Web Appendix to

Dynamic clustering of multivariate panel data*

Igor Custodio João,[a] *André Lucas,*[a]

Julia Schaumburg,[a] *Bernd Schwaab,*[b]

[a] Vrije Universiteit Amsterdam and Tinbergen Institute

[b] European Central Bank, Financial Research

*Author information: Igor Custodio João, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, Email: i.custodiojoao@vu.nl. André Lucas, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, Email: a.lucas@vu.nl. Julia Schaumburg, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, Email: j.schaumburg@vu.nl. Bernd Schwaab, European Central Bank, Kaiserstrasse 29, 60311 Frankfurt, Germany, Email: bernd.schwaab@ecb.int. 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.

A Derivation of the parameter updating equations

A.1 Time-varying mean dynamics

This appendix derives the updating equation (12) for the time-varying cluster 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

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

We now turn to the scaling matrix for the time-varying means. 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

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

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

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

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

A.2 Time-varying scale matrix dynamics

This appendix derives the transition equation (13) for the time-varying cluster scale matrices $\boldsymbol{\Sigma}_{jt}$, given by

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

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 $\text{vec}(ABC) = (C' \otimes$

$A)\mathrm{vec}(B)$, we obtain

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

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 \mathrm{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 \mathrm{vec}(\boldsymbol{\Sigma}_{jt})'}\tfrac{1}{2}\mathrm{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 \mathrm{vec}(\boldsymbol{\Sigma}_{jt})'}\mathrm{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 \Bigg\{\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] + \\
&\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 \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]\Bigg\} \\
&= \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

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

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

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

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 equal 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 $\tau_{ij,1|1}^{k}$ 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 version 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.

A.4 Forecasting simulation algorithm

Obtaining h-step-ahead forecasts of either the time varying parameters $\boldsymbol{\mu}_{jt}$, $\boldsymbol{\Sigma}_{jt}$, $\boldsymbol{\Pi}_t$, or the cluster membership probabilities $\tau_{ij,t|t}$ with the current model is straightforward. The following algorithm can be used.

- Given the estimated static parameters $\hat{\theta}$ and the last observation vectors $\boldsymbol{y}_{i,T}$, the time-varying parameters $\boldsymbol{\mu}_{j,T+1}$, $\boldsymbol{\Sigma}_{j,T+1}$, $\boldsymbol{\Pi}_T$ and $\boldsymbol{\Pi}_{T+1}$ are perfectly predictable, and therefore known (as the model is observation-driven in this respect);

- using $\boldsymbol{\Pi}_T$ and the latent cluster indicators $c_{i,T}$, the next latent cluster indicators $c_{i,T+1}^{(s)}$ can be simulated from a multinomial distribution; alternative to the cluster assignments $c_{i,T}$, one can simulate the cluster IDs $c_{i,T}$ at time T from the filtered probabilities $\tau_{ij,T|T}$;

- given the simulated $c_{i,T+1}^{(s)}$, and the means $\boldsymbol{\mu}_{j,T+1}$ and scale matrices $\boldsymbol{\Sigma}_{j,T+1}$, it is then possible to simulate $\boldsymbol{y}_{i,T+1}^{(s)}$ for $i = 1, \ldots, N$;

- from these simulated $\boldsymbol{y}_{i,T+1}^{(s)}$, we can obtain (simulated) $\tau_{ij,T+1|T+1}^{(s)}$ as well as updated $\boldsymbol{\mu}_{j,T+2}^{(s)}$, $\boldsymbol{\Sigma}_{j,T+2}^{(s)}$; note that $\boldsymbol{\Pi}_{T+1}$ is fully known given the data up to time T due to the lag structure of the model; this allows us to simulate new cluster assignments $c_{i,T+2}^{(s)}$ via the multinomial distribution, followed by new simulated $\boldsymbol{y}_{i,T+2}^{(s)}$ from the appropriate t distributions, followed by new $\boldsymbol{\mu}_{j,T+3}^{(s)}$, $\boldsymbol{\Sigma}_{j,T+3}^{(s)}$, and $\boldsymbol{\Pi}_{T+2}^{(s)}$ values, and so on;

- these steps can be repeated H times to obtain draws of $\boldsymbol{y}_{i,T+H}^{(s)}$ and their by-products $\tau_{ij,T+h|T+h}^{(s)}$ for $h = 1, 2, \ldots, H$;

- repeating the process, say, $s = 1, \ldots, 10,000$ times allows us to forecast the distribution of $\tau_{ij,T+h|T}$ from the simulated $\tau_{ij,T+h|T+h}^{(s)}$.

B Additional simulation figures

This appendix discusses additional figures associated with our simulation experiments in Section 3. A key outcome is that the performance of our dynamic clustering method is consistently better when using a heterogeneous mean adjustment parameter $\boldsymbol{A}_1$, instead of a homogeneous mean adjustment parameter. For the first set of DGPs (DGP 1), the improvements are particularly large if $\sigma_2^2 = 8$. This is intuitive, as these settings imply stronger parameter heterogeneity. Figure B.1 suggests that, indeed, for this setting the estimated main diagonal elements of $\boldsymbol{A}_1$ differ the most from the common parameter a_1 in the homogeneous case.

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

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 largest.

