Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis Econometrics

High-Dimensional Markov-Switching Vector Autoregression Spillover Networks

Author: Boyen Pronk (501149)

Supervisor:	Dr. H.J.W.G. Kole
Second assessor:	Prof. Dr. D.J.C. van Dijk
Date:	24th June 2024

Abstract

The generalised impulse response function of a Markov-switching vector autoregression (MS-VAR) is used to construct a time-varying spillover index (SI). The MS-VAR is estimated by means of maximum likelihood subject to adaptive elastic net penalisation with a view to the construction of the SI for the high-dimensional global bank stock return volatility dataset of Demirer, Diebold, Liu and Yilmaz (2018). It is found that the parameter estimates of the MS-VAR are preferred over those of a vector autoregression based on an appropriate information criterion. The time-variation in the SI of the MS-VAR, however, mainly consists of switching between the full-sample SIs of the prevailing regimes. It is recommended to combine the MS-VAR with rolling windows to obtain the dynamic SI. To advance the analysis and interpretation of the spillover networks that are hereby obtained, I apply network-theoretic methods to inquire into regional patterns of spillovers and the evolution of the network structure over time. A community detection algorithm is used to determine clusters of nodes. It is found that the MS-VAR leads to markedly different networks that can be linked to the differences in the parameter estimates over the regimes. When aggregating the banks to obtain a network at the level of the country by means of multiple shock impulse response functions, similar results are obtained. Finally, I attempt to improve the predictive performance of spillover networks by means of graph embeddings. It is found that the graph embeddings are competitive with the SI as features in a classifier.

Keywords: Spillover index, Markov-switching vector autoregression, Adaptive elastic net, Bank network, Graph embedding, Systemic event prediction.

List of Abbreviations

ADF-GLS Augmented Dickey-Fuller-generalised least squares

AIC Akaike information crtierion

AICc Corrected Akaike information criterion

BIC Bayesian information criterion

DBSCAN Density-based spatial clustering of applications with noise

DDLY Demirer, Diebold, Liu and Yilmaz

EDF Effective degrees of freedom

EM Expectation-maximisation

GFEVD Generalised forecast error variance decomposition

GIRF Generalised impulse response function

GLASSO Graphical least absolute shrinkage and selection operator

KL Kullback-Leibler

LASSO Least absolute shrinkage and selection operator

ML Maximum likelihood

MS-VAR Markov-switching vector autoregression

MSC Markov-switching criterion

MSIRF Multiple shock impulse response function

OLS Ordinary least squares

PCD Pathwise coordinate descent

SE Systemic event

SI Spillover index

t-SNE t-distributed stochastic neighbour embedding

VAR Vector autoregression

Contents

1 Introduction

The spillover index (SI), developed by Diebold and Yilmaz (2009, 2012), is one of the main metrics employed in the analysis of connectedness between financial and economic time series. Empirically, it has been applied extensively, among others, to volatility connectedness of financial institutions (Diebold & Yilmaz, 2014; Diebold & Yilmaz, 2015b; Demirer, Diebold, Liu & Yilmaz, 2018), commodity returns (Diebold, Liu & Yilmaz, 2017) and international market indices (Beraich, Amzile, Laamire, Zirari & Fadali, 2022), return connectedness of asset classes (Bouri, Cepni, Gabauer & Gupta, 2021), cryptocurrencies (Kumar, Iqbal, Mitra, Kristoufek & Bouri, 2022), exchange rates (Antonakakis, Chatziantoniou & Gabauer, 2020), oil markets (Zhang & Wang, 2014), implied volatility (Kae-Yih, 2023), housing prices (Gabauer, Gupta, Marfatia & Miller, 2024) and in the variety of settings of Diebold and Yilmaz (2015a). Methodologically, it has been extended by means of a spectral representation of the SI (Baruník & Křehlík, 2018), by modelling the conditional quantiles of time series rather than their conditional means (Bouri, Lucey, Saeed & Vinh Vo, 2020) and by means of refining the definition of the SI (Lastrapes & Wiesen, 2021).

This thesis will focus on another aspect, namely, the time-variation in the SI. A recurring theme in many of the aformentioned applications is the emphasis on periods in which connectedness increases. These coincide with uncertain and volatile episodes in the economy and in financial markets such as the great financial crisis, the COVID-19 pandemic and the Russo-Ukrainian war. In the development of the SI, as well as in other metrics of connectedness (e.g. Billio, Getmansky, Lo & Pelizzon, 2012; Dungey & Martin, 2007), the ability to describe such periods is considered to be of importance. By inducing dynamics in the SI, the connectedness of the system can be monitored in these periods.

These dynamics are conventionally obtained by estimating a vector autoregression (VAR) using a rolling window of observations. Alternatively, time-varying parameter VARs have been proposed by Antonakakis et al. (2020) to better capture dynamics in the SI and for volatility connectedness, the DCC-GARCH has been proposed by Gabauer (2020).[1] These methods are suited for smooth changes in the parameters over time (Granger, 2008).[2] Consequently, changes in the SI over time are often gradual. Yet, it is desirable that the SI swiftly responds to changes in underlying connectedness.[3] Moreover, the number of observations decreases drastically when using rolling windows.

For both financial (Ang & Timmermann, 2012) and economic time series (Hamilton, 2016), changes are often of a more abrupt nature. Regime-switching models, in which the process governing the regimes is typically described by a Markov process, are frequently used to model such changes in the parameters (Guidolin, 2011). BenSaïda, Litimi and Abdallah (2018) and Kim and Lee (2023) to my knowledge are the only applications of a Markov-switching VAR (MS-VAR) to obtain regime-specific SIs.[4] Thus, the use of regime-switching is rare and moreover

[1] DCC-GARCH stands for dynamic conditional correlation generalised autoregressive conditional heteroskedasicity.

[2] Cf. Diebold & Yilmaz (2015a), pp. 22-23.

[3] Cf. Korobilis & Yilmaz (2018), pp. 2.

[4] Regime-switching models for volatility spillovers through volatility models have been in place for longer, e.g.

has been limited to low-dimensional data. Recently, Kole and Van Dijk (2023) have derived a closed-form expression of the generalised impulse response function (GIRF) of the MS-VAR, which takes account of the regimes over time in the determination of the impulse response. This enables going beyond regime-specific SIs and allows for time-variation in the SI.

A drawback of the SI is that high-dimensional systems can lead to inaccurate parameter estimates due to the parametrisation of a VAR. For the estimation of large systems, containing dozens of variables or more, recourse must be had to the methods of high-dimensional VARs. This holds a fortiori for MS-VARs, as these require parameter estimates for every regime. The least absolute shrinkage and selection operator (LASSO) and the elastic net are popular methods to obtain high-dimensional SIs. Examples are Demirer et al. (2018), Yi, Xu and Wang (2018), Bostanci and Yilmaz (2020), Gabauer et al. (2024) and Chen and Schienle (2022), who also extend the methodology to a vector error correction model in which the LASSO selects the cointegration relations.[5] The inclusion of more variables into the system is not only of interest because such variables could be informative per se, but also because this enables the description of more extensive networks. For example, Gabauer et al. (2024) describe housing price connectedness for the states of the United States and, precisely this level of aggregation being of interest, this necessitates a high-dimensional system. The aforementioned applications however, do not provide for in-built, i.e. without resorting to a rolling window, dynamics in the parameter estimates.

To that end, recent contributions in the area of high-dimensional MS-VARs are useful, with different forms of penalised maximum likelihood (ML) estimation having been developed by Monbet and Ailliot (2017), Maung (2023) and Chavez-Martinez, Agarwal, Khalili and Ahmed (2023) respectively. The main contribution of this thesis will therefore be the application of an MS-VAR with adaptive elastic net penalisation to obtain SIs, which enables both the estimation of high-dimensional systems, as well as the incorporation of abrupt changes in model parameters. In an application to the global bank stock return volatility dataset of Demirer et al. (2018), it will be seen that the use of an MS-VAR estimated by means of penalised ML is preferred to that of a VAR based on an appropriate information criterion. The SI of the MS-VAR will exhibit time-variation, but it mostly coincides with the full-sample SI of the prevailing, inferred regime, although time-variation stemming from the forecast error variance is also present. The bootstrap method of Choi and Shin (2020) will be applied to the SI of the VAR to construct confidence intervals thereof. Using this method, it will also be shown that directional spillovers differ significantly across regions.

The second contribution of this thesis pertains to the networks defined by the generalised forecast error variance (GFEVD) and their analysis. Node centrality scores will be used to extend the analysis of regional spillovers to an inquiry into the centrality of regions over time, which

Baele (2005). Another more recent and closely related example is the work of Kangogo and Volkov (2022), who use the historical decomposition of an MS-VAR.

[5] Another method to deal with the problem of parameter proliferation in the VAR that has been applied in this context is the global VAR (Greenwood-Nimmo, Nguyen & Shin, 2021). The measurement of volatility spillovers by conditional correlations can also be subjected to regularisation with the sparse multivariate GARCH models of Dhaene, Sercu and Wu (2022). To my knowledge, time-varying parameter VARs have not been extended to the high-dimensional case, as the Markov chain Monte Carlo methods involved are too burdensome computationally.

differ in a more pronounced fashion than the spillovers. Chan-Lau (2018) introduced the use of community detection algorithms to spillover networks and in this thesis, I will use node embeddings for this purpose, a new method in this context. These are used in a full-sample analysis of communities across regimes and it will be seen how the parameter estimates affect the network structure. The effect of aggregating the individual banks by means of multiple shock impulse response functions (MSIRF), introduced by Van der Zwan (2023), into a country network will also be explored. It will be seen that the networks contain clear, interpretable clusters, but that these break down for different regimes in the MS-VAR. The evolution of the network structure over time will also be considered, as in Isogai (2017) who obtained clusters of networks over time. To that end, I propose the use of graph embeddings and it is found that the networks are similar to such an extent that no clusters are found.

Finally, graph embeddings will be used to inquire into the predictive power of spillover networks and a comparison will be made with that of the SI, which has been used for these purposes in Korobilis and Yilmaz (2018) and Arsov, Canetti, Kodres and Mitra (2013). The graph embeddings and the SI are used as features in a logistic regression model by means of which trading days are classified as constituting a systemic event (SE). A comparison of these models shows that the graph embeddings can be fruitfully applied as features for these purposes and that they are competitive with the SI, both with respect to in-sample fit, as well as with respect to classification performance.

The remainder of this thesis will be structured as follows. In Section 2, I will discuss the VAR and the MS-VAR and their estimation subject to adaptive elastic net penalisation. The criteria which are used to select specifications of the respective models are discussed as well. In Section 3, I will discuss the SI and the GIRF which is its main building block. The bootstrap method, as well as the network-theoretic methods and the node and graph embedding algorithms employed will also be discussed. In Section 4, I will discuss the global bank stock return volatility dataset that is used in an empirical application of the aforementioned methods. In Section 5, I will first discuss the obtained model specifications and their estimation results, followed by the obtained SIs. Furthermore, I will present an analysis of the obtained networks and I will discuss prediction results obtained by means of graph embeddings. In Section 6, I will summarise the findings of this thesis and draw conclusions therefrom.

2 Markov-Switching Vector Autoregression

2.1 Model Formulation

In the following description of the MS-VAR, the notation follows Kole and Van Dijk (2023) with some adjustments. Let y_t be a k-dimensional vector of time series and s_t be the prevailing regime at a time t. If there are M different regimes, s_t can take values $1, \ldots, M$ and the general MS(M)-VAR(p) can be formulated as follows:

$$y_t = c_{s_t} + \Phi_{1,s_t} y_{t-1} + \ldots + \Phi_{p,s_t} y_{t-p} + u_t, \quad u_t \sim \mathcal{N}(\mathbf{0}, \Sigma_{s_t}), \quad t = p+1, \ldots, T \tag{2.1}$$

where c_{s_t} is a vector of intercepts. The reduced form error terms u_t can be decomposed as $\Lambda_{s_t}\varepsilon_t$, such that $\Lambda_{s_t}\Lambda'_{s_t} = \Sigma_{s_t} = \text{Var}[y_t|s_t,\ y_{t-1},\ \ldots,\ y_{t-p}]$, the conditional variance of y_t. Λ_{s_t} can be identified using a Cholesky decomposition. The structural error terms ε_t are serially uncorrelated, $\mathbb{E}[\varepsilon_t\varepsilon'_{t+s}] = \mathbf{O},\ s \neq 0$. The VAR, which corresponds to equation (2.1) with $M = 1$, is assumed to be stable. This entails that all of the roots of $|\mathbf{I}_k - \sum_{l=1}^{p} \Phi_l z| = 0$, an equation in z with $\mathbf{I}_k$ being the k-dimensional identity matrix, are outside the complex unit circle. This implies that an infinite-order moving average representation of the VAR exists, which is required for the evaluation of its GIRF. This condition is equivalent to the spectral radius of the autoregressive parameter matrix of its VAR(1) representation not being greater than or equal to one. For the MS-VAR, a similar restriction on the spectral radius holds (Kole & Van Dijk, 2023).[6]

It is generally not desirable to labour under the assumption of Gaussian error terms for a linear VAR in practice, as it is not plausible. This lead for example Diebold and Yilmaz (2014) to apply natural logarithms to the volatilities of United States financial institution stock returns. The MS-VAR could provide another advantage. As such transformations might not be necessary, more information is retained in the data, although in this thesis, the MS-VAR will be applied to the same data as the benchmark VAR. The MS-VAR can generate skewed, leptokurtic distributions of y_t and conditional heteroskedasticity, Gaussianity of the error terms within regimes notwithstanding (Krolzig, 1997). Moreover, Kole and Van Dijk (2023) show that the Gaussian MS-VAR corresponds to a non-Gaussian linear VAR. Although even within regimes this assumption need not necessarily hold, this assumption is much weaker than for a VAR. It is moreover used very frequently in the literature on the MS-VAR (Maung, 2023). Therefore, as well as with a view to evaluating the GIRF, I proceed with the assumption of Gaussianity.

The process that governs the regimes is modelled as an irreducible, ergodic, first-order Markov chain with transition matrix $\mathbf{P}$, where $\mathbf{P}_{i,j} = p_{i,j} = \mathbb{P}[s_t = i|s_{t-1} = j]$. The prevailing regime can be described by $\boldsymbol{\xi}_t = (\mathbb{1}(s_t = 1),\ \ldots,\ \mathbb{1}(s_t = M))$, where $\mathbb{1}(s_t = m)$ is an indicator function that is equal to 1 if $s_t = m$ and 0 otherwise. The irreducibility and ergodicity of the Markov chain implies that there exists a vector π, the elements of which are $(\pi)_m = \mathbb{P}[s_t = i]$, the ergodic probabilities of the chain such that $\pi = \mathbf{P}\pi$. Moreover, the Markov chain described above is homogeneous. In principle, it is possible to model s_t using a non-homogenous Markov chain, but this is outside the scope of this thesis.[7]

Next, unless stated otherwise, the intercepts, the autoregressive parameters and the covariance matrix are regime-dependent and governed by the same regime. In addition, I will consider the use of multiple, independent Markov chains.[8] For example, let $\{s_t^{(1)}\}_{t=1}^{T}$ and $\{s_t^{(2)}\}_{t=1}^{T}$ be two independent, homogeneous, irreducible, ergodic Markov chains with state space $\mathcal{S}_i$ and transition probability matrix $\mathbf{P}_i$, $i = 1,\ 2$. These can be modelled as one Markov chain $\{s_t\}_{t=1}^{T}$, which has a state space $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$ and a transition probability matrix $\mathbf{P}_1 \otimes \mathbf{P}_2$, where $\times$ is the Cartesian

[6] This condition for the MS-VAR is included in the discussion of the GIRF of the MS-VAR in Appendix D.

[7] Bazzi, Blasques, Koopman and Lucas (2017) incorporate time-varying parameters using generalised autoregressive score parameter updates and the transition probabilities obtained differ non-trivially over time. Yet, their inferred regimes do not markedly differ from those that are obtained using constant probabilities.

[8] This can be extended to dependent Markov chains using the methodology of Catania (2022). However, this is quite involved and will not be pursued in this thesis.

product and $\otimes$ is the Kronecker product (Hamilton & Lin, 1996). The obtained Markov chain is still homogeneous, irreducible and ergodic. The use of multiple chains can pertain both to different sets of parameters, as well as to different equations of the system. In the latter case, $(c_1,\ \Phi_1),\ \ldots,\ (c_p,\ \Phi_p)$ are different for each regime, but their rows are subject to parameter restrictions across regimes.

2.2 Model Estimation

2.2.1 Single Regime

Estimation of high-dimensional VARs requires specific methods to deal with the problem of parameter proliferation that these models are subject to. First, I discuss estimation of the VAR, by means of ordinary least squares (OLS) with adaptive elastic net penalisation, a regularisation method also used by Demirer et al. (2018), who use equally weighted LASSO- and Ridge-penalties. The adaptive elastic net, due to Zou and Zhang (2009), can deal with multicollinearity better than the adaptive LASSO and asymptotically has the oracle property, shown for the VAR by Furman (2014). This requires a set of preliminary, consistent parameter estimates that can be obtained by Ridge-penalised OLS. Weights are then obtained for $(\Phi_l)_{i,j}$ by setting $w_{i,j;l} = (|\hat{\beta}^{\text{Ridge}}_{i,j;l}|)^{-\gamma}$, $\gamma > 0$. As the dimension of the number of predictors is finite, Zou and Zhang (2009) describe that any positive value of γ can be chosen and following these authors, I set $\gamma = 1$. Moreover, the use of the adaptive weights can also function to render the specification of a specific sparsity pattern superfluous, as these per se provide for variable-specific penalisation in a data-driven manner.[9]

For the objective function of the VAR, note that equation-by-equation estimation is equivalent to system-based, simultaneous estimation. The imposed penalisation is also identical for both equation-by-equation and system-based estimation and does not have a differential effect on the parameter estimates. The objective function can thus be constructed as follows, letting $\mathbf{\Phi} = (\Phi_1, \ldots, \Phi_p)$ and $\beta = (c, \mathbf{\Phi})$:

$$\begin{aligned}\hat{\beta}_i = \underset{\beta_i}{\operatorname{argmin}} \sum_{t=p+1}^{T} (y_{i,t} - c_i - \sum_{l=1}^{p} (\Phi_l)_i' y_{t-l})' (y_{i,t} - c_i - \sum_{l=1}^{p} (\Phi_l)_i' y_{t-l}) \\ + \lambda_i \Big[(1-\alpha_i) \|(\mathbf{\Phi})_i\|_2 + \alpha_i (W_i \odot \|(\mathbf{\Phi})_i\|_1) \Big],\ \ i = 1,\ \ldots,\ k\end{aligned} \tag{2.2}$$

where β_i, $(\Phi_l)_i$ and $(\mathbf{\Phi})_i$ are row i of matrices β, Φ_l and $\mathbf{\Phi}$ respectively, W_i is the kp-dimensional vector of adaptive elastic net weights, $i = 1,\ \ldots,\ k$, $\odot$ is the Hadamard product and $\| \bullet \|_p$ is the L_p-norm. Thus, the parameters are shrunk towards zero. The reason for this is that the LASSO aims to select parameters of variables that are relevant to explain the variation in the dependent variable. If a variable lacks such explanatory value, the parameter is set to zero and the variable is effectively excluded from the model. In the context of a VAR, this interpretation

[9] Cf. Chavez-Martinez et al. (2023), pp. 555. Other forms of penalisation, such as those discussed in Nicholson, Matteson and Bien (2017), can be applied if desirable. Relevant examples are group-sparsity and lag-sparsity. The former arises when the variables are part of groups that are specified a priori. Then, the parameters of variables that are part of another group can be penalised more severely. Lag-sparsity penalises parameters of higher-order lags more severely than those of lower-order lags.

in accordance with the notion of Granger-causality; if the parameter(s) of a variable are all equal to zero, this variable does not Granger-cause the dependent variable. The constant terms are not penalised, which is usual practice since it is unrealistic to assume sparsity thereof. Finally, I condition on the first p observations $y_1, \ldots, y_p$, which also is usual practice.

Note that the penalised OLS estimate is equivalent to the penalised ML estimate under the assumption of Gaussian error terms and the fact that Σ is non-diagonal is without repercussions, as shown in Hamilton (1994, pp. 293 et seq.).[10] Moreover, this is the case for arbitrary forms of heteroskedasticity, i.e. time-variation in Σ.

To estimate Σ, the graphical LASSO (GLASSO) of Friedman, Hastie and Tibshirani (2008) will be used. The GLASSO shrinks off-diagonal elements of the precision matrix $\Omega = \Sigma^{-1}$ to zero, which is suitable for high dimensionality. The GLASSO consists of solving the following optimisation problem:

$$\hat{\Omega} = \underset{\Omega}{\operatorname{argmax}} \ \log|\Omega| - \operatorname{tr}(\mathbb{S}\Omega) - \rho\|\Omega - \operatorname{diag}(\Omega)\|_1 \tag{2.3}$$

where $\operatorname{tr}(\bullet)$ denotes the trace of a matrix and $\mathbb{S}$ is the Gaussian ML estimate of the covariance matrix of u_t, i.e.

$$\mathbb{S} = \frac{1}{T-p}\sum_{t=p+1}^{T}(y_t - c - \sum_{l=1}^{p}\Phi_l y_{t-l})(y_t - c - \sum_{l=1}^{p}\Phi_l y_{t-l})' \tag{2.4}$$

Thus, $\hat{\Omega}$ can be obtained by plugging in $\hat{\beta}$ into equation (2.4) and then solving equation (2.3). Demirer, Diebold, Liu and Yilmaz (2018, pp. 5) explicitly refrain from shrinking the error term covariance matrix, as they: "... are not necessarily comfortable with the standard 'statistical' shrinkage directions (e.g., toward zero).". Under multivariate normality, $\Omega_{i,j} = 0$ implies that the error terms of variables i and j are independent conditional on the other variables (Friedman et al., 2008). In the context of a VAR, this is a very relevant shrinkage direction.

Parameter estimates can be obtained by means of the pathwise coordinate descent (PCD) algorithm of Friedman, Hastie, Höfling and Tibshirani (2007), implemented in the R package `glmnet` (Friedman, Hastie & Tibshirani, 2010). The GLASSO algorithm is implemented in the R package `glasso` of Friedman et al. (2008). Details on these algorithms are included in Appendices A and B respectively.

2.2.2 Markov-Switching Regimes: Basis

Next, I describe penalised ML estimation of the MS-VAR using the adaptive elastic net. The ML estimate of an MS-VAR is consistent (Douc, Moulines & Rydén, 2004). This result has been extended by Kasahara and Shimotsu (2019) to the case when some of the transition probabilities

[10] If the error terms are not Gaussian, then the OLS estimates are asymptotically not efficient, although they remain unbiased. In that case, one could resort to generalised least squares estimation. This aspect seldomly receives attention in the literature on the SI. The work of Ando, Greenwood-Nimmo and Shin (2022) is an exception. These authors model contemporaneous correlation between the error terms of different variables by means of another alternative, a common factor error structure.

are equal to zero. Chavez-Martinez et al. (2023) build on Douc et al. (2004) and show that the ML estimator is consistent for penalisation by means of the adaptive LASSO. The adaptive elastic net also satisfies the assumptions on the penalty function under which the penalised ML estimate is consistent. Moreover, they prove that the adaptive LASSO has the oracle property for a suitable choice of the penalty parameter and, under their assumptions, the same holds for the adaptive elastic net.

To obtain the ML estimate, a penalised expectation maximisation (EM) algorithm will be used. For the MS-VAR this algorithm is favourable from a computational perspective compared to numerical optimisation, as well as with respect to its properties regarding convergence (Krolzig, 1997). This holds true a fortiori for the high-dimensional case, with Monbet and Ailliot (2017), Maung (2023) and Chavez-Martinez et al. (2023) each employing a version of the EM-algorithm.

In describing the EM-algorithm, I depart from the conditional likelihood $f(y_{p+1:T}|y_{1:p},\ s_p;\ \boldsymbol{\theta})$. The subscript $t:q$ indicates that observations t to q of the variable are considered jointly and $\boldsymbol{\theta} = [\text{vec}(c_1,\ \ldots,\ c_M,\ \Phi_{1,1},\ \ldots,\ \Phi_{p,1},\ \ldots,\ \Phi_{1,M},\ \ldots,\ \Phi_{p,M},\ \Sigma_1, \ldots,\ \Sigma_M),\ p_{1,1},\ \ldots,\ p_{M,1},\ \ldots, p_{1,M},\ \ldots, p_{M,M}]$. The following derivations are due to Chavez-Martinez et al. (2023). Like these authors, I condition on s_p, the state of period p, as this simplifies matters and, following Douc et al. (2004), the parameters that maximise the likelihood conditioned on s_p are asymptotically equivalent to those that maximise the likelihood function that is not conditioned as such. First, by the law of total probability it holds that

$$f(y_{p+1:T}|y_{1:p},\ s_p;\ \boldsymbol{\theta}) = \sum_{s_T=1}^{M} \cdots \sum_{s_{p+1}=1}^{M} \left(\prod_{t=p+1}^{T} p_{s_{t-1},s_t} \right) \left(\prod_{t=p+1}^{T} \phi(y_t; \mu_{t,s_t}, \Sigma_{s_t}) \right) \tag{2.5}$$

where $\phi(\bullet; \mu_\bullet, \Sigma_\bullet)$ is a multivariate Gaussian density function for which $\mu_{t,s_t} = c_{s_t} + \sum_{l=1}^{p} \Phi_{l,s_t} y_{t-l}$. Let $\ell(\boldsymbol{\theta};\ s_p) = \log f(y_{p+1:T}|y_{1:p},\ s_p;\ \boldsymbol{\theta})$, $\Phi = (\Phi_{1,1},\ \ldots,\ \Phi_{p,1},\ \ldots,\ \Phi_{1,M},\ \ldots,\ \Phi_{p,M})$ and W be a $(k \times Mkp)$-dimensional matrix containing the adaptive weights, again obtained by preliminary Ridge-estimates. Then, the penalty for the autoregressive parameters and for the off-diagonal terms of the precision matrices respectively is defined as follows:

$$\mathcal{P}_\Phi = \lambda \left[(1-\alpha) \|\Phi\|_2 - \alpha (W \odot \|\Phi\|_1) \right] \qquad \mathcal{P}_\Omega = \rho \sum_{m=1}^{M} \|\Omega_m - \text{diag}(\Omega_m)\|_1$$

Then, the penalised ML estimate is formulated as:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\, \ell(\boldsymbol{\theta}; s_p) - \mathcal{P}_\Phi - \mathcal{P}_\Omega \tag{2.6}$$

The EM-algorithm obtains $\hat{\boldsymbol{\theta}}$ as follows. First, the complete data conditional likelihood is equal to $f_c(y_{p+1:T},\ \boldsymbol{\xi}_{p+1:T}|y_{1:p},\ s_p;\ \boldsymbol{\theta})$ and $\ell_c(\boldsymbol{\theta};\ s_p) = \log f(y_{p+1:T},\ \boldsymbol{\xi}_{p+1:T}|y_{1:p},\ s_p;\ \boldsymbol{\theta})$, with the subscript c denoting complete data. For this complete data log-likelihood it holds that

$$\ell_c(\boldsymbol{\theta};\ s_p) = \sum_{i,j=1}^{M} \sum_{t=p+1}^{T} \left(\boldsymbol{\xi}_{t-1,j} \boldsymbol{\xi}_{t,i} p_{i,j} \right) + \sum_{m=1}^{M} \sum_{t=p+1}^{T} \left[\boldsymbol{\xi}_{t,i} \log \phi(y_t; \mu_{t,m}, \Sigma_m) \right] \tag{2.7}$$

where $\mu_{t,m} = y_t - (c_m + \sum_{l=1}^{p} \Phi_{l,m} y_{t-l})$. The E-step of the algorithm can then be calculated as follows:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(n)}) &= \mathbb{E}_{\boldsymbol{\xi}_{p+1:T}}[\ell_c(\boldsymbol{\theta};\ s_p)|y_{1:T},\ \boldsymbol{\theta}^{(n)}] \\ &= \sum_{i,j=1}^{M} \sum_{t=p+1}^{T} \hat{\boldsymbol{\xi}}_{t|T,i,j}^{(n)} p_{i,j} + \sum_{m=1}^{M} \sum_{t=p+1}^{T} \left[\frac{\hat{\boldsymbol{\xi}}_{t|T,m}^{(n)}}{2} \left(|\log \Omega_m| - (y_t - \mu_{t,m})' \Omega_m (y_t - \mu_{t,m}) \right) \right] \end{aligned} \tag{2.8}$$

where the superscript $(\bullet)$ denotes the iteration of the algorithm. $\hat{\boldsymbol{\xi}}_{t|T,i,j}^{(n)}$ and $\hat{\boldsymbol{\xi}}_{t|T,m}^{(n)}$ can be obtained by means of the subsequently running the Hamilton (1989) filter and the Kim (1994) smoother. For details thereof, the reader is referred to Appendix C. The algorithm proceeds with the M-step as follows:

$$\boldsymbol{\theta}^{(n+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(n)}) - \mathcal{P}_\Phi - \mathcal{P}_\Omega \tag{2.9}$$

where inclusion of the penalty in equation (2.9) is possible following Green (1990). For this, define $\Phi_m = (\Phi_{1,m},\ \ldots,\ \Phi_{p,m})$ and $\boldsymbol{\beta}_m = (c_m,\ \Phi_m)$. The optimisation problem for the M-step is separable in $\mathbf{P}$ and $(\boldsymbol{\beta}_m,\ \Omega_m)$. Moreover, due to equation (2.9) being a summation in the regimes, for the problem is separable in the regimes for the constant terms, the autoregressive parameters and the precision matrices. For the transition probabilities this leads to the following respective updates, where the separate cases arise due to conditioning on s_p,

$$\hat{p}_{i,j}^{(n+1)} = \begin{cases} \dfrac{\sum_{t=p+1}^{T} \hat{\boldsymbol{\xi}}_{t|T,i,s_p}^{(n)}}{\sum_{m=1}^{M} \sum_{t=p+1}^{T} \hat{\boldsymbol{\xi}}_{t|T,m,s_p}^{(n)}}, & i = 1,\ \ldots,\ M,\ j = s_p \\ \dfrac{\sum_{t=p+2}^{T} \hat{\boldsymbol{\xi}}_{t|T,i,j}^{(n)}}{\sum_{m=1}^{M} \sum_{t=p+2}^{T} \hat{\boldsymbol{\xi}}_{t|T,m,j}^{(n)}}, & i,\ j = 1,\ \ldots,\ M,\ j \neq s_p \end{cases} \tag{2.10}$$

Then, for each regime $m = 1,\ \ldots,\ M$, the constant terms, the autoregressive parameters and the precision matrix can be updated by solving the following problem:

$$(\boldsymbol{\beta}_m^{(n+1)},\ \Omega_m^{(n+1)}) = \underset{(\boldsymbol{\beta}_m, \Omega_m)}{\operatorname{argmax}}\ \log|\Omega_m| - \operatorname{tr}(\mathbb{S}_{\boldsymbol{\beta}_m} \Omega_m) - \mathcal{P}_{\Phi_m} - \mathcal{P}_{\Omega_m} \tag{2.11}$$

in which the first two terms correspond to the second term of equation (2.9) and can be rewritten as such due to the within-regime Gaussianity using the properties of the trace. The separability in regimes also allows for regime-specific penalisation λ_m, α_m and ρ_m. $\mathbb{S}_{\boldsymbol{\beta}_m}$ corresponds to the sample covariance matrix of the error term of the regime m where each observation is weighted by its smoothed probability to be in that regime

$$\mathbb{S}_{\boldsymbol{\beta}_m} = \frac{\sum_{t=p+1}^{T} \hat{\boldsymbol{\xi}}_{t|T,m} \big((y_t - \mu_{t,m})(y_t - \mu_{t,m})' \big)}{\sum_{t=p+1}^{T} \hat{\boldsymbol{\xi}}_{t|T,m}} \tag{2.12}$$

As in Maung (2023), the updates of equation (2.11) can be obtained sequentially. I first update

the constant terms and the autoregressive parameters as follows:

$$\boldsymbol{\beta}_m^{(n+1)} = \underset{\boldsymbol{\beta}_m}{\operatorname{argmin}} \frac{\sum_{t=p+1}^{T} \hat{\boldsymbol{\xi}}_{t|T,m}(y_t - c_m - \sum_{l=1}^{p} \Phi_{l,m} y_{t-l})' \Omega_m^{(n)} (y_t - c_m - \sum_{l=1}^{p} \Phi_{l,m} y_{t-l})}{\sum_{t=p+1}^{T} \hat{\boldsymbol{\xi}}_{t|T,m}} + \mathcal{P}_{\Phi_m} \tag{2.13}$$

It can be seen that this is of the form of the ML estimate of constant terms and autoregressive parameters for a Gaussian error term distribution, such that the inclusion of $\Omega_m^{(n)}$ does not detract from this. The parameters can therefore again be estimated on an equation-by-equation basis, with the observations weighted accordingly.

