jt}^{-1}\right)'\right) \\
&= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \text{vec}\left(-\tfrac{1}{2}\left(\boldsymbol{\Sigma}_{jt}'\right)^{-1} + \tfrac{1}{2}\left(\boldsymbol{\Sigma}_{jt}'\right)^{-1} w_{ij,t}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)'\left(\boldsymbol{\Sigma}_{jt}'\right)^{-1}\right) \\
&= \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \text{vec}\left(-\tfrac{1}{2}\boldsymbol{\Sigma}_{jt}^{-1} + \tfrac{1}{2}\boldsymbol{\Sigma}_{jt}^{-1} w_{ij,t}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)'\boldsymbol{\Sigma}_{jt}^{-1}\right) \\
&= \tfrac{1}{2}\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \text{vec}\left(\boldsymbol{\Sigma}_{jt}^{-1}\left(w_{ij,t}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)' - \boldsymbol{\Sigma}_{jt}\right)\boldsymbol{\Sigma}_{jt}^{-1}\right) \\
&= \tfrac{1}{2}\left(\boldsymbol{\Sigma}_{jt}\otimes\boldsymbol{\Sigma}_{jt}\right)\cdot\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \text{vec}\left(w_{ij,t}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_{jt}\right)' - \boldsymbol{\Sigma}_{jt}\right), \qquad \text{(A.9)}
\end{aligned}$$

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}$$

Inserting the score and the scaling matrix into (A.8) we obtain

$$\begin{aligned}\text{vec}(\boldsymbol{\Sigma}_{j,t+1}) = \text{vec}(\boldsymbol{\Sigma}_{jt}) + \boldsymbol{A}_2 \left(\tfrac{1}{2} \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \left(\boldsymbol{\Sigma}_{jt}^{-1} \otimes \boldsymbol{\Sigma}_{jt}^{-1} \right) \right)^{-1} \times \\ \left(\tfrac{1}{2} \left(\boldsymbol{\Sigma}_{jt} \otimes \boldsymbol{\Sigma}_{jt} \right) \cdot \sum_{i=1}^{N} \tau_{ij,t|t} \cdot \text{vec} \left(w_{ij,t} \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right) \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right)' - \boldsymbol{\Sigma}_{jt} \right) \right) \\ = \text{vec}(\boldsymbol{\Sigma}_{jt}) + \boldsymbol{A}_2 \frac{\sum_{i=1}^{N} \tau_{ij,t|t} \cdot \text{vec} \left(w_{ij,t} \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right) \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right)' - \boldsymbol{\Sigma}_{jt} \right)}{\sum_{i=1}^{N} \tau_{ij,t|t}}. \end{aligned} \tag{A.10}$$

Unvectorizing (A.10), we obtain the scale matrix transition equation

$$\boldsymbol{\Sigma}_{j,t+1} = \boldsymbol{\Sigma}_{jt} + \boldsymbol{A}_2 \frac{\sum_{i=1}^{N} \tau_{ij,t|t} \left[w_{ij,t} \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right) \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{jt} \right)' - \boldsymbol{\Sigma}_{jt} \right]}{\sum_{i=1}^{N} \tau_{ij,t|t}}. \tag{A.11}$$

A.3 Initialization of the time-varying parameters

The cluster probabilities $\tau_{ij,1|1}$, the cluster means $\boldsymbol{\mu}_{j1}$, and the cluster scale 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, 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 approaches with time-varying parameters could be applied to all data $t = 1, \ldots, T$, such as e.g. approach proposed by 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 scale 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 can be used.

The initial $\tau_{ij,1|1}$ can be replaced by the filtered $\tau_{ij,1|1}$ from (7) once a first estimate of parameters $\boldsymbol{\theta}$ is available. Parameters $\boldsymbol{\theta}$ 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.