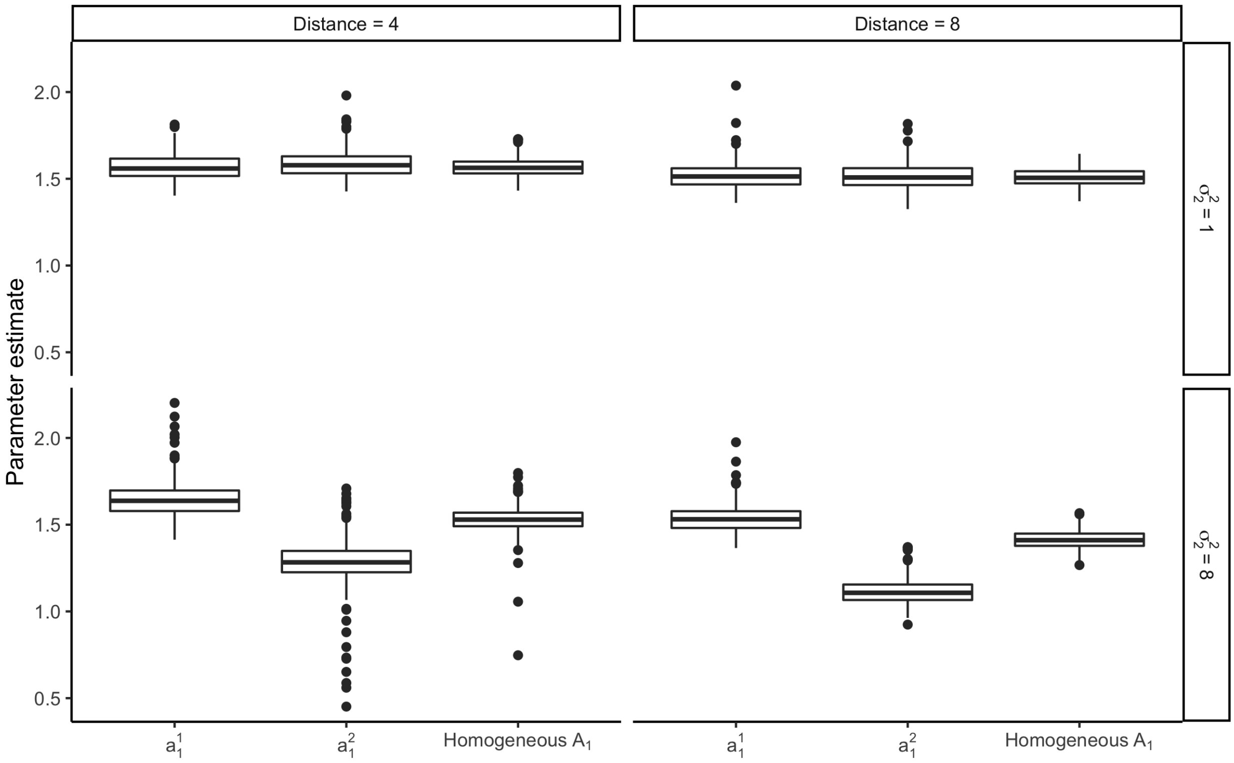


Figure B.2 presents the outcomes from ten simulation runs for each simulation setting. In the high-variance cases, the estimated means are noticeably more scattered, and more so when using a homogeneous $\boldsymbol{A}_1$. These differences are more subtle in the low-variance settings. Out of the ten 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 right corner plot in Figure B.2). When using a heterogeneous $\boldsymbol{A}_1$ parameter, the estimated means evolves around the correct ellipse, even though the generated data are the same.

Figure B.3 presents the estimated means for ten simulations of DGP 2, colored by their classification accuracy. The accuracy is shown to improve the further the means drift apart. Figure B.4 plots the corresponding MSEs of the estimated means. When the means start apart from each other, we can identify the clusters via the covariances in the second half of the sample, bringing the MSEs down. When the means start at the same position and move apart in the second half of the sample, the different covariances do not lead to systematic performance improvements.

Figure B.2: Mean estimates from a sample of 10 simulations in each settings, DGP 1.

