Granger Causality and Regime Inference in Bayesian Markov-Switching VARs☆

Matthieu Droumaguet[b], Anders Warne[a], Tomasz Woźniak[c,b]

[a]*Directorate General Research, European Central Bank, Kaiserstrasse 29, 60311 Frankfurt am Main, Germany*
[b]*Department of Economics, European University Institute, Via della Piazzuola 43, 50133 Florence, Italy*
[c]*Department of Economics, University of Melbourne, Arts West Building, Victoria 3010, Australia*

Abstract

Recent economic developments have shown the importance of spillover and contagion effects in financial markets as well as in macroeconomic reality. Such effects should potentially be modeled and analyzed taking into account such properties of time series as changes of parameter values over time and heteroskedasticity. We derive restrictions for Granger noncausality within the framework of Markov-switching Vector Autoregressive Models. Due to the complicated structure and the nonlinearity of the resulting restrictions, classical tests have limited use. We, therefore, choose a Bayesian approach to evaluate the hypotheses of noncausality. The inference consists of a novel Block Metropolis-Hastings sampling algorithm for estimation of the restricted models, and of standard methods of computing the Posterior Odds Ratio. As an empirical illustration we analyze the system of the US money and income variables. We found that while the the past information about the money aggregate M1 is dispensable for the forecasting of the conditional mean of income, it crucial for predicting the next period's state of the economy.

Keywords: Granger Causality, Regime Inference, Markov-switching Models, Posterior Odds Ratio, Block Metropolis-Hastings Sampling
JEL classification: C11, C12, C32, C53, E32

1. Introduction

The concept of Granger causality was introduced by Granger (1969) and Sims (1972). One variable does not Granger-cause some other variable, if past and current information about the former cannot improve the forecast of the latter. Knowledge of Granger causal relations allows a researcher to formulate an appropriate model and obtain a good forecast of values of interest. But what is even more important, a Granger causal relation, once established, informs us that past observations of one variable have a significant effect on the forecast value of the other, delivering crucial information about the relations between economic variables.

Note that this concept refers to the forecasting of variables, in contrast to the causality concept based on *ceteris paribus* effects attributed to Rubin (1974) (for the comparison of the two concepts used in econometrics, see e.g. Lechner, 2011). We also underline that in general Granger causality does not relate to any causal relation implied by structural economic theories either. Such correspondence has only been shown for linear Gaussian models by White & Lu (2010).

☆This paper was presented at the Econometric Society Australasian Meeting 2012 in Melbourne and at the 22nd EC^2 Conference: *Econometrics for Policy Analysis: after the Crisis and Beyond* in Florence in December 2011, as well as during seminars at the Australian National University, the National Bank of Poland and Deutsche Bundesbank. The authors thank Joshua Chan, Peter R. Hansen, Andrzej Kocięcki, Helmut Lütkepohl, Massimiliano Marcellino, Mateusz Pipień, John Stachurski, Rodney Strachan for their useful comments on the paper.

Email addresses: matthieu.droumaguet@eui.eu (Matthieu Droumaguet), anders.warne@ecb.europa.eu (Anders Warne), tomasz.wozniak@unimelb.edu.au (Tomasz Woźniak)
URL: http://www.texlips.net/awarne/ (Anders Warne), http://bit.ly/tomaszw (Tomasz Woźniak)

Among the parametric time series models that have been analyzed for Granger causality of different types are: a family of Vector Autoregressive Moving Average (VARMA) models (see Boudjellaba, Dufour & Roy, 1994, and references therein), the Logistic Smooth Transition Vector Autoregressive (LST-VAR) model (Christopoulos & León-Ledesma, 2008), some models from the family of Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models (Comte & Lieberman, 2000; Woźniak, 2011; Woźniak, 2012). Note that all these works analyzed *one period ahead* Granger noncausality (see Lütkepohl, 1993; Lütkepohl & Burda, 1997; Dufour, Pelletier & Renault, 2006, for h periods ahead inference in VAR models).