For the error term covariance matrix, the update is obtained by plugging in $\hat{\boldsymbol{\xi}}_{t|T}^{(n+1)}$ and $\boldsymbol{\beta}_m^{(n)}$ into $\mu_{t,m}$ in equation (2.16) to obtain $\mathbb{S}_{\boldsymbol{\beta}_m}^{(n)}$. It then follows that:

$$\Omega_m^{(n+1)} = \underset{\Omega_m}{\operatorname{argmax}} \ \log|\Omega_m| - \operatorname{tr}(\mathbb{S}_{\boldsymbol{\beta}_m}^{(n+1)} \Omega_m) - \mathcal{P}_{\Omega_m} \tag{2.14}$$

This is a more general GLASSO estimate of equation (2.3).

For the initialisation of the EM-algorithm, I use the parameters of a Ridge-penalised VAR. Then, the elements of the error term covariance matrices are scaled by a factor of $1.1^{(m-1)}$ for $m = 1, \ldots, M$, leading to higher and lower volatility regimes where the correlation structure is preserved. When $M = 2$, $c_{i,2} = c_{i,1} + 0.25\sigma_{i,i}$, $i = 1, \ldots, k$. When $M > 2$, I do not opt for different initial intercepts, to accomodate e.g. a higher volatility regime in which the level of the series is lower. For any value of M, I do not scale the autoregressive parameter matrices as this could lead to erratic autoregressive dynamics, such as by inducing severe within-regime instability. Next is the transition probability matrix. Let $\boldsymbol{\iota}$ be a conformable vector with 1 as each element. Then, $\mathbf{P}$ is initialised as $0.8\mathbf{I} + \frac{0.2}{M}\boldsymbol{\iota}\boldsymbol{\iota}'$. $\hat{\boldsymbol{\xi}}_{0|0}$ is initialised as $M^{-1}\boldsymbol{\iota}$, the effect of which asymptotically is negligible (Chavez-Martinez et al., 2023). Note that this is different from conditioning on s_p, which pertains to the parameter estimates through the likelihood function, whereas this pertains to the distribution of s_p, which affects the Hamilton filter. After the first iteration of the algorithm, the Hamilton filter can be initialised by means of $\hat{\boldsymbol{\xi}}_{0|T}$ of the previous iteration, as in BenSaïda et al. (2018). For any value of M, s_p is set to 1. This can be used to estimate the parameters using a Ridge-penalised EM-algorithm. The Ridge-estimates are then used to determine the adaptive weights and to initialise the algorithm for each combination of penalty parameters. Convergence is attained when the relative increase in the expectation of the complete data log-likelihood is less than 1‰, i.e:

$$\frac{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(n+1)}) - Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(n)})}{|Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(n)})|} < 0.001$$

For a converged EM-algorithm, the expected complete data log-likelihood, evaluated in $\boldsymbol{\theta}$ is approximately equal to the log-likelihood. As a backstop, a maximum of 250 iterations is employed. In the implementation, conversion was always obtained in a few dozen iterations. To ameliorate the computational burden, the smoothed probabilities are obtained only for the parameters for the value of λ that obtain the highest likelihood.

2.2.3 Markov-Switching Regimes: Extensions

I consider the extension mentioned in Section 2.1, namely, the use of multiple Markov chains. First, I discuss the case of the constant terms and autoregressive parameters on the one hand and the error term covariance matrix on the other hand being driven by their own processes. This case will hereinafter be referred to as Extension I. Under this extension, if these parameters do not switch simultaneously, less regimes are required to capture their dynamics, leading to more efficient parameter estimates. For the sake of clarity, I restrict the exposition to that of an MS(2)-VAR(p), but it trivially generalises to an MS(M)-VAR(p). The MS(2)-VAR(p) of Extension I is marked by two $\boldsymbol{\beta}$ and two Σ, which together form an MS(4)-VAR(p). Assume without loss of generality that the combination of parameters and regimes is as follows:

$s_t = 1$:	$s_t = 2$:	$s_t = 3$:	$s_t = 4$:
$\boldsymbol{\beta}_1$, Σ_1	$\boldsymbol{\beta}_1$, Σ_2	$\boldsymbol{\beta}_2$, Σ_1	$\boldsymbol{\beta}_2$, Σ_2

From equation (2.8) it is apparent that the update for $\boldsymbol{\beta}_1$ in the EM-algorithm is then given by

$$\begin{aligned}\boldsymbol{\beta}_1^{(n+1)} = \underset{\boldsymbol{\beta}_1}{\operatorname{argmin}} \sum_{t=p+1}^{T} \Bigg[& \hat{\boldsymbol{\xi}}_{t|T,1} (y_t - c_1 - \sum_{l=1}^{p} \Phi_{l,1} y_{t-l})' \Omega_1^{(n)} (y_t - c_1 - \sum_{l=1}^{p} \Phi_{l,1} y_{t-l}) \\ & + \hat{\boldsymbol{\xi}}_{t|T,2} (y_t - c_1 - \sum_{l=1}^{p} \Phi_{l,1} y_{t-l})' \Omega_2^{(n)} (y_t - c_1 - \sum_{l=1}^{p} \Phi_{l,1} y_{t-l}) \Bigg] + \mathcal{P}_{\Phi_1}\end{aligned} \tag{2.15}$$

and the update for $\boldsymbol{\beta}_2$ is obtained mutatis mutandis. This corresponds to the maximisation of a Gaussian likelihood of a sample of size $2(T-p)$. The two sub-samples of size $T-p$ have the same values for their respective observations, but are weighted by the smoothed probabilities to be in regimes 1 and 2 for the first and second subsample respectively. Moreover, the error term covariance matrix is heteroskedastic. Thus, the update for $\boldsymbol{\beta}_1$ can again be obtained on an equation-by-equation basis. Σ_1 can then be updated by obtaining the corresponding precision matrix through solving the GLASSO with the following sample covariance matrix:

$$\mathbb{S}_1 = \frac{\sum_{t=p+1}^{T} \hat{\boldsymbol{\xi}}_{t|T,1} \big((y_t - \mu_{t,1})(y_t - \mu_{t,1})' \big) + \hat{\boldsymbol{\xi}}_{t|T,3} \big((y_t - \mu_{t,3})(y_t - \mu_{t,3})' \big)}{\sum_{t=p+1}^{T} \hat{\boldsymbol{\xi}}_{t|T,1} + \hat{\boldsymbol{\xi}}_{t|T,3}} \tag{2.16}$$

and the update for Σ_2 is obtained mutatis mutandis.

Next, I consider the case of different equations of the system being governed by their own Markov chain. This extension will hereinafter be referred to as Extension II. Again, the exposition is restricted to two-state Markov-chains and can be trivially generalised to M-state Markov chains. This extension is motivated by the data that will be used in the empirical application. Baele (2005) considers the possibility that stock returns and volatilites of the United States and from those of European countries are governed by their own regimes. Within the context of the SI, Diebold and Yilmaz (2015b) found that spillovers in stock return volatilites of financial institutions changed directions in the course of the great recession, first mainly originating from the United States, then bidirectional and finally mainly originating from Europe. Explicitly modelling different regimes for the different regions could therefore lead to a more accurate representation of the underlying connectedness.

In the global bank stock return volatility dataset, the banks are from 29 different countries, with 40 banks being from European countries (including Russia and Turkey), 24 from the Americas (United States, Canada and Brazil), and 32 from Asia.[11] The above idea finds support in Figure 1, which plots the volatility synchronicity for each of these regions, defined by Isogai (2014) as the proportion of stocks for which the volatility at a period exceeds its respective 95^{th} empirical quantile. Hence, Extension I will be employed in which the banks are partitioned in European, American and Asian banks such that the corresponding constant terms and rows of the autoregressive matrices are governed by independent, two-state Markov processes, leading to an eight-regime model.

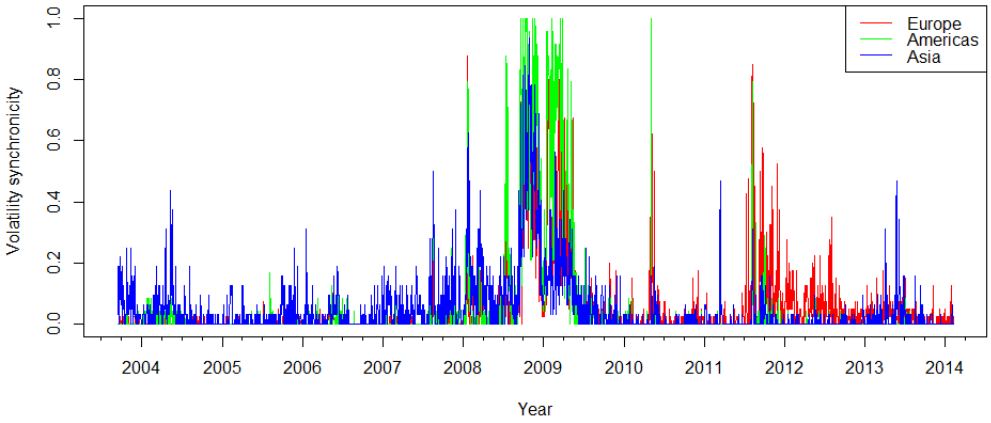


Figure 1: The volatility synchronicity of Europe, the Americas and Asia.

So, although in the following I describe the case in which there are three independent Markov chains, the approach can be generalised to any number of processes. Assume without loss of generality that the first process governs equations 1 to k_1, the second process governs equations k_1+1 to k_2 and the third one governs equations k_2+1 to k, where k_2-k_1, $k-k_2 > 1$. This entails that $\boldsymbol{\beta}$ can be partitioned in three different sets of rows. Denote by $\beta_{i:j}$ the $[(j-i+1)\times k]$-dimensional matrix of which the rows correspond to rows i to j from $\boldsymbol{\beta}$. Define $\beta_1 = \beta_{1:k_1}$, $\beta_2 = \beta_{k_1+1:k_2}$ and $\beta_3 = \beta_{k_2+1:k}$. Thus, $\boldsymbol{\beta}$ can be partitioned as $(\beta_1,\ \beta_2,\ \beta_3)'$. Since each of these vectors is governed by independent, two-state Markov chains, we can assume without loss of generality that the combination of equations and regimes is as follows:

$s_t = 1$:	$s_t = 2$:	$s_t = 3$:	$s_t = 4$:
$\beta_{1;1},\ \beta_{2;1},\ \beta_{3;1},\ \Sigma_1$	$\beta_{1;2},\ \beta_{2;1},\ \beta_{3;1},\ \Sigma_2$	$\beta_{1;1},\ \beta_{2;2},\ \beta_{3;1},\ \Sigma_3$	$\beta_{1;1},\ \beta_{2;1},\ \beta_{3;2},\ \Sigma_4$
$s_t = 5$:	$s_t = 6$:	$s_t = 7$:	$s_t = 8$:
$\beta_{1;2},\ \beta_{2;2},\ \beta_{3;1},\ \Sigma_5$	$\beta_{1;2},\ \beta_{2;1},\ \beta_{3;2},\ \Sigma_6$	$\beta_{1;1},\ \beta_{2;2},\ \beta_{3;2},\ \Sigma_7$	$\beta_{1;2},\ \beta_{2;2},\ \beta_{3;2},\ \Sigma_8$

where the second subscript pertains to the regime. This is a special case of the general MS-VAR with parameters $(\boldsymbol{\beta}_m,\ \Sigma_m)$ for $m = 1,\ \ldots,\ M$, for which we know that equation-by-equation

[11] One bank is from South Africa, which for the purposes of this application will be considered part of Asia.

OLS maximises the likelihood. Namely, this is the same as the update of equation (2.13) with a parameter restriction across regimes, which allows for summing over the regimes for which the parameters are the same. Therefore, it holds that the update for row i of $\beta_{1;1}$ is obtained by solving

$$\underset{c_1,\mathbf{a}_1}{\operatorname{argmin}} \sum_{t=p+1}^{T} \sum_{m\in I} \hat{\boldsymbol{\xi}}_{t|T,m}(y_t - c_1 - \sum_{l=1}^{p}\sum_{j=1}^{k} a_{i,j;l;1} y_{j,t-l})^2 + \mathcal{P}_{\mathbf{a}_1} \tag{2.17}$$

where $\mathbf{a}_1$ is a vector containing the i-th rows of the autoregressive parameter matrices for regimes 1, 3, 4 and 7 and $I = \{1,\ 3,\ 4,\ 7\}$ and the third subscript of $a_{i,j;l;1}$ denotes that this parameter belongs to the first regime for this equation. The updates for other combinations of the process of which row i can be part and the regime can be obtained mutatis mutandis. Σ_m can be estimated as in equation (2.14).

2.3 Model Selection

2.3.1 Single Regime

For the VAR, the lag-order p and penalty parameters λ and ρ and the elastic net parameter α need to be selected. In the following, define $T' = T - p$ and

$$K = \sum_{n=1}^{k} K_n + \sum_{i=1}^{k}\sum_{j=i+1}^{k} \mathbb{1}(\Omega_{i,j} \neq 0)$$

as the effective degrees of freedom (EDF), with K_n denoting the EDF of the autoregressive parameters for the equation of variable n. K does not include the constant terms and the diagonal elements of the error term precision matrix. The precision matrix is used, rather than the covariance matrix, because the GLASSO determines the number of freely estimated parameters of the former. Off-diagonal elements of the error term precision matrix are only counted once, since it is symmetric. For the LASSO, the EDF is equal to $|\mathcal{A}|$, the cardinality of the active set, i.e. the number of coefficients that is not set to zero (Zou, Hastie & Tibshirani, 2007). Hence, the number of off-diagonal elements of the error term precision matrix that are not set to zero are counted in the overall EDF. For Ridge-regression, the EDF can be obtained as

$$\operatorname{tr}\big(X(X'X + \lambda(1-\alpha)\mathbf{I})^{-1}X'\big)$$

where X is a design matrix which for the VAR is equal to $y'_{1:T-p}$. The effective degrees of freedom of the elastic net can be obtained as a combination of the two (Zou & Hastie, n.d.):

$$K_n = \operatorname{tr}\big(X_{\mathcal{A},n}(X'_{\mathcal{A},n}X_{\mathcal{A},n} + \lambda(1-\alpha)\mathbf{I})^{-1}X'_{\mathcal{A},n}\big),\quad n = 1,\ \ldots,\ k \tag{2.18}$$

where $X_{\mathcal{A},n}$ is the design matrix for the equation of variable n with the variables that are part of the active set. To estimate the lag order, the corrected Akaike information criterion (AICc) of Hurvich and Tsai (1989) will be applied.[12] The AIC aims to select the lag order that minimises

[12] Nicholson, Wilms, Bien and Matteson (2020) show by means of simulations that for different sparsity patterns, the lag orders can be more accurately recovered by means of the penalised estimation of the autoregressive parameters in accordance with the purported pattern than when using either the AIC or the Bayesian information

the Kullback-Leibler (KL) divergence over the lag order and is asymptotically efficient. However, the estimated lag order of the AIC is biased and tends to overfit models by selecting higher lag orders. Yet, Gonzalo and Pitarakis (2002) show that for $k = 10$, the AIC does not tend to overfit for moderate sample sizes and performs best. However, the AIC deteriorates when $pk >> T$. The AICc corrects for the bias of the AIC and is better suited for high-dimensional data. The formulation in terms of the likelihood function is taken from Burnham and Anderson (2004) and is as follows:

$$\text{AICc} = -2 \log f(y_{p+1:T};\ \boldsymbol{\theta}) + 2K + \frac{2K(K+1)}{T' - K - 1} \tag{2.19}$$

For λ, Maung (2023) applies the adjusted Bayesian information criterion (BIC) due to Wang, Li and Leng (2009). This adjustment is made to accommodate a diverging number of parameters. For the MS-VAR, this is not necessary, however, as the number of parameters is large and increases quadratically in the dimension of the system, but does not increase with the sample size. Although an increase in sample size could entail selection of additional lags into the model, the dimension of the predictor space itself is fixed. Thus, I follow Zou and Zhang (2009), who use the BIC to tune the adaptive elastic net. The tuning strategy will be through a grid search over α and λ. The grid of alpha will consist of $\{0.50,\ 0.75,\ 1.00\}$. For λ, a grid of 9 values is considered. For each equation, the grid starts at a value of λ for which all but one of the estimated parameters are set to zero, $\lambda_{\max}$, and decreases linearly on the logarithmic scale to $0.0001\lambda_{\max}$. The 9 values correspond to the values of the grid that are equidistant with respect to their position on the grid.[13] For example, for a grid of 40 values, this entails that the values of λ at 10%, 20%, ..., 90% of the length of the grid will be used, i.e. the values at the 4^{th}, 8^{th}, ..., 36^{th} indices of the grid are used. In the following, unless stated otherwise, when I refer to the n-th value of λ, I refer to the value at index n of the grid. This procedure will be repeated for all candidate values of the lag order. The BIC, which is due to Schwarz (1978), is as follows:

$$\text{BIC} = -2 \log f(y_{p+1:T};\ \boldsymbol{\theta}) + K \log T' \tag{2.20}$$

Based on computational considerations, ρ will be set to a single value, namely $0.1 \cdot 0.9 \max(\mathbb{S})_{i \neq j}$, i.e. to a fraction of the largest off-diagonal value of the sample covariance matrix. Preliminary estimates of the VAR indicated that low values of ρ generally led to lower values of the BIC. The formula to determine a concrete value of ρ is based on that used by Mazumder and Hastie (2012).

Based on the application, a preliminary selection can be made of the set of models that will be subjected to the procedure applied above. Demirer et al. (2018), like Diebold and Yilmaz (2014), Diebold and Yilmaz (2015a), Diebold and Yilmaz (2015b) and Diebold et al. (2017) use a VAR(3). Yi et al. (2018) also use a VAR(3). Bostanci and Yilmaz (2020) do this too and perform a sensitivity analysis of the SI to the lag order by using a VAR(2) and a VAR(4) as alternatives and report that the SIs are not sensitive to the lag order selection. BenSaïda et al. (2018) on the other hand, choose a VAR(1) Therefore, for the VAR, the lag orders considered

criterion (BIC). For the element-wise sparsity pattern employed in this thesis, this is not the case however.

[13] The `glmnet` package supplies the user with parameter estimates at a grid containing a relatively large number of values of λ at a high speed. The reason that a subset of these values are chosen is that other computations, such as obtaining the log-likelihood, are slower.

will be one to four; providing for a different number of specifications including those used most often.

2.3.2 Markov-Switching Regimes

For Markov-switching models, specific information criteria have been developed that facilitate simultaneous estimation of p and M.[14] Although Psaradakis and Spagnolo (2006) report favourable results for the use of the AIC and the BIC, Smith, Naik and Tsai (2006) report that the AIC tends to underestimate M and Psaradakis and Spagnolo (2003), in a study on determining M, rather than M and p jointly, found that the BIC tends to underestimate M. Smith et al. (2006) developed the Markov-switching criterion (MSC) that minimises the KL-divergence for this class of models and show that it performs well in a variety of settings; for differing number of regimes and sample sizes. The MSC is as follows:

$$\text{MSC} = -2\log f(y_{p+1:T};\ \boldsymbol{\theta}) + \sum_{m=1}^{M} \frac{\hat{T}_m(\hat{T}_m + MK)}{\hat{T}_m - MK - 2} \tag{2.21}$$

where $\hat{T}_m = \sum_{t=p+1}^{T} \hat{\boldsymbol{\xi}}_{t|T,m}$.

For the MS-VAR, the value of α will be used that was selected for the VAR to limit the computational burden. Then, for each regime m, a grid search will be performed for three values of λ. The highest and lowest values are obtained by means of those that performed best for the VAR and those that performed best in the Ridge-penalised MS-VAR estimates, with the second value chosen as the mean of the two values. For ρ, the same formula is used as for the VAR. For each regime, the EDF can be established and the total EDF will consist of the sum of these individual EDFs.

Taken together, the total number of parameters is counted as follows

$$K = kM + M(M-1) + \sum_{m=1}^{M}\sum_{n=1}^{k} K_{m,n} + \sum_{m=1}^{M}\sum_{i=1}^{k}\sum_{j=i+1}^{k} \mathbb{1}\Big((\Omega_m)_{i,j} \neq 0\Big) \tag{2.22}$$

where

$$K_{n,m} = \text{tr}\big(X_{\mathcal{A},n,m}(X'_{\mathcal{A},n,m}X_{\mathcal{A},n,m} + \lambda_m(1-\alpha)\mathbf{I})^{-1}X'_{\mathcal{A},n,m}\big),\ n = 1,\ \ldots,\ k,\ m = 1,\ \ldots,\ M \tag{2.23}$$

Thus, the active set is now defined over each regime. The term kM arises because for the MS-VAR the number of intercepts is not constant across models. The term $M(M-1)$ accounts for the number of freely estimated transition probabilities.

Finally, I consider the values of M between which a selection shall be made. BenSaïda et al. (2018) use an MS(2)-VAR(1) based on theoretical considerations, although they report lower BIC values for three, four and five regime VAR(1) models, although most of the decrease in the

[14] Cavicchioli (2014) has proven that an MS(M)-VAR(p) admits a VAR-moving average representation with lag orders $p*$ and $q*$ for the AR and MA components respectively. Then, $\hat{p}*$ can be used as an upper bound for p (Guidolin, 2011). Unfortunately, VAR-moving average estimation is not suitable for high-dimensional data, as penalised estimation would be prohibitively complicated.

BIC was due to the inclusion of a second regime and it barely decreased when including a fourth and fifth regime. Kangogo and Volkov (2022) use an MS(2)-VAR(5) for a system of 32 variables which is estimated by unpenalised ML. Based on these results and given the dimension of the dataset, up to four lags and four regimes will be considered for the MS-VAR.

3 Spillover Index

3.1 Generalised Impulse Response Function

After having obtained parameter estimates, it is possible to determine the GIRF. The GIRF, developed by Koop, Pesaran and Potter (1996) and Pesaran and Shin (1998), can be used to determine the effect of a shock in one of the variables on the other variables of the system, while taking account of the contemporaneous correlation of this shock with those of the other variables of the system. The GIRF is defined by the difference of the conditional expectation of the system perturbed by an impulse at time $t+h$ and the conditional expectation of the system at time $t+h$ unperturbed by the impulse. Letting $\nu_{j,t}$ be an impulse to variable j at time t, the GIRF is as follows

$$\mathrm{GI}_y(h,\ \nu_{j,t},\ \mathcal{I}_{t-1}) = \mathbb{E}[y_{t+h}|\nu_{j,t},\ \mathcal{I}_{t-1}] - \mathbb{E}[y_{t+h}|\mathcal{I}_{t-1}], \quad h = 0,\ 1,\ \ldots \tag{3.1}$$

in which $\nu_{j,t}$ is typically set to $\sigma_{j,j}^{\frac{1}{2}}$, the standard deviation of the error term corresponding to the j-th variable. Accordingly, the standardised GIRF can be defined as $\Psi_{y_j}(h) = \mathrm{GI}_y(h,\ \sigma_{j,j}^{\frac{1}{2}},\ \mathcal{I}_{t-1})$. This leads to the following standardised GIRF for the VAR with Gaussian error terms:

$$\Psi_{y_j} = \sigma_{j,j}^{-\frac{1}{2}} A_h \Sigma e_j \tag{3.2}$$

where A_h is the h-th of the impulse response matrices, which corresponds to the h-th moving average term in the moving average representation of the VAR. These matrices are recursively defined as $A_i = \sum_{l=1}^{i} A_{i-l}\Phi_i$ for $i = 1,\ 2,\ \ldots$, with $A_0 = \mathbf{I}_k$ and $\Phi_i = 0$ for $i > p$ and e_j is the j-th basis vector of $\mathbb{R}^k$ which functions as a selection vector. It is the use of Gaussian error terms that leads to this closed form solution and as a consequence, the use of Gaussianity is ubiquitous in the literature on the SI. Namely, for other distributions, one must resort to simulations in evaluating equation (3.1), which is cumbersome due to the sheer number of calculations that must be made in determining the SI. For the MS-VAR, the derivation of the GIRF is due to Kole and Van Dijk (2023). Its derivation is quite involved and for that reason, the main steps of their derivation are reproduced in Appendix D. Here, it is relevant to note that for the MS-VAR, I also standardise the GIRF of a variable j by specifying $\nu_{j,t} = (\mathrm{Var}[y_{j,t}|\mathcal{I}_{t-1}])^{\frac{1}{2}}\ \forall\ t$.

The GIRF is a special case of the MSIRF. Namely, the MSIRF allows for the specification of simultaneous shocks to n variables in the system, $1 \leqslant n \leqslant k$ (Van der Zwan, 2023). Letting $\mathcal{M}$ be the set of shocked variables, this entails the substitution of an n-dimensional vector $\nu_{\mathcal{M},t}$ for the scalar valued $\nu_{j,t}$ in equation (3.1). For a VAR, Van der Zwan (2023) shows by means of simulations how this simultaneous specification leads to different impulse responses than the

aggregation of the separate GIRFs. For the VAR, the MSIRF is as follows:

$$\mathrm{GI}_{\mathcal{M}}(h,\ \boldsymbol{\nu}_{\mathcal{M},t},\ \mathcal{I}_{t-1}) = A_h \Sigma \mathbf{N}(\mathbf{N}'\Sigma\mathbf{N})\boldsymbol{\nu}_{\mathcal{M},t} \tag{3.3}$$

where $\mathbf{N}$ is a matrix of dimension $k \times n$, the columns of which correspond to the basis vectors of $\mathbb{R}^k$ that correspond to the elements of $\mathcal{M}$.[15] Similar to the GIRF, the standardised MSIRF $\Psi_{\mathcal{M}}$ is obtained by setting the elements $\boldsymbol{\nu}_{\mathcal{M},t}$ to the corresponding error term standard deviations.

The GIRF can be used to construct the GFEVD. Pesaran and Shin (1998) have defined it by means of the ratio of the cumulative standardised GIRFs of a variable j to the h-step ahead mean squared forecast error of a variable i, i.e.

$$\delta_{i,j}^{PS}(h) = \frac{\sigma_{i,i}^{-1}\sum_{l=0}^{h}(e_i A_l \Sigma e_j)^2}{\sum_{l=0}^{h}(e_i' A_l \Sigma A_l' e_i)},\quad j = 1,\ \ldots,\ k \tag{3.4}$$

Because Σ in general is non-diagonal, $\sum_{j=1}^{k}\delta_{i,j}^{PS}(h) \neq 1$. This is a consequence of the denominator of equation (3.4) being inherited from the FEVD, i.e. from the decomposition for a model with orthogonal(ised) errors. Specifically, for the FEVD, the denominator consists of the sum of the individual squared impulse response functions (Lütkepohl, 2005). For a linear, Gaussian VAR, this coincides with the h-step ahead mean squared forecast error of variable i. To correct this, Lanne and Nyberg (2016) propose the following definition of the GFEVD, which will be used in this thesis:

$$\delta_{i,j}(h) = \frac{\sum_{l=0}^{h}(\Psi_{\mathbf{y}_j})_i^2}{\sum_{n=1}^{k}\sum_{l=0}^{h}(\Psi_{\mathbf{y}_n})_i^2},\quad j = 1,\ \ldots,\ k,\quad \mathbf{y}_j = y_j,\ \tilde{Y}_j \tag{3.5}$$

where $\tilde{Y}$ corresponds to $\tilde{\mathbf{y}}$ of equation (7) of Kole and Van Dijk (2023).

In addition to improving the interpretation of $\delta_{i,j}(h)$, as this GFEVD by construction sums to 1, this could ameliorate the concerns about overstating connectedness that are described by Wiesen, Beaumont, Norrbin and Srivastava (2018). Specifically, although the GFEVD within the meaning of Pesaran and Shin (1998) is able to identify the structural error terms, in determining the contribution of the shocks in variable j to the forecast error variance of variable i it does not take account that these shocks are in general correlated with those of other variables. If these are positively correlated, then they are overcounted. In a comparison of the two specifications of the GFEVD, Chan-Lau (2017) found that the contributions of individual financial institutions differed markedly, although this was not the case for results aggregated over individual institutions. Moreover, he judged the corrected GFEVD to more adequately identify the riskiest financial institutions as measured by the SIs based on stock returns.

The MSIRF can be used in the construction of the GFEVD for a higher level of aggregation. For example, individual variables can be part of sectors, indices, asset classes, markets, countries or regions. Under the assumption that a shock occurs at this level, a GFEVD can be constructed that takes this into account. To that end, partition the variables $i = 1,\ \ldots,\ k$ into $\mathcal{M}_1,\ \ldots \mathcal{M}_r$, where r is the number of units at the considered level of aggregation, i.e. $r \leqslant k$. Then, I define

[15] The MSIRF can also be derived for the MS-VAR. However, as it is not required for this thesis, this is not pursued.

the GFEVD based on the MSIRF, or GFEVD$_{\mathcal{M}}$, as follows:

$$\delta_{\mathcal{M}_i,\mathcal{M}_j}(h) = \frac{\sum_{l=0}^{h}(\Psi_{\mathcal{M}_j})'_{\mathcal{M}_i}(\Psi_{\mathcal{M}_j})_{\mathcal{M}_i}}{\sum_{q=1}^{r}\sum_{l=0}^{h}(\Psi_{\mathcal{M}_q})'_{\mathcal{M}_i}(\Psi_{\mathcal{M}_q})_{\mathcal{M}_i}}, \quad j = 1, \ \ldots, \ r \tag{3.6}$$

where $(\Psi_{\mathcal{M}_j})_{\mathcal{M}_i}$ is a vector consisting of the elements of $\Psi_{\mathcal{M}_j}$ corresponding to the variables in $\mathcal{M}_i$. The GFEVD$_{\mathcal{M}}$ uses the property of the MSIRF that it takes the correlation of shocks in the variables of $\mathcal{M}_j$ into account, which is neglected when using $(\Psi_{\mathcal{M}})_i = \sum_{q\in\mathcal{M}}(\Psi_q)_i$ (Van der Zwan, 2023). The GFEVD$_{\mathcal{M}}$ can be considered the MSIRF-analogue to the GFEVD of Lanne and Nyberg (2016) and similarly, $\sum_{j=1}^{r}\delta_{\mathcal{M}_i,\mathcal{M}_j}(h) = 1$.

3.2 Total and Directional Spillovers

The GFEVD is the main building block of the SI, defined by Diebold and Yilmaz (2009, 2012).

$$S(h) = \frac{\sum_{\substack{i,j=1\\ i\neq j}}^{k}\delta_{i,j}(h)}{\sum_{i,j=1}^{k}\delta_{i,j}(h)} \cdot 100 \tag{3.7}$$

$$S(h)_{i\leftarrow\bullet} = \frac{\sum_{\substack{j=1\\ i\neq j}}^{k}\delta_{i,j}(h)}{\sum_{j=1}^{k}\delta_{i,j}(h)} \cdot 100, \ i = 1, \ \ldots, \ k \tag{3.8}$$

$$S(h)_{j\rightarrow\bullet} = \frac{\sum_{\substack{i=1\\ i\neq j}}^{k}\delta_{i,j}(h)}{\sum_{i=1}^{k}\delta_{i,j}(h)} \cdot 100, \ j = 1, \ \ldots, \ k \tag{3.9}$$

Equation (3.7) denotes the total spillover, or the SI. Equations (3.8)-(3.9) denote the directional spillovers, the spillover from all other variables to a variable i and the spillover from a variable j to all other variables.[16] The information can be summarised in matrix form:

$$\Delta(h) = \begin{pmatrix} \delta_{1,1}(h) & \delta_{1,2}(h) & \ldots & \delta_{1,k}(h) \\ \delta_{2,1}(h) & \delta_{2,2}(h) & \ldots & \delta_{2,k}(h) \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{k,1}(h) & \delta_{k,2}(h) & \ldots & \delta_{k,k}(h) \end{pmatrix} \tag{3.10}$$

It can be seen that $S(h)_{i\leftarrow\bullet}$ and $S(h)_{j\rightarrow\bullet}$ correspond to scaled off-diagonal row and column sums of $\Delta(h)$ respectively. Mutatis mutandis, the same definitions can be applied to the GFEVD$_{\mathcal{M}}$, thereby obtaining an aggregate shock SI, or SI$_{\mathcal{M}}$.