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

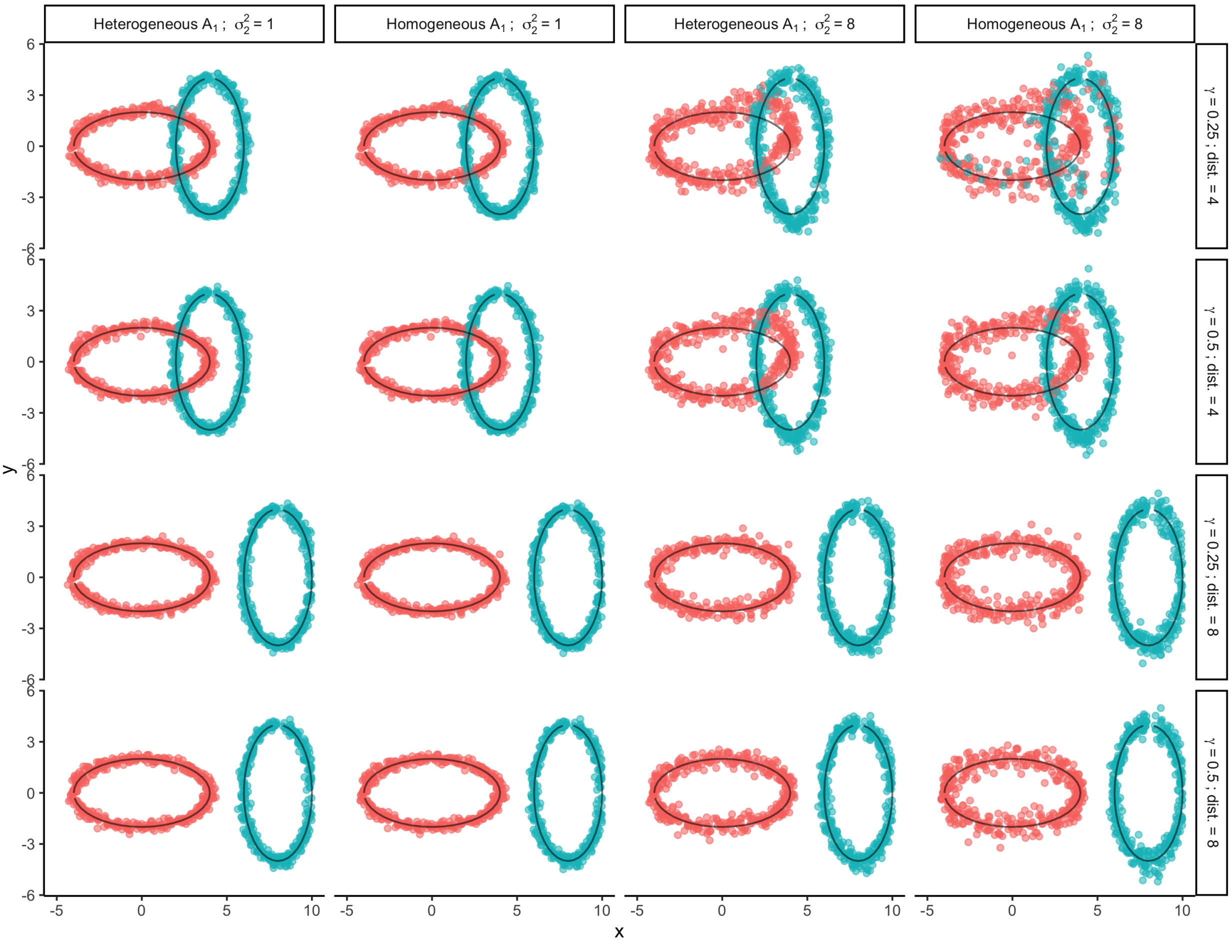


Figure B.3: Estimated means and the rate of correct classification, DGP 2.

The estimated time-varying means from a subset of 10 simulations of this DGP are plotted with colors representing the rate of correct classification at each point in time in each simulation. The true and estimated number of clusters is always two. In all settings the true means move from the top to the bottom, such that in rows 2 and 4 they start apart, while in rows 1 and 3 they start on the same spot.