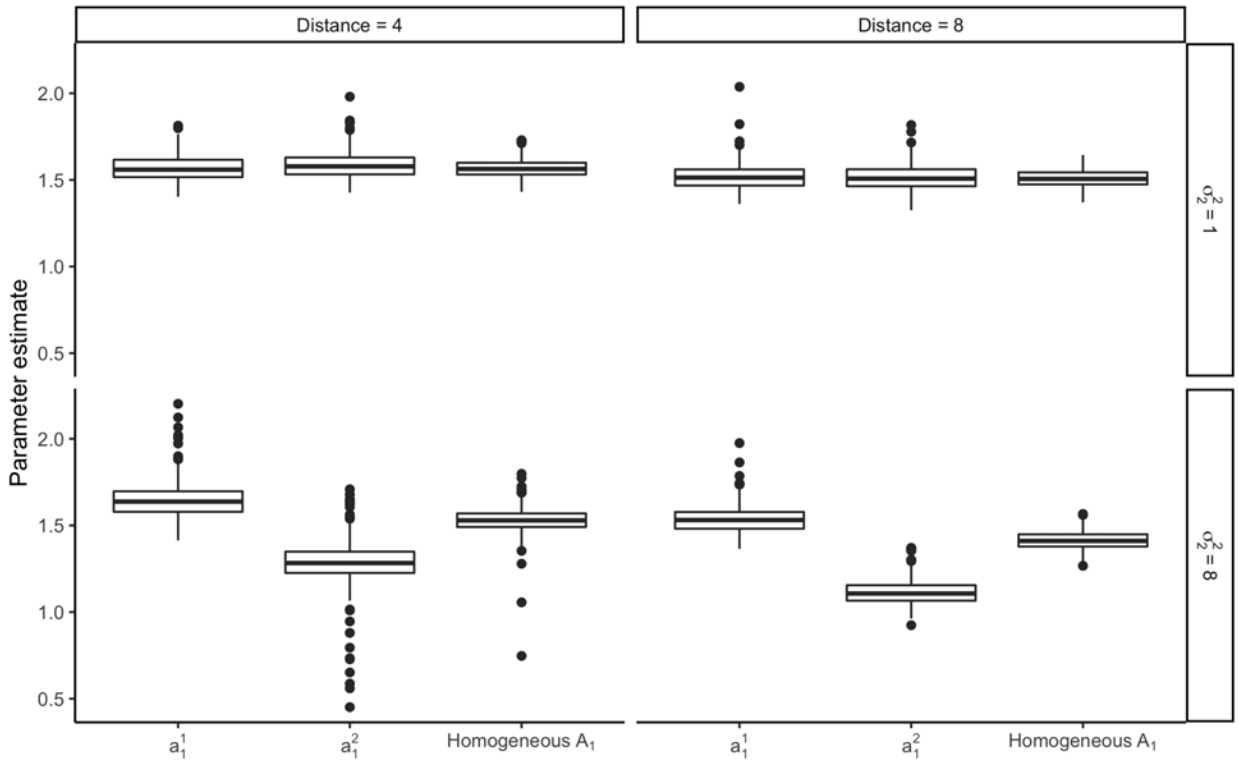
B Additional simulation figures

The performance of our dynamic clustering method is consistently better when using a heterogeneous mean adjustment parameter $\boldsymbol{A}_1$, instead of a homogeneous one. However, the improvement is particularly large if $\sigma_2^2 = 8$. This is intuitive, as these settings imply stronger parameter heterogeneity. Figure B.1 demonstrates that indeed, it is for this setting that the estimated main diagonal elements of $\boldsymbol{A}_1$ differ the most from the common parameter a_1 in the homogeneous case.

Figure B.2 presents the outcomes from ten simulation runs for each simulation setting. In the high-variance cases, the estimated means are noticeably more scattered, and more so when using an homogeneous $\boldsymbol{A}_1$. This difference is more subtle in the low-variance settings. Out of the ten simulations, only once is an estimated mean grossly misplaced. This occurs for the most challenging setting with a homogeneous $\boldsymbol{A}_1$ parameter, $\sigma_2^2 = 8$, $\gamma = 0.25$, and dist. $= 4$ (the upper

Figure B.1: $\boldsymbol{A}_1$ parameter estimates in different simulation settings.