In this work, we make Granger-causal inference on the Markov-switching Vector Autoregressive (MS-VAR) models. Considering the econometric specification of these models, in which the values of the parameters change over time according to a hidden Markov process with a discrete state space, the analysis is split in two parts. First, we make the regime inference on the hidden Markov process. We derive conditions under which one variable is dispensable for the one-period-ahead forecast of the regime of a second variable. Further, we derive the parameter conditions for Granger noncausality when the parameter changes are treated entirely as endogenous. Such analysis overcomes the stereotypical view on the MS-VAR models that treats them as a way of modeling the parameters of VAR model changing over time. We show that, since the parameter changes are modeled endogenously, the regime inference constitute an integral part of the Granger causality analysis.

The derived restrictions on the parameters of the model may be tested. They have a form of several sets of potentially nonlinear restrictions. Therefore, they imply, first of all, that a hypothesis on noncausality is represented by several restricted models, and secondly, that each of the model is restricted by nonlinear restrictions on the parameters. Both of these features cause problem with classical testing of the hypotheses.

The second contribution of this work is a Bayesian inference that allows the testing of all the hypotheses of Granger noncausality, as well as of the independence of the hidden Markov processes. None of the existing classical solutions, that we describe in Section 5, to the problem of testing such restrictions on parameters is easily applicable to Markov-switching VAR models. The proposed approach consists of a Bayesian estimation of the unrestricted model, allowing for Granger causality, and of the restricted models, where the restrictions represent hypotheses of noncausality and regime independence. For this purpose, we construct a novel block Metropolis-Hastings sampling algorithm that allows for restricting the models. Having estimated the models, we compare competing hypotheses, represented by the unrestricted and the restricted models, with standard Bayesian methods using Posterior Odds Ratios and Bayes factors.

The main advantage of our approach is that we can test a hypothesis represented by several restricted models (with nonlinear restrictions). Thus, the analysis of causal relations between variables is profound and potentially informative. Other advantages include an effect of adopting Bayesian inference. First, the Posterior Odds Ratio method gives arguments *in favour of* the hypotheses, as posterior probabilities of the competing hypotheses are compared. In consequence, all the hypotheses are treated symmetrically. Secondly, our estimation procedure combines and improves the existing algorithms restricting the models, but it also preserves the possibility of using different methods for computing the marginal density of data necessary to compute the Posterior Odds Ratio.

As potential applications of the testing procedure, we indicate macroeconomic as well as financial time series. In particular, recent financial turmoil and the following global recession are interesting periods for analysis. There exist many applied studies presenting evidence that these events have the nature of switching the regime. Taylor & Williams (2009), on the example of Libor-OIS and Libor-Repo spreads, being an approximation for counterpart risk, present how different the perception of the risk by agents on the financial market was, first, starting from August 2007 and then, even more, from October 2008. Further, Diebold & Yilmaz (2009) show how different behaviors characterize return spillovers and volatility spillovers for stock exchange markets. These two studies clearly indicate that the financial data should be analyzed in terms of Granger causality with a model that allows for changes in regimes, such as a Markov-switching model.

For macroeconomic time series, the motivation for using Markov-switching models comes mainly from the business cycle analysis, as in Hamilton (1989). It is important to know whether variables have different impacts on other variables during the expansion and recession periods. Still, allowing for higher number of states than two may allow a more detailed analysis of the interactions between variables within the cycles.

Psaradakis, Ravn & Sola (2005) used the Markov-switching VAR models to analyze, the so called temporary Granger causality within the Money-Output system. They condition their causality analysis on realizations of the Hidden-Markov process. Our approach consists of choosing a Markov-switching VAR model specification which is best supported by the data, and then restricting it according to the derived restrictions. This approach takes into account the two sources of relations between the variables: first, having a source in linear relations modeled with the VAR model, and second, taking into consideration the fact that all of the variables are used to forecast the future probabilities of the states. Therefore, in the setting considered in the present paper, Granger noncausality is not conditioned on the past realizations of the hidden Markov process.