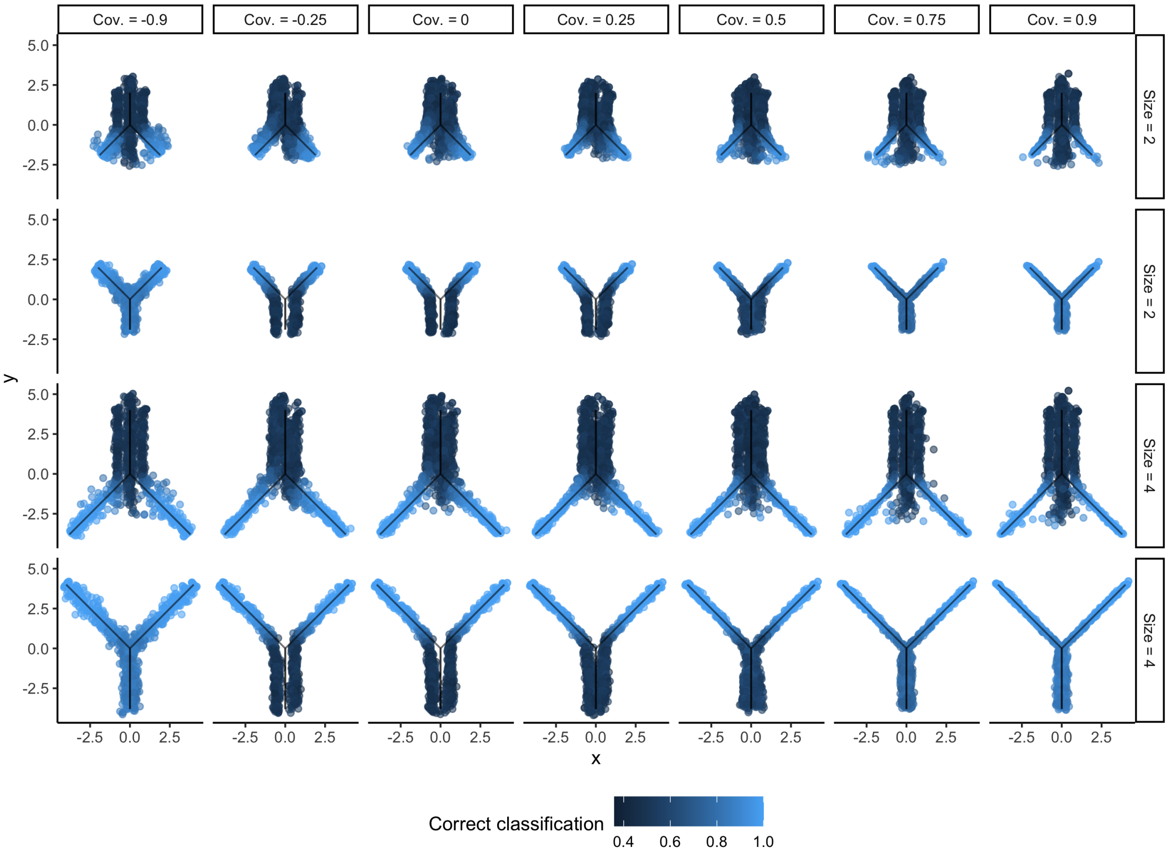
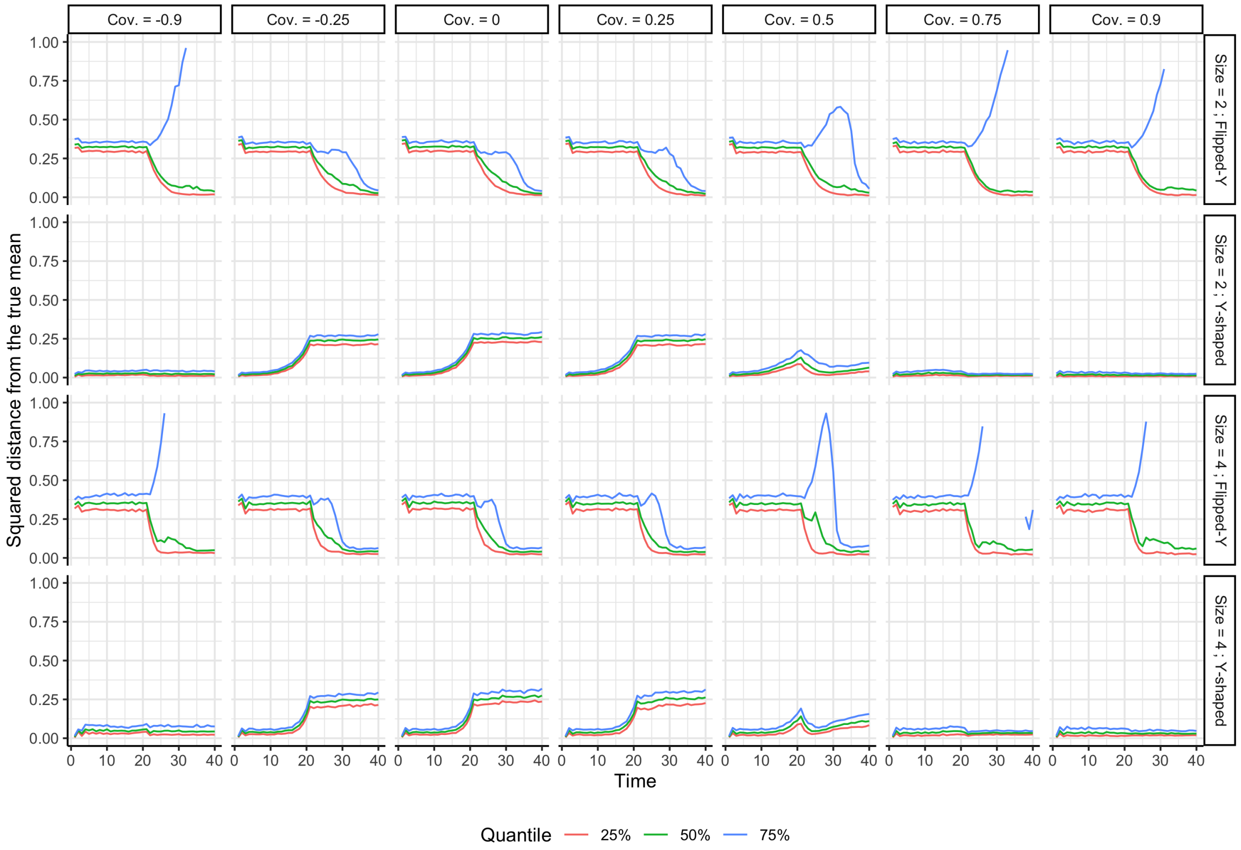


Figure B.4: MSE of the estimated means, DGP 2.

The quartiles of the MSE of the estimated means over time are plotted for each corresponding setting of Figure B.3, calculated over 250 simulations. In the "flipped-Y" settings (rows 1 and 3), our method performs poorly at the beginning, and increased covariance differences (higher Cov) only increase the spread of the estimates in the second half of the sample. In the "Y-shape" settings (rows 2 and 4), the method benefits from identification through the covariances, producing accurate tracking result even as the means overlap in the second half of the sample.



C Data details

This appendix provides additional details about the data used in Section 4.

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.