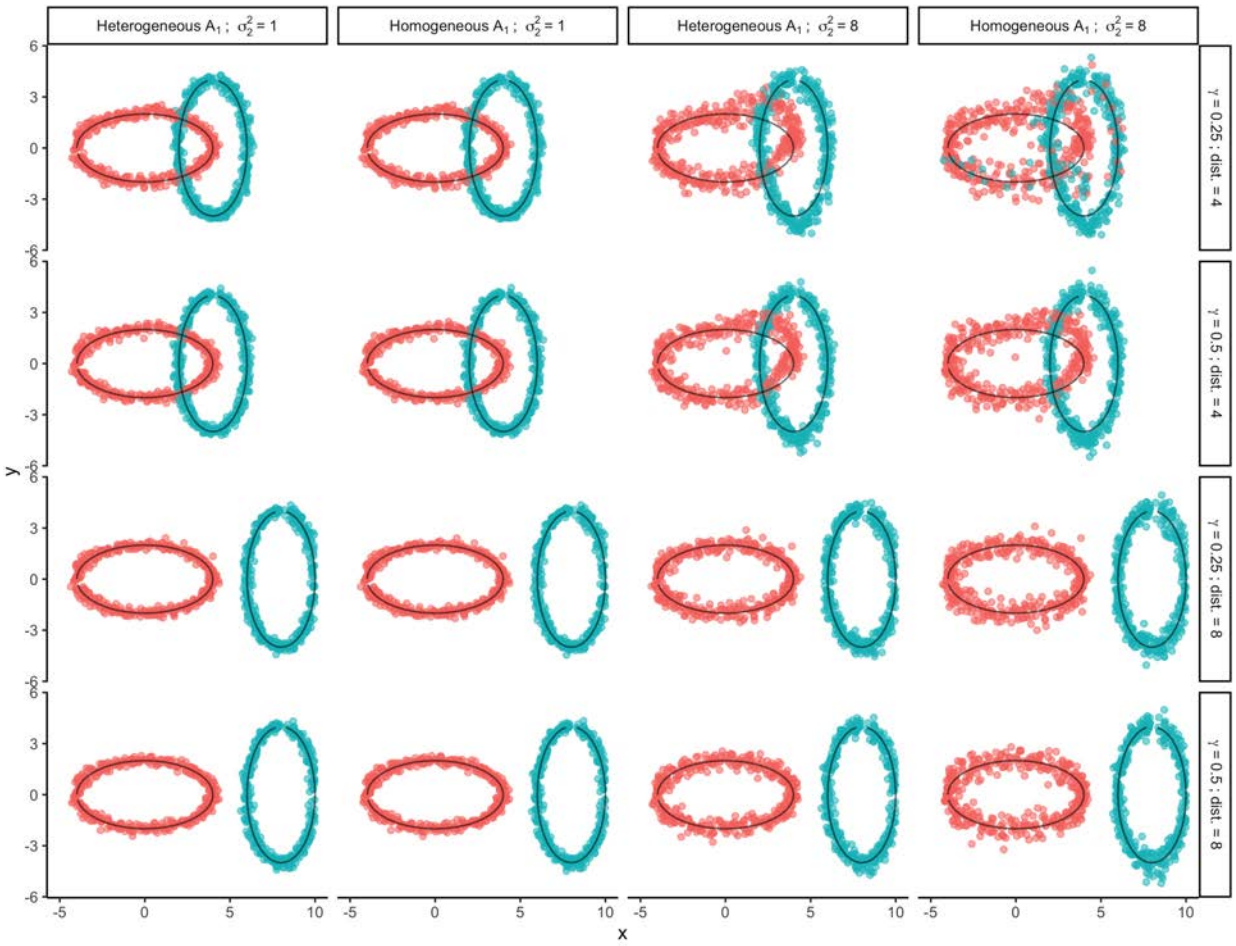
Settings where the parameter heterogeneity is strongest ($\sigma_2^2 = 8$) see a departure in the estimates of $\boldsymbol{A}_1$ between the homogeneous and heterogeneous cases. In these settings, the performance improvement afforded by an heterogeneous $\boldsymbol{A}_1$ parameter is also the greatest.



right corner plot in Figure B.2). When using a heterogeneous $\boldsymbol{A}_1$ the estimated means evolves around the correct ellipse, even though the generated data are the same.