The remaining part of the paper is organized as follows. In Section 2 we present the model and the Bayesian estimation of the unrestricted model. The notation, the definitions for Granger noncausality and regime independence, and the model for the causality inference are presented in Section 3. Section 4 introduces the restrictions for the considered relations between variables. Section 5 presents discussion and critique of classical methods of testing restrictions for Granger noncausality in different multivariate models. The discussion is followed by a proposal of solution of the testing problem. First, the Posterior Odds Ratio is defined, and then the algorithm for estimating the restricted models is discussed. The block Metropolis-Hastings algorithm is described in detail in Section 6. Section 7 gives empirical illustration of the methodology, using the example of the money-income system of variables in the USA. The data support the hypothesis of Granger noncausality from money to income. However, money in found important for forecasting the future state of the economy. Section 8 concludes. All the proofs are presented in the Mathematical Appendix A, whereas the details of the estimation results are reported in Appendix B and in Appendix C.

2. A Markov-Switching Vector Autoregressive Model

Model. Let $\mathbf{y} = (y_1, \ldots, y_T)'$ denote a time series of T observations, where each y_t is a N-variate vector for $t \in \{1, \ldots, T\}$, taking values in a sampling space $\mathbf{Y} \subset \mathbb{R}^N$. $\mathbf{y}$ is a realization of a stochastic process $\{Y_t\}_{t=1}^T$. We consider a class of parametric finite Markov mixture distribution models in which the stochastic process Y_t depends on the realizations, s_t, of a hidden discrete stochastic process S_t with finite state space $\{1, \ldots, M\}$. Such a class of models has been introduced in time series analysis by Hamilton (1989). Conditioned on the state, s_t, and realizations of $\mathbf{y}$ up to time $t-1$, $\mathbf{y}_{t-1}$, y_t follows an independent identical normal distribution. A conditional mean process is a Vector Autoregression (VAR) model in which an intercept, μ_{s_t}, as well as lag polynomial matrices, $A_{s_t}^{(i)}$, for $i = 1, \ldots, p$, and covariance matrices, Σ_{s_t}, depend on the state $s_t = 1, \ldots, M$.

$$y_t = \mu_{s_t} + \sum_{i=1}^{p} A_{s_t}^{(i)} y_{t-i} + \epsilon_t, \tag{1}$$

$$\epsilon_t \sim i.i.\mathcal{N}(\mathbf{0}, \Sigma_{s_t}), \tag{2}$$

for $t = 1, \ldots, T$. We set the vector of initial values $\mathbf{y}_0 = (y_{p-1}, \ldots, y_0)'$ to the first p observations of the available data.

S_t is assumed to be an irreducible aperiodic Markov chain starting from its ergodic distribution $\pi = (\pi_1, \ldots, \pi_M)$, such that $\Pr(S_0 = i|\mathbf{P}) = \pi_i$. Its properties are sufficiently described by the $(M \times M)$ transition probabilities matrix:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1M} \\ p_{21} & p_{22} & \cdots & p_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M1} & p_{M2} & \cdots & p_{MM} \end{bmatrix},$$

in which an element, p_{ij}, denotes the probability of transition from state i to state j, $p_{ij} = \Pr(s_{t+1} = j|s_t = i)$. The elements of each row of matrix $\mathbf{P}$ sum to one, $\sum_{j=1}^{M} p_{ij} = 1$.