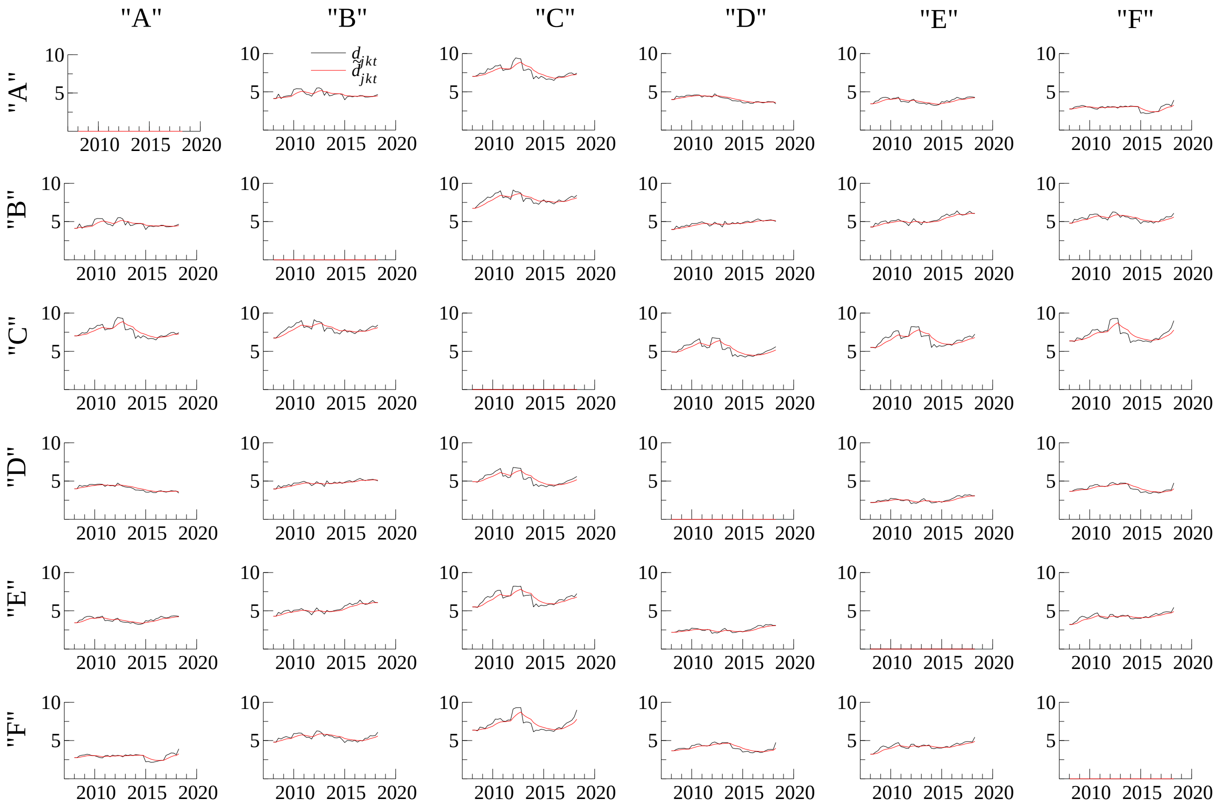
D Additional results

D.1 Cluster distances

Figure D.1 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: 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.



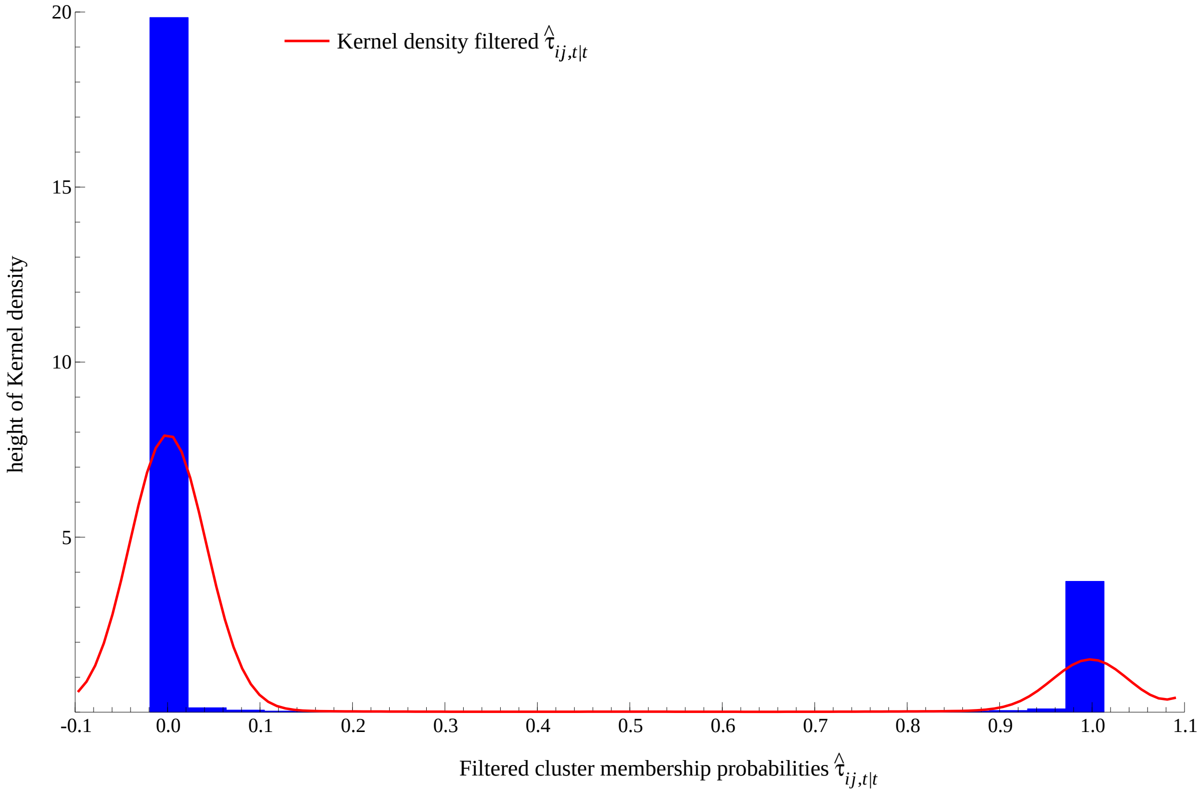
D.2 Filtered cluster probabilities $\hat{\tau}_{ij,t|t}$

Figure D.2 plots a histogram (blue bars) and a fitted Kernel density (red solid line) of the filtered cluster (membership) probabilities $\hat{\tau}_{ij,t|t}$. The histogram and Kernel density are drawn for $(T-1) \times N$ = 41 × 299 estimates of $\tau_{ij,t|t}$, omitting the initial allocation at $t = 1$.