From the construction of $S(h)$ it is apparent that dynamics in the SI for the MS-VAR are introduced by means of the time-variation in Ψ which stems from the difference in the parameters over the regimes and the inference about the prevailing regimes and the process that generates it. On the contrary, for a given sample, $S(h)$ is constant if constructed using a VAR. As mentioned above, time-variation in this case is induced by means of a rolling window. Following Demirer et al. (2018), I set the window size to 150 days. A sensitivity analysis for the rolling window size is performed and is included in Appendix G. In general, it can be noted that this size implies a

[16] This definition deviates from those of Diebold and Yilmaz (2012) who use $\sum_{i,j=1}^{k}\delta_{i,j}(h)$ in the denominator. My definition expresses the directional spillovers as a total of its corresponding row- or column sum, which is more interpretable. Moreover, as discussed in Section 3.4, it has the interpretation of a node degree.

trade-off between estimation uncertainty in the model parameters and oversmoothing of $S(h)$. In the high-dimensional case especially, one should proceed with caution as to not set the size of the window too small. For each rolling window, I will use the value of α that was obtained for the entire sample based to limit the computational burden. For λ, I will perform a new grid search in which the value of λ that is obtained for the full sample is the lowest one. Of this new grid, four lower values, equidistant on the grid will be considered in addition. The reason for this is that the rolling window size is relatively small. Accordingly, because of estimation uncertainty, it is wise to consider higher values of λ. Of this new grid, the highest value of λ is ubiquitously selected, supporting this idea.

Following Demirer et al. (2018), I set h to 10. The value of h affects $S(h)$ due to the possibility that shocks in one variable affect another variable with a lag and of this lag not being equal for different variables in the system (Diebold & Yilmaz, 2015a). Hence, as a sensitivity analysis, I also set $h = 7$ and $h = 12$. The differences obtained were negligible and not visually discernible. As it by now is explained that the SI is dependent on h, it will be dropped from the notation of the SI henceforth.

3.3 Bootstrapped Confidence Intervals

Values of S for the VAR will in general differ from those of the MS-VAR. Greenwood-Nimmo and Tarassow (2022) note that, although asymptotic distributions of impulse response functions have been derived for the VAR by Lütkepohl (1990), the rolling window necessarily is of a limited length. Moreover, the asymptotic distributions hold for impulse response functions of an orthogonal(ised) VAR and these do not necessarily correspond to those for GIRFs. To qualify differences in S the bootstrap-based procedure of Choi and Shin (2020) can be used, which can be summarised as follows.[17]

1. Given a window size w, estimate the model for each rolling sample $r = 1, \ldots, R$. Store $\hat{\boldsymbol{\beta}}^{(r)}$ and $\hat{\mathbf{u}}^{(r)}$ where $\hat{\mathbf{u}}$ are the w vectors of estimated residuals.

2. For each r generate $b = 1, \ldots, B$ bootstrap samples of size w by using the initial values of rolling sample r, the resampled residuals $\mathbf{u}^{*(r,b)}$ and $\hat{\boldsymbol{\beta}}^{(r)}$.

3. For each b, obtain estimates $\boldsymbol{\beta}^{*(r,b)}$ and $\Sigma^{*(r,b)}$. Calculate and store $S^{*(r,b)}$, $i,\ j = 1, \ldots, k$.

4. For each r, estimate
$$se^*(S^{*(r)}) = \left(\frac{1}{w}\sum_{b=1}^{B}(S^{*(r,b)} - \bar{S}^{*(r)})^2\right)^{\frac{1}{2}}$$
where $\bar{S}^{*(r)}$ is the sample mean of the SI of rolling window r over all bootstrap samples.

5. For each r, construct a $(1-a)\%$ confidence interval for $\hat{S}^{(r)}$ as $\hat{S}^{(r)} \pm z_\alpha se^*(S^{*(r)})$, where

[17] An alternative for this procedure could be the use of Bayesian methods, where the SIs can be based on samples from the posterior distribution of parameters. This idea has been applied by Shapovalova and Eichler (2023), who use a particle Markov Chain Monte Carlo method to estimate a multivariate stochastic volatility model in which the logarithm of volatilites are modelled using a VAR and obtain highest posterior density intervals. Their methods however, are limited to the low-dimensional case, both due to the computational aspects, as well as due to the lack of shrinkage priors.

z_α is the critical value of a standard Gaussian distribution at the significance level a.

Choi and Shin (2020) show by means of simulations that the coverage for large values of T and under a correct specification of the lag order is satisfactory, with most coverage rates being around 92%. The bootstrap samples can be generated by means of a moving block bootstrap scheme used by Brüggemann, Jentsch and Trenkler (2016), which is as follows:

1. Fit a VAR(p) and obtain $\hat{c},\ \hat{\Phi}_1,\ \ldots,\ \hat{\Phi}_p$. Use the estimated parameters to obtain the residuals $\hat{u}_t = y_t - \hat{c} - \sum_{l=1}^{p} \hat{\Phi}_l y_{t-l},\ t = p+1,\ \ldots,\ T$.

2. Set a block length $l < T - p$ and let $N = \lceil \frac{T-p}{l} \rceil$. Define $(k \times l)$-dimensional blocks $B_i = (\hat{u}_i,\ \ldots,\ \hat{u}_{i+l}),\ i = p+1,\ \ldots,\ T-l$. Sample $i_j,\ j = 0,\ \ldots,\ N-1$ from a uniform distribution with support set $\{p+1,\ \ldots,\ T-l\}$. Lay blocks $B_{i_0},\ \ldots,\ B_{i_{N-1}}$ together and discard the final $Nl - (T-p)$ columns of the concatenated matrix.

3. The columns of the obtained matrix are $\hat{u}^*_{p+1},\ \ldots,\ \hat{u}^*_T$. Centre these according to the rule

$$u^*_{jl+s} = \hat{u}^*_{jl+s} - \frac{1}{T-p-l+1} \sum_{r=0}^{T-p-l} \hat{u}_{s+r}$$

   for $s = p+1,\ \ldots,\ p+l$ and $j = 0,\ \ldots,\ N-1$.

4. Set $y^*_1,\ \ldots,\ y^*_p = 0$. Generate $y^*_t = \hat{c} + \sum_{l=1}^{p} \hat{\Phi}_l y^*_{t-l},\ t = p+1,\ \ldots,\ T$.

For the block length, I apply the estimator of Politis and White (2004) to each individual series of a rolling window sample and set l to the median of these values. The bootstrap estimates of the parameters hereby obtained are consistent under error terms that are not independently and identically distributed, but serially uncorrelated and subject to α-mixing conditions, an example of which is conditional heteroskedasticity. Moreover, Furman (2014) has shown that the residual-based bootstrap leads to consistent estimates for the VAR with adaptive elastic net penalisation for independently and identically distributed error terms. Although Choi and Shin (2020) recommend the use of a residual bootstrap over a moving block bootstrap, the possible presence of regime-switching renders the above procedure preferable. Because of limited computational resources, B will be set to 100.

3.4 Spillover Networks

Diebold and Yilmaz (2014) introduced the notion that Δ defines a weighted, directed graph. Specifically, let $G = (V,\ E)$ be a graph, where V is the set of nodes and E is the set of edges. Then, Δ' is its adjacency matrix A. Two main aspects of economic and financial networks are the centrality of individual nodes and the communities that are formed by the nodes. First, I consider centrality. Chan-Lau (2018, pp. 473) describes centrality as capturing "too-connected-to-fail" risk. $S_{i\leftarrow\bullet}$ and $S_{j\rightarrow\bullet}$ respectively correspond to the in- and out-degree of the nodes, which for spillover networks are weighted. Another measure that readily lends itself to application of weighted, directed graphs is eigenvector centrality. The eigenvector centrality assumes that the centrality of each node is proportional to the centrality of each of its neighbours, weighted by the edges. (Bloch, Jackson & Tebaldi, 2023). The eigenvector centrality of node i is defined as

follows:

$$c_i = \frac{1}{\lambda_{\max}} \sum_{j=1}^{k} A_{i,j} c_j \tag{3.11}$$

where $\lambda_{\max}$ is the largest eigenvalue of A. The vector of node centralities can thus be determined as $A\mathbf{c} = \lambda_{\max}\mathbf{c}$, where $\mathbf{c}$ is normalised, i.e. $\|\mathbf{c}\| = 1$. As A is a stochastic matrix, c_i also has the interpretation of the unconditional probability of being in node i when considering the network as a Markov chain with A containing the transition probabilities.

Centrality measures will be used in the visualisation of the spillover networks and as input for an analysis of the regional pattern of spillovers. This analysis will consist of rank-regressions of centrality measures on a dummy variable containing the region of the bank. These will be performed for each period t. These regressions are of the following form:

$$R(r_i) = \beta_0 + \beta_1 \mathbb{1}(i \in EU) + \beta_2 \mathbb{1}(i \in AM) + \epsilon_i,\ i = 1,\ \ldots,\ k,\ r_i = c_i,\ S_{i\to\bullet} \tag{3.12}$$

where $R(r_i)$ is the centrality rank of bank i, EU is the set of European banks and AM is the set of American banks. In recent work, Chetverikov and Wilhelm (2023) have derived the consistency of the OLS estimator of $(\beta_0,\ \beta_1,\ \beta_2)$ and its asymptotic normality, as well as a consistent estimator of its asymptotic variance. These results are implemented in the R package `csranks`.

Next, I consider community detection. Community detection entails finding groupings in the nodes of the networks. This concept can be operationalised as finding a partition of the graph such that the obtained groupings contain nodes that are most similar to each other as measured by the elements of A (Schaub, Delvenne, Rosvall & Lambiotte, 2017). Using the obtained communities, banks that are "too-important-to-fail" can be identified (Chan-Lau, 2018, pp. 473). Demirer et al. (2018) apply the ForceAtlas2 algorithm due to Jacomy, Venturini, Heymann and Bastian (2014) to visualise the obtained network, which indicates the existence of country- and region-based clusters.

However, such a cluster assignment is solely based on visual imputation, which can be problematic. First, the output of the ForceAtlas2 algorithm is dependent on the initial positions provided. Secondly, the clusters might not be adequately demarcated. For example, it is not clear whether the banks in the middle of the network in Figure 2 of Demirer et al. (2018) should be considered as a single cluster. By means of community detection, such clusters can be made explicit. Specifically, I will use the node2vec-spectral clustering algorithm of Hu, Liu, Li and Liang (2020). The use of spectral clustering is convenient for node embeddings, as spectral clustering can be used to simultaneously reduce the dimension of the embeddings and perform clustering of the nodes. The algorithm consists of two components.

Node embeddings are obtained using the node2vec algorithm of Grover and Leskovec (2016), from which the following exposition is taken. $f : V \to \mathbb{R}^{\mathbf{d}}$ is a function that maps nodes to $\mathbf{d}$-dimensional vectors of real numbers. The objective is to obtain vectors such that the Euclidean distance between them is small if the corresponding nodes are close to each other in the network. To determine, for each node $u \in V$, what nodes it is close to, the network is sampled. These

samples provide the network neighbourhood $N(u) \subset V$. The objective function can thus be formulated as:

$$\underset{f}{\operatorname{argmax}} \sum_{u \in V} \log \mathbb{P}[N(u)|f(u)] \tag{3.13}$$

To make this problem tractable it is assumed that observing a neighbourhood node is independent of observing another neighbourhood node, conditional on $f(u)$. It then holds that $\mathbb{P}[N(u)|f(u)] = \prod_{n \in N(u)} \mathbb{P}[n|f(u)]$. Then, the probability to observe a node of this neighbourhood is modelled by means of the softmax function of the dot products of the nodes of the neighbourhood

$$\mathbb{P}[n|f(u)] = \frac{\exp\{f(n)'f(u)\}}{\sum_{v \in V} \exp\{f(v)'f(u)\}}$$

The objective function (3.13) can then be reformulated as

$$\underset{f}{\operatorname{argmax}} \sum_{u \in V} \left[-\log\left(\sum_{v \in V} \exp\{f(v)'f(u)\} \right) + \sum_{n \in N(u)} f(n)'f(u) \right] \tag{3.14}$$

which can be optimised in f, i.e. in the elements of the vectors in $\mathbb{R}^{\mathbf{d}}$, which is done using stochastic gradient descent.

The embeddings are dependent on the samples. These are generated for each node by means of second-order random walks. The unnormalised probability of transitioning from node $v_{i-1} = x$ to node $v_i = y$, in which $v_{i-2} = w$ is given by $p_{y,x} = A_{x,y}[p^{-1}\mathbb{1}(y = w) + \mathbb{1}(y \neq w)]$. The parameter p, through which the random walk becomes of the second order, can be set to disincentivise the random walk revisiting w, but will, following Hu et al. (2020), be set to 1.[18] The dimension $\mathbf{d}$ will be set to k, the number of walks will be set to 100 and the walk length will be set to 80. The algorithm is implemented in the `node2vec` package in Python.

The embeddings can be collected in a matrix $\mathcal{E}$ of dimension $|V| \times \mathbf{d}$. From $\mathcal{E}$, a similarity matrix $\mathbf{S}$ using the radial basis kernel is constructed as follows:

$$\mathbf{S}_{i,j} = \exp\left(-\frac{\sum_{q=1}^{\mathbf{d}} (\mathcal{E}_{i,q} - \mathcal{E}_{j,q})^2}{2\sigma^2} \right)$$

in which σ^2 governs the degree to which the similarity decreases as the squared Euclidean distance between two embeddings increases. The use of the radial basis kernel is appropriate as the differences between the embeddings are on the same scale across the dimensions of the embeddings. Then, construct the diagonal matrix D for which $D_{i,i} = \sum_{j=1}^{k} \mathbf{S}_{i,j}$. The Laplacian matrix is then defined as $L = D - \mathbf{S}$ and is normalised as $\tilde{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$. Normalisation is performed as this entails finding clusters that both maximise within-cluster similarity and minimise between-cluster similarity (Von Luxburg, 2007). Clustering is then performed using the K-means algorithm on the n normalised eigenvectors of $\tilde{L}$ that correspond to the lowest eigenvalues, $n << \mathbf{d}$. Clustering will be performed for a grid of values for σ^2. To choose the

[18] The node2vec algorithm also contains a parameter q by which the unnormalised transition probability will be divided if the length of the shortest path between y and w is two, which is not applicable to spillover networks as the corresponding graph is complete.

value of σ^2, I will use the weighted cut (WC), which corresponds to the objective function of spectral clustering. The WC is generalised by Meilă and Pentney (2007) to weighted, directed graphs and is defined as follows:

$$\text{WC}(\mathcal{C}) = \sum_{c} \sum_{c' \neq c} \frac{\sum_{i \in C_c} \sum_{j \in C_{c'}} A_{i,j}}{\sum_{i \in C_c} \sum_{j \in V} A_{i,j}}$$

Put differently, a set of clusters $\mathcal{C}$ should be obtained that minimises over the clusters C_c the sum of outgoing weights to $\mathcal{C} \backslash C_c = C_{c'}$ as a fraction of total outgoing weights. To compare the clusters of the VAR with those of the MS-VAR, the modularity, generalised to weighted and directed matrices will be used to gauge the degree to which clusters are formed across regimes. Higher values of the modularity are obtained if the total weight of edges in a community is larger than the number of edges that would have been obtained under a random graph in which the weight of the edge between nodes i and node j is equal to the average of their respective in- and out-degree, which would entail the absence of communities. Moreover, the modularity can be determined for each period, indicating how the degree to which communities form develops over time. The following definition of the modularity is taken from Molnár, Márton, Horvát and Ercsey-Ravasz (2024):

$$Q = \frac{1}{\mathbf{A}} \sum_{i,j \in V} \left[A_{i,j} - \frac{S_{i \to \bullet} S_{\bullet \leftarrow j}}{\mathbf{A}} \right] \mathbb{1}(c_i = c_j) \tag{3.15}$$

where $\mathbf{A} = \sum_{i,j=1}^{k} A_{i,j}$ and the indicator function is equal to 1 if nodes i and j share their cluster membership.

Finally, the use of graph embeddings will be considered. Graph embeddings generalise the notion of node embeddings discussed above to graphs. Specifically, the graph2vec algorithm of Narayanan et al. (2017) will be used, on which the following exposition is based. Let $\mathcal{G} = \{G_1, \ldots G_{T_{\max}}\}$ be the graphs that correspond to the spillover networks. If the MS-VAR is used, $T_{\max} = T - p$ and for the rolling window VAR, $T_{\max} = T - p - w$, where w is the rolling window size. For each node v_i of graph G_t, the algorithm obtains subgraphs of degree d rooted at this node, $sg_{v_i}^{(d)}$. Similar to how node2vec obtains embeddings through maximising the similarity of two nodes in the embedding space if they are frequently observed in the same neighbourhood, graph2vec aims to construct the matrix of embeddings $\mathbb{G}$ to maximise the probability of observing $sg_{v_i}^{(d)}$, conditional on $\mathbb{G}$:

$$-\log \sum_{G \in \mathcal{G}} \sum_{v \in V} \sum_{d=0}^{D} \mathbb{P}[sg_v^{(d)} | \mathbb{G}] \tag{3.16}$$

As in node2vec, the problem of maximizing this probability is made tractable by the asssumption of conditional independence, implicit in equation (3.16) being a summation and by assuming symmetry in the embedding space. As a consequence of the latter, similarity can be expressed by means of the inner product, as in the node2vec algorithm. Rather than exhaustively optimising over the set of all possible subgraphs of $\mathcal{G}$, SG, a negative sampling strategy is employed for every G_i. Let $\mathbf{C}$ be the set of rooted subgraphs of G_i. This strategy then entails that the embeddings will be updated using a sample $\mathbf{C}' \subset SG$, $|\mathbf{C}'| = n$, $n << |SG|$, $\mathbf{C}' \cap \mathbf{C} = \varnothing$. Therefore, if G_j contains subgraphs that are very similar to those of G_i, the embedding of G_i

becomes more similar to that of G_j by virtue of how the objective function is defined. As with the node embeddings, spectral clustering can be applied to obtain clusters of the networks, which are indicative of sub-periods in which the network structure is (relatively) similar. The implementation of the algorithm in the `karateclub` Python library has been applied with its default parameters.

3.5 Systemic Event Prediction

Finally, the link between bank spillover networks and systemic risk will be explored. Diebold and Yilmaz (2014) describe how the row- and column-sums of Δ for a set of financial institutions are closely related to, respectively, the marginal expected shortfall and the system-wide value at risk conditional on distress of individual institutions, which are oft-used risk measures. This motivates the use of the Δ and the SI in the context of systemic risk. Korobilis and Yilmaz (2018) use the lagged SI obtained using bank stock return volatilities as an independent variable in a logistic regression and obtain McFadden R^2 values of 0.2-0.4 in one-day ahead predictions of the occurrence of an SE, which is the case when at day t the daily stock returns of more than 25% of banks are lower than the 5^{th} percentile of their respective empirical distributions. This is used to construct the SE index, which equals 1 for periods in which an SE occurs and 0 otherwise. The SE index for this sample is displayed in Figure 2. The SE index is equal to 1 for 137 periods, mostly in the period of 2008-2012; the great financial crisis, the great recession and the Euro-crisis. The following model is based on that of Korobilis and Yilmaz (2018):

$$\mathbb{P}[SE_t = 1] = \frac{\exp\Big\{\beta_0 + \beta_1 SE_{t-l} + \beta_2 SI_{t-l} + \beta_3 (SI_{t-l} - \hat{SI}_{t-l})\Big\}}{1 + \exp\Big\{\beta_0 + \beta_1 SE_{t-l} + \beta_2 SI_{t-l} + \beta_3 (SI_{t-l} - \hat{SI}_{t-l})\Big\}} \tag{3.17}$$

for a certain lag order l. For the value of the lag order, I consider $l \in \{1,\ 2,\ 3,\ 4,\ 5,\ 10,\ 22\}$, which respectively correspond to one to four trading days, one to two trading weeks and one trading month in advance. $SI_{t-l} - \hat{SI}_{t-l}$ is an addition to the model of Korobilis and Yilmaz (2018) and denotes the deviation of the SI from its trend value. The trend is estimated using a linear spline. This is motivated by the idea that disrupting events could lead the SI to be temporarily higher and that it is under these conditions that SEs become more probable. The deviation from the trend level captures this.

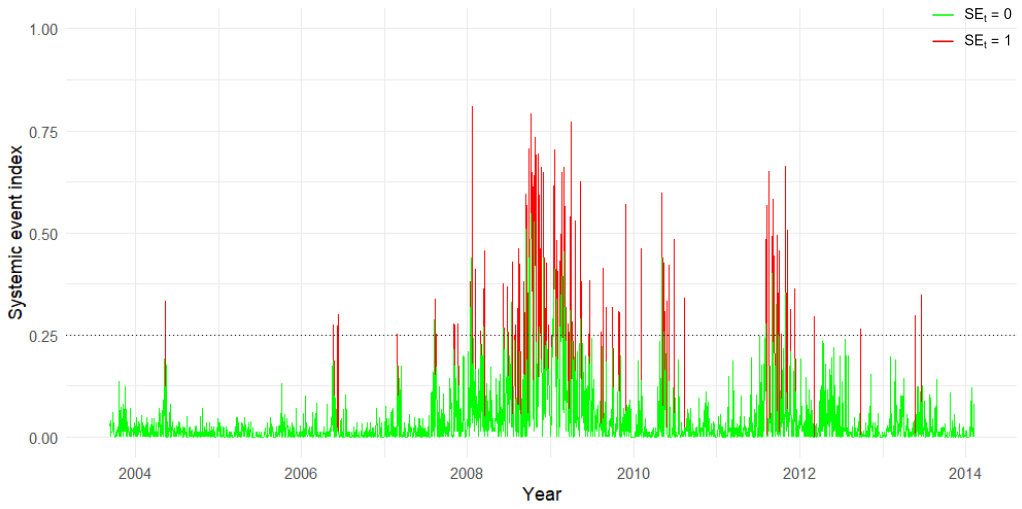


Figure 2: The systemic event index at a threshold of 0.25.

The SI effectively is a compressed form of the information contained in the spillover network. By means of graph embeddings, this information can be used directly. Although the embeddings are an approximation of the network, it could be the case that less information is lost than for compression to the SI. Thus the logistic regression can then be estimated with the graph embeddings as features, i.e. through the substitution of $\beta'(\mathbb{G})_{t-l}$ for $\beta_2 SI_{t-l}$.

I will compare the models with respect to the McFadden R^2. This will mainly serve to determine whether the graph embeddings carry more information than the SI. Secondly, the estimated probabilities will be used to predict whether observations are SEs. Although classification can be considered of interest per se, it can also indicate whether the graph embeddings, due to their high dimensionality, are not liable to overfit. To this end, the logistic regression will be estimated for a training set consisting of 70% of the observations, randomly chosen. The training set will be used to estimate the logistic regression for both models, whereas the test set, consisting of the remaining 30% of observations, will serve to determine a threshold. If $\hat{p}_t$ is below (above) the threshold, $\hat{SE}_t = 0$ $(= 1)$. To account for possible overfitting, the estimation of the logistic regression on the graph embeddings will also performed subject to elastic net penalisation. To obtain the values of the penalty parameters, 10-fold cross validation will be performed. A complication with classification is that the SE index is quite unbalanced, being equal to 1 for approximately 5.4% of the observations. Therefore, I will use the F-score to evaluate the class assignments. The F-score is the harmonic mean of the precision, the ratio of true positives (TP) to the sum of true positives and false positives (FP), and the recall, the proportion of true positives to positives. Formulated, in terms of the elements of the confusion matrix, the F-score is formulated as follows:

$$\mathrm{F} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

where FN is the number of false negatives. The value of the F-score will be determined by means of the test set. As another means to deal with class imbalance, I will allocate higher weights to the likelihood contributions of observations that are a systemic event, thereby attributing higher importance to these observations in estimation. This will induce parameter estimates

that lead to higher estimated probabilities for observations of an SE, which could lead to better class assignments. Weights of 1, 2, 4 and 8 will be applied.

4 Bank Stock Return Volatility Data

I use the dataset of Demirer et al. (2018), which consists of the volatility of the stock returns of 96 banks. Hereinafter, I refer to these authors as DDLY. Thus, $k = 96$. DDLY collected stock return data from Thomson-Reuters from September 12$^{\text{th}}$, 2003 to February 7$^{\text{th}}$, 2014. Thus, $T = 2,676$. The banks are the largest in the world, as measured by total assets, with publicly traded stocks. All of the stocks have been publicly traded throughout the entire sample. Specifically, daily opening, closing, high and low prices have been used for the following range-based volatility estimate of Garman and Klass (1980):

$$\begin{aligned}\hat{v}_{k,t}^2 = {} & 0.511(H_{k,t} - L_{k,t})^2 - 0.019[(C_{k,t} - O_{k,t})(H_{k,t} + L_{k,t} - 2O_{k,t}) \\ & - 2(H_{k,t} - O_{k,t})(L_{k,t} - O_{k,t})] - 0.383(C_{k,t} - O_{k,t})^2\end{aligned} \tag{4.1}$$

where $O_{k,t}$, $C_{k,t}$, $H_{k,t}$ and $L_{k,t}$ are the natural logarithms of the opening, closing, high and low prices of the stock of bank k at day t. DDLY argue in favour of this range-based estimate based on the results of Alizadeh, Brandt and Diebold (2002), who find that such estimates are efficient in the context of stochastic volatility models. Molnár (2012) inquires into range-based volatility estimates for estimating daily volatility and finds that the Garman-Klass estimator performs best. The volatility data can be found at `http://qed.econ.queensu.ca/jae/datasets/demirer001/`. An overview of the country of each bank, its market capitalisation, its total assets, its bank code and its Reuters ticker can be found on the same web page. The dataset of DDLY does not contain the underlying daily returns which are used in the construction of the SE index. These have been downloaded separately from Yahoo Finance for 86 banks and from `Investing.com` for five banks. For the remaining five banks, the daily returns were unavailable or available in limited quantity. Details on the collection of the stock return data and the missing banks are included in Appendix E.

Transforming the volatility series by taking natural logarithms yields approximate normality based on histograms and quantile-quantile plots, although some right-skewness occurs frequently. These series are used for both the VAR and the MS-VAR. The conditional heteroskedasticity of the underlying stock returns is apparent. Next to that, it holds that the partial autocorrelations are significant at lower-order lags. Moreover, the autocorrelation function decays slowly, ostensibly at a sub-exponential rate, indicating possible long memory. Diebold and Inoue (2001) show that Markov-switching models (without autoregressive terms), even though they are integrated of order zero, can generate data that even in large samples are difficult to distinguish from fractionally integrated data. This holds especially if $p_{i,i}$ is close to 1 for all i, which is often the case in empirical applications of Markov-switching models (Guidolin, 2011).

The null-hypotheses of augmented Dickey-Fuller tests with generalised least squares detrending (ADF-GLS) for a unit root are rejected for 75 of these series at a significance level of 5% and for 47 of these at a significance level of 1%. For 21 of the series, the null-hypothesis is not rejected.

More details on these tests are included in Appendix F. The unit root tests were performed with the alternative hypothesis of a model with a constant term and no deterministic trend.[19] The unit root tests not rejecting the null-hypothesis for some series does not violate the assumption of stability of the VAR. For the estimated parameter matrices, which are discussed in Section 5.1, the spectral radius is less than 1. For the MS-VAR, the spectral radius of the relevant matrix also is less than 1.

Figure 3, the colour-scale of which is centred at the median of pairwise sample correlations, indicates that the elastic net might be preferable over the LASSO. An interesting observation follows from the fact that the banks are ordered by total assets. Namely, the strongest correlation seems to be between the volatilities of the largest banks. Therefore, if the LASSO haphazardly chooses between two banks because of multicollinearity, it is likely to do so for two important banks, the coefficients of which are likely to be important in the (MS-)VAR. Note also that the volatilities of the largest banks seem to most correlated with those of other banks in general. Contrary to many applications of the elastic net, the data will not be standardised. This is because the volatilities are all defined on the same scale, meaning that the absence of scale invariance of the LASSO-penalty will not be to the detriment of the parameter estimates.

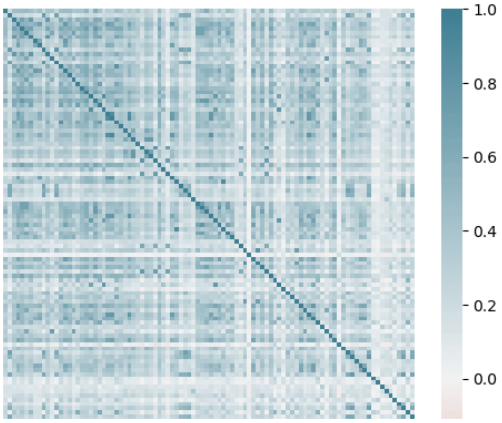


Figure 3: Visualisation of the sample correlation matrix.

5 Results

5.1 Estimation Results

First, I discuss the parameter estimates of the VAR. The information criteria for the different specifications are reported in Table 1. Whereas the BIC has the lowest value for the VAR(1), the AICc is lower for the other models. An inquiry into the EDF reveals that for lag orders 2 to 4, the LASSO-penalty imposes a very sparse $\mathbf{\Phi}$, such the model reduces to a first-order autoregressive model, with a few equations containing some distributed lags in addition. Therefore, the AICc can hardly be considered a estimator of the lag-order and the selection between the VAR(1) and the 'VAR(4)' effectively becomes a matter of selecting the penalty parameters, for which the BIC

[19] Note that the absence of a deterministic trend is without prejudice to the results of the ADF-GLS tests. On the contrary, this entails a less accurate model under the alternative hypothesis and although this is to the detriment of the power of the test, it nevertheless rejects the null-hypotheses for most of the series

is applied. Moreover, for lower values of λ the log-likelihood decreases markedly. For the VAR(1), however, this is not the case, and the log-likelihood is much flatter for different values of λ, which is positive from the perspective of model stability. Furthermore, the autoregressive structure selected is much richer, which is of interest as our area of application is the SI. Therefore, I opt for the VAR(1). The difference between the VAR(1) and the VAR(2)-VAR(4) could perhaps be explained by the adaptive elastic net weights. Namely, for the VAR(1) the Ridge-penalised VAR favoured very mild penalisation, whereas for the higher order VARs the strictest penalisation was favoured. As a robustness check, the adaptive elastic net weights of the VAR(1) were used for the VAR(2)-VAR(4), but these did not yield an improvement of the BIC and still led to the aforementioned sparsity in the autoregressive coefficients.

Table 1: Model Selection Criteria for the VAR

Model	AICc	BIC	ℓ	K	α
VAR(1)	588,293	557,671	-269,310	2,414	1.00
VAR(2)	564,539	569,738	-279,669	1,318	0.50
VAR(3)	564,268	569,463	-279,579	1,306	0.50
VAR(4)	563,984	569,174	-279,471	1,297	0.50

Notes: ℓ is the log-likelihood, K is the EDF and α is the selected elastic net parameter. K is rounded to the nearest integer.

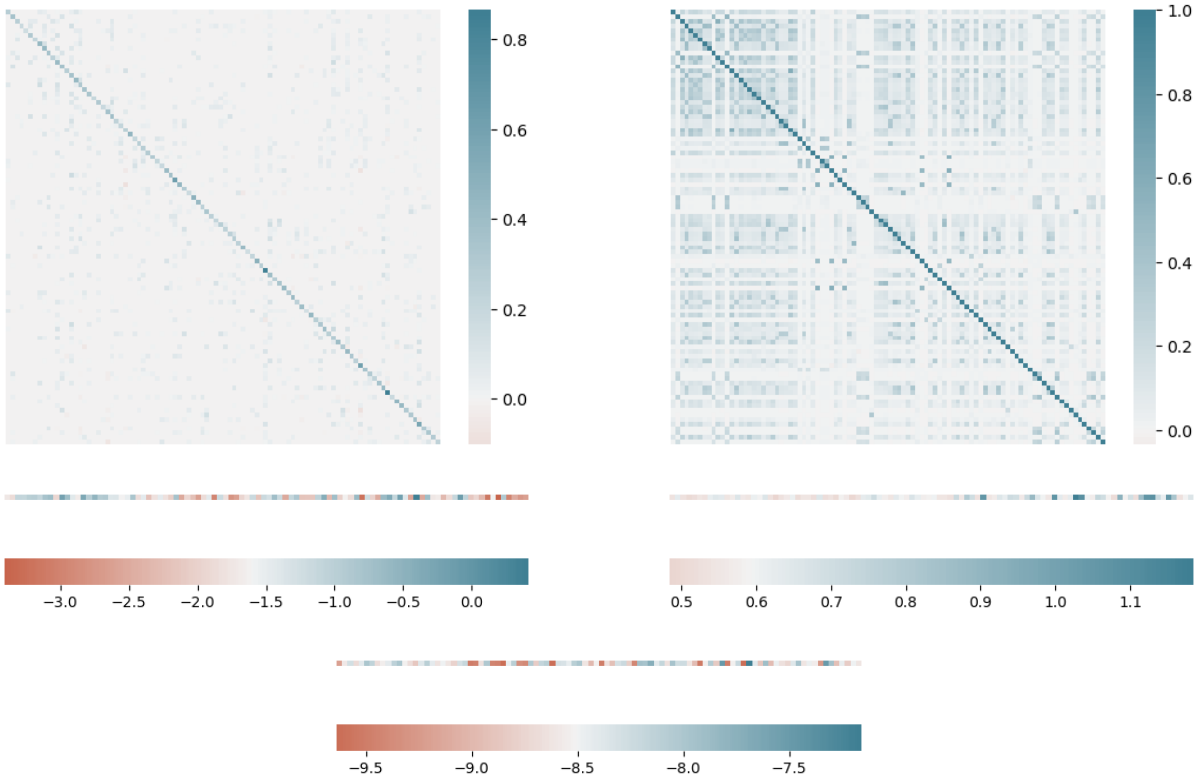


Figure 4: Estimated autoregressive parameters Φ (top-left), error term correlation matrix $\mathbf{R}$ (top-right), constant terms (bottom-left), error term variances (bottom-right) and unconditional mean (bottom) for the VAR(1).