The model, conditioned on the state s_t, models a current vector of observations, y_t, with an intercept, μ_{s_t}, and a linear function of its lagged values up to p periods backwards. The linear relation is captured by matrices of the lag polynomial $A^{(i)}_{s_t}$, for $i = 1, \ldots, p$. The parameters of the VAR process, as well as the covariance matrix Σ_{s_t}, change with time, t, according to discrete valued hidden Markov process, s_t. Such a formulation of the model is called, according to the taxonomy of Krolzig (1997), MSIAH-VAR(p). These changes in parameter values introduce nonlinear relationships between variables. Consequently, the inference about interactions between variables must consider the linear and nonlinear relations; this is the subject of the analysis in Section 4.

Complete-data likelihood function. Let $\theta \in \Theta \subset \mathbb{R}^k$ be a vector of size k, collecting parameters of the transition probabilities matrix $\mathbf{P}$ and all the state-dependent parameters of the VAR process, θ_{s_t}: μ_{s_t}, $A^{(i)}_{s_t}$, Σ_{s_t}, for $s_t = 1, \ldots, M$ and $i = 1, \ldots, p$. As stated by Frühwirth-Schnatter (2006), the complete-data likelihood function is equal to the joint sampling distribution $p(\mathbf{S}, \mathbf{y}|\theta)$ for the complete data $(\mathbf{S}, \mathbf{y})$ given θ, where $\mathbf{S} = (s_1, \ldots, s_T)'$. This distribution is now considered to be a function of θ for the purpose of estimating the unknown parameter vector θ. It is further decomposed into a product of a conditional distribution of $\mathbf{y}$ given $\mathbf{S}$ and θ, and a conditional distribution of $\mathbf{S}$ given θ:

$$p(\mathbf{S}, \mathbf{y}|\theta) = p(\mathbf{y}|\mathbf{S}, \theta)p(\mathbf{S}|\theta). \tag{3}$$

The former is assumed to be a conditional normal distribution function of ϵ_t, for $t = 1, \ldots, T$, given the states, s_t, with the mean equal to a vector of zeros and Σ_{s_t} as the covariance matrix:

$$p(\mathbf{y}|\mathbf{S}, \theta) = \prod_{t=1}^{T} p(y_t|\mathbf{S}, \mathbf{y}^{t-1}, \theta) = \prod_{t=1}^{T} (2\pi)^{-K/2} |\Sigma_{s_t}|^{-1/2} \exp\left\{-\frac{1}{2}\epsilon_t' \Sigma_{s_t}^{-1} \epsilon_t\right\}. \tag{4}$$

The form of the latter comes from the assumptions about the Markov process and is given by:

$$p(\mathbf{S}|\theta) = p(s_0|\mathbf{P}) \prod_{i=1}^{M} \prod_{j=1}^{M} p_{ij}^{N_{ij}(\mathbf{S})}, \tag{5}$$

where $N_{ij}(\mathbf{S}) = \#\{s_{t-1} = j, s_t = i\}$ is a number of transitions from state i to state j, $\forall i, j \in \{1, \ldots, M\}$.

A convenient form of the complete-data likelihood function (3) results from representing it as a product of $M + 1$ factors. The first M factors depend on the state-specific parameters, θ_{s_t}, and the remaining one depends on the transition probabilities matrix, $\mathbf{P}$:

$$p(\mathbf{y}, \mathbf{S}|\theta) = \prod_{i=1}^{M} \left(\prod_{t:s_t=i} p(y_t|\mathbf{y}^{t-1}, \theta_i) \right) \prod_{i=1}^{M} \prod_{j=1}^{M} p_{ij}^{N_{ij}(\mathbf{S})} p(s_0|\mathbf{P}). \tag{6}$$

Classical estimation of the model consists of the maximization of the likelihood function with e.g. the EM algorithm (see Krolzig, 1997; Kim & Nelson, 1999b). For the purpose of testing Granger-causal relations between variables, we propose, however, the Bayesian inference, which is based on the posterior distribution of the model parameters θ. (For details of a standard Bayesian estimation and inference on Markov-switching models, the reader is referred to Frühwirth-Schnatter, 