Figure B.2: Mean estimates from a sample of 10 simulations in each settings.

The time-varying means are well estimated in almost all cases, especially so when using an hetergogeneous $\boldsymbol{A}_1$. Higher variance in the second variable makes the problem visibly more challenging. Nevertheless, only once we see an estimated mean tracking the wrong ellipse.



C Data details

We consider all banks at their highest level of consolidation (the bank holding group level). In addition, however, we also include the largest subsidiaries of these bank holding groups if sufficient data are available. Most banks are located in the euro area (55%) and the European Union (73%). European non-E.U. banks (27%) in our sample are located in Norway and Switzerland, among other countries. Banks that were acquired between 2008Q1 and 2018Q2, or ceased to operate for other reasons during that time, are excluded from the analysis. Acquiring banks remain in the sample, and can undergo a business model transition as a result.

Table C.1 lists our indicators $d = 1, \ldots, D$ used in our empirical clustering exercise. Our multivariate panel data is unbalanced. While many banks in our sample report at a quarterly frequency, other banks report only semi-annually or annually. We remove such missing observations by substituting the most recently available observation for that variable and bank. Other ways of handling missing values, for example through interpolation, have negligible effects on our clustering outcomes and transition estimates, particularly when the econometric model allows for variable- and cluster-specific adjustment speeds when updating the cluster means.

Approximately one in two banks in our sample (48%) reports quarterly, with the remainder reporting semi-annually or annually (52%). It is mostly the large, listed banks that report at a quarterly frequency, while small, unlisted banks report semi-annually or annually. The former tend to be allocated to clusters A to D, while the latter tend to be allocated to clusters E and F (domestic diversified lenders and domestic retail lenders). Substituting the most recently available observation for a missing observation could artificially increase the persistence of that variable. As a result, our handling of missing observations, owing to differences in reporting frequency, may be a contributing factor why different adjustment speeds across clusters lead to a significantly better model fit; see model specification M5 in Table 2.

The share of missing observations is also not constant across the variables as listed in Table C.1. Missing observations are relatively more prevalent for more disaggregated breakdowns of accounting items (for example, the division of total loans into domestic and foreign loans, or of

total loans into retail and corporate loans), while simpler, non-disaggregated ratios are less subject to missing observations (for example, the ratio of total loans to total assets, or of CET1-capital to total assets). It is particular the former, more disaggregated ratios that are required to allocate banks into economically meaningful business model groups. When allowing for different adjustment speeds across variables, differences in persistence from this source are accommodated as well.

Table C.1: Indicator variables

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

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.

D Additional results

D.1 Cluster medians, scale matrices, and distances

Figure D.1 reports the filtered cluster medians for the twelve indicator variables as listed in Table C.1. The cluster medians coincide with the cluster means as modeled (14) provided no nonlinear transformation has been used for variable $d = 1, \ldots, D$, see the last column in Table C.1.

Figure D.2 plots the filtered component-specific time-varying standard deviations $\hat{\sigma}_{j,t|t}(d) = \left(\hat{\mathbf{\Sigma}}_{j,t|t}(d,d)\right)^{\frac{1}{2}}$ for variables $d = 1, \ldots, 12$. The off-diagonal elements of $\mathbf{\Sigma}_{jt}$ are not reported. 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 D.3 plots the time-varying cluster distance functions d_{jkt} and $\tilde{d}_{jkt}$ as given in (4) and (2′) – (4), respectively. The cluster distances display a modest degree of time-variation. While the filtered cluster distances are symmetric, the HMM transition matrix $\mathbf{\Pi}_t$ need not be; see (3) and (3′).