The estimated parameters are visualised in Figure 4. Of the $96^2 = 9,612$ possible autoregressive parameters, 1,404 are selected into the model. The distributed lags selected mostly correspond to banks that are of a similar size, as elements closer to the diagonal of the autoregressive parameter matrix are more often selected. The error term correlation matrix, which is displayed as it is more insightful than the covariance matrix, is quite similar to the sample correlation matrix of Figure 3. Of the $96 \cdot 95/2 = 4,560$ pairwise correlations, 1,010 exceed 0.1 in absolute value. The estimated error term variances tend to be higher for the smaller banks. For the unconditional means, this is not the case. The estimated residuals for each individual bank are used to test for normality. The null-hypothesis of a Jarque-Bera test is not rejected at a level of 5% for only 12 of the 96 banks. Nevertheless, histograms of the residuals show that for the vast majority of banks, the empirical distributions of the residuals can be approximated by normal distributions quite well.

Next are the results for the MS-VAR. The information criteria for the different specifications of the MS-VAR are reported in Table 2. Here, based on the MSC, the preference for a lag order of 1 is more pronounced. As in BenSaïda et al. (2018), higher values of M yield better models. Although the MS(4)-VAR(1) has an EDF of 18,053, approximately thrice that of the MS(2)-VAR(1) (6,173) and almost twice that of the MS(3)-VAR(1) (10,839), the log-likelihood increased markedly, from -260,977 and -256,239 for the MS(2)- and MS(3)-VAR(1) respectively, to -248,165 for the MS(4)-VAR(1). For comparison, the MSC of the VAR(1) is equal to 591,181, which forms evidence for the presence of regime-switching. Moreover, Smith et al. (2006) show that the MSC is not liable to spuriously select regime-switching models. The BIC values for the MS-VAR(1) models on the other hand, are higher, at 570,669, 598,016 and 638,798, for the MS(2)-, MS(3)- and MS(4)-VAR(1) respectively, in line with the results of Psaradakis and Spagnolo (2003) in the sense that the BIC is liable to select a value of M that is too low.

Table 2: MSC values for the MS-VAR

	$p = 1$	$p = 2$	$p = 3$	$p = 4$
$M = 2$	518,130	536,354	536,158	535,971
$M = 3$	509,470	534,734	534,568	535,583
$M = 4$	493,504	534,817	534,728	534,200

The parameter estimates for the MS(4)-VAR(1) are included in Figure 5 and the inferred regimes as per the highest smoothed probability are included in Figure 6. The first regime resembles the VAR(1) most with respect to the autoregressive parameters, error term correlation matrix and error term variances as it is the regime that is estimated to prevail most frequently. The number of selected autoregressive parameters is 1,706. The unconditional mean of this regime is highest, entailing that this regime can be considered a high-volatility regime for bank stock returns.[20] It also has the highest error term variances. The first regime predominantly prevails later in the sample, after the onset of the great financial crisis and is marked by the richest correlation

[20] The unconditional means were determined in accordance with the results of Kole and Van Dijk (2023), instead of calculating $(\mathbf{I} - \Phi_m)^{-1} c_m$ for each regime m.

structure of error terms, with 1,341 pairwise correlations being larger than 0.1 in absolute value. As the VAR most resembles this regime, it can thus be stated the years 2008-2012 are the most informative period of the sample in the sense that it most strongly affects full-sample parameter estimates.

The second regime predominantly occurs earlier in the sample, before the onset of the crisis, with similar parameters to the first regime, yet seemingly 'calmer'. 2,517 autoregressive parameters are selected in this regime. The variances are similar to that of the first regime and the correlation structure is similar, though sparser, with 514 pairwise correlations being larger than 0.1 in absolute value. The unconditional means also are somewhat lower than those of the first regime.

The third and fourth regimes are marked by much richer autoregressive structures, in which the own lagged volatility moreover is relatively less important than in the first two regimes. 4,413 and 3,815 parameters are respectively selected into the model. This entails a higher degree of interdependence among bank stock return volatilities in these regimes. The correlation structure of the error terms is also different in these regimes, which seems to be more dispersed, rather than concentrated around the large banks. The number of pairwise correlations exceeding 0.1 in absolute value is similar to that of the to that of the VAR, amounting to 1,079 and 1,099 for the third and fourth regime respectively. One reason for these results for the third and fourth regimes could be that the same degree of regularisation is applied across regimes, meaning that, owing to the fewer observations in which regimes three and four prevail, the applied penalisation, in relative terms, is less severe for these regimes. However, robustness checks for two higher values of λ for the third and fourth regimes do not decrease the MSC. The third and fourth regimes also have lower unconditional means and error term variances than the first two regimes and can hence be considered low-volatility regimes.[21]

$\hat{\mathbf{P}}$ and $\hat{\pi}$, which are included below as equation (5.1), show that the first regime is highly persistent. Together with Figure 6, which shows that the first regime occurs for long periods after 2008, accounts for the high estimated unconditional probability of this regime. The third and fourth regime being more likely to switch to the second regime and not to the first, is likely the consequence of the earlier years of the sample, in which the process mainly switched between the second, third and fourth regimes. Only in the years of 2006 and 2007 does the process switch between all regimes.

The regimes are identified extremely well, with just one period in which the maximum smoothed probability does not exceed 0.99. This also is a consequence of the high-dimensionality of the data; if regime switching indeed is part of the data-generating process, it becomes exponentially more manifest in the relative likelihoods of the regimes if the dimension increases. Moreover, the most likely a posteriori sequence of regimes estimated by the Viterbi algorithm, as described by Franke (2012) for Markov-switching autoregressive models, is equal to the sequence of regimes for which the smoothed probability is highest.

[21] This does not preclude the possibility of the unconditional variances being higher in these regimes. Determining the unconditional variances for this application involves solving a system of 36.864 equations in 36.864 variables, which proved to be too memory-intensive, even when applying sparse matrix objects and solvers.

As there are four regimes, there are 384 individual series of regimes for which a Jarque-Bera test is employed by assigning every period to the regime for which is smoothed probability was highest. For 193 of these series, the null-hypothesis of the test is not rejected at a significance level of 5%, indicating that within-regime normality is a more appropriate assumption than overall normality. Again, histograms show for the majority of banks that the empirical distributions of the residuals can be approximated quite well by normal distributions, although this is the case to a somewhat lesser extent for the third and fourth regimes.

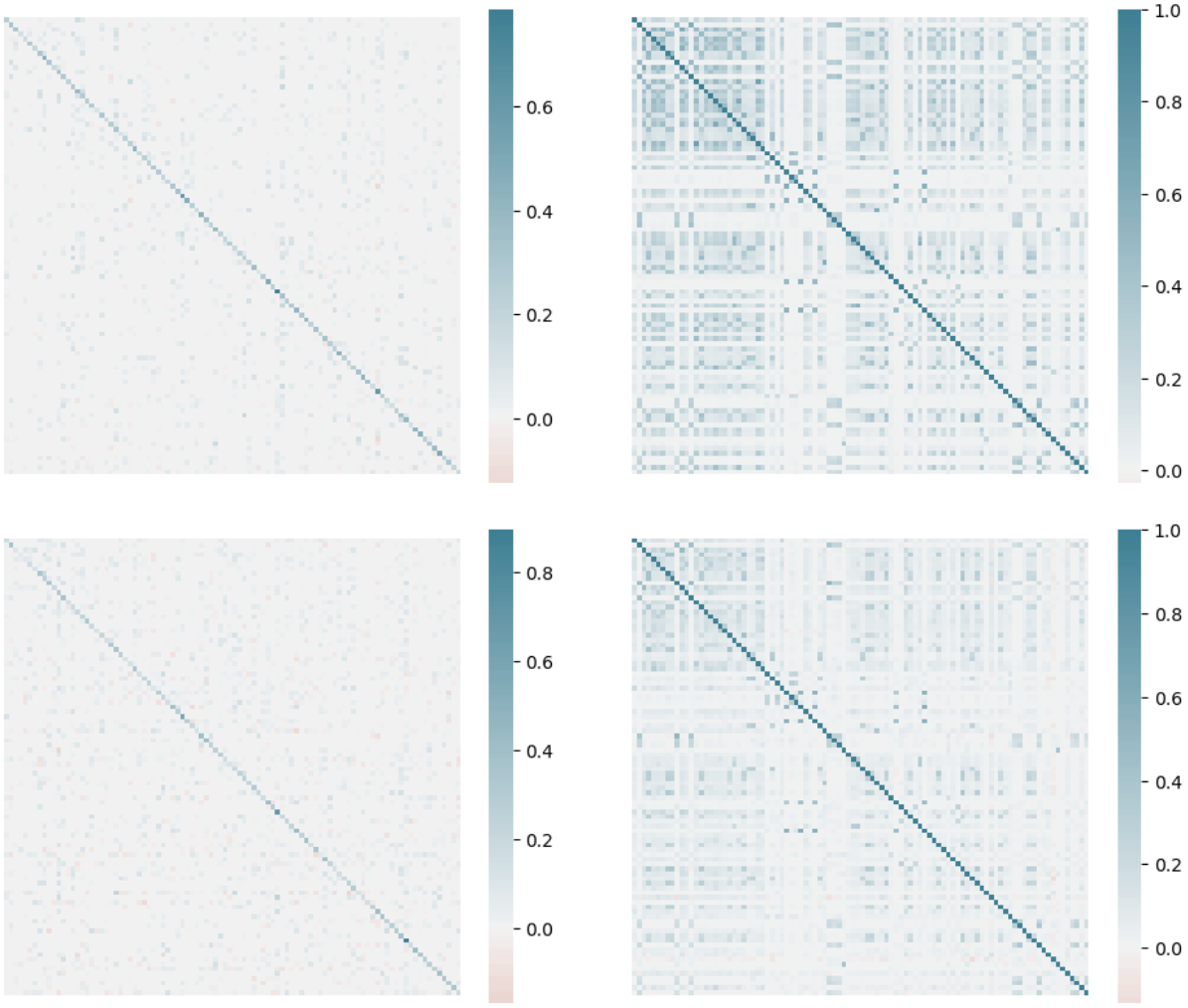


Figure 5: (1/2). Estimated autoregressive parameters (Φ_1 top-left and Φ_2 bottom-left) and error term correlation matrices ($\mathbf{R}_1$ top-right and $\mathbf{R}_2$ bottom-right).

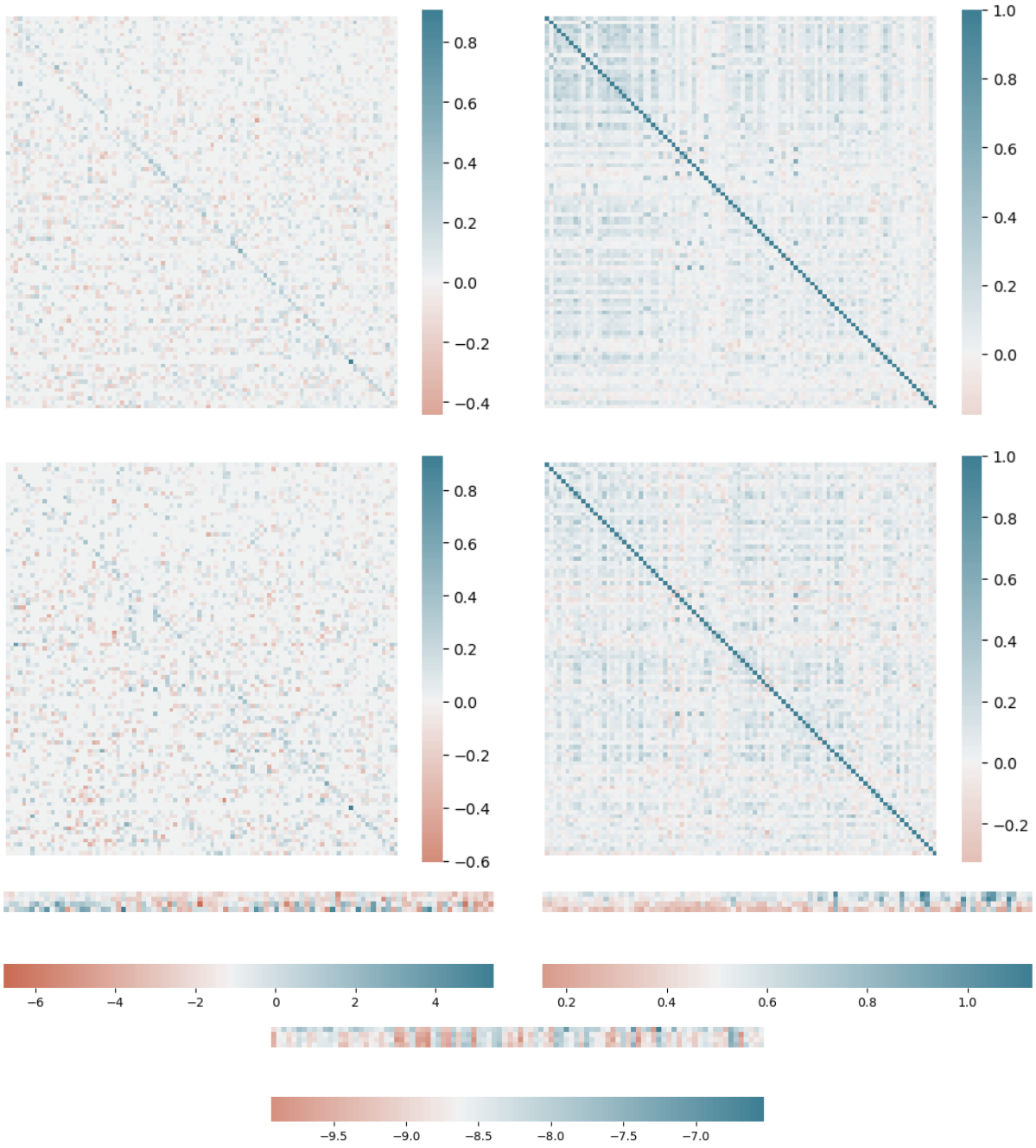


Figure 5: (2/2). Estimated autoregressive parameters (Φ_3 top-left and Φ_4 middle-left) and error term correlation matrices ($\mathbf{R}_3$ top-right and $\mathbf{R}_4$ middle-right), constant terms (bottom-left), error term variances (bottom-right) and unconditional means (bottom).

$$\hat{\mathbf{P}} = \begin{pmatrix} 0.952 & 0.071 & 0.086 & 0.010 \\ 0.032 & 0.793 & 0.297 & 0.405 \\ 0.010 & 0.066 & 0.476 & 0.288 \\ 0.006 & 0.070 & 0.141 & 0.207 \end{pmatrix} \quad \hat{\pi} = \begin{pmatrix} 0.617 \\ 0.273 \\ 0.069 \\ 0.041 \end{pmatrix} \tag{5.1}$$

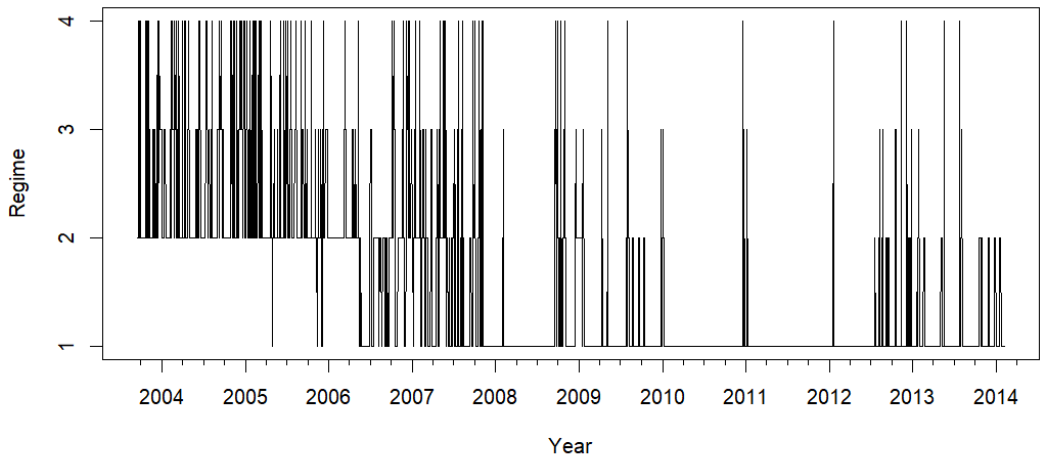


Figure 6: Regimes by highest smoothed probability.

Finally, I briefly discuss the two extensions. Extension II is estimated for $p = 1$ and for two regimes per set of variables, which leads to $M = 8$. Estimation of more regimes per set of variables is not pursued, as this would either entail the estimation of 27 error term covariance matrices, or the imposition of the restriction that the error term covariance matrix is (partly) equal across one or more of these regimes. Extension II yields an MSC of 504,759, an improvement over the MS(2)-VAR(1) and lower than the MS(3)-VAR(1) as well. Extension I is estimated for two regimes for the constant terms and the autoregressive parameters on the one hand and the error term covariance matrices on the other hand. This leads to $M = 4$, although vis-à-vis the MS(2)-VAR(1) only 10 additional transition probabilities are freely estimated. Extension I yields an MSC of 524,277, higher than that of the MS(2)-VAR(1). Extension II has also been estimated for four regimes for each set of parameters, leading to $M = 16$, but yields an even higher MSC of 609,754. More details on the results of Extension II are included in Appendix G.

5.2 Connectedness Results

Now follows the discussion of the results pertaining to the spillovers. Incorporated in the discussion will be the spillovers based on the rolling window VAR(1) and the MS(4)-VAR(1). Evaluating the GIRFs for the MS-VAR has a very high computational burden and has therefore not been pursued for the extensions. In Figure 7 the dynamic SI is plotted for the considered models. The vertical lines correspond to the dates of important events that are, except for the ninth one, adopted from Korobilis and Yilmaz (2018) and Bostanci and Yilmaz (2020).[22] These events consist of the following:

1. 26$^{\text{th}}$ of December 2004. Indian Ocean earthquake and tsunami.

2. 7$^{\text{th}}$ of July 2005. London terrorist attacks.

[22] For a further historical discussion of the events of the sample period and their connection to the SI for individual banks, see Diebold and Yilmaz (2015b).

3. 10th of May 2006. Federal Open Market Committee increases federal funds rate. Unwinding of carry trades.

4. 27th of February 2007. Subprime mortgage lenders file for bankruptcy. Dow Jones drops 416 points.

5. 20th of July 2007. Asset-backed commercial paper market collapses.

6. 15th of September 2008. Lehman Brothers files for bankruptcy. On the 16th of September 2008, Reserve Primary Fund 'breaks the buck', triggering large withdrawals from money market funds.[23] On the 17th of September 2008, AIG is bailed out by the Federal Reserve.

7. 23rd of December 2009. Moody's, as the final one of the Big Three, downgrades the credit rating of the Greek government. Onset of the Euro-crisis.

8. 6th of May 2010. Flash crash on United States stock markets.

9. 5th of August 2011. Downgrade of United States federal government credit ratings.[24]

10. 27th of July 2012. Mario Draghi's speech on the Euro-crisis.

11. 19th of June 2013. Ben Bernanke's press conference on tapering of asset purchases.

The rolling window SI is able to adequately capture the stark increase in system-wide connectedness as the consequence of the listed events. This is notwithstanding the sparsity of the autoregressive coefficients; as for the rolling windows $k >> T$, there are not enough observations to (roughly, as the parameters are likely unstable over time) recover the autoregressive structure of Figure 4. The cycles that are triggered by these events are indicative of periods of increased connectedness vis-à-vis the trend that last around 50 days. This can be seen from the increases of the 10th of May 2006 and the 5th of August 2011, after which no significant increases took place for at least 200 days and it can be seen that after this period, the SI is back to its trend level as the rolling window leaves the event behind.

With respect to the obtained trend, it is similar to that obtained by DDLY and more pronounced than that of the smaller network of Diebold and Yilmaz (2015b). Compared to DDLY, the differences in the SI are larger, with their values ranging between 55 and 90, whereas mine range from 30 to 90. This could be the consequence of using the GFEVD of Lanne and Nyberg (2016), such that for low values of the SI, connectedness is overstated when using the GFEVD of Pesaran and Shin (1998).

Although the listed events are informative, they usually are not stand-alone. The increase in the SI near the end of the second quarter of 2009 could not be attributed to a specific event. Such a rise was also not found by DDLY. However, I do not find it likely that this increase was spurious. June of 2009 was marked by multiple important events that are liable to increase the SI. On the 11th of June, the swine flu outbreak was declared a pandemic. Secondly, it

[23] Brewster, D. (2008, 17th September). Fear of money market funds 'breaking the buck'. *Financial Times*. Retrieved from `https://www.ft.com/content/696e3dc0-84e4-11dd-b148-0000779fd18c`

[24] Brandimarte, W & Bases, D. (2011, 7th August). United States loses prized AAA credit rating from S&P. *Reuters*. Retrieved from `https://www.reuters.com/article/us-usa-debt-downgrade-idUSTRE7746VF20110807/`

was around this period that it became known that a large number of economies entered or had entered recessions. Another example is the period of September 2008. Although the bankruptcy of Lehman Brothers proved to be important for the SI, it is unlikely that the increase can be attributed in its entirety to this and to the other events listed under 6. For example, on September 26$^{\text{th}}$ of 2008, Washington Mutual filed for bankruptcy. In early October of 2008, the United Kingdom nationalised the Royal Bank of Scotland, the stock price of which had fallen by two-thirds from its September 2008 high, for at least £20,000,000,000.[25] Similar actions were undertaken in other countries. The SI reaching its highest value in the end of 2009 is the consequence of this constellation of events.

In Figure 7, the SI of the MS-VAR is also displayed. In conjunction with Figure 6 it can be seen that this SI is mainly driven by the regime switches of the MS-VAR. As a consequence, it is most similar to the SI of the rolling window VAR for the periods in which the process is in the first regime, the regime that prevails most frequently. As a consequence of the first regime being quite persistent, the SI is relatively flat in that period. This is also the period in which the SI of the MS-VAR least often is significantly different from that of the VAR, whereas this is ubiquitously the case before 2006 and after 2012.

For an MS-VAR in which the regimes are clearly identified, if an event is not accompanied by a sufficiently large increase in the forecast error variance as to trigger a regime switch or at least entail such a switch to become more likely, it will not be picked up by the SI. This lead the SI to completely ignore the flash crash of May 2010 and the downgrade of the United States credit rating. Conversely, not every regime-switch can be explained by an event. Although the SI shows marked increases quickly, after the occurrence of events 2, 3, 6 and 7, i.e. within 5 days, there are myriad of such increases throughout the sample, meaning that these events often do not stand out.

Moreover, even if a regime change is triggered, the time-variation in the parameters of the MS-VAR is not of the same nature as that induced by the use of a rolling window. The time-variation is more discrete, with the SI mostly at or near the unconditional SI of the inferred regime. All in all, I find that the MS-VAR does not seem fit to result in a dynamic SI that can function properly as an interpretative tool of bank connectedness over time. The SI of the MS-VAR instead should be considered as a more suitable measure of full-sample connectedness, as will be seen in Section 5.3. To obtain a counterpart to the SI of the rolling window VAR, the MS-VAR can be combined with a rolling window, as in BenSaïda et al. (2018) and together with the time-variation in the GIRF, this could lead to promising results. Namely, owing to the time-variation of the conditional mean and more importantly, of the forecast error variance and the GIRF, it is able to produce values of the SI that are higher than the unconditional, full-sample SI of each regime. For example, this occurs at the day of the bankruptcy of Lehman-Brothers. Therefore, this source of time-variation can be a useful addition to that obtained by means of rolling windows. For the current dataset however, more extensive computational resources are required for the realisation thereof.

[25] Waerden, G. (2008, 13$^{\text{th}}$ October). British government unveils £37bn banking bail-out plan. *Guardian*. Retrieved from `https://www.theguardian.com/business/2008/oct/13/marketturmoil-creditcrunch`

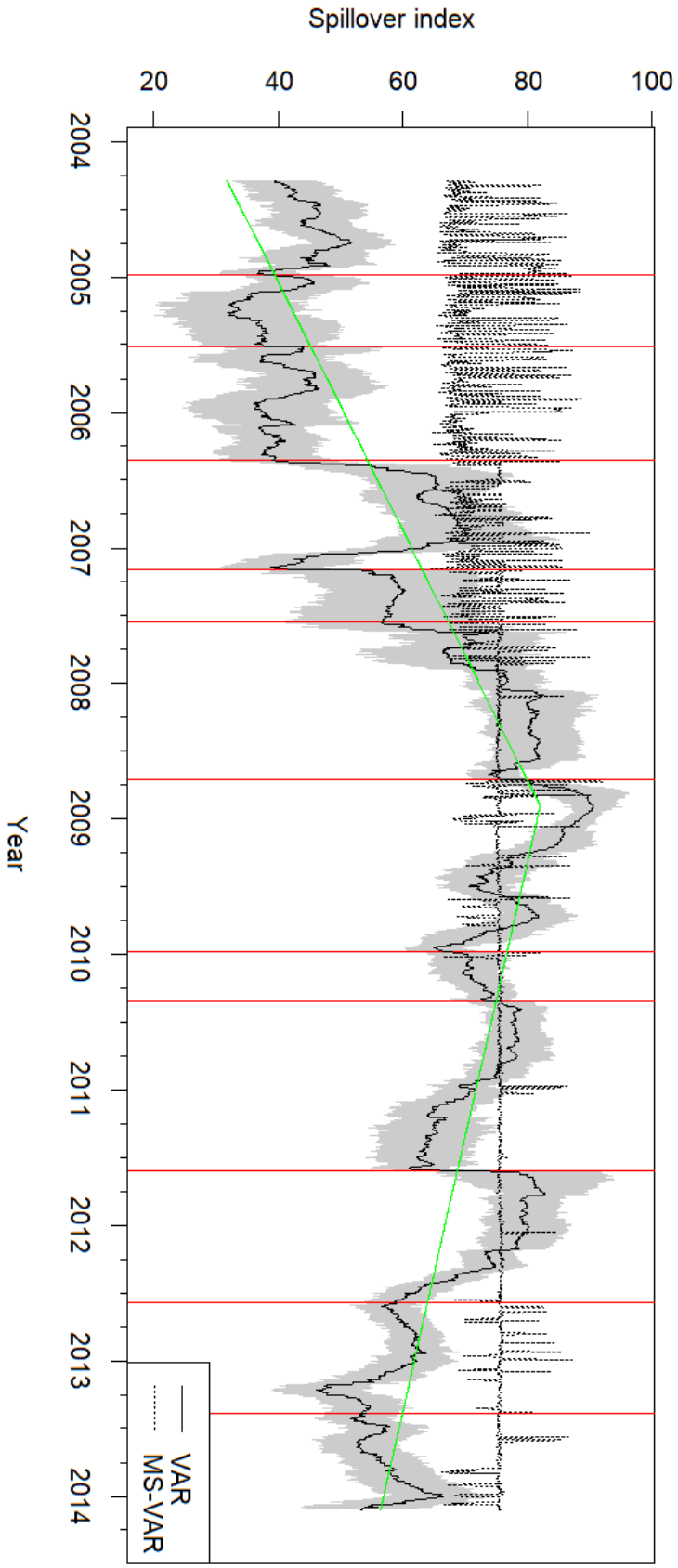


Figure 7: SI of the VAR using a 150-day rolling window with 95% confidence intervals and of the MS-VAR. The red lines correspond to dates of the listed events. The green line is a trend line obtained using a linear spline with a knot at the date of the global maximum.

In Figures 8 and 9, regional directional spillovers are displayed. These are constructed for each bank by taking column and row sums respectively of the index of the bank in Δ respectively of the elements that correspond to banks that do not share its region as a proportion of the total column of row sum. For example, a value of 40 for the 'from' ('to') entails that for this region, 40% of the outgoing (incoming) spillovers are directed towards (received from) the other regions. The regional spillovers follow the general SI. This implies that increasing connectedness is accompanied by increased connectedness across regions. Moreover, whereas the 'to' spillovers are similar, the 'from' spillovers are significantly different across regions. As in Diebold and Yilmaz (2015b), the American 'from' spillovers are relatively (and absolutely) the highest during the great financial crisis. The European 'from' spillovers are relatively high during the Euro-crisis, in 2010-2011. At the 5^{th} of August 2011, the downgrade of the United States credit rating again leads to a relatively high American 'from' spillover, as well as its highest, vis-à-vis the other regions, 'to' spillover.

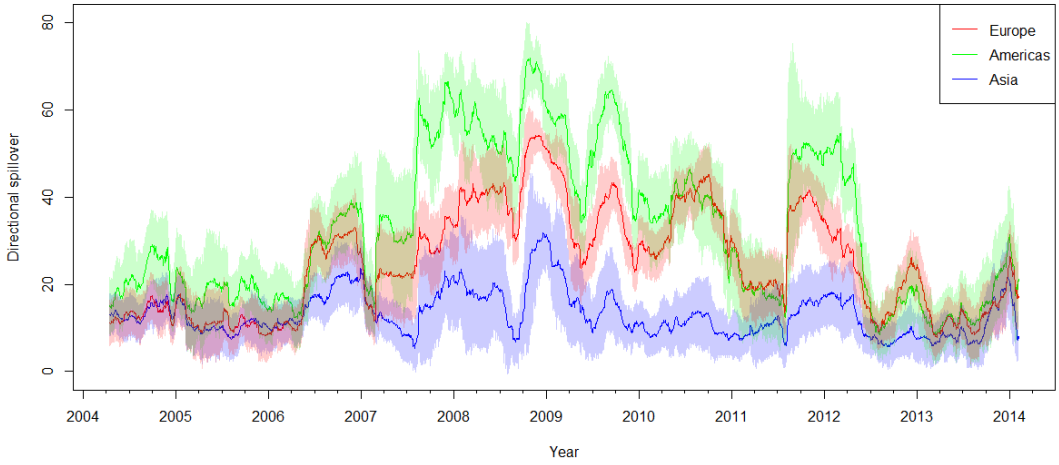


Figure 8: Directional 'from' spillovers with 95% confidence intervals per region.

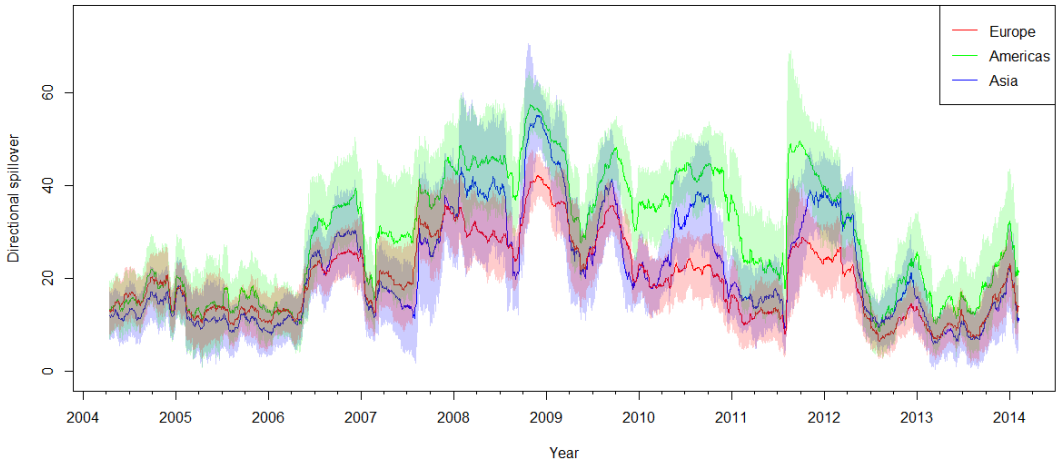


Figure 9: Directional 'to' spillovers with 95% confidence intervals per region.