The histogram and associated Kernel estimate is clearly bimodal. The $\hat{\tau}_{ij,t|t}$ are either close to zero, or close to one, with little in between. If one $\hat{\tau}_{ij,t|t}$ is close to one, then five other $\hat{\tau}_{ij,t|t}$ need to be close to zero (given J=6), explaining the relative heights of the blue bars. 91.3% of all $\hat{\tau}_{ij,t|t}$s are either below 0.1 or above 0.9. 87.6% are either below 0.05 or above 0.95. As a result, most banks are fairly unequivocally allocated to one cluster at any time.

Figure D.2: Kernel density estimate and histogram of filtered $\tau_{ij,t|t}$

Kernel density estimate (red) and histogram (blue) of filtered cluster probabilities $\tau_{ij,t|t}$.

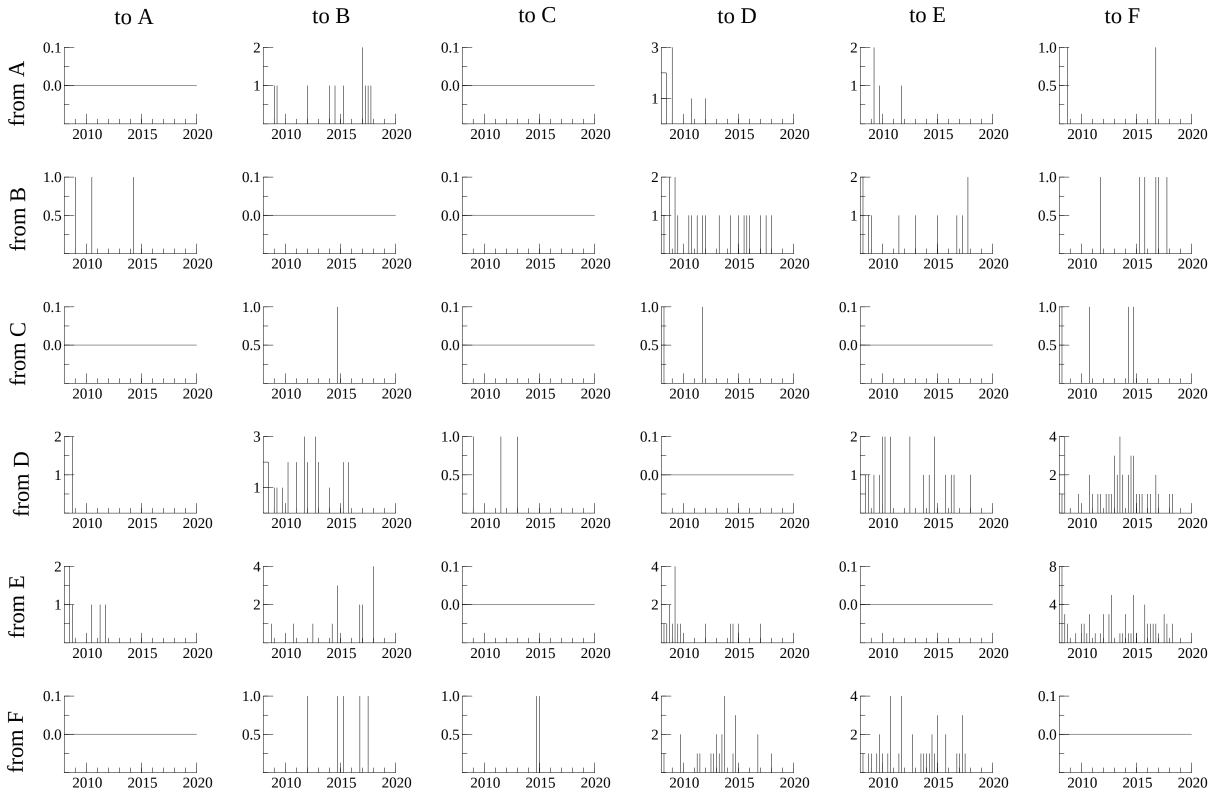


D.3 Cluster transitions over time

Figure D.3 plots the number of estimated transitions from cluster j (rows) to k (columns) at any time. Most transitions take place between ‘nearby’ clusters, e.g. between A and B, B and D, D and E, and E and F. The composition of Cluster C is relatively stable in the sense that it receives few inflows (five in total between 2008 and 2018 from D and F) and outflows (seven in total into B, D, and F).

Figure D.3: Cluster transitions and popularity

Number of transitions from cluster j (rows) to cluster k (columns) over time. Bank i is assigned to the cluster j^* at time t for which the filtered cluster probability is maximal, i.e., $j^* = \arg\max_j \tau_{ij,t|t}$.