Figure D.1: Time-varying cluster medians

Filtered cluster medians for twelve indicator variables; see Table C.1. The cluster medians coincide with the cluster means unless the variable is transformed; see the last column of Table C.1. The cluster mean estimates are based on a t-mixture model with $J = 6$ clusters and time-varying cluster means $\boldsymbol{y}_{jt}$ and scale matrices $\boldsymbol{\Sigma}_{jt}$. We distinguish A: market-oriented universal banks (black line), B: international diversified banks (red line), C: fee-focused retail lenders (blue line), D: international corporate lenders (green line), E: domestic diversified lenders (purple dashed line), and F: domestic retail lenders (green dashed line).

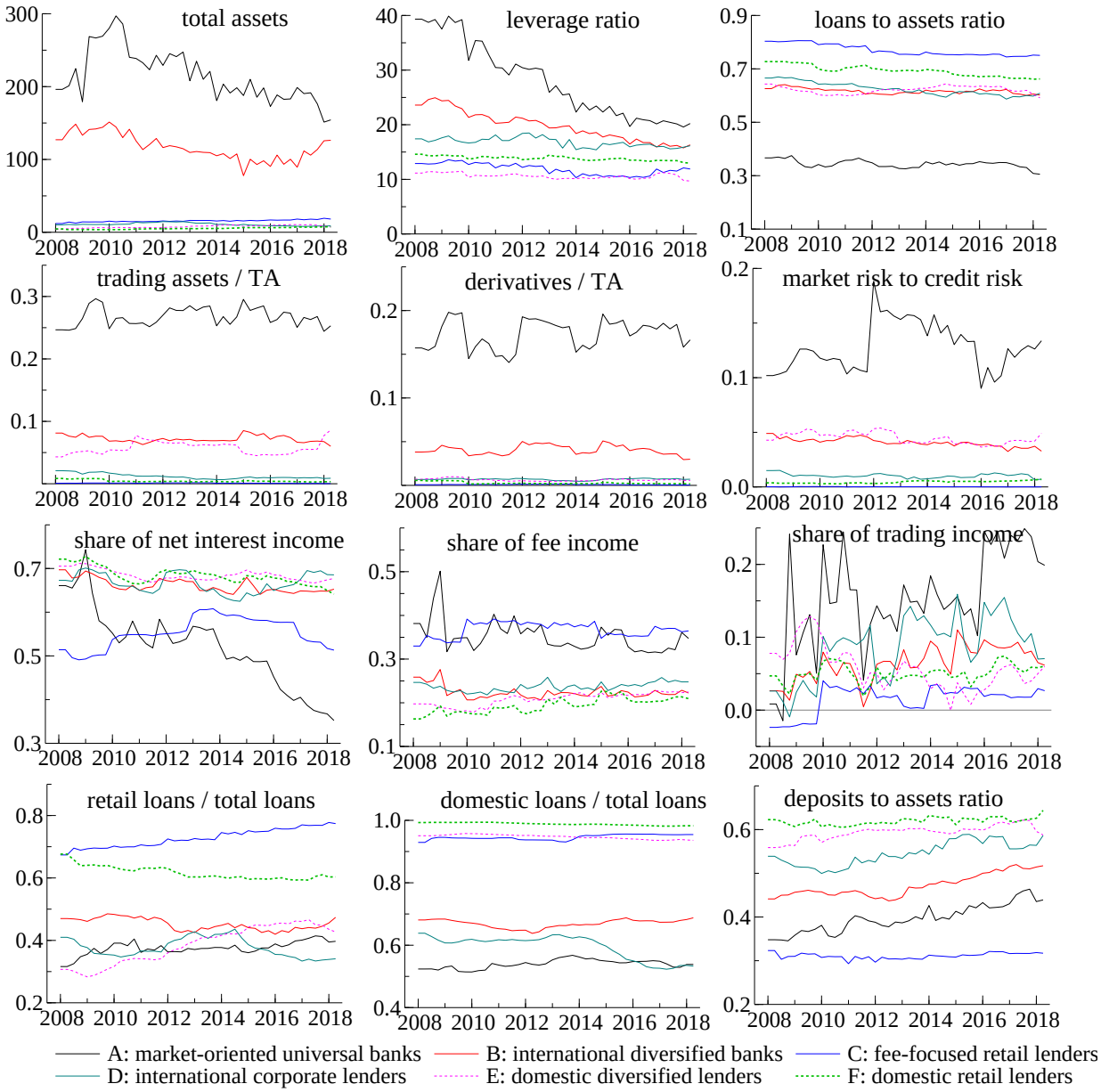


Figure D.2: Time-varying standard deviations

Filtered time-varying standard deviations $\hat{\sigma}_{j,t|t}(d) = \left(\hat{\mathbf{\Sigma}}_{j,t|t}(d,d)\right)^{\frac{1}{2}}$ for variables $d = 1, \ldots, 12$. Each panel contains 12 standard deviation estimates over time, corresponding to the variables listed in Table C.1. The standard deviation estimates refer to model specification M5 in Table 2.

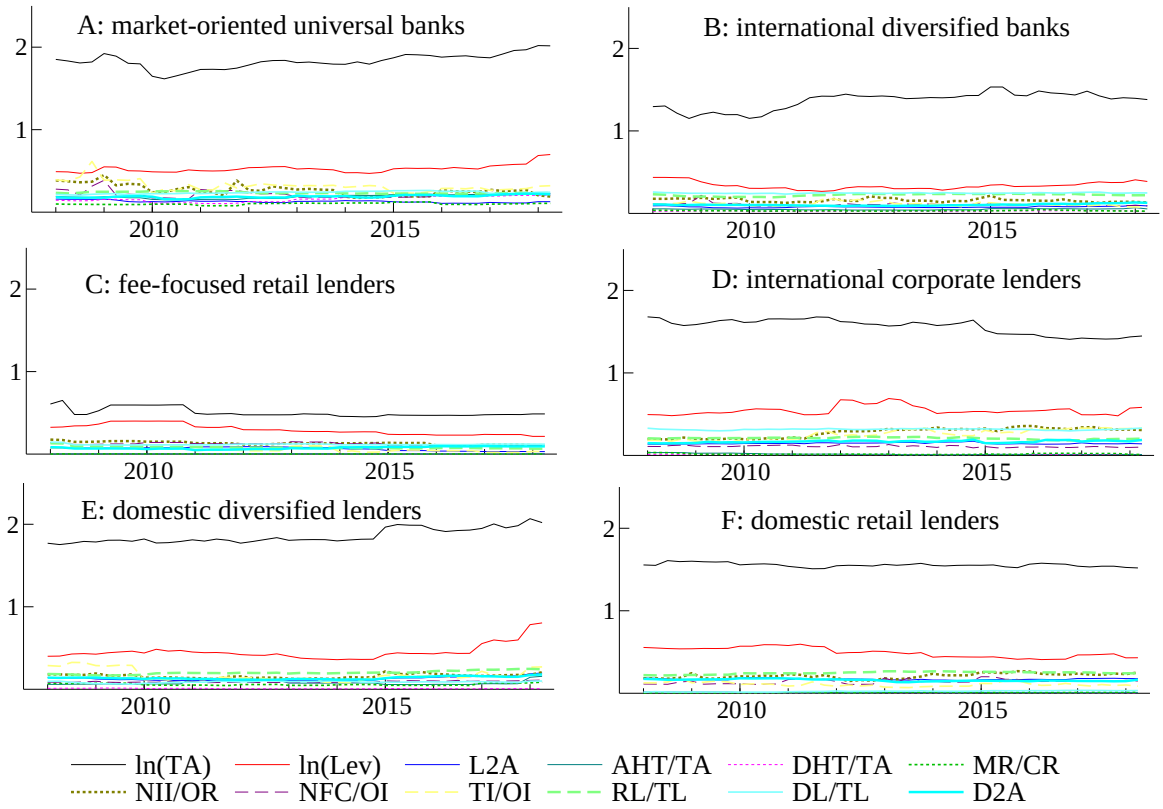
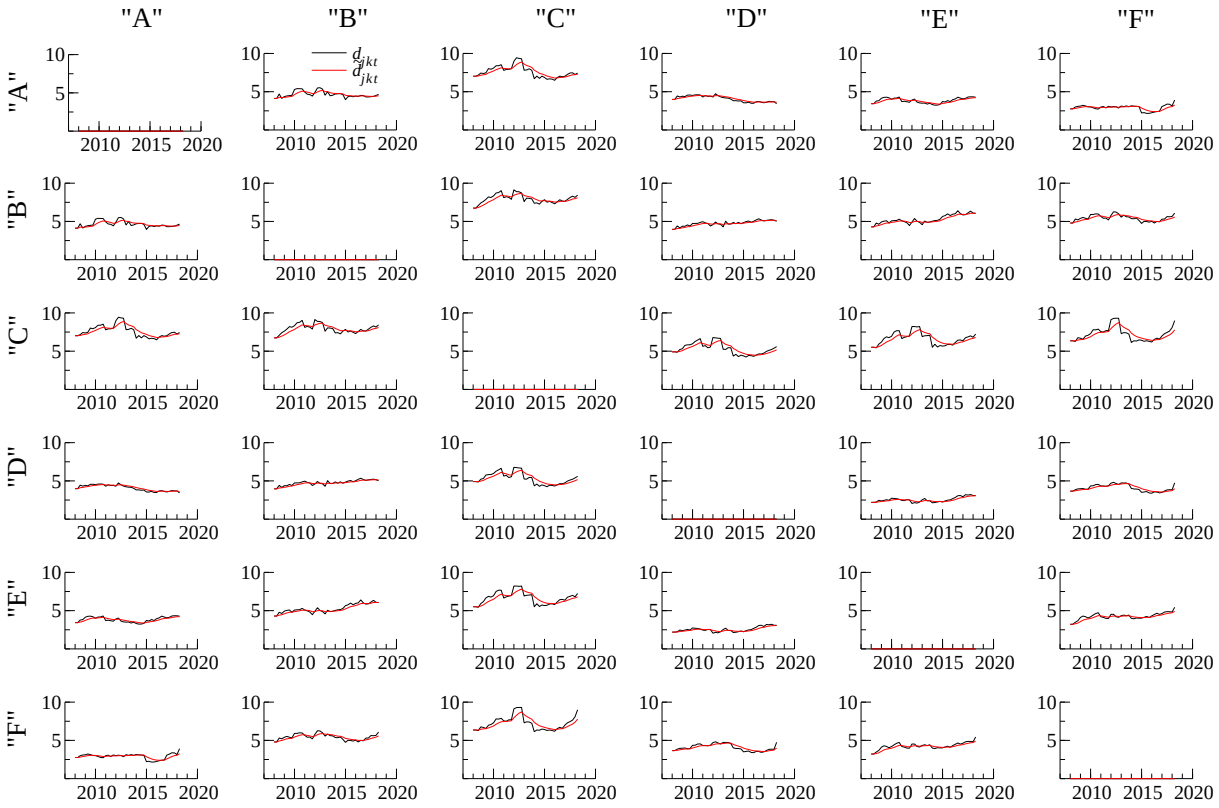


Figure D.3: Filtered cluster distances

Time-varying cluster distances d_{jkt} (black lines) and $\tilde{d}_{jkt}$ (red lines) as given in (4) and (2′) – (4), respectively. The time-varying parameter estimates refer to model specification M5 in Table 2.

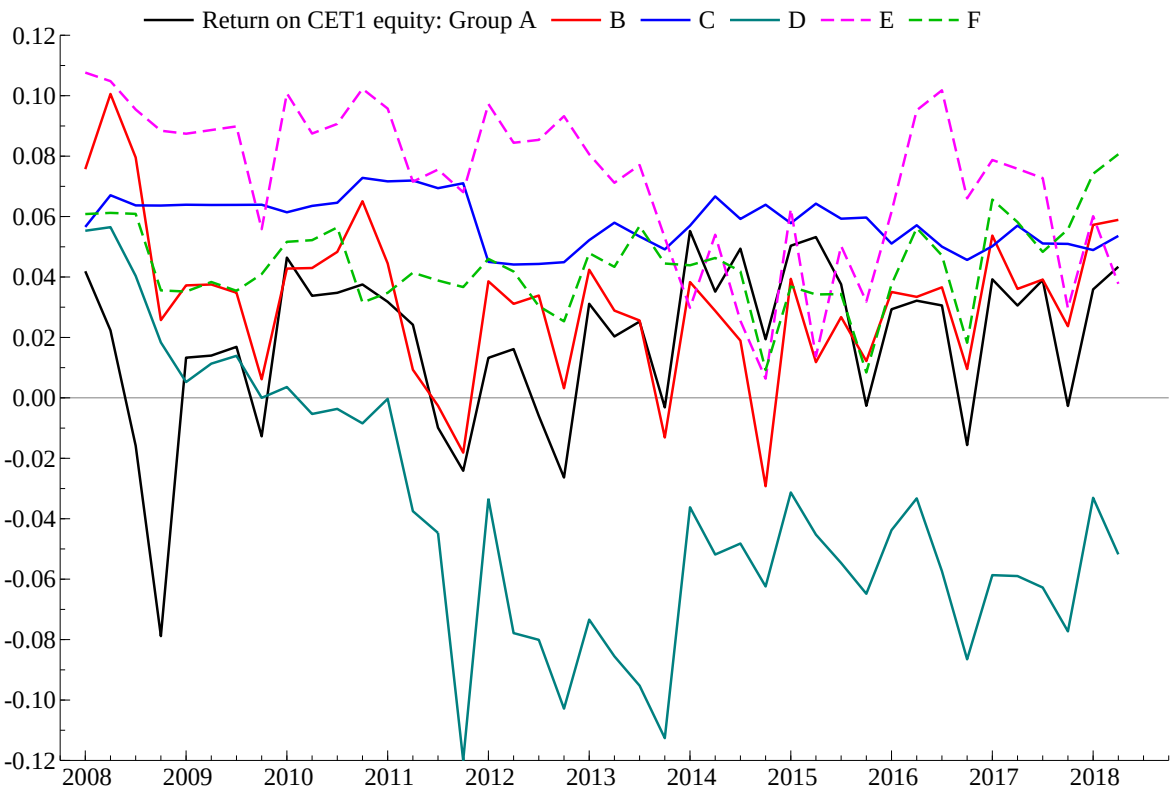


D.2 Bank group profitability

The cluster transitions underlying Figures 3 – 5 are in part explained by differences in bank profitability. Figure D.4 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 C.1. European banks' ROE tend to vary between approximately -2% and 12% over time. Banks assigned to cluster D 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, contributing to observed migrations out of group D into other bank business model groups.

Figure D.4: Bank profitability

Average return on CET1 equity (ROE) for banks in each business model group A to F. At any quarter t bank-specific observations ROE_{it}s are weighted by the conditional probability $\tau_{ij,t|t}$ that bank i belongs to cluster j.



References

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

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.

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.

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

Acknowledgements

This work was supported by the Dutch National Science Foundation (NWO) [406.18.EB.011 to I.C.J. and A.L., VI.VIDI.191.169 to J.S]. The views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the European Central Bank.

Igor Custodio Joao
Vrije Universiteit Amsterdam and Tinbergen Institute, Amsterdam, The Netherlands; email: i.custodiojoao@vu.nl

André Lucas
Vrije Universiteit Amsterdam and Tinbergen Institute, Amsterdam, The Netherlands; email: a.lucas@vu.nl

Julia Schaumburg
Vrije Universiteit Amsterdam and Tinbergen Institute, Amsterdam, The Netherlands; email: j.schaumburg@vu.nl

Bernd Schwaab
European Central Bank, Frankfurt am Main, Germany; email: bernd.schwaab@ecb.europa.eu

Postal address 60640 Frankfurt am Main, Germany
Telephone +49 69 1344 0
Website www.ecb.europa.eu

PDF ISBN 978-92-899-4763-3 ISSN 1725-2806 doi:10.2866/657440 QB-AR-21-068-EN-N