It is apparent that Asia is a net receiver of spillovers, whereas its outgoing spillovers are mostly confined to banks within the region. This also holds for the 2004 earthquake. This implies that Asia is relatively unimportant in the spillover network, something that is also apparent from the absence of more events that are linked to increases in the SI. For example, the Tōkohu earthquake and tsunami of the 11^{th} of March 2011 is barely discernible in Figure 7.

5.3 Network Analysis

First, I present the network that is defined by the full-sample GFEVD for the VAR and the four networks that are defined by each regime of the MS(4)-VAR(1). To that end, spectral clustering on the node embeddings of these networks is performed with six clusters. This number of clusters is based on Figure 2 of DDLY. Namely, their results indicate that there is a big American-European cluster, that there are distinctive Chinese and Japanese clusters and that there are three peripheral clusters.

After performing spectral clustering, two-dimensional representations of the node embeddings are obtained by means of t-distributed stochastic neighbour embedding (t-SNE).[26] t-SNE aims to construct these representations by minimising the KL-divergence between the $\mathbf{d}$-dimensional node embeddings and the two-dimensional reductions, which is a function of the two sets of pairwise similarities of the observations, one for each space (Van der Maaten & Hinton, 2008). The pairwise similarities define for each observation a probability distribution over the other observations being a neighbour. By preserving the structure of pairwise similarities as well as possible, the KL-divergence of the corresponding probability distributions is minimised. As a consequence, the relative positions of the representations in $\mathbb{R}^2$ are indicative of those in $\mathbb{R}^{\mathbf{d}}$. Hence, the representations can be used to initialise the position of the nodes in the ForceAtlas2 algorithm to obtain a layout of the network. For this, I use the implementation of the algorithm in the `ForceAtlas2` package in R.

The obtained clusters are checked for outliers. As the K-means algorithm assigns every bank to a cluster, banks that do not fit well with any cluster are nevertheless assigned to the cluster to which their distance is lowest. To deal with these outliers, clusters are again formed by means of density-based spatial clustering of applications with noise (DBSCAN) (Ester, Kriegel, Sander & Xu, 1996). This generally preserves the K-means cluster assignments, but assigns banks that are too dissimilar from the other banks to a class of noise points. To improve the visualisation of the obtained layout, loops, i.e. edges from a node to the same node, as well as edges corresponding to $\delta_{i,j} < 0.01$, i.e. if bank j contributes less than 1% to the sum of squared impulse responses of bank i, are removed.

In Figure 10, the obtained network of the VAR is displayed. The colours of the nodes correspond to the DBSCAN cluster assignments. The clusters are contingent on the value of the distance parameter. This parameter was chosen as to reduce the number of noise points, while retaining as much as possible clusters that are visually coherent.[27] For larger values of the distance

[26] Strictly speaking, t-SNE also produces embeddings, but to prevent confusion with the node embeddings, I speak of representations.

[27] Note that visual coherence is not necessary for the clusters to be coherent in the space of eigenvectors that

parameter, Singapore and Malaysia, which each have two banks in the sample that are relatively close to each other, would have been assigned to a cluster, but other banks would have been (spuriously) assigned to the Americas-Europe cluster. Conversely, a smaller value of the distance parameter would have led to the United States banks being identified as a separate cluster, but would have entailed (spuriously) assigning other banks to the noise class. The node size is determined by adding the standardized eigenvector centrality score to a constant. Hence, larger nodes are more central ones, although differences in size are of a qualitative, rather than quantitative nature.

The main result of the VAR network is that clusters in the layout space conform to the statistical clusters, i.e. to clusters that are based on feature representations of the nodes. Moreover, as in DDLY, the clusters strongly correspond to countries and regions. Here too, the chief cluster is a combination of multiple sub-clusters that are closely connected and highly integrated, as many of the edges are not pruned. The sub-clusters themselves also correspond to countries and regions, with United States, Canadian, Brazilian, Scandinavian and Southern European clusters clearly discernible. The Irish, Greek, Finnish and Austrian banks are located on the periphery of the cluster, but are still included. The Russian and Turkish banks, although ostensibly close to the European banks, are too distant and are considered noise points.

The other clusters found are similar to those of DDLY. Contrary to their layout, this layout indicates the network to be much more of a hub-spoke network, rather than a bimodal network in the sense of there being two major groups of clusters. Not only is the Americas-Europe cluster in the centre of the layout space, the overwhelming majority of non-pruned edges, besides those within clusters, are those from the peripheral clusters to the US sub-cluster. The nodes of the Americas-Europe cluster and specifically those of the US sub-cluster, also have the highest eigenvector centrality scores.

Figure 11 displays the networks that are defined by $(\Phi_m,\ \Sigma_m)$, $m = 1,\ \ldots,\ 4$. To be precise, these are the networks that are obtained conditional on regime m prevailing indefinitely. Nevertheless, they are indicative of the differences in the parameter estimates over the regimes and show how the SI of the MS-VAR, absent time-variation through rolling windows, on average interpolates between these network structures based on the inference about the regime process.

Analogous to the parameter estimates, the network topology of the VAR is most similar to that of the the first regime, in which the United States banks are central in the network in a cluster together with the European banks. Possibly due to the more accurate parameter estimates, the DBSCAN algorithm now also assigns the Singaporean and Malaysian banks to their own respective clusters. The network topology of the second regime is also similar, with the major difference that the United States banks have much lower eigenvector centrality scores in the network, although the ForceAtlas2 algorithm still positions these banks at the centre of the layout space and there are still many non-pruned links. This could be the reason that the Canadian banks are now identified as their own cluster. It can also be seen as a specific effect of the more general tendency of banks in the Americas-Europe cluster to be more distant from

are used for spectral clustering, as t-SNE and the ForceAtlas2 algorithm approximate the spatial position of the nodes in the layout.

each other. In this regard, a connection can be made with the parameter estimates of the second regime, which entail a sparser structure of the autoregressive parameter matrix.

In Table 3, which contains statistics that describe the networks, the value of the SI, the modularity and the number of non-pruned edges for the first and second regime are most similar to those of the network of the VAR. This can be expected, as these are the regimes that prevail most frequently. These values respectively being lower, higher and lower for the second regime is in accordance with the above observations.

As with the estimated parameters, it holds that the starkest differences are observed in the third and fourth regimes. The obtained clusters of the first two regimes break down in these regimes, where now the networks are not like hub-and-spoke networks, but consist of a centre with a closely connected periphery which manifests itself visually as a spherical layout. For these regimes, a larger distance parameter was also used as to prevent the DBSCAN algorithm from assigning all banks to the noise class. For the third regime, there is still a clearly discernible Americas-Europe cluster, as well as Japanese and Chinese clusters. In the fourth regime, these clusters break down too and only one cluster remains, which lacks a country- or region-specific interpretation but mostly contains points that are at the centre of the layout space. These differences between the third and the fourth regime are similar to those of the parameter estimates, in the sense that although both regimes display a higher degree of interdependence between banks, it is the fourth regime for which this tendency is most apparent.

These results are supported by the values of the SI, the modularity and the number of non-pruned edges for these regimes. Namely, these statistics are respectively higher, lower and higher for regimes three and four. Moreover, the manifestly lower value of the modularity for regime four compared to regime three is in accordance with their observed cluster assignments. The higher degree of interdependence as apparent from the parameter estimates thus seems to translate to a more connected network of banks in which there is a lower degree to which banks tend to form clusters.

These results also have implications for the dynamic SI of the MS-VAR. Namely, they support the previously mentioned observation that the the time-variation in the parameters of the MS-VAR is not of the same degree as is induced by the use of a rolling window. Although the third and fourth regimes have the highest SI, for the periods in which the SI is highest overall the inference about the regime process entails that it is the first regime that prevails in these periods.

Table 3: Summary statistics of the networks.

	VAR	Regime 1	Regime 2	Regime 3	Regime 4
S	69.28	74.80	60.31	80.16	87.89
Q	0.37	0.32	0.34	0.23	0.13
$\lvert\boldsymbol{\delta}\rvert$	2,077	2,266	1,732	2,910	2,855

Notes: S is the spillover index, Q is the modularity of the K-means cluster assignments and $\boldsymbol{\delta}$ is the set of elements of Δ that exceed 0.01, excluding the diagonal elements.

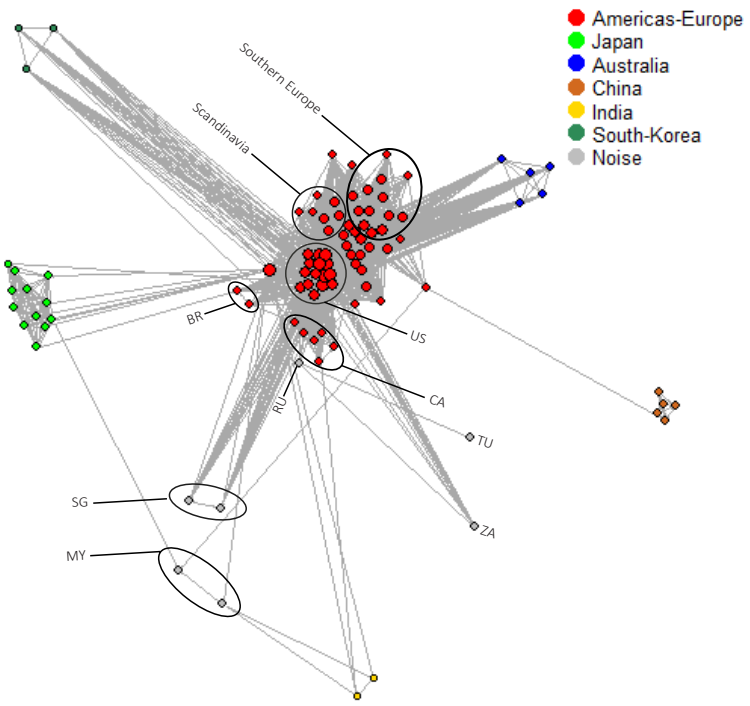
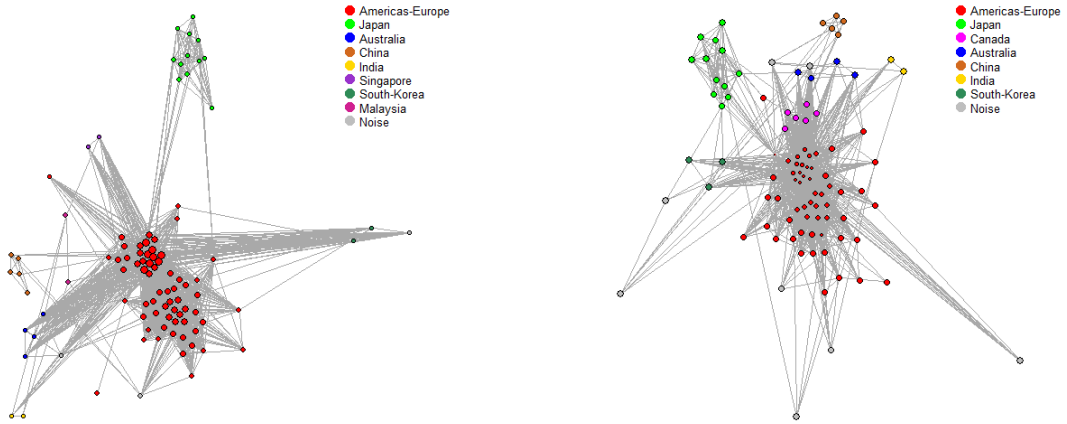
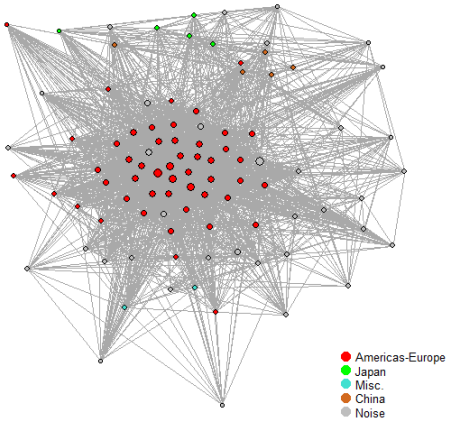


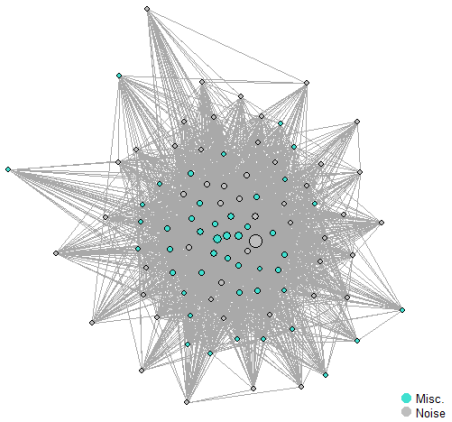
Figure 10: Bank network layout for the VAR. Colours correspond to DBSCAN cluster assignments. The two-letter acronyms are ISO 3166-1 alpha-2 country codes.



Regime 1 Regime 2

Figure 11: (1/2). Bank network layouts for the MS-VAR. Colours correspond to DBSCAN cluster assignments.





Regime 3 Regime 4

Figure 11: (2/2). Bank network layouts for the MS-VAR. Colours correspond to DBSCAN cluster assignments. Misc. corresponds to a cluster which does not have a clear country- or region-specific interpretation.

It is clear from the foregoing that the level of aggregation of the country is highly relevant for the global bank network. Not only are banks of the same countries closely connected in the network, events that are relevant for spillovers do not infrequently occur at the level of the countries. A prime example of this are bailouts, which were enacted at the country level in the course and the aftermath of the great financial crisis.[28] This motivates the use of the MSIRF at the level of the country and the $\text{GFEVD}_{\mathcal{M}}$ can be obtained by means of equation (3.6) by partitioning the 96 banks into sets that are simultaneously shocked that correspond to the 29 countries.

The networks thus obtained are displayed in Figure 12 for the VAR and in Figure 13 for the MS-VAR. Again, for the visualisations clusters are obtained by means of the DBSCAN algorithm. Here as well, the distance parameters were chosen to balance the number of noise points with the retention of coherent clusters. To be consistent with the networks of the individual banks, loops and edges that correspond to $\delta_{\mathcal{M}_i,\mathcal{M}_j} < 0.01$, are removed from the visualisation.

An important result of the country-level $\text{GFEVD}_{\mathcal{M}}$, both for the VAR as well as for the four regimes of the MS-VAR, is that the spillovers are dominated by those from the United States, with the column means of $\Delta_{\mathcal{M}}$ for the United Stated being equal to 0.80, 0.77, 0.82, 0.81 and 0.90 respectively, i.e. on average, the United States contributes to around 80% of the total squared MSIRF of a country. To highlight this feature, the edges corresponding to United States 'from' spillovers are coloured in a darker shade of grey. Only for the VAR and for the first two regimes of the MS-VAR are there two countries, Japan and China, of which the corresponding diagonal element of $\Delta_{\mathcal{M}}$ is largest. The dominance of the United States in the spillovers also

[28] See for example for United States banks: Bailout Recipients. (2022, 18th August). *ProPublica*. Retrieved from `https://projects.propublica.org/bailout/list/index`

has repercussions for the eigenvector centrality ranks, with the loading of the United States on the eigenvector being 50-100 times larger than the loading of the second-most central country. Hence, to keep the visualisation of eigenvector centrality through node size feasible, the node size is determined by means of adding the standardised logarithms of the eigenvector centrality scores to the same constant as was used for the bank networks. As a consequence, relative differences in the node size of the country networks are exponentially larger than for the bank networks.

Another main result is that the country-structure of the bank networks carries over to the country networks. Again, there is a clear centre consisting of the Americas and the European countries, with the Asian countries at the periphery. As a consequence of the clusters of Asian banks being within-country clusters, all of the Asian countries are classified as noise points. Qualitatively, the differences in the bank networks over the regimes also carry over to the country networks. Namely, the network topology of the country network of the VAR is most similar to that of the first and second regimes. For the third and fourth regimes, more points are classified as noise points and in the fourth regime, the cluster also lacks a clear regional interpretation.

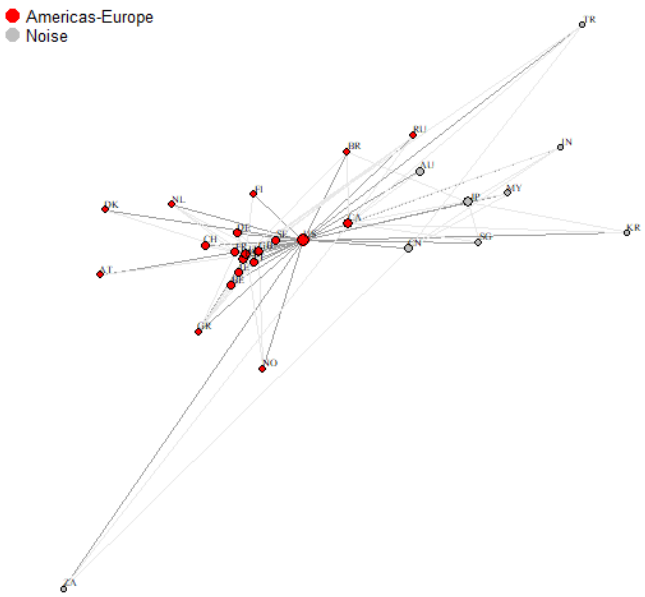
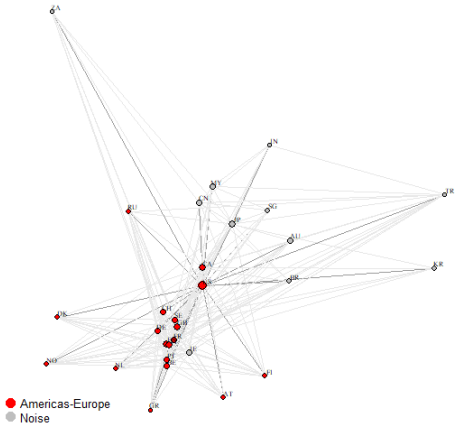
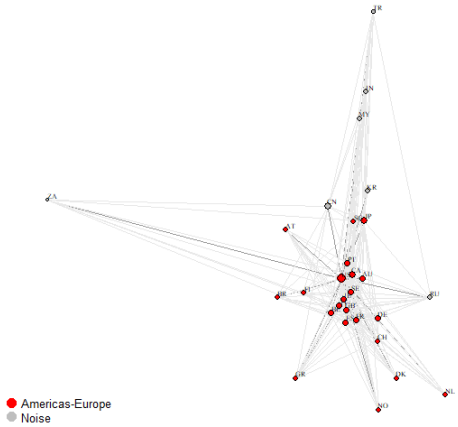


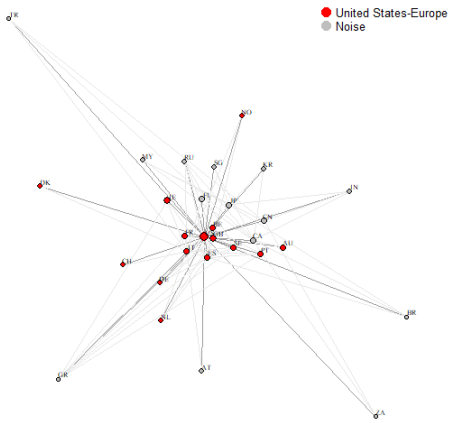
Figure 12: Country network layout for the VAR. Colours correspond to DBSCAN cluster assignments. The two-letter acronyms are ISO 3166-1 alpha-2 country codes.

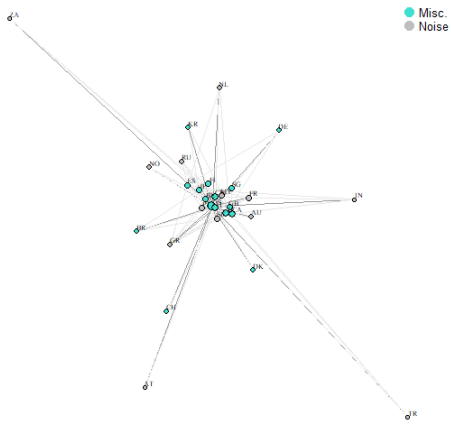




Regime 1 Regime 2

Figure 13: (1/2). Country network layouts for the MS-VAR. Colours correspond to DBSCAN cluster assignments. The two-letter acronyms are ISO 3166-1 alpha-2 country codes.





Regime 3 Regime 4

Figure 13. (2/2). Country network layouts for the MS-VAR. Colours correspond to DBSCAN cluster assignments. The two-letter acronyms are ISO 3166-1 alpha-2 country codes. Misc. corresponds to a cluster which does not have a clear region-specific interpretation.

The statistics for the country networks, which are included in Table 4, are in accordance with these observations, with the SI and the modularity increasing, respectively decreasing over the regimes. For the number of pairwise spillovers exceeding 0.01, the results of the bank networks carry over, but only partly, as for the first regime this number is relatively high and for the

fourth regime this number is relatively low. However, this could also have been the consequence of the United States having very low and high 'from' spillovers in these respective regimes. As a consequence, a larger (smaller) proportion of the sum of squared MSIRFs was to be divided among the remaining countries for the first (fourth) regime, which could have been a reason for these values of $|\boldsymbol{\delta}|$. If the threshold is lowered to 0.001, then again the third and fourth regimes have the highest values of $|\boldsymbol{\delta}|$.

Table 4: Summary statistics of the networks.

	VAR	Regime 1	Regime 2	Regime 3	Regime 4
$S_{\mathcal{M}}$	86.09	87.28	90.17	80.16	95.01
Q	0.130	0.102	0.127	0.078	0.014
$\lvert\boldsymbol{\delta}\rvert$	97	140	67	120	93

Notes: $S_{\mathcal{M}}$ is the aggregate shock spillover index, Q is the modularity of the K-means cluster assignments and $\boldsymbol{\delta}$ is the set of elements of Δ that exceed 0.01, excluding the diagonal elements.

Figures 14 and 15 display the average estimated centrality rank for the three regions with 95% confidence intervals for the eigenvector centrality and the weighted out-degree respectively. The rank regressions lead to more interpretable results with respect to the position of the regions in the network of the three regions than when only using the directional spillovers. Namely, during the financial crisis, the American banks are significantly more central than the other banks. During the Euro-crisis, the European banks become significantly more central than the other banks. Moreover, some of the listed events directly affect the centrality of the network. The increase of the federal funds rate and the unwinding of carry trades on the 10^{th} of May 2006 saw a significant increase, respectively decrease, in the centrality of the European and American banks. The collapse of the asset-backed commercial paper market was accompanied by a significant increase in the centrality of the American banks. The accompanying decrease in the centrality of Asian banks is the consequence of centrality ranks being defined on an ordinal scale. For the Asian banks, the observation that few events of Asian origin affect the SI and the low 'from' spillovers finds additional support in the estimated ranks, which are consistently higher and significantly so in the years of 2007-2013, which where marked by the highest SI.

The estimated centrality ranks can also be connected to the network layouts. Namely, the estimated average ranks of the European and American banks being lowest in relative terms vis-à-vis that of the Asian banks in the period of 2008-2012 is in accordance with the layout of the first regime, which prevailed in these years. Furthermore, the regressions reveal that the Asian banks were more central in the earlier observations. This is in accordance with the layout of the second regime, which prevailed most frequently for these observations.

A connection can also be made with the directional spillovers. Namely, the ranks of the European and American banks seem to be inversely related to the 'from' spillovers, i.e. if the banks of a region are more central in a network, the outgoing spillovers from the banks of this region to a larger extent tend to be received by banks outside of this region. This connection is made apparent in Table 5, which for the European and American banks shows the positive correlation between the centrality of the banks of the region and the 'from' spillovers. For the Asian banks,

it shows a converse relationship; if the Asian banks are less central in the network, their 'to' spillovers tend to be higher.

Table 5: Rank correlation of regional directional spillovers with centrality ranks.

	$S_{\text{Region}\rightarrow\bullet}$			$S_{\text{Region}\leftarrow\bullet}$		
	EU	AM	AS	EU	AM	AS
EC	-0.42	-0.32	0.06	-0.13	-0.10	0.52
OD	-0.51	-0.30	0.11	-0.21	-0.07	0.57

Notes: EC is the eigenvector centrality, OD is the out-degree, EU is Europe, AM stands for the Americas and AS is Asia. The rank correlation used is Kendall's τ.

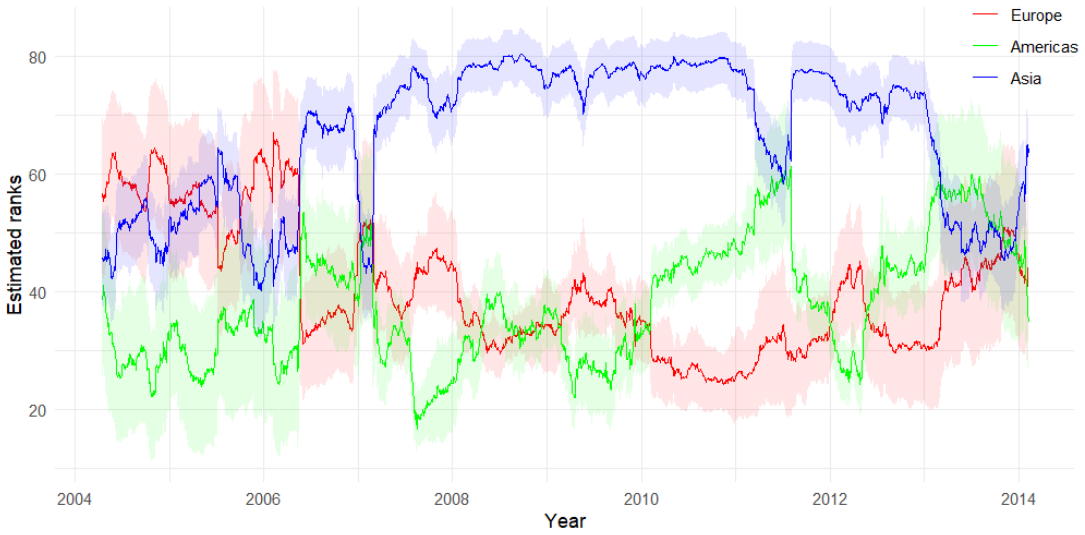


Figure 14: Estimated average centrality rank per region with 95% confidence intervals for the eigenvector centrality. A lower rank entails higher centrality.

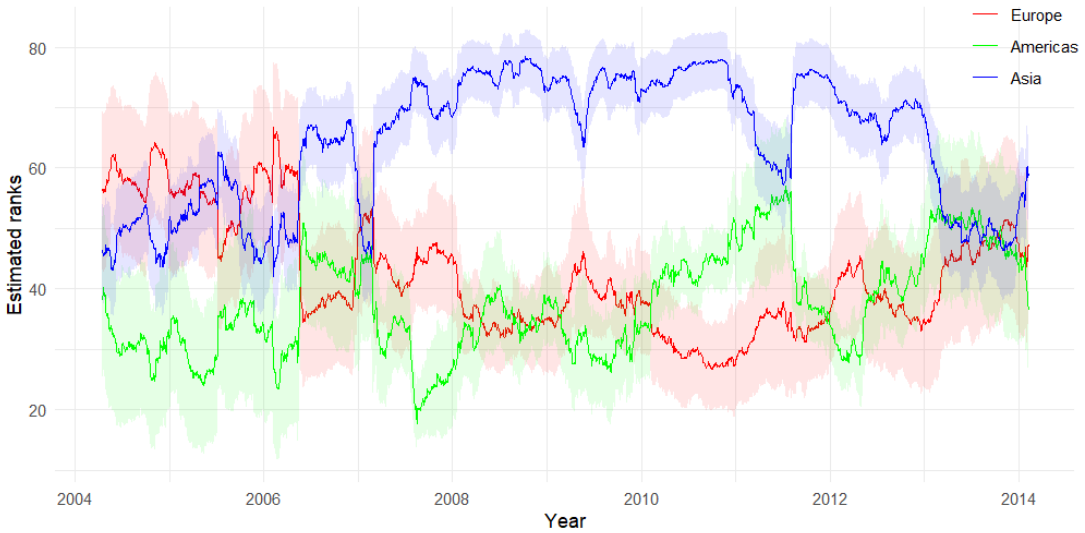


Figure 15: Estimated average centrality rank per region with 95% confidence intervals for the weighted out-degree. A lower rank entails higher centrality.

The modularity of the network over time for the obtained K-means cluster assignments is displayed in Figure 16. The degree to which communities are present in the network is inversely related to the total connectedness. The sample correlation between the modularity and the SI is -0.65. This entails that for periods in which the total connectedness increases, the intensified connections transcend the prevailing communities, instead of manifesting themselves within the existing communities. Put differently, the network itself becomes more global during periods of increased connectedness. This connection between the modularity and the SI was already observed for the network structure over the regimes and it supports the more general idea that for networks that are more connected, the degree to which clusters form tends to be lower.

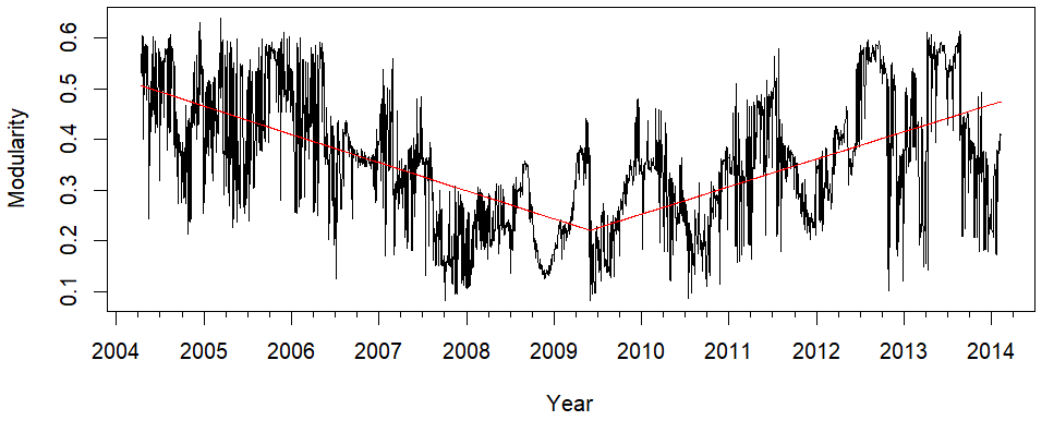


Figure 16: Modularity of the K-means cluster assignments over time. The red line is a trend line obtained using a linear spline with a knot at the date of the global minimum.

To finalise the network analysis, I discuss the results for the graph embeddings. The cluster assignments for two clusters do not exhibit any discernable pattern. Figure 17 visualises the embeddings using t-SNE. It can be seen that the embeddings are part of one group. Thus, no clusters are found. Clustering has also been performed with four clusters and as expected, this result remains unchanged. The same holds when running the graph2vec algorithm with a different set of hyperparameters.[29] Thus, although it is likely, based on the above results pertaining to the modularity, that there are marked differences in the network structure over time, these do not lead to dissimilarities in the embedding space to such an extent that they can be detected by the clustering algorithm. I do not find it likely that this is the case because of an insufficient degree of time-variation in the SI. Namely, the difference between the values of the modularity over time are larger than those across the regimes and it can be seen that the network structures were markedly different across regimes. Hence, it can be concluded either that differences in the network structure that strike me as being marked, are not quantitatively so, or that the graph embeddings are inadequate feature representations of the graphs. In the following, it will be seen that the former is more likely.

[29] Specifically, for this different set of parameters the number of Weisfeiler-Lehman iterations is set to 4, $\mathbf{d}$ is set to $k = 96$, the down-sampling parameter is set to 0.0001 and the number of epochs is set to 20.

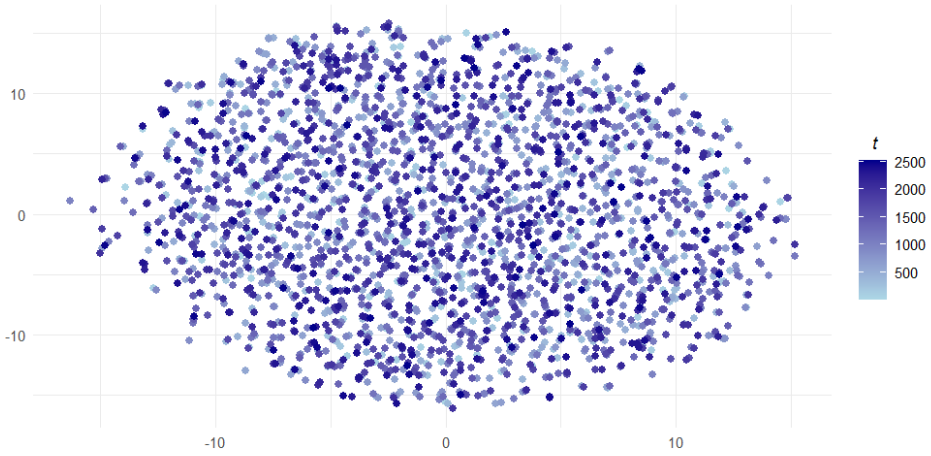


Figure 17: Two-dimensional t-SNE of the graph embeddings for each period t.

5.4 Prediction Results

Table 6 displays the results of the SE prediction. From the McFadden R^2, it can be seen that the use of graph embeddings yield a substantial improvement, except for the model estimated at a lag order of 10. In-sample, the embeddings thus carry more information than the SI. When estimation is regularised, hereinafter referred to as $\mathbb{G}_\alpha$, the in-sample fit remains similar to that of the logistic regression with the SI as an independent variable, again with the exception for the model estimated at a lag order of 10.

Table 6: Systemic event prediction results.

	McFadden R^2			F-score			Weight			Threshold		
l	SI	$\mathbb{G}$	$\mathbb{G}_\alpha$	SI	$\mathbb{G}$	$\mathbb{G}_\alpha$	SI	$\mathbb{G}$	$\mathbb{G}_\alpha$	SI	$\mathbb{G}$	$\mathbb{G}_\alpha$
1	0.242	**0.340**	0.239	0.407	0.262	**0.427**	1	1	4	0.15	0.18	0.34
2	0.226	**0.385**	0.249	0.406	0.323	**0.416**	4	2	4	0.60	0.57	0.31
3	0.209	**0.340**	0.203	0.373	0.277	**0.392**	8	2	4	0.62	0.43	0.33
4	0.212	**0.334**	0.218	0.340	0.248	**0.375**	8	8	2	0.57	0.72	0.18
5	0.217	**0.327**	0.206	0.353	0.270	**0.394**	2	4	8	0.35	0.65	0.51
10	**0.480**	0.351	0.181	**0.667**	0.296	0.405	1	1	1	0.60	0.63	0.10
22	0.175	**0.298**	0.169	**0.402**	0.318	0.383	8	1	8	0.41	0.28	0.42

Notes: l is the lag order. SI, $\mathbb{G}$ and $\mathbb{G}_\alpha$ respectively correspond to the logistic regression with the SI, the graph embeddings and the graph embeddings subject to elastic net penalisation as independent variables. The McFadden R^2 is based on the model estimated on the entire sample. The F-score is obtained on the test set. The models that were used for classifications have had observations corresponding to SEs weighted in estimation in accordance with columns 8-10 vis-à-vis non-SE observations. The threshold is the fitted probability below which an observation was classified as a non-SE. **Bold** entries denote that the model of the corresponding column is the best for the lag order of the corresponding row.

Next are the classification results. To prima facie gauge the suitability of a logistic regression model, a comparison is made with a simple classification rule. Namely, the F-score has been

determined for the rule that classifies a trading day as a systemic event if at least one of the previous $\mathbf{L}$ days have been a systemic event for $\mathbf{L} = 1,\ 2,\ \ldots,\ 10$. Even though $\mathbf{L}$ was selected based on the entire sample, the F-score obtained was 0.329, indicating that both the SI and $\mathbb{G}_\alpha$ are useful for classification. The F-scores show how $\mathbb{G}_\alpha$ are competitive with the SI-based logistic regression model, outperforming the SI for the lower lag-orders. The weights indicate that it is useful to assign more importance to SE-observations in estimation and that this can lead to improved out-of-sample classification performance. The thresholds show that the estimated probabilities should be subject to a different interpretation when used as a classifier; for the SI, in general higher probabilities are required to signal an SE.

For $l = 1$, the estimated probabilities, thresholds and SE index are visualised in Figure 18. It can be seen that most of the correctly classified SEs are those in the years of 2008-2012, when they occur more frequently and when the SI is higher. It can be seen that for $\mathbb{G}_\alpha$, $\hat{p}_t$ follows a much smoother pattern over time. Table 7 shows that the estimated probabilities of the SI are more closely associated with both the proportion of banks that experience a very low stock return, as well as with the occurrence of an SI than those of $\mathbb{G}_\alpha$. From this perspective, the estimated probabilities with the SI can more readily be interpreted as a gauge for the possibility of systemic distress. Similar figures for the other values of l are included in Appendix G.

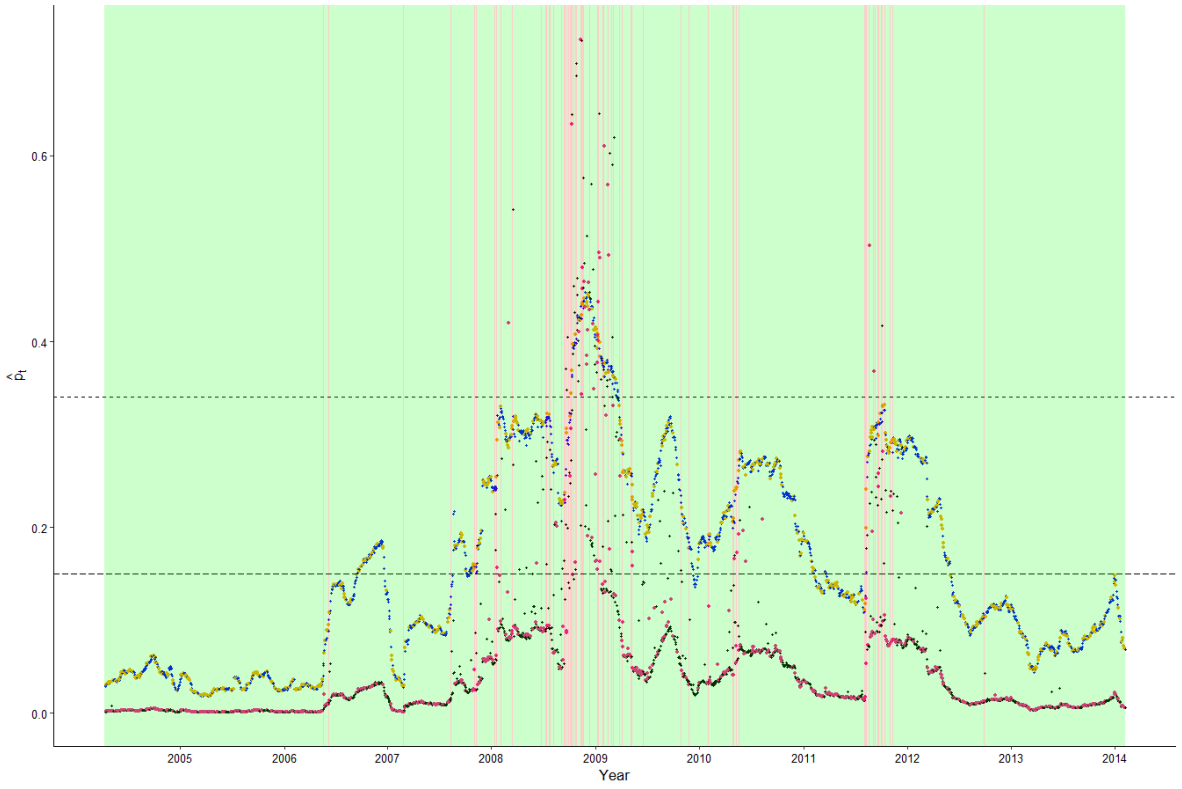


Figure 18: Estimated probabilities and classifications of an SE for $l = 1$. The background is red for periods that constitute an SE and green otherwise. The dashed and long-dashed lines are the thresholds for the SI and $\mathbb{G}_\alpha$ respectively. Black and blue dots are estimated probabilities that correspond to an observation of the training set for the SI and $\mathbb{G}_\alpha$ respectively. The pink and orange dots are estimated probabilities that correspond to an observation of the test set for the SI and $\mathbb{G}_\alpha$ respectively.

Table 7: Correlation of $\hat{p}_t$ with the SE index.

	SI		$\mathbb{G}_\alpha$	
	Index	SE	Index	SE
$l = 1$	0.54	0.40	0.43	0.31
$l = 2$	0.48	0.37	0.41	0.30
$l = 3$	0.45	0.33	0.43	0.31
$l = 4$	0.46	0.33	0.42	0.30
$l = 5$	0.49	0.37	0.40	0.28
$l = 10$	0.61	0.69	0.40	0.28
$l = 22$	0.42	0.30	0.37	0.25

Notes: SI and $\mathbb{G}_\alpha$ respectively correspond to the logistic regression with the SI and the graph embeddings subject to elastic net penalisation as independent variables. Index denotes the continuous valued SE index and SE is the categorical variable obtained by means of thresholding the SE index at 0.25. For the correlation of $\hat{p}_t$ with SE, the point biserial correlation coefficient is used.

6 Conclusions

This thesis inquired into the use of an MS-VAR subject to adaptive elastic net penalisation to estimate the SI and the spillover network of the high-dimensional global bank stock return volatility dataset of Demirer et al. (2018). In addition, extensions were considered that consisted of the use of multiple Markov chains that together govern the regime process. For the volatility dataset, evidence of regime switching is found through markedly lower values of the MSC for the MS-VAR than for a VAR subject to adaptive elastic net penalisation which was employed as a benchmark model. This led to the selection of an MS(4)-VAR(1) model. Moreover, for a lag order of two, an improvement in the MSC can be obtained by letting the constant terms and autoregressive parameters that respectively correspond to European, American and Asian banks be governed by their respective Markov chains. For future research, such extensions could be further explored. Another possibility is to inquire into the use of Bayesian methods. For example, Sugita (2022) used the stochastic search variable selection prior in the estimation of an MS-VAR.

The dynamic SI is conventionally obtained by means of a rolling window VAR. The SI of the rolling window VAR displays time-variation that can be linked to overall trends of increasing and decreasing connectedness over the course of the great financial crisis, the great recession and the Euro-crisis, and also captures important financial and economic events that have ramifications for the connectedness of bank stock return volatility. The MS-VAR is less suited for the construction of a dynamic SI, as its time-variation is predominantly driven by (the possibility of) regime switching. The comparison with the SI of the rolling window VAR shows that the time-variation induced by regime switching is not of the same nature as that induced by the rolling window. Therefore, for future research I recommend the combination of an MS-VAR with a rolling window for the construction of a dynamic SI, which could improve the SI by using the time-variation in the forecast error variance of the MS-VAR. The main limitation of this approach is the computational burden involved. However, it should be feasible if more computational resources

are available, as well as with a more efficient implementation of the EM-algorithm and of the GIRF.

In a full-sample analysis of the global bank spillover network, the MS-VAR reveals the difference in the network structure over the regimes. The regimes with rich autoregressive structures translate to networks that are more connected and for which the community structure, consisting of country- and region-specific clusters, breaks down. This analysis was extended by constructing networks at the level of the country by means of the MSIRF and similar results were obtained. It was found that the United States is pivotal in the country network. The full-sample analysis also shows the viability of spectral clustering in combination with K-means clustering and DBSCAN, of node embeddings for the purposes of community detection. Moreover, this enables a data-driven initialisation of layout-generating algorithms such as ForceAtlas2. Thus, I recommend the use of this procedure for community detection in networks and their visualisation.

For the SI of the rolling window VAR, a bootstrap method was applied to obtain confidence intervals. It has thereby been shown that significant differences exist in the the 'from' spillovers, which are highest for the Americas and Europe, indicating that these banks transmit volatility shocks to other regions to a larger extent than Asian banks. This idea is supported by means of rank regressions on the centrality ranks of regions in the network as measured by the eigenvector centrality and the out-degree. The European and American banks throughout the sample are significantly more central than Asian banks, with the American banks being most central before the great financial crisis and the the European banks being most central during the Euro-crisis. Finally, by means of the modularity it has been shown how the degree to which banks in the network from clusters evolves over time and that the modularity is inversely correlated with the SI.

A graph embedding algorithm has been applied to obtain feature representations of the networks. The application of spectral clustering to these embeddings indicates the absence of clusters of networks over time. Plotting the t-SNE reveals the absence of clusters in the networks in general, conditional on the embeddings. The graph embeddings were also used as features in a logistic regression to model the probability of the occurrence of an SE and a comparison was made with a logistic regression model that contained the SI as a feature. Both with respect to the McFadden R^2, as well as with respect to classification performance, the graph embeddings are competitive with the SI. Therefore, graph embeddings can fruitfully function as features in downstream learning or modelling tasks. For future research it would be interesting consider node embeddings as features in supervised learning or modelling tasks as well.

References

Alizadeh, S., Brandt, M. W. & Diebold, F. X. (2002). Range-Based Estimation of Stochastic Volatility Models. *Journal of Finance*, *57*(3), 1047–1091.

Ando, T., Greenwood-Nimmo, M. & Shin, Y. (2022). Quantile Connectedness: Modeling Tail Behavior in the Topology of Financial Networks. *Management Science*, *68*(4), 2401–2431.

Ang, A. & Timmermann, A. (2012). Regime Changes and Financial Markets. *Annual Review of Financial Economics*, *4*(1), 313–337.

Antonakakis, N., Chatziantoniou, I. & Gabauer, D. (2020). Refined Measures of Dynamic Connectedness based on Time-Varying Parameter Vector Autoregressions. *Journal of Risk and Financial Management*, *13*(4), 84.

Arsov, I., Canetti, E., Kodres, L. E. & Mitra, S. (2013, May). *Near-Coincident Indicators of Systemic Stress* (Working Paper No. 115). International Monetary Fund. Retrieved from `https://www.imf.org/en/Publications/WP/Issues/2016/12/31/Near-Coincident-Indicators-of-Systemic-Stress-40551`

Baele, L. (2005). Volatility Spillover Effects in European Equity Markets. *Journal of Financial and Quantitative Analysis*, *40*(2), 373–401.

Banerjee, O., El Ghaoui, L. & d'Aspremont, A. (2008). Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. *Journal of Machine Learning Research*, *9*, 485–516.

Baruník, J. & Křehlík, T. (2018). Measuring the Frequency Dynamics of Financial Connectedness and Systemic Risk. *Journal of Financial Econometrics*, *16*(2), 271–296.

Bazzi, M., Blasques, F., Koopman, S. J. & Lucas, A. (2017). Time-Varying Transition Probabilities for Markov Regime Switching Models. *Journal of Time Series Analysis*, *38*(3), 458–478.

BenSaïda, A., Litimi, H. & Abdallah, O. (2018). Volatility spillover shifts in global financial markets. *Economic Modelling*, *73*, 343–353.

Beraich, M., Amzile, K., Laamire, J., Zirari, O. & Fadali, M. A. (2022). Volatility Spillover Effects of the US, European and Chinese Financial Markets in the Context of the Russia-Ukraine Conflict. *International Journal of Financial Studies*, *10*(4), 95.

Billio, M., Getmansky, M., Lo, A. W. & Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, *104*(3), 535–559.

Bloch, F., Jackson, M. O. & Tebaldi, P. (2023). Centrality measures in networks. *Social Choice and Welfare*, *61*(2), 413–453.

Bostanci, G. & Yilmaz, K. (2020). How connected is the global sovereign credit risk network? *Journal of Banking & Finance*, *113*, 105761.

Bourette Sicotte, X. (2018, June). *Lasso regression: derivation of the coordinate descent update rule.* Retrieved from `https://xavierbourretsicotte.github.io/lasso_derivation.html`

Bouri, E., Cepni, O., Gabauer, D. & Gupta, R. (2021). Return connectedness across asset classes around the COVID-19 outbreak. *International Review of Financial Analysis*, *73*, 101646.

Bouri, E., Lucey, B., Saeed, T. & Vinh Vo, X. (2020). Extreme spillovers across Asian-Pacific currencies: A quantile-based analysis. *International Review of Financial Analysis*, *72*, 101605.

Brüggemann, R., Jentsch, C. & Trenkler, C. (2016). Inference in VARs with conditional heteroskedasticity of unknown form. *Journal of Econometrics*, *191*(1), 69–85.

Burnham, K. P. & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304.

Catania, L. (2022, November). *Multiple Chains Markov Switching Vector Autoregression.* Retrieved from `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3662346`

Cavicchioli, M. (2014). Determining the Number of Regimes in Markov Switching VAR and VMA Models. *Journal of Time Series Analysis*, *35*(2), 173–186.

Chan-Lau, J. A. (2017, May). *Variance Decomposition Networks: Potential Pitfalls and a Simple Solution* (Working Paper No. 107). International Monetary Fund. Retrieved from `https://www.imf.org/en/Publications/WP/Issues/2017/05/04/Variance-Decomposition-Networks-Potential-Pitfalls-and-a-Simple-Solution-44883`

Chan-Lau, J. A. (2018). Systemic centrality and systemic communities in financial networks. *Quantitative Finance and Economics*, *2*(2), 468–496.

Chavez-Martinez, G., Agarwal, A., Khalili, A. & Ahmed, S. E. (2023). Penalized Estimation of Sparse Markov Regime-Switching Vector Auto-Regressive Models. *Technometrics*, *65*(4), 553–563.

Chen, S. & Schienle, M. (2022). Large spillover networks of nonstationary systems. *Journal of Business & Economic Statistics*, 1–15.

Chetverikov, D. & Wilhelm, D. (2023, October). *Inference for Rank-Rank Regressions.* Retrieved from `https://arxiv.org/abs/2310.15512`

Choi, J.-E. & Shin, D. W. (2020). Bootstrapping volatility spillover index. *Communications in Statistics - Simulation and Computation*, *49*(1), 66–78.

Demirer, M., Diebold, F. X., Liu, L. & Yilmaz, K. (2018). Estimating global bank network connectedness. *Journal of Applied Econometrics*, *33*(1), 1–15.

Dhaene, G., Sercu, P. & Wu, J. (2022). Volatility spillovers: A sparse multivariate GARCH approach with an application to commodity markets. *Journal of Futures Markets*, *42*(5), 868–887.

Diebold, F. X. & Inoue, A. (2001). Long memory and regime switching. *Journal of Econometrics*, *105*(1), 131–159.

Diebold, F. X., Liu, L. & Yilmaz, K. (2017, August). *Commodity Connectedness* (Working Paper No. 23685). National Bureau of Economic Research. Retrieved from `http://www.nber.org/papers/w23685`

Diebold, F. X. & Yilmaz, K. (2009). Measuring Financial Asset Return and Volatility Spillovers, with Application to Global Equity Markets. *Economic Journal*, *119*(534), 158–171.

Diebold, F. X. & Yilmaz, K. (2012). Better to give than to receive: Predictive directional measurement of volatility spillovers. *International Journal of Forecasting*, *28*(1), 57–66.

Diebold, F. X. & Yilmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics*, *182*(1), 119–

134.

Diebold, F. X. & Yilmaz, K. (2015a). *Financial and Macroeconomic Connectedness: A Network Approach to Measurement and Monitoring*. Oxford University Press, USA.

Diebold, F. X. & Yilmaz, K. (2015b). Trans-Atlantic Equity Volatility Connectedness: US and European Financial Institutions, 2004-2014. *Journal of Financial Econometrics*, *14*(1), 81-127.

Douc, R., Moulines, E. & Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Annals of Statistics*, *32*(5), 2254-2304.

Dungey, M. & Martin, V. L. (2007). Unravelling Financial Market Linkages during Crises. *Journal of Applied Econometrics*, *22*(1), 89–119.

Elliott, G., Rothenberg, T. J. & Stock, J. H. (1996). Efficient Tests for an Autoregressive Unit Root. *Econometrica*, *64*(4), 813–836.

Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the second international conference on knowledge discovery and data mining* (p. 226–231).

Franke, J. (2012). Markov Switching Time Series Models. In T. S. Rao, S. S. Rao & C. R. Rao (Eds.), *Handbook of Statistics. Time Series Analysis: Methods and Applications* (Vol. 30, pp. 99–122). Elsevier.

Friedman, J., Hastie, T., Höfling, H. & Tibshirani, R. (2007). Pathwise Coordinate Optimization. *Annals of Applied Statistics*, *1*(2), 302-332.

Friedman, J., Hastie, T. & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1-22.

Furman, Y. (2014, June). *VAR Estimation with the Adaptive Elastic Net.* Retrieved from `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2456510`

Gabauer, D. (2020). Volatility impulse response analysis for DCC-GARCH models: The role of volatility transmission mechanisms. *Journal of Forecasting*, *39*(5), 788–796.

Gabauer, D., Gupta, R., Marfatia, H. A. & Miller, S. M. (2024). Estimating US housing price network connectedness: Evidence from dynamic Elastic Net, Lasso, and ridge vector autoregressive models. *International Review of Economics & Finance*, *89*, 349–362.

Garman, M. B. & Klass, M. J. (1980). On the Estimation of Security Price Volatilities from Historical Data. *Journal of Business*, 67–78.

Gonzalo, J. & Pitarakis, J.-Y. (2002). Lag length estimation in large dimensional systems. *Journal of Time Series Analysis*, *23*(4), 401–423.

Granger, C. W. (2008). Non-Linear Models: Where Do We Go Next - Time Varying Parameter Models? *Studies in Nonlinear Dynamics & Econometrics*, *12*(3).

Green, P. J. (1990). On Use of the EM Algorithm for Penalized Likelihood Estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *52*(3), 443–452.

Greenwood-Nimmo, M., Nguyen, V. H. & Shin, Y. (2021). Measuring the connectedness of the global economy. *International Journal of Forecasting*, *37*(2), 899–919.

Greenwood-Nimmo, M. & Tarassow, A. (2022). Bootstrap-based probabilistic analysis of spillover scenarios in economic and financial networks. *Journal of Financial Markets*, *59*, 100661.

Grover, A. & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 855–864).

Guidolin, M. (2011). Markov Switching Models in Empirical Finance. In D. Drukker (Ed.), *Missing Data Methods: Time-Series Methods and Applications* (Vol. 27, pp. 1–86). Emerald Group Publishing Limited.

Hamilton, J. D. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, 357–384.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hamilton, J. D. (2016). Macroeconomic Regimes and Regime Shifts. In J. B. Taylor & H. Uhlig (Eds.), *Handbook of Macroeconomics* (Vol. 2, pp. 163–201). Elsevier.

Hamilton, J. D. & Lin, G. (1996). Stock market volatility and the business cycle. *Journal of Applied Econometrics*, *11*(5), 573–593.

Hastie, T., Tibshirani, R. & Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Vol. 2). Springer.

Hu, F., Liu, J., Li, L. & Liang, J. (2020). Community detection in complex networks using Node2vec with spectral clustering. *Physica A: Statistical Mechanics and its Applications*, *545*, 123633.

Hurvich, C. M. & Tsai, C.-L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, *76*(2), 297–307.

Isogai, T. (2014). Clustering of Japanese stock returns by recursive modularity optimization for efficient portfolio diversification. *Journal of Complex Networks*, *2*(4), 557–584.

Isogai, T. (2017). Dynamic correlation network analysis of financial asset returns with network clustering. *Applied Network Science*, *2*(8).

Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLOS One*, *9*(6), e98679.

Kae-Yih, T. (2023). The international spillover behaviour of implied volatilities and forecasting ability of spillover indices. *Applied Economics*, *55*(48), 5719–5735.

Kangogo, M. & Volkov, V. (2022). Detecting signed spillovers in global financial markets: A Markov-switching approach. *International Review of Financial Analysis*, *82*, 102161.

Kasahara, H. & Shimotsu, K. (2019). Asymptotic properties of the maximum likelihood estimator in regime switching econometric models. *Journal of Econometrics*, *208*(2), 442–467.

Kim, C. (1994). Dynamic linear models with Markov-switching. *Journal of Econometrics*, *60*(1-2), 1–22.

Kim, Y. M. & Lee, S. (2023). Spillover shifts in the FX market: Implication for the behavior of a safe haven currency. *North American Journal of Economics and Finance*, *65*, 101885.

Kole, E. & Van Dijk, D. (2023). Moments, shocks and spillovers in Markov-switching VAR models. *Journal of Econometrics*, *236*(2), 105474.

Koop, G., Pesaran, M. H. & Potter, S. M. (1996). Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics*, *74*(1), 119–147.

Korobilis, D. & Yilmaz, K. (2018, January). *Measuring Dynamic Connectedness with Large Bayesian VAR Models.* Retrieved from `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3099725`

Krolzig, H.-M. (1997). *Markov-Switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis.* (Lecture notes in Economics and Mathematical Systems Vol. 454). Springer.

Kumar, A., Iqbal, N., Mitra, S. K., Kristoufek, L. & Bouri, E. (2022). Connectedness among major cryptocurrencies in standard times and during the COVID-19 outbreak. *Journal of International Financial Markets, Institutions and Money*, *77*, 101523.

Lanne, M. & Nyberg, H. (2016). Generalized Forecast Error Variance Decomposition for Linear and Nonlinear Multivariate Models. *Oxford Bulletin of Economics and Statistics*, *78*(4), 595–603.

Lastrapes, W. D. & Wiesen, T. F. (2021). The joint spillover index. *Economic Modelling*, *94*, 681–691.

Lütkepohl, H. (1990). Asymptotic Distributions of Impulse Response Functions and Forecast Error Variance Decompositions of Vector Autoregressive models. *Review of Economics and Statistics*, 116–125.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis.* Springer Science & Business Media.

Maung, G. Y. K. (2023). *Essays on High-Dimensional Econometrics.* University of Rochester.

Mazumder, R. & Hastie, T. (2012). The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, *6*, 2125-2149.

Meilă, M. & Pentney, W. (2007). Clustering by weighted cuts in directed graphs. In *Proceedings of the 2007 SIAM International Conference on Data Mining* (pp. 135–144).

Molnár, B., Márton, I.-B., Horvát, S. & Ercsey-Ravasz, M. (2024). Community detection in directed weighted networks using Voronoi partitioning. *Scientific Reports*, *14*(1), 8124.

Molnár, P. (2012). Properties of range-based volatility estimators. *International Review of Financial Analysis*, *23*, 20–29.

Monbet, V. & Ailliot, P. (2017). Sparse vector Markov switching autoregressive models. Application to multivariate time series of temperature. *Computational Statistics & Data Analysis*, *108*, 40–51.

Narayanan, A., Chandramohan, M., Venkatesan, R., Chen, L., Liu, Y. & Jaiswal, S. (2017, July). *graph2vec: Learning Distributed Representations of Graphs.* Retrieved from `https://arxiv.org/abs/1707.05005`

Ng, S. & Perron, P. (2001). Lag Length Selection and the Construction of Unit Root Tests with Good Size and Power. *Econometrica*, *69*(6), 1519–1554.

Nicholson, W. B., Matteson, D. S. & Bien, J. (2017). VARX-L: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting*, *33*(3), 627–651.

Nicholson, W. B., Wilms, I., Bien, J. & Matteson, D. S. (2020). High Dimensional Forecasting

via Interpretable Vector Autoregression. *Journal of Machine Learning Research*, *21*(166), 1–52.

Pesaran, H. H. & Shin, Y. (1998). Generalized impulse response analysis in linear multivariate models. *Economics Letters*, *58*(1), 17–29.

Politis, D. N. & White, H. (2004). Automatic Block-Length Selection for the Dependent Bootstrap. *Econometric Reviews*, *23*(1), 53–70.

Psaradakis, Z. & Spagnolo, N. (2003). On the Determination of the Number of Regimes in Markov-Switching Autoregressive Models. *Journal of Time Series Snalysis*, *24*(2), 237–252.

Psaradakis, Z. & Spagnolo, N. (2006). Joint Determination of the State Dimension and Autoregressive Order for Models with Markov Regime Switching. *Journal of Time Series Analysis*, *27*(5), 753–766.

Schaub, M. T., Delvenne, J.-C., Rosvall, M. & Lambiotte, R. (2017). The many facets of community detection in complex networks. *Applied Network Science*, *2*(4).

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 461–464.

Shapovalova, Y. & Eichler, M. (2023). Measuring and Quantifying Uncertainty in Volatility Spillovers: A Bayesian Approach. *Data Science in Science*, *2*(1), 2176379.

Smith, A., Naik, P. A. & Tsai, C.-L. (2006). Markov-switching model selection using Kullback-Leibler divergence. *Journal of Econometrics*, *134*(2), 553–577.

Sugita, K. (2022). Time Series Forecasting Using a Markov Switching Vector Autoregressive Model with Stochastic Search Variable Selection Method. In N. N. Thach, V. Kreinovich, D. T. Ha & N. D. Trung (Eds.), *Financial Econometrics: Bayesian Analysis, Quantum Uncertainty and Related Topics* (pp. 147–170). Springer.

Van der Maaten, L. & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(11), 2579-2605.

Van der Zwan, T. (2023, November). *Multiple Shock Impulse Response Functions.* Retrieved from `https://sites.google.com/view/terrivanderzwan/research`

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*, 395–416.

Wang, H., Li, B. & Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *71*(3), 671–683.

Wiesen, T. F., Beaumont, P. M., Norrbin, S. C. & Srivastava, A. (2018). Are generalized spillover indices overstating connectedness? *Economics Letters*, *173*, 131–134.

Yi, S., Xu, Z. & Wang, G.-J. (2018). Volatility connectedness in the cryptocurrency market: Is Bitcoin a dominant cryptocurrency? *International Review of Financial Analysis*, *60*, 98–114.

Zhang, B. & Wang, P. (2014). Return and volatility spillovers between China and world oil markets. *Economic Modelling*, *42*, 413–420.

Zou, H. & Hastie, T. (n.d.). *Regularization and Variable Selection via the Elastic net.* Retrieved from `https://hastie.su.domains/TALKS/enet_talk.pdf`

Zou, H., Hastie, T. & Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *Annals*

of Statistics, *35*(5), 2173-2192.

Zou, H. & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, *37*(4), 1733-1751.

A Pathwise Coordinate Descent

The following is based on Friedman et al. (2010). The PCD algorithm, which is a cyclical coordinate descent algorithm, exploits the fact that for the univariate case, the elastic net has a closed form solution. Because in general, the predictors are not uncorrelated, one can iteratively apply such closed form solutions to regressions of the partial residuals on the predictor currently considered. Consider a regression of a univariate dependent variable y_t on a k-dimensional vector of independent variables x_t, $t = 1, \ldots, T$. Let β_0 be a constant and β be the slope coefficients corresponding to x_t. These are estimated by means of the adaptive elastic net as follows:

$$(\hat{\beta}_0, \hat{\beta}) = \underset{(\beta_0,\ \beta)}{\operatorname{argmin}} \ \frac{1}{2T} \sum_{t=1}^{T} (y_t - \beta_0 - x_t'\beta)^2 + \lambda \sum_{j=1}^{k} (\frac{1-\alpha}{2}\beta_j^2 + w_j \alpha |\beta_j|) \tag{A.1}$$

where w_j is the adaptive weight corresponding to β_j. The additional fractions in the sum of squared residuals and in the Ridge penalty are without loss of generality and are included for purposes of the derivation. Given a set of estimates for the constant and for all slope coefficients except for variable j, plug these into equation (A.1). Now, I focus on the terms containing β_j, which, when opening the brackets, gives

$$\begin{aligned} \hat{\beta}_j = \underset{\beta_j}{\operatorname{argmin}} \ \frac{1}{2T} \sum_{t=1}^{T} & -2y_t\beta_j x_{j,t} + 2\beta_0\beta_j x_{j,t} + \beta_j x_{j,t} \sum_{i \neq j} \beta_i x_{i,t} + \beta_j^2 x_{j,t}^2 \\ & + \lambda(\frac{1-\alpha}{2}\beta_j^2 + w_j\alpha|\beta_j| \end{aligned} \tag{A.2}$$

Denoting the objective function by O and taking the derivative with respect to β_j yields

$$\begin{aligned} \frac{dO}{d\beta_j} = \frac{1}{T} \sum_{t=1}^{T} & -y_t x_{j,t} + \beta_0 x_{j,t} + x_{j,t} \sum_{i \neq j} \beta_i x_{i,t} + \beta_j x_{j,t}^2 \\ & + \lambda(1-\alpha)\beta_j + w_j \alpha \frac{d}{d\beta_j}|\beta_j|) \end{aligned} \tag{A.3}$$

Define $\tilde{y}_t^{(j)} = \tilde{\beta}_0 + \sum_{i \neq j} x_{t,j}\tilde{\beta}_i$, the fitted value of y_t for the given estimates of the other parameters. Because β_j is non-differentiable at the origin, the three cases of the subderivative of O are as follows (Bourette Sicotte, 2018):

$$\frac{dO}{d\beta_j} = \begin{cases} T^{-1}\sum_{t=1}^{T} x_{j,t}(\tilde{y}_t^{(j)} - y_t) + \beta_j\big(\lambda(1-\alpha) + \sum_{t=1}^{T} x_{j,t}^2\big) - w_j\lambda\alpha, \ \text{if } \beta_j < 0 \\ \big[T^{-1}\sum_{t=1}^{T} x_{j,t}(\tilde{y}_t^{(j)} - y_t) - w_j\lambda\alpha, \ T^{-1}\sum_{t=1}^{T} x_{j,t}(\tilde{y}_t^{(j)} - y_t) + w_j\lambda\alpha\big], \ \text{if } \beta_j < 0 \\ T^{-1}\sum_{t=1}^{T} x_{j,t}(\tilde{y}_t^{(j)} - y_t) + \beta_j\big(\lambda(1-\alpha) + \sum_{t=1}^{T} x_{j,t}^2\big) + w_j\lambda\alpha, \ \text{if } \beta_j > 0 \end{cases} \tag{A.4}$$

Equating the subderivative to zero and solving for β_j provides the following update

$$\tilde{\beta}_j \leftarrow \frac{S(T^{-1}\sum_{t=1}^{T} x_{j,t}(y_t - \tilde{y}_t^{(j)}),\ w_j\lambda\alpha)}{\sum_{t=1}^{T} x_{t,j}^2 + \lambda(1-\alpha)} \tag{A.5}$$

where $S(z,\ \gamma)$ is the soft-thresholding operator which is defined as follows:

$$S(z,\ \gamma) = \text{sign}(z)(|z| - \gamma)_+ = \begin{cases} z - \gamma, \text{if } z > 0 \text{ and } \gamma < |z| \\ z + \gamma, \text{if } z < 0 \text{ and } \gamma < |z| \\ 0 \text{ if } \gamma > |z| \end{cases}$$

Moreover, PCD uses 'warm starts', i.e. it starts at a sufficiently high value of λ for which $\beta = \mathbf{0}$ and decreases λ until the PCD has converged to a non-zero vector solution. For the subsequent λ, this solution is used to initialise the algorithm for that value of λ which drastically speeds up the computations (Hastie, Tibshirani & Friedman, 2009).

B Graphical Least Absolute Shrinkage and Selection Operator

Below, the objective function of the GLASSO is repeated for convenience

$$\underset{\Omega}{\operatorname{argmax}} \log|\Omega| - \operatorname{tr}(\mathbb{S}\Omega) - \rho\|\Omega - \operatorname{diag}(\Omega)\|_1 \tag{B.1}$$

where $\mathbb{S}$ is the Gaussian ML estimate of the error term covariance matrix, which is calculated by means of previously obtained parameter estimates and can be treated as given. Notation for the regimes is suppressed for convenience. The following exposition of the GLASSO is based on Friedman et al. (2008). First, let V be the current estimate of Σ. The GLASSO corresponds to a LASSO-penalised regression over each column of V. To derive the algorithm, first note that $V\Omega = \mathbf{I}_k$, where Ω is the current estimate of the precision matrix, which can be expanded as

$$\begin{pmatrix} V_{1,1} & v_{1,2} \\ v_{1,2}' & v_{2,2} \end{pmatrix} \begin{pmatrix} \Omega_{1,1} & \omega_{1,2} \\ \omega_{1,2}' & \omega_{2,2} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{k-1} & \mathbf{0} \\ \mathbf{0}' & 1 \end{pmatrix} \tag{B.2}$$

Setting the subgradient of (B.1) equal to zero yields

$$V - \mathbb{S} - \rho\Gamma = \mathbf{O} \tag{B.3}$$

where $V = \frac{d}{d\Omega}\log|\Omega|$, $\mathbb{S} = \frac{d}{d\Omega}\operatorname{tr}(\mathbb{S}\Omega)$ and $(\Gamma)_{i,j} = \operatorname{sign}(\omega_{i,j})$ if $\omega_{i,j} \neq 0$ and $(\Gamma)_{i,j} = 0$ if $\omega_{i,j} = 0$ The upper-right block of equation (B.3) is equivalent to

$$v_{1,2} - s_{1,2} - \rho\gamma_{1,2} = \mathbf{0} \tag{B.4}$$

where $\mathbf{0}$ is of dimension $k-1$. Banerjee, El Ghaoui and d'Aspremont (2008) show for the solution of $v_{1,2}$ that it satisfies the following problem

$$v_{1,2} = \underset{u}{\operatorname{argmin}}\, u' V_{1,1}^{-1} u : \|u - s_{1,2}\|_\infty \leqslant \rho \tag{B.5}$$

Then, they show that solving this problem is equivalent to solving the following problem, which is the dual problem of (B.5)

$$\min_{\beta} \frac{1}{2}\|V_{1,1}^{\frac{1}{2}}(\beta - s_{1,2})\|_2^2 + \rho\|\beta\|_1 \tag{B.6}$$

This dual problem corresponds to a LASSO-penalised regression and this observation forms the basis of the GLASSO algorithm, which is as follows:

1. Initialise $W = \mathbb{S} + \rho\mathbf{I}$
2. Loop over each column j and permute V such that column (row) j is placed at the position of $v_{1,2}$ ($v_{1,2}'$) in the decomposition of V in equation (B.2) and set $v_{2,2}$ to $(V)_{j,j}$ respectively.
3. Solve problem (B.6) for $j = 1, \ldots, k$ using PCD. Set $v_{1,2} = V_{1,1}\hat{\beta}$
4. Repeat steps 2 and 3 until convergence.

C Hamilton Filter and Kim Smoother

Denote by $\hat{\boldsymbol{\xi}}_{t|q} = (\mathbb{P}[s_t = 1|\mathcal{I}_q],\ \ldots,\ \mathbb{P}[s_t = M|\mathcal{I}_q])$ the estimated probabilities of being in the respective regimes, conditional on the information set $\mathcal{I}$ at time q, $q = 0,\ \ldots,\ T$, $t = 1,\ \ldots,\ T$. Then, the filter of Hamilton (1989) provides, given initial values $\hat{\boldsymbol{\xi}}_{0|0}$ the following expressions that can be used to recursively estimate in a forwards manner the probabilities of being in the respective regimes for $t = 1,\ \ldots,\ T$:

$$\hat{\boldsymbol{\xi}}_{t|t-1} = \mathbf{P}\hat{\boldsymbol{\xi}}_{t-1|t-1} \tag{C.1}$$

$$\hat{\boldsymbol{\xi}}_{t|t} = \frac{\hat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta_t}}{\boldsymbol{\iota}'(\hat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta_t})} \tag{C.2}$$

In equation (C.2), $(\boldsymbol{\iota})_m = 1,\ m = 1,\ \ldots,\ M$ and $\boldsymbol{\eta_t} = [f(y_t|s_t = 1,\ \mathcal{I}_{t-1};\ \boldsymbol{\theta}),\ \ldots,\ f(y_t|s_t = M,\ \mathcal{I}_{t-1};\ \boldsymbol{\theta})]$, a vector containing conditional densities of y_t for the different regimes. Subsequently, given $\hat{\boldsymbol{\xi}}_{T|T}$ it is possible using the smoother of Kim (1994) to recursively estimate in a backwards manner the probabilities of being in the respective regimes for $t = p+1,\ \ldots,\ T-1$ conditional on $\mathcal{I}_T$:

$$\hat{\boldsymbol{\xi}}_{t|T} = \hat{\boldsymbol{\xi}}_{t|t} \odot \left[\mathbf{P}'(\hat{\boldsymbol{\xi}}_{t+1|T} \oslash \hat{\boldsymbol{\xi}}_{t+1|t})\right] \tag{C.3}$$

where $\oslash$ denotes the Hadamard division. The notation of equations (C.1)-(C.3) is based on that of Hamilton (1994) and their derivations can be found ibidem. Finally, define $\tilde{\mathbf{P}}_t$ to be an $(M \times M)$-dimensional matrix such that $(\tilde{\mathbf{P}}_t)_{i,j} = \mathbb{P}[s_t = i, s_{t-1} = j|\mathcal{I}_T]$. It then holds that

$$\tilde{\mathbf{P}}_t = \mathbf{P} \odot (\hat{\boldsymbol{\xi}}_{t|T}\hat{\boldsymbol{\xi}}'_{t-1|t-1}) \oslash (\hat{\boldsymbol{\xi}}_{t|t-1}\boldsymbol{\iota}'),\quad t = 1,\ \ldots,\ T \tag{C.4}$$

The following derivation of $\tilde{\mathbf{P}}_t$ corresponds to that of E. Vladimirov (personal communication, November 16, 2023).

$$\begin{aligned}
\mathbb{P}[s_t = i, s_{t-1} = j|\mathcal{I}_T] &= \mathbb{P}[s_{t-1} = j|s_t = i,\ \mathcal{I}_T]\mathbb{P}[s_t = i|\mathcal{I}_T] \\
&= \mathbb{P}[s_{t-1} = j|s_t = i,\ \mathcal{I}_{t-1}]\mathbb{P}[s_t = i|\mathcal{I}_T] \\
&= \frac{\mathbb{P}[s_t = i, s_{t-1} = j|\mathcal{I}_{t-1}]\mathbb{P}[s_t = i|\mathcal{I}_T]}{\mathbb{P}[s_t = i|\mathcal{I}_{t-1}]} \\
&= \frac{p_{i,j}\mathbb{P}[s_{t-1} = j|\mathcal{I}_{t-1}]\mathbb{P}[s_t = i|\mathcal{I}_T]}{\mathbb{P}[s_t = i|\mathcal{I}_{t-1}]} \\
&= \frac{\mathbf{P}_{i,j}\hat{\boldsymbol{\xi}}_{t|T,i}\hat{\boldsymbol{\xi}}_{t-1|t-1,j}}{\hat{\boldsymbol{\xi}}_{t|t-1,i}} \\
&= \frac{[\mathbf{P} \odot (\hat{\boldsymbol{\xi}}_{t|T}\hat{\boldsymbol{\xi}}'_{t-1|t-1})]_{i,j}}{\hat{\boldsymbol{\xi}}_{t|t-1,i}} \\
&= [\mathbf{P} \odot (\hat{\boldsymbol{\xi}}_{t|T}\hat{\boldsymbol{\xi}}'_{t-1|t-1}) \oslash (\hat{\boldsymbol{\xi}}_{t|t-1}\boldsymbol{\iota}')]_{i,j}
\end{aligned}$$

where the first and third equalities use the definition of conditional probability, the second equality uses the Markov property, the fourth equality uses conditional independence and the definition of the Markov chain and the remaining equalities use the definition of predicted, updated and smooth probabilities respectively.

D Generalised Impulse Response Function of the Markov-Switching Vector Autoregression

The following is reproduced from Kole and Van Dijk (2023, pp. 4-13), but for an MS(M)-VAR(p). To that end, define the following matrices as in Lütkepohl (2005) to rewrite equation (2.1) into an MS-VAR(1).

$$Y_t = \begin{pmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{pmatrix}, \ \mathbf{c}_{s_t} = \begin{pmatrix} c_{s_t} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}, \ \Theta_{s_t} = \begin{pmatrix} \Phi_{1,s_t} & \Phi_{2,s_t} & \dots & \Phi_{p-1,s_t} & \Phi_{p,s_t} \\ \mathbf{I}_k & \mathbf{O} & \dots & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_k & \dots & \mathbf{O} & \mathbf{O} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{O} & \mathbf{O} & \dots & \mathbf{I}_k & \mathbf{O} \end{pmatrix} \text{ and } U_t = \begin{pmatrix} u_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}$$

$$Y_t = \mathbf{c}_{s_t} + \Theta_{s_t} Y_{t-1} + U_t, \tag{D.1}$$

where $\mathbb{E}[U_t U_t']$ is a $(kp \times kp)$-dimensional matrix of zeroes, except for the upper left $k \times k$ block, which consists of Σ_{s_t}. Next, define for matrices of arbitrary dimension $B_1, \ B_2, \ \dots, \ B_m$

$$\text{bdiag}_{i=1}^{m}(B_i) = \begin{pmatrix} B_1 & \mathbf{O} & \dots & \mathbf{O} \\ \mathbf{O} & B_2 & \dots & \mathbf{O} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{O} & \mathbf{O} & \dots & B_m \end{pmatrix}$$

Now, define $Y_t^* = \boldsymbol{\xi}_t \otimes Y_t$ and $\tilde{Y}_t = (Y_t^{*\prime}, \ \boldsymbol{\xi}_t')'$. The Markov chain can be written as a VAR(1) as $\boldsymbol{\xi}_t = \mathbf{P}\boldsymbol{\xi}_{t-1} + v_t$, where v_t is a martingale difference sequence. It then holds that

$$Y_t^* = CP\boldsymbol{\xi}_{t-1} + \boldsymbol{\Theta}(P \otimes \mathbf{I}_k) Y_{t-1}^* + \varepsilon_t^* \tag{D.2}$$

$$\varepsilon_t^* = \Lambda(P \otimes \mathbf{I}_k)(\boldsymbol{\xi}_{t-1} \otimes \varepsilon_t) + Cv_t + \boldsymbol{\Theta}(v_t \otimes Y_{t-1}) + \Lambda(v_t \otimes \varepsilon_t) \tag{D.3}$$

where $C = \text{bdiag}_{i=1}^{M}(\boldsymbol{c}_i)$, $\boldsymbol{\Theta} = \text{bdiag}_{i=1}^{M}(\Theta_i)$ and $\Lambda = \text{bdiag}_{i=1}^{M}(\Lambda_i)$. Furthermore, define $\tilde{Y}_t = (Y_t^{*\prime}, \ \boldsymbol{\xi}_t')'$ and $\tilde{\varepsilon}_t = (\varepsilon_t^{*\prime}, \ v_t')'$. Then:

$$\tilde{Y}_t = \tilde{\boldsymbol{\Theta}} \tilde{Y}_{t-1} + \tilde{\varepsilon}_t \tag{D.4}$$

where

$$\tilde{\boldsymbol{\Theta}} = \begin{pmatrix} \boldsymbol{\Theta}(\mathbf{P} \otimes \mathbf{I}_{kp}) & C\mathbf{P} \\ \mathbf{O} & \mathbf{P} \end{pmatrix}$$

This allows for the formulation of the GIRF of the MS-VAR in which the shocks $\nu_{j,t}$ are quantified in terms of the difference with the conditional expectiation, i.e. $\nu_{j,t} = y_{j,t} - \mathbb{E}[y_{j,t}|\mathcal{I}_{t-1}]$. The GIRF then is defined as

$$\text{GI}_{\tilde{Y}}(h, \ \nu_{j,t}, \ \mathcal{I}_{t-1}) = \tilde{\Theta}^h \begin{pmatrix} \mathbb{E}[\varepsilon_t^* | y_{j,t}, \ \mathcal{I}_{t-1}] \\ \mathbb{E}[v_t | y_{j,t}, \ \mathcal{I}_{t-1}] \end{pmatrix} \tag{D.5}$$

where $\mathbb{E}[\varepsilon_t^*|y_{j,t},\ \mathcal{I}_{t-1}] = C\mathbb{E}[v_t|y_{j,t},\ \mathcal{I}_{t-1}] + \Theta(E[v_t|y_{j,t},\ \mathcal{I}|_{t-1}] \otimes Y_{t-1}) + \Lambda E[\boldsymbol{\xi}_t \otimes \varepsilon_t|y_{j,t},\ \mathcal{I}_{t-1}]$, $\mathbb{E}[v_t|y_{j,t},\ \mathcal{I}_{t-1}] = \dfrac{\mathbf{f} \odot \hat{\boldsymbol{\xi}}_{t|t-1}}{\mathbf{f}'\hat{\boldsymbol{\xi}}_{t|t-1}} - \hat{\boldsymbol{\xi}}_{t|t-1}$ where $\mathbf{f} \in \mathbb{R}^M : (\mathbf{f})_m = \phi\Big(y_{j,t};\ (\mu_{t,m})_j,\ \sigma_{j,j;m}\Big)$, where

$$E[\boldsymbol{\xi}_t \otimes \varepsilon_t|y_{j,t},\ \mathcal{I}_{t-1}] = \begin{pmatrix} (\hat{\boldsymbol{\xi}}_{t|t})_1 \mathbb{E}[\varepsilon_t, |y_{j,t},\ s_t = 1,\ \mathcal{I}_{t-1}] \\ \vdots \\ (\hat{\boldsymbol{\xi}}_{t|t})_M \mathbb{E}[\varepsilon_t, |y_{j,t},\ s_t = M,\ \mathcal{I}_{t-1}] \end{pmatrix}$$

$\mathbb{E}[\varepsilon_t, |y_{j,t},\ s_t = m,\ \mathcal{I}_{t-1}] = \Lambda_m^{-1}[\sigma_{j,j;m}^{-1}\Big(y_{j,t} - (\mu_{t,m})_j\Big)\Sigma_m e_j]$. $\mathrm{GI}_{\tilde{Y}}(h,\ \nu_{j,t},\ \mathcal{I}_{t-1})$ is a vector of dimension $kpM + M$. The information that pertains to Y_t can be obtained by means of the matrix $\tilde{G}_Y = (G_Y,\ \mathbf{O}_{kp\times M})$, where $G_Y = \boldsymbol{\iota}_M' \otimes \mathbf{I}_{kp}$, the subscripts of which denote the dimensions and, using that $Y_t = \tilde{G}_Y \tilde{Y}_t$, the standardised GIRF is obtained as

$$\Psi_{Y_j} = \tilde{G}_Y \mathrm{GI}_{\tilde{Y}}(h,\ \mathrm{Var}[y_{j,t}|\mathcal{I}_{t-1}]^{\frac{1}{2}},\ \mathcal{I}_{t-1}) \tag{D.6}$$

of which the first k elements are selected. Thus, as for the VAR, the standardised GIRF is the GIRF evaluated in a shock of the square root of its one-step ahead forecast error variance. For the MS-VAR, the forecast error variance and the conditional expectation of Y_j are time-dependent. Next to the differences in the parameters over the regimes, this also induces time-variation in the GIRF. For convenience of the notation, I now shift t forward one period. The conditional expectation can be retrieved as the j-th element of $\tilde{G}_Y \mathbb{E}[\tilde{Y}_{t+h}|\mathcal{I}_t]$, where

$$\mathbb{E}[\tilde{Y}_{t+h}|\mathcal{I}_t] = \tilde{\Theta}^h \begin{pmatrix} \hat{\boldsymbol{\xi}}_{t|t} \otimes Y_t \\ \hat{\boldsymbol{\xi}}_{t|t} \end{pmatrix} \tag{D.7}$$

in which h is set to 1. For the forecast error variance, define $Z_t = Y_t \otimes Y_t$, $Z_t^* = \boldsymbol{\xi}_t \otimes Z_t$ and $\tilde{Z}_t = (Z_t^{*\prime},\ Y_t^{*\prime},\ \boldsymbol{\xi}_t')'$. It then holds, first, that

$$Z_t = \boldsymbol{\gamma}_{s_t} + \omega_{s_t} + \boldsymbol{\Psi}_{s_t} Y_{t-1} + \Upsilon_{s_t} Z_{t-1} + \zeta_t \tag{D.8}$$

where $\boldsymbol{\gamma}_{s_t} = \mathbf{c}_{s_t} \otimes \mathbf{c}_{s_t}$, $\omega_{s_t} = \mathrm{vec}(\Sigma_{s_t})$, $\boldsymbol{\Psi}_{s_t} = \Theta_{s_t} \otimes \mathbf{c}_{s_t} + \mathbf{c}_{s_t} \otimes \Theta_{s_t}$, $\Upsilon_{s_t} = \Theta_{s_t} \otimes \Theta_{s_t}$ and

$$\begin{aligned} \zeta_t = {} & (\Lambda_{s_t} \otimes\ \mathbf{c}_{s_t} + \mathbf{c}_{s_t} \otimes\ \Lambda_{s_t})\varepsilon_t + (\Lambda_{s_t} \otimes\ \Theta_{s_t})(\varepsilon_t \otimes\ Y_{t-1}) + (\Theta_{s_t} \otimes\ \Lambda_{s_t})(Y_{t-1} \otimes\ \varepsilon_t) \\ & + (\Lambda_{s_t} \otimes\ \Lambda_{s_t})\Big(\varepsilon_t \otimes\ \varepsilon_t - \mathrm{vec}(\mathbf{I}_{kp})\Big) \end{aligned}$$

secondly, that

$$Z_t^* = (\boldsymbol{\Gamma} + \boldsymbol{\Omega})\mathbf{P}\boldsymbol{\xi}_{t-1} + \boldsymbol{\Psi}(\mathbf{P} \otimes \mathbf{I}_k) Y_{t-1}^* + \Upsilon(\mathbf{P} \otimes \mathbf{I}_{k^2}) Z_{t-1}^* + \zeta_t^* \tag{D.9}$$

where $\mathbf{\Gamma} = \text{bdiag}_{i=1}^{M}(\boldsymbol{\gamma}_i)$, $\mathbf{\Omega} = \text{bdiag}_{i=1}^{M}(\omega_i)$, $\mathbf{\Psi} = \text{bdiag}_{i=1}^{M}(\mathbf{\Psi}_i)$, $\Upsilon = \text{bdiag}_{i=1}^{M}(\Upsilon_i)$ and

$$\begin{aligned}\zeta_t^* = (\mathbf{\Gamma} + \mathbf{\Omega})v_t + \mathbf{\Psi}(v_t \otimes Y_{t-1}) + \Upsilon(v_t \otimes Z_{t-1}) + \text{bdiag}_{i=1}^{M}(\Lambda_i \otimes \mathbf{c}_i + \mathbf{c}_i \otimes \Lambda_i)(\boldsymbol{\xi}_t \otimes \varepsilon_t) \\ + \text{bdiag}_{i=1}^{M}(\Lambda_i \otimes \Theta_i)(\boldsymbol{\xi}_t \otimes \varepsilon_t \otimes Y_{t-1}) + \text{bdiag}_{i=1}^{M}(\Theta_i \otimes \Lambda_i)(\boldsymbol{\xi}_t \otimes Y_{t-1} \otimes \varepsilon_t) \\ + \text{bdiag}_{i=1}^{M}(\Lambda_i \otimes \Lambda_i)\Big(\boldsymbol{\xi}_t \otimes [\varepsilon_t \otimes \varepsilon_t - \text{vec}(\mathbf{I}_{(kp)^2})]\Big)\end{aligned}$$

and finally that

$$\tilde{Z}_t = \tilde{\Upsilon}\tilde{Z}_{t-1} + \tilde{\zeta}_t \tag{D.10}$$

where

$$\tilde{\Upsilon} = \begin{pmatrix} \Upsilon(\mathbf{P} \otimes \mathbf{I}_{(kp)^2}) & \mathbf{\Psi}(\mathbf{P} \otimes \mathbf{I}_{kp}) & (\mathbf{\Gamma} + \mathbf{\Omega})\mathbf{P} \\ \mathbf{O} & \mathbf{\Theta}(\mathbf{P} \otimes \mathbf{I}_{kp}) & C\mathbf{P} \\ \mathbf{O} & \mathbf{O} & \mathbf{P} \end{pmatrix}, \ \tilde{\zeta}_t = \begin{pmatrix} \zeta_t^* \\ \varepsilon_t^* \\ v_t \end{pmatrix}$$

in which the spectral radius of $\Upsilon(\mathbf{P} \otimes \mathbf{I}_{(kp)^2})$ is restricted to be less than 1 if the MS-VAR is to be stable. Next, define $\tilde{H}_Z$ and $\tilde{H}_Y$ such that $Z_t = \tilde{H}_Z\tilde{Z}_t$ and $Y_t = \tilde{H}_Y\tilde{Z}_t$, i.e. $\tilde{H}_Z = (H_Z, \ \mathbf{O}_{(kp)^2 \times M(kp+1)})$, where $H_Z = \boldsymbol{\iota}_M' \otimes \mathbf{I}_{(kp)^2}$ and $\tilde{H}_Y = (\mathbf{O}_{kp \times M(kp)^2}, \ \tilde{G}_Y)$. It then holds that

$$\text{vec}(\text{Var}[Y_{t+h}|\mathcal{I}_t]) = \tilde{H}_Z\mathbb{E}[\tilde{Z}_{t+h}|\mathcal{I}_t] - \tilde{H}_Y\mathbb{E}[\tilde{Z}_{t+h}|\mathcal{I}_t] \otimes \tilde{H}_Y\mathbb{E}[\tilde{Z}_{t+h}|\mathcal{I}_t] \tag{D.11}$$

where

$$\mathbb{E}[\tilde{Z}_{t+h}|\mathcal{I}_t] = \tilde{\Upsilon}\begin{pmatrix} \hat{\boldsymbol{\xi}}_{t|t} \otimes Y_t \otimes Y_t \\ \hat{\boldsymbol{\xi}}_{t|t} \otimes Y_t \\ \hat{\boldsymbol{\xi}}_{t|t} \end{pmatrix}$$

Taking the $[(j-1)(kp+1)+1]$-th element of $\text{vec}(\text{Var}[Y_{t+h}|\mathcal{I}_t])$ and setting $h = 1$ yields $\text{Var}[y_{j,t+1}|\mathcal{I}_t]$, the required one-step ahead forecast error variance.

E Stock Return Data for the Systemic Event Index

The online data appendix of DDLY contains the Reuters tickers of the included banks. Based on these tickers, the Yahoo Finance tickers were obtained that correspond to the same listing and daily returns were downloaded using the R package `yfR`.[30] For 86 of the 96 banks, the data were available in sufficient quantity. For two of these banks, Banco Bradesco (Brazil) and Woori Finance Holdings (South Korea), the listing of the New York Stock Exchange was taken instead of the local one due to data availability. In addition to these 86 banks, the returns for five banks were downloaded manually from `Investing.com`. These consist of Credit Suisse Group (Switzerland), Sberbank Rossii (Russia), SunTrust Banks (United States), Türkiye İş Bankası (Turkey) and Shizuoka Bank (Japan).

For each trading day, banks for which the returns are available are used in determining the value of the SE index. A daily return of exactly zero was treated as a missing data point, as this mostly coincided with opening, closing and high prices staying the same vis-à-vis those of the previous day or with a trading volume of zero. Although, e.g. the cancellation of a trading day need not be uninformative for a systemic event, its quantification in this regard is ambiguous at best and hence not pursued. Missing data points to some extent tend to occur together, which could indicate some sample selection bias.

The banks for which data were unavailable (or only available in limited quantity) were Bank of Yokohama (Japan), Pohjola Bank (Finland), Dexia (Belgium), Banco Popular Español (Spain) and Banco Espírito Santo (Portugal), each of which are/were relatively small banks with respect to total assets. Dexia experienced severe stress during the financial crisis and received state aid in the end of 2008, which sparked an extensive series of restructurings, further aid and bailouts that dismantled a major part of the bank.[31] As a result, the stock price was reduced to near zero and trading activity became highly irregular from mid-2012 onwards in the Yahoo Finance data. Bank of Yokohama was, to my knowledge, not subject to significant events, at least compared to the other banks, during the sample period. The same holds for Pohjola Bank. Banco Popular Español got acquired in 2017 by Banco Santander after the bank was unable to deal with the aftermath of the financial crisis.[32] Banco Espírito Santo collapsed in August of 2014 after the unravelling of dubious financial structures of the owners.[33] Thus, the five banks that are not included in the construction of the SE index are not expected to have had their lowest returns simultaneously, thus somewhat alleviating concerns regarding sample selection bias with respect to which banks are part of the sample.

[30] Perlin, M.S. (2023, 16th February). *yfR: Downloads and Organizes Financial Data from Yahoo Finance.* Retrieved on 2024, 5th May from `https://rdrr.io/cran/yfR/`

[31] See Pignal, S. (2011, 10th October). Dexia break-up deal reached. *Financial Times.* Retrieved from `https://www.ft.com/content/5235cfb9-d12f-38c2-9a5a-372b51ee961c` for a timeline.

[32] Banco Popular fails and is bought by Santander. (2017, 10th June). *Economist*, *423*(9044), 73 et seq.

[33] See Kowsmann, P., Enrich, D. & Patrick, M. (2014, 12th August). Behind the Collapse of Portugal's Espírito Santo Empire. *Wall Street Journal.* Retrieved from `https://www.wsj.com/articles/behind-the-collapse-of-portugals-espirito-santo-empire-1407879423` for a timeline.

F Unit Root Tests of the Logarithms of the Volatility Series

Ng and Perron (2001) show that combining an augmented Dickey-Fuller (ADF) test on data that are locally detrended using generalised least squares (GLS) estimates of deterministic components, introduced by Elliott, Rothenberg and Stock (1996), in combination with a modified version of the AIC criterion (MAIC) for selecting the lag order of the ADF-GLS test regression yields a power and size that are desirable compared to other unit-root tests.

The following exposition of the ADF-GLS test is based on Ng and Perron (2001). This test and the MAIC are derived under Gaussian error terms. Let x_t be a univariate time series $\{x_t\}_{t=0}^{T}$ and define $(x_1^{\bar{\alpha}},\ x_t^{\bar{\alpha}}) = (x_1,\ (1-\bar{\alpha}L)x_t)$, $t = 1,\ \ldots,\ T$, where $\bar{\alpha} = 1 + \frac{\bar{c}}{T-1}$ and L is the lag operator. For the ADF-based GLS test, $\bar{c}$ is set to -13.5 based on the results of Elliott et al. (1996). The GLS detrended series is defined as $\tilde{y}_t = y_t - \hat{\psi}' z_t$, where $\hat{\psi}$ minimises $(y^{\bar{\alpha}} - \psi' z^{\bar{\alpha}})'(y^{\bar{\alpha}} - \psi' z^{\bar{\alpha}})$, where z_t is a vector of deterministic components, here a constant term. The ADF-GLS test then is a t-test on the coefficient ρ with the null-hypothesis that it is equal to zero in the following regression

$$\Delta_1 \tilde{y}_t = +\rho \tilde{y}_{t-1} + \sum_{i=1}^{p} \delta_i \Delta_1 \tilde{y}_{t-i} + \varepsilon_t \tag{F.1}$$

where Δ_1 is the first-difference operator. p is chosen to minimise

$$\text{MAIC} = \log\left(\hat{\sigma}^2\right) + \frac{2(p + \hat{\sigma}^{-1}\hat{\rho}\sum_{t=p_{\max}+1}^{T} \tilde{y}_{t-1}^2)}{T - p_{\max}} \tag{F.2}$$

where $\hat{\sigma}^2 = (T - p_{\max})^{-1} \sum_{t=p_{\max}+1}^{T} \hat{\varepsilon}_t^2$ and $p_{\max}$ is the highest considered lag order which is set a priori. $p_{\max}$ is determined according to the formula $\lfloor 12(0.01T)^{\frac{1}{4}} \rceil$ as in Ng and Perron (2001), which is equal to 27 for the dataset used in this thesis.

G Supplementary Results

The results in this Appendix are listed by the subsection to which they are supplementary.

Section 5.1

Figure 19 displays the estimated parameters for the MS-VAR in which the three groups of rows of the autoregressive parameter matrices corresponding to the regions are governed by their own Markov chains. It can be seen that Φ_1 corresponds to the first two autoregressive matrices of the MS(4)-VAR(1) and that Φ_2 corresponds to the latter two. For the error term correlation matrices, it can be seen that the dispersed structure, also seen for the third and fourth of the error term correlation matrices of the MS(4)-VAR(1), is accompanied by the American and Asian banks being in their second, low-volatility regime. For the constant terms, the difference between regimes is most apparent for the American banks, as can be seen in the third, fifth, seventh and eighth row of the constant terms. Moreover, the first regime is now not the regime for which the error term variances are clearly higher. The estimated transition probabilities and unconditional probabilities are included below as equation (G.1) and the inferred regimes are displayed in Figure 20, which qualitatively are very similar to those of the MS(4)-VAR(1).

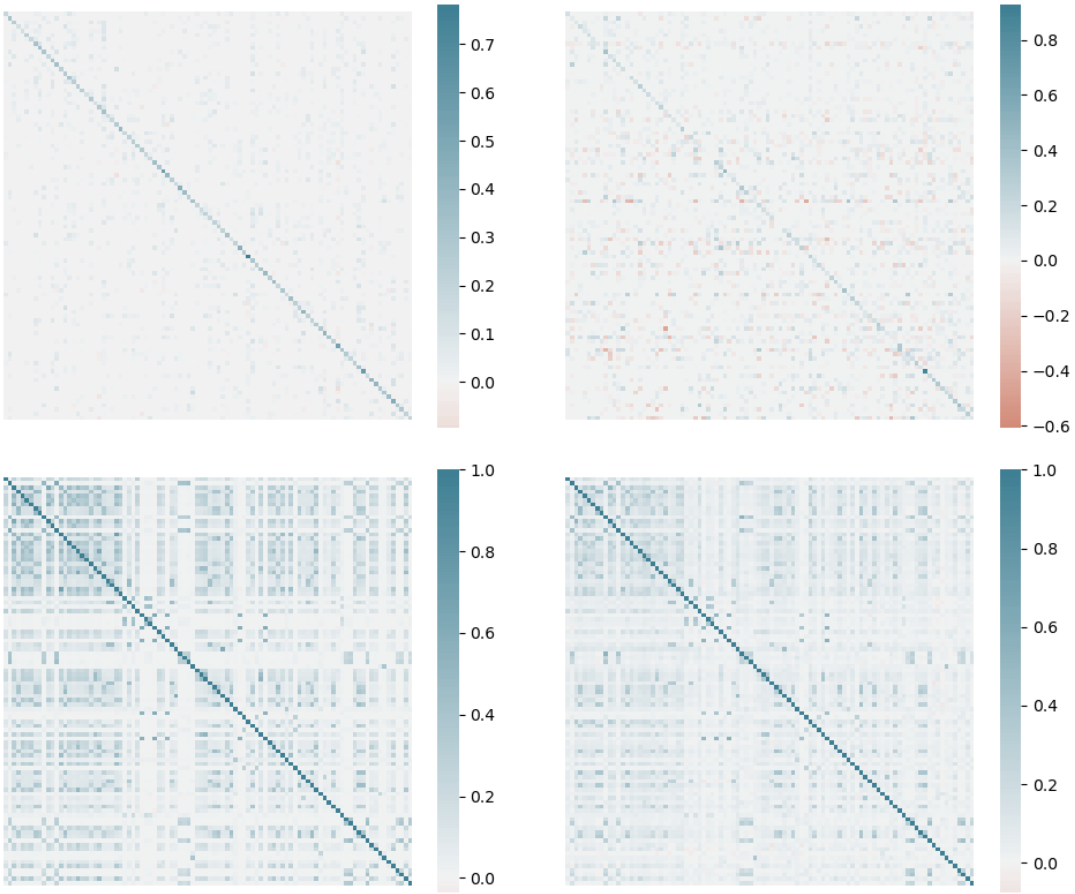


Figure 19: (1/2). Estimated autoregressive parameters (Φ_1 top-left and Φ_2 top-right) and error term correlation matrices ($\mathbf{R}_1$ middle-left and $\mathbf{R}_2$ middle-right).

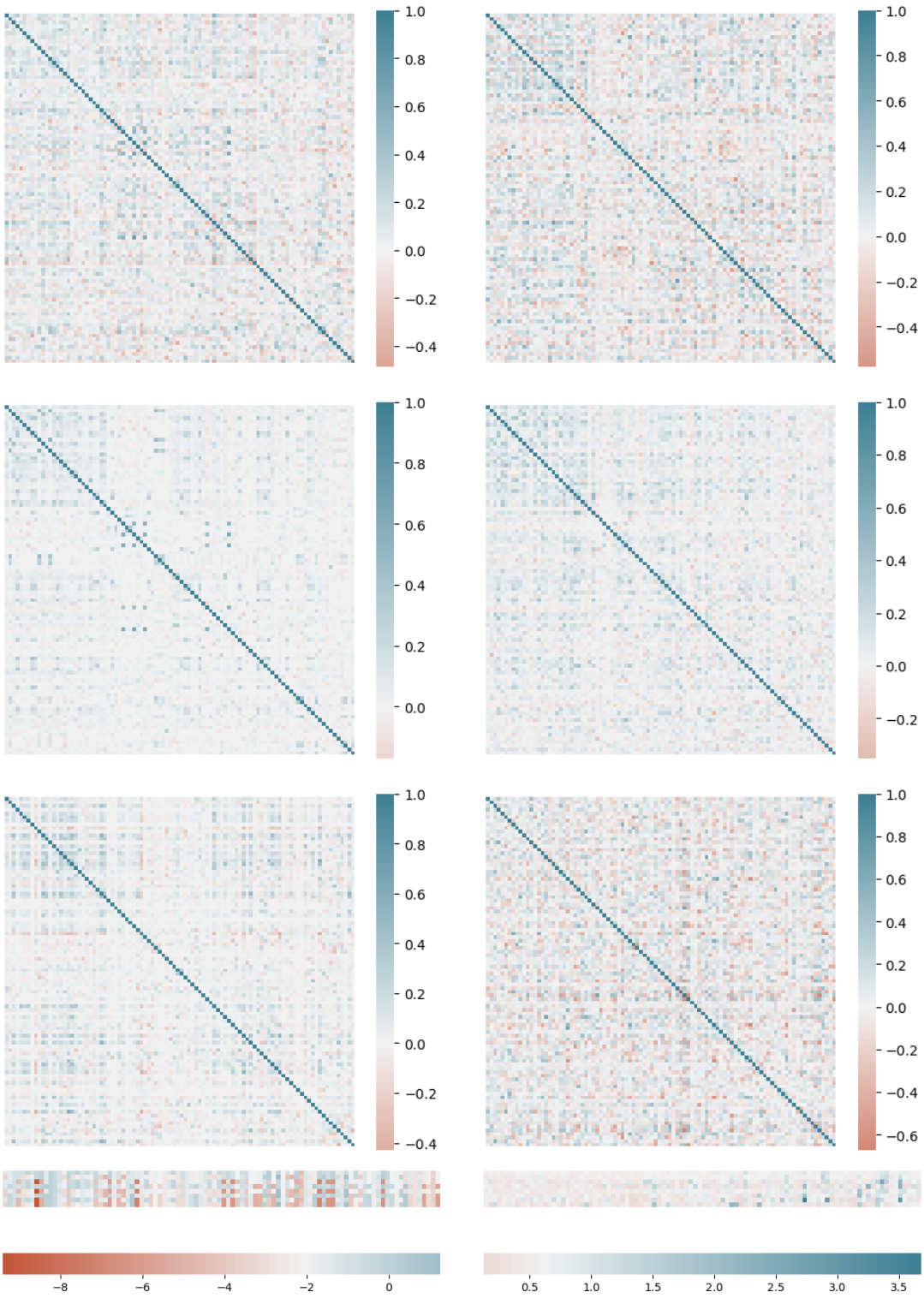


Figure 19: (2/2). Estimated error term correlation matrices ($\mathbf{R}_3$ top-left, $\mathbf{R}_4$ top-right, $\mathbf{R}_5$ upper middle-left, $\mathbf{R}_6$ upper middle-right, $\mathbf{R}_7$ lower middle-left, $\mathbf{R}_8$ lower middle-right), constant terms (bottom-left) and error term variances (bottom-right).

$$\hat{\mathbf{P}} = \begin{pmatrix} 0.996 & 0.004 & 0.000 & 0.000 & 0.004 & 0.038 & 0.031 & 0.000 \\ 0.003 & 0.926 & 0.077 & 0.000 & 0.046 & 0.179 & 0.031 & 0.500 \\ 0.000 & 0.001 & 0.692 & 0.125 & 0.012 & 0.026 & 0.031 & 0.000 \\ 0.000 & 0.003 & 0.000 & 0.438 & 0.021 & 0.013 & 0.031 & 0.000 \\ 0.000 & 0.022 & 0.077 & 0.250 & 0.808 & 0.218 & 0.031 & 0.000 \\ 0.001 & 0.030 & 0.000 & 0.125 & 0.042 & 0.487 & 0.031 & 0.250 \\ 0.001 & 0.007 & 0.115 & 0.063 & 0.046 & 0.026 & 0.031 & 0.250 \\ 0.000 & 0.006 & 0.038 & 0.000 & 0.021 & 0.013 & 0.031 & 0.000 \end{pmatrix} \quad \hat{\pi} = \begin{pmatrix} 0.628 \\ 0.235 \\ 0.009 \\ 0.081 \\ 0.027 \\ 0.011 \\ 0.004 \end{pmatrix} \tag{G.1}$$

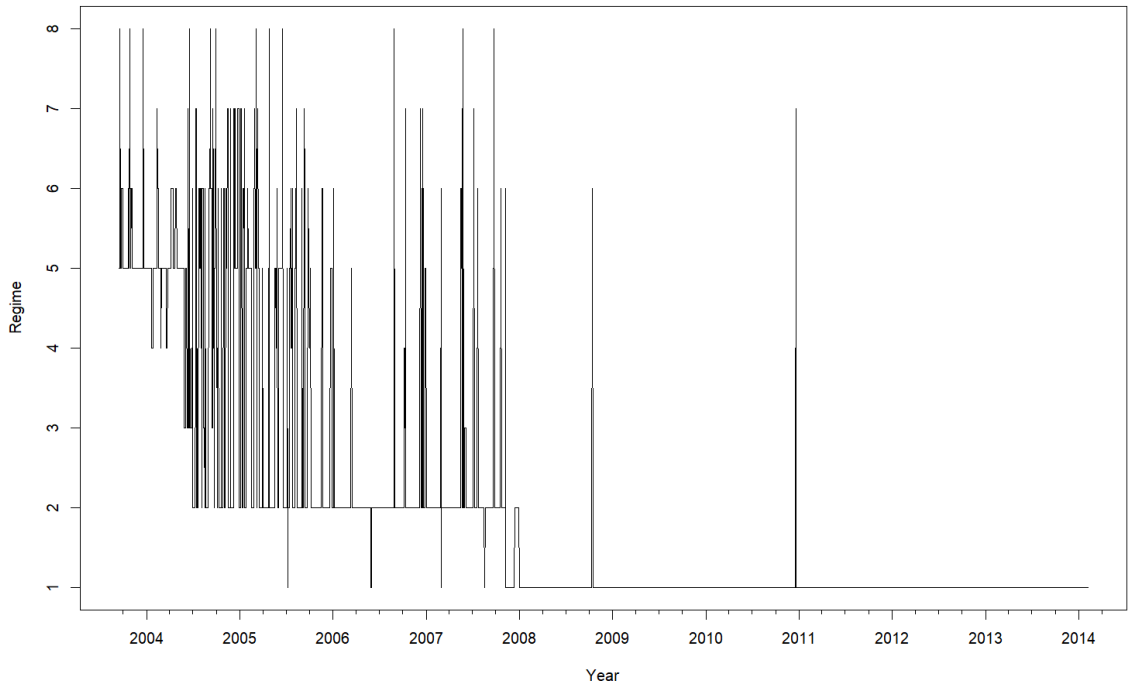


Figure 20: Regime by highest smoothed probability of Extension II.

Section 5.2

Next, I discuss the effects of a different size of the rolling window. It can be seen that qualitatively, the trajectory of the SI remains the same. As expected, the VAR with a rolling window size of 100 is most responsive and returns to its trend level most swiftly. This can be seen best after the increase in the federal funds rate of the 10^{th} of May 2006, where the rolling window with a size of 150 days and 200 days, the SI decreases 50 days and 100 days later respectively than the 100-day SI. In general, the other SIs are within the 95% confidence intervals, indicative of the results being quantitatively similar too.

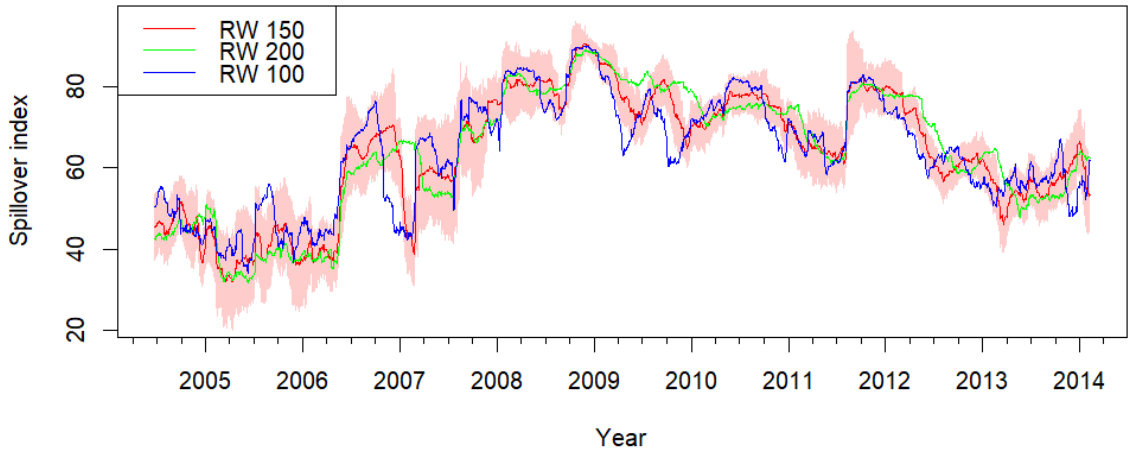
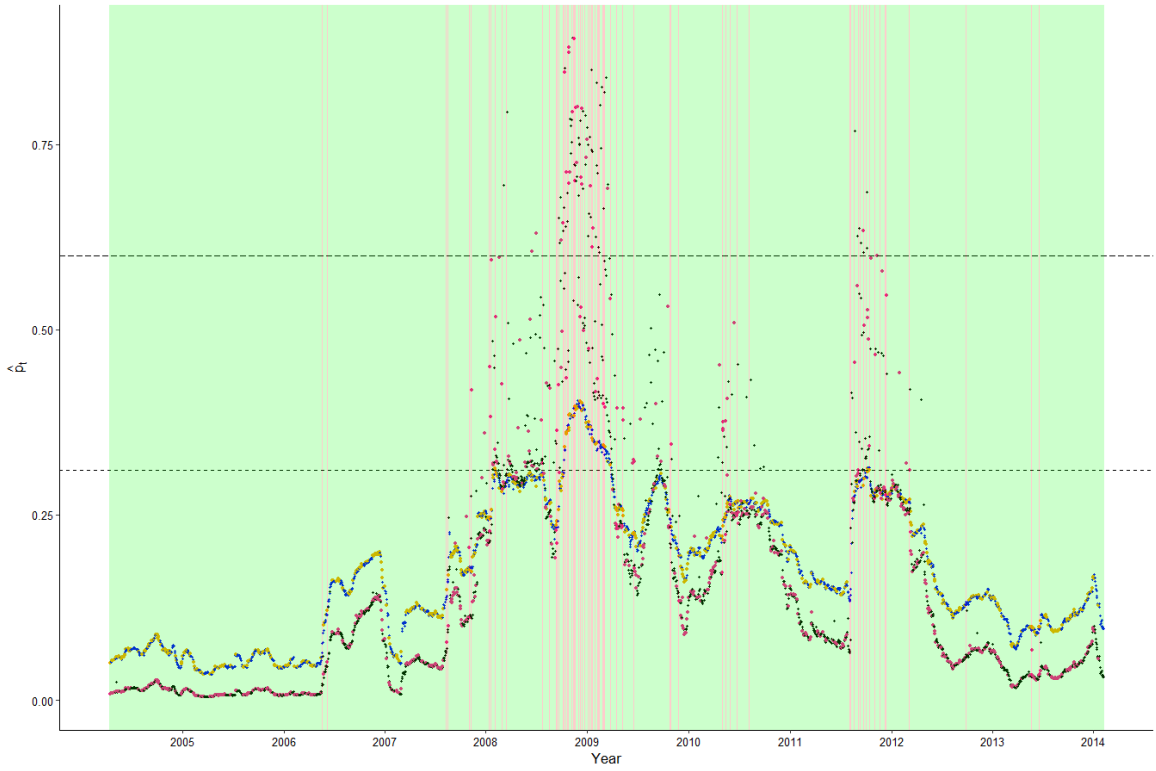


Figure 21: Dynamic SI for multiple sizes of the rolling window (RW) with 95% confidence intervals for the 150-day RW.

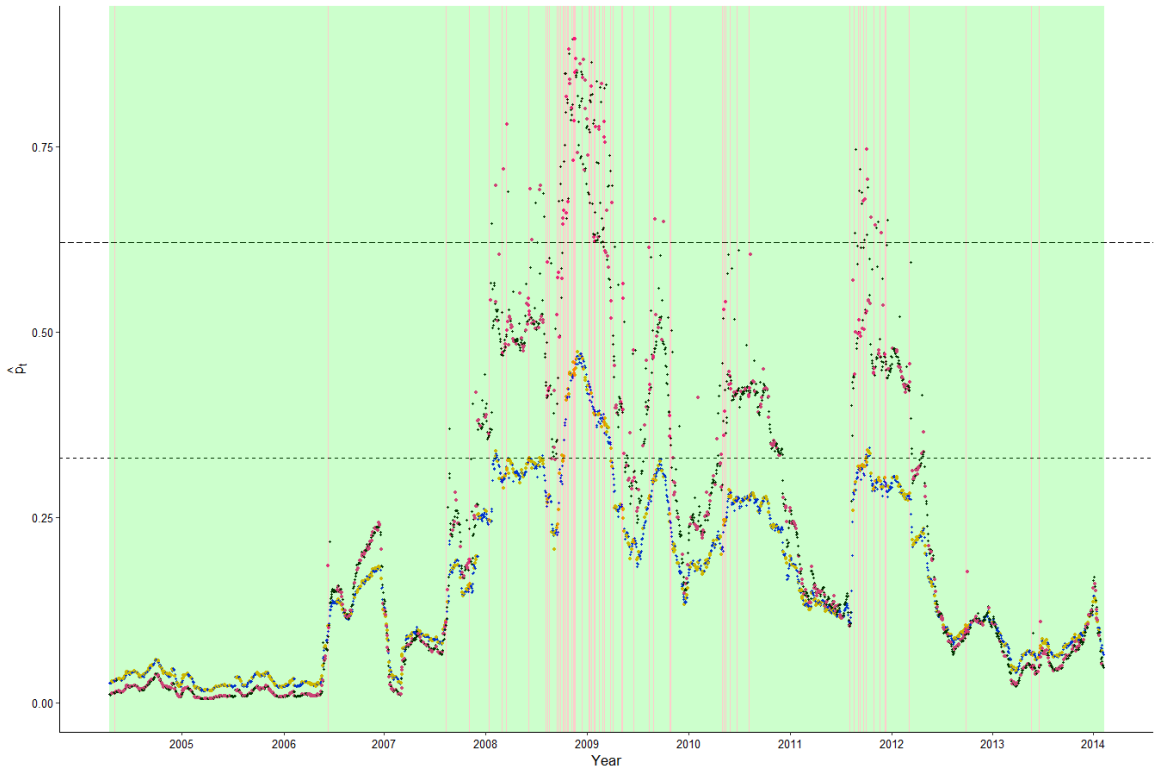
Section 5.4

For the other values of l, the estimated probabilities, thresholds and SE index are visualised in the plots of Figure 22.



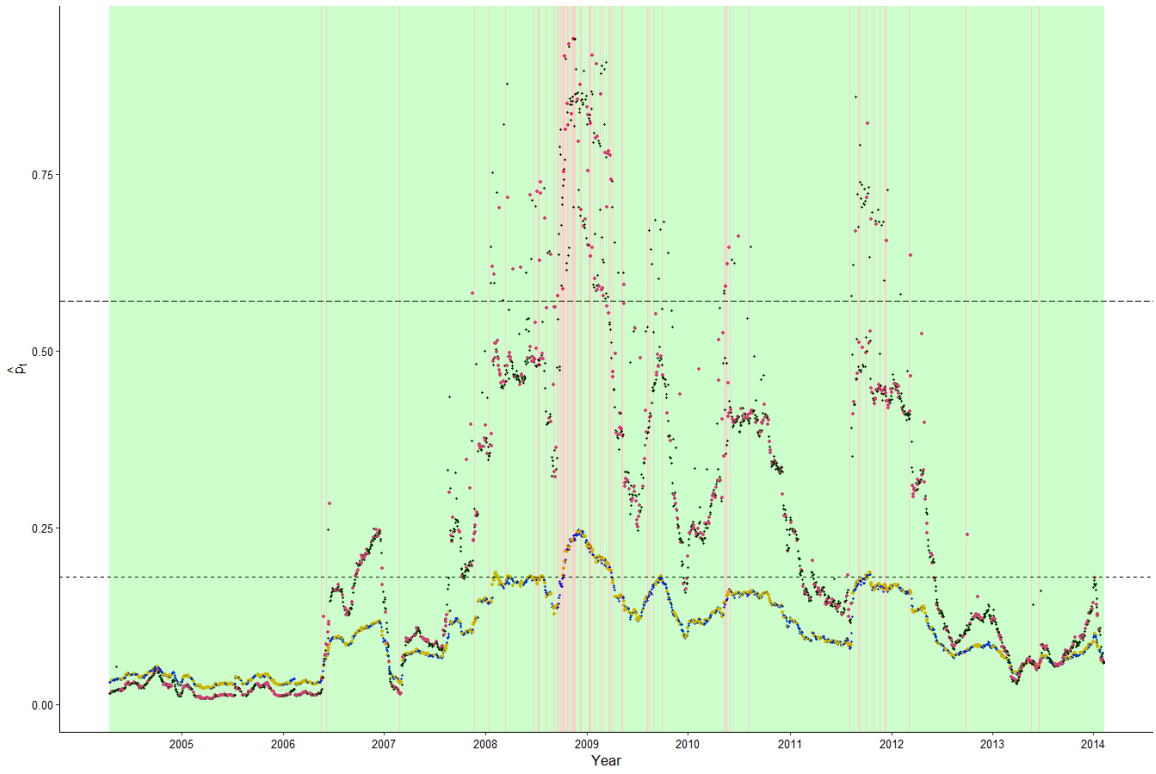
$l = 2$

Figure 22: (1/6). Estimated probabilities and classifications of an SE where the value of l is indicated at each plot. The background is red for periods that constitute an SE and green otherwise. The dashed and long-dashed lines are the thresholds for the SI and $\mathbb{G}_\alpha$ respectively. Black and blue dots are estimated probabilities that correspond to an observation of the training set for the SI and $\mathbb{G}_\alpha$ respectively. The pink and orange dots are estimated probabilities that correspond to an observation of the test set for the SI and $\mathbb{G}_\alpha$ respectively.



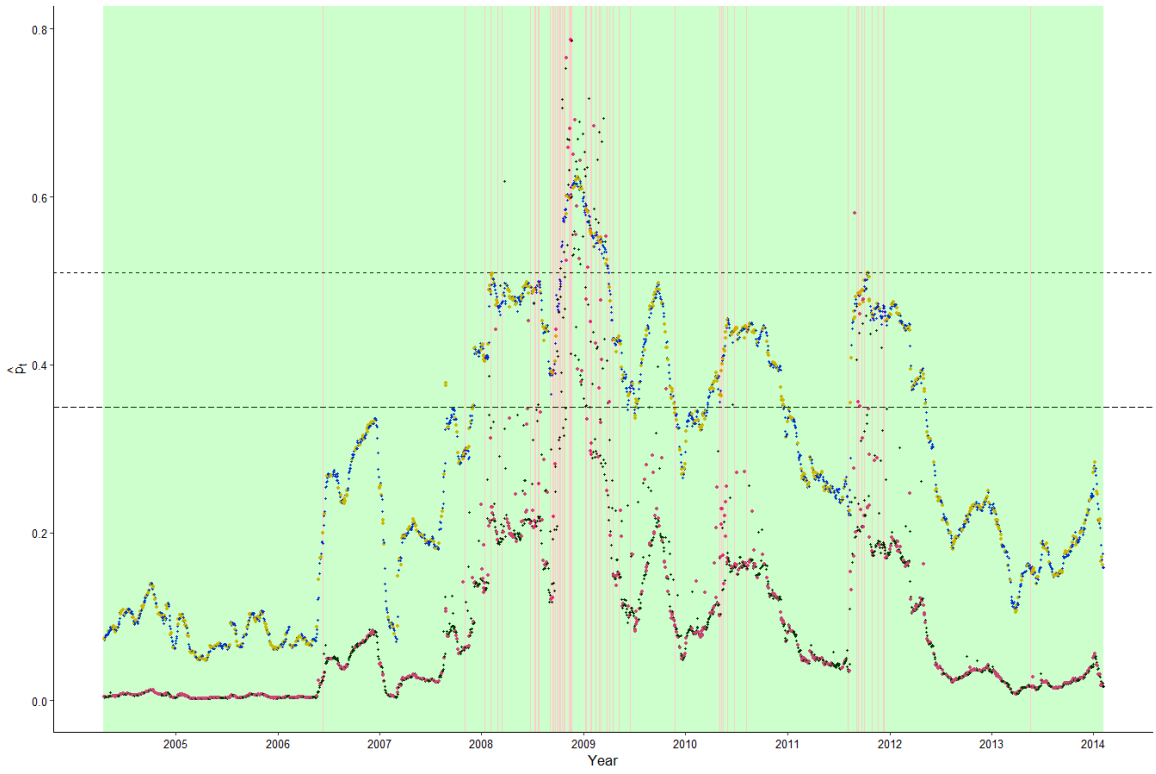
$l = 3$

Figure 22: (2/6). Estimated probabilities and classifications of an SE where the value of l is indicated at each plot. The background is red for periods that constitute an SE and green otherwise. The dashed and long-dashed lines are the thresholds for the SI and $\mathbb{G}_\alpha$ respectively. Black and blue dots are estimated probabilities that correspond to an observation of the training set for the SI and $\mathbb{G}_\alpha$ respectively. The pink and orange dots are estimated probabilities that correspond to an observation of the test set for the SI and $\mathbb{G}_\alpha$ respectively.



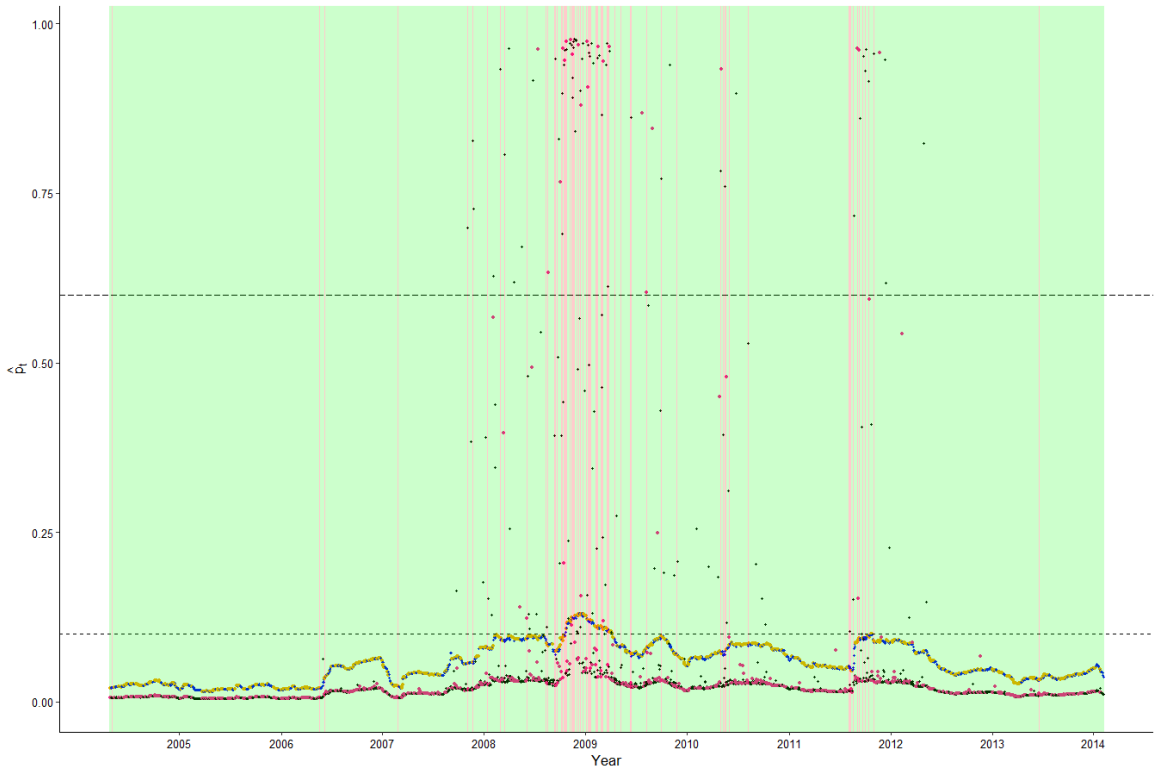
$l = 4$

Figure 22: (3/6). Estimated probabilities and classifications of an SE where the value of l is indicated at each plot. The background is red for periods that constitute an SE and green otherwise. The dashed and long-dashed lines are the thresholds for the SI and $\mathbb{G}_\alpha$ respectively. Black and blue dots are estimated probabilities that correspond to an observation of the training set for the SI and $\mathbb{G}_\alpha$ respectively. The pink and orange dots are estimated probabilities that correspond to an observation of the test set for the SI and $\mathbb{G}_\alpha$ respectively.



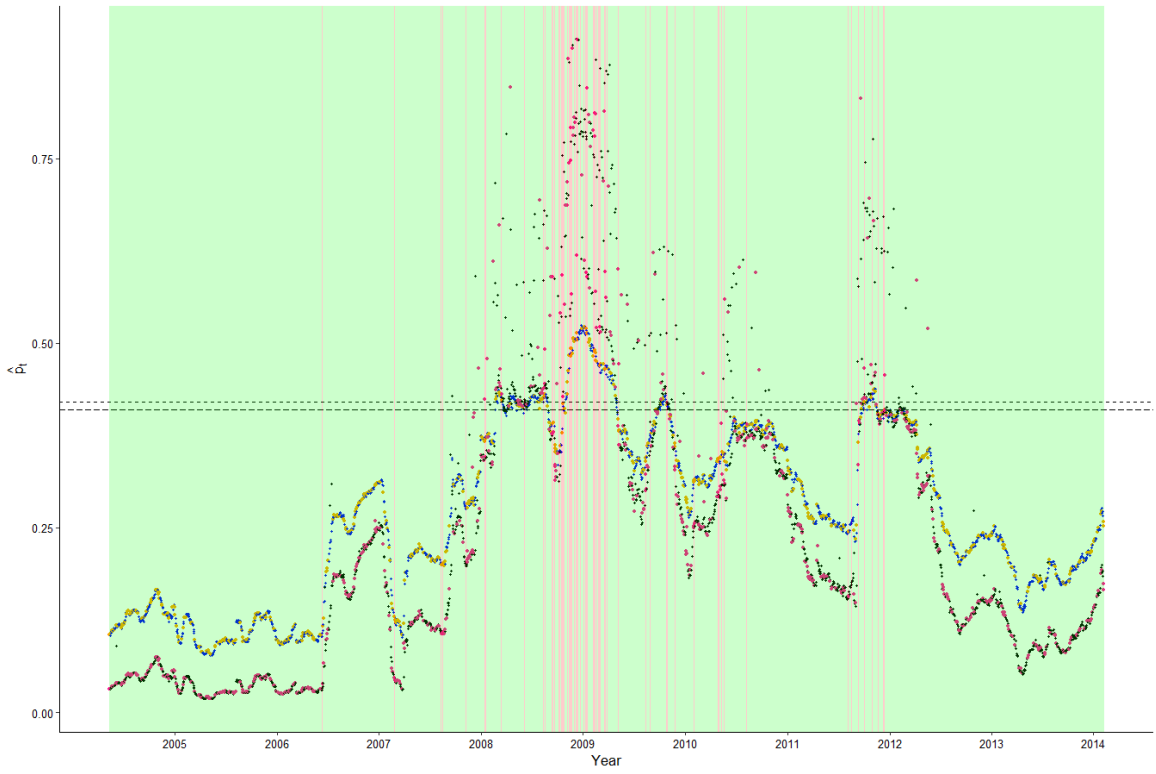
$l = 5$

Figure 22: (4/6). Estimated probabilities and classifications of an SE where the value of l is indicated at each plot. The background is red for periods that constitute an SE and green otherwise. The dashed and long-dashed lines are the thresholds for the SI and $\mathbb{G}_\alpha$ respectively. Black and blue dots are estimated probabilities that correspond to an observation of the training set for the SI and $\mathbb{G}_\alpha$ respectively. The pink and orange dots are estimated probabilities that correspond to an observation of the test set for the SI and $\mathbb{G}_\alpha$ respectively.



$l = 10$

Figure 22: (5/6). Estimated probabilities and classifications of an SE where the value of l is indicated at each plot. The background is red for periods that constitute an SE and green otherwise. The dashed and long-dashed lines are the thresholds for the SI and $\mathbb{G}_\alpha$ respectively. Black and blue dots are estimated probabilities that correspond to an observation of the training set for the SI and $\mathbb{G}_\alpha$ respectively. The pink and orange dots are estimated probabilities that correspond to an observation of the test set for the SI and $\mathbb{G}_\alpha$ respectively.



$l = 22$

Figure 22: (6/6). Estimated probabilities and classifications of an SE where the value of l is indicated at each plot. The background is red for periods that constitute an SE and green otherwise. The dashed and long-dashed lines are the thresholds for the SI and $\mathbb{G}_\alpha$ respectively. Black and blue dots are estimated probabilities that correspond to an observation of the training set for the SI and $\mathbb{G}_\alpha$ respectively. The pink and orange dots are estimated probabilities that correspond to an observation of the test set for the SI and $\mathbb{G}_\alpha$ respectively.