D.4 Cluster medians and scale matrices

Figure D.4 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.5 plots the filtered component-specific time-varying standard deviations $\hat{\sigma}_{j,t|t}(d) = \left(\hat{\boldsymbol{\Sigma}}_{j,t|t}(d,d)\right)^{\frac{1}{2}}$ for variables $d = 1, \ldots, 12$. The off-diagonal elements of $\boldsymbol{\Sigma}_{jt}$ 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.4: 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).

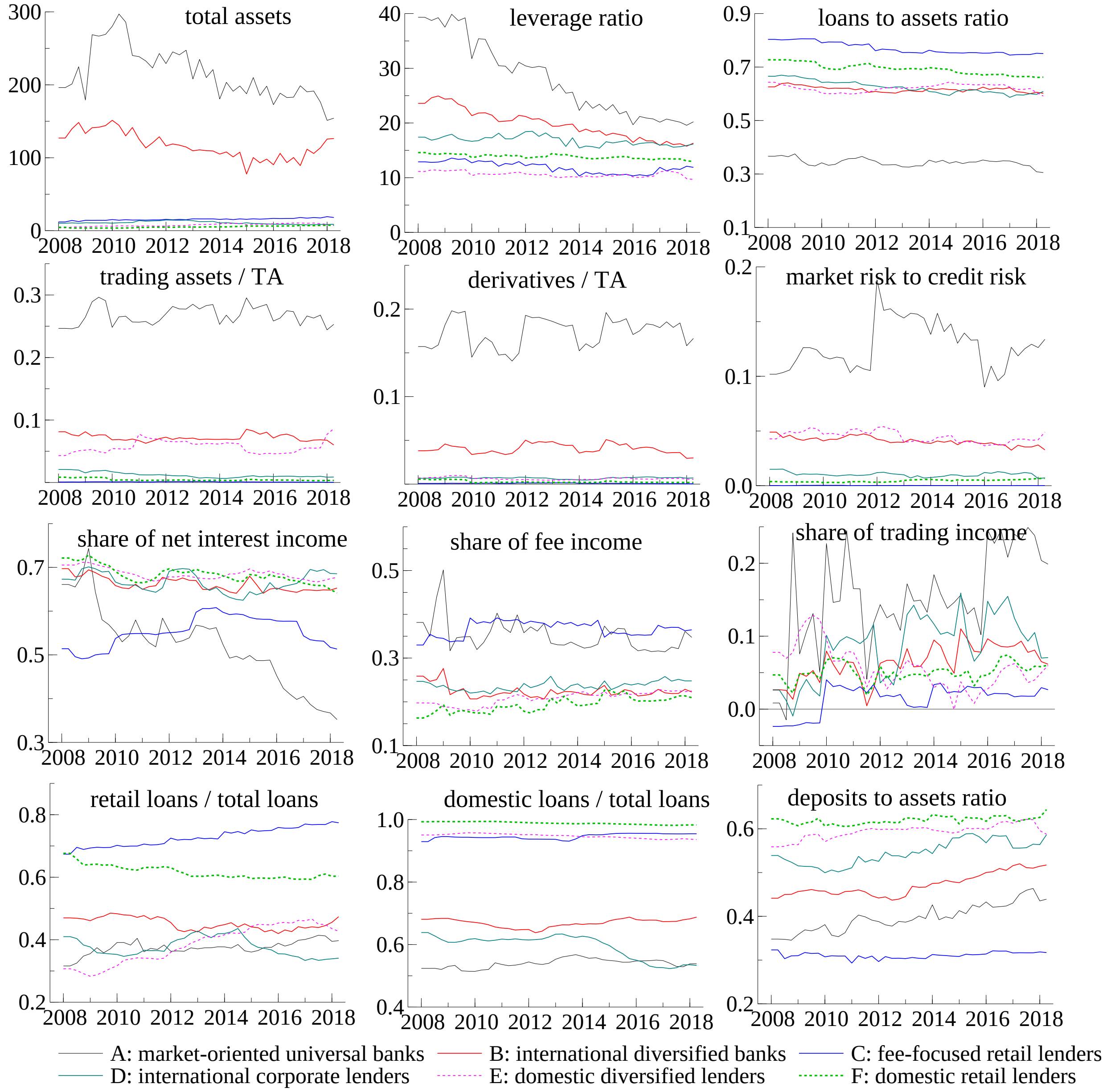
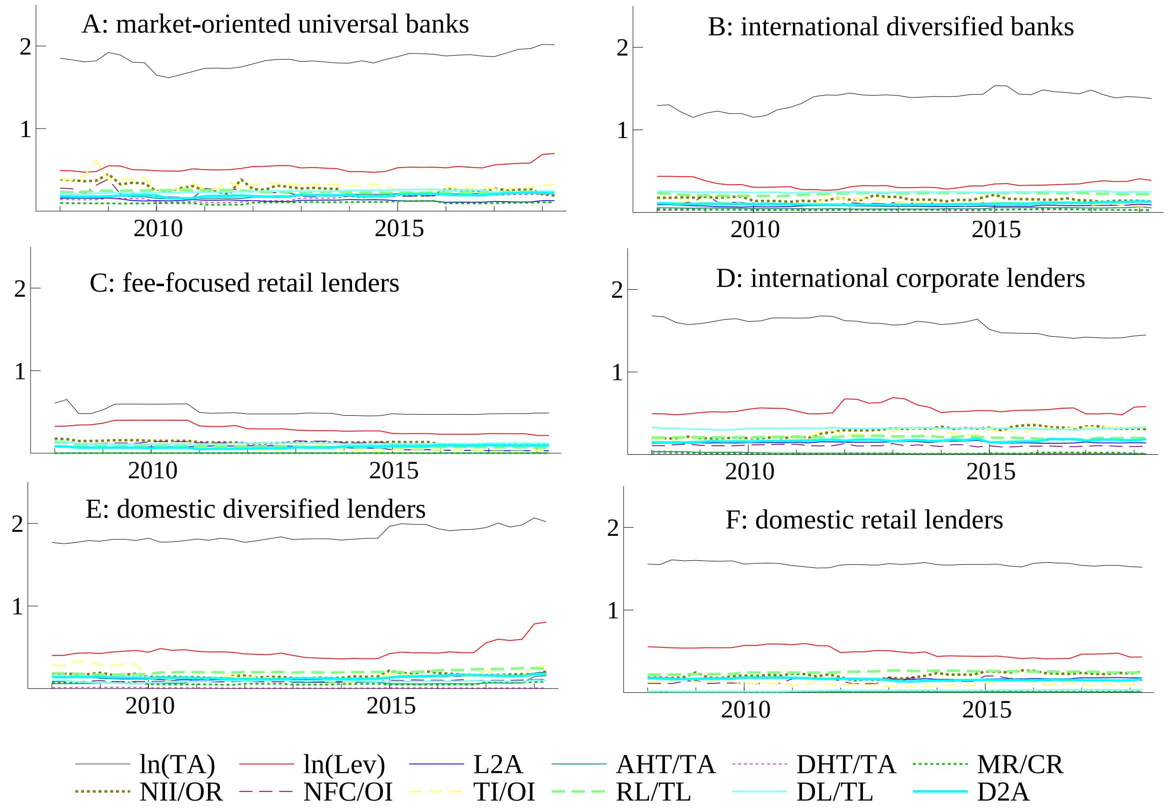


Figure D.5: Time-varying standard deviations

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.

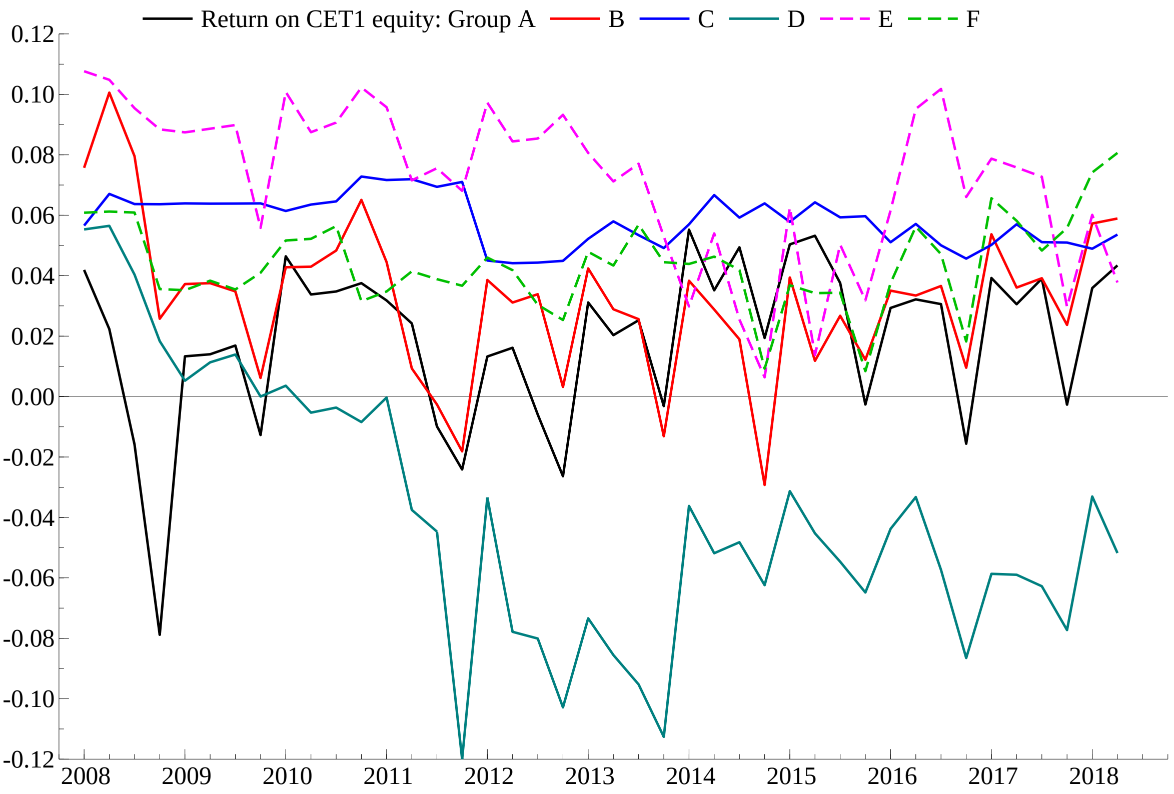


D.5 Bank group profitability

The cluster transitions underlying Figures 5 – 6 are in part explained by differences in bank profitability. Figure D.6 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.6: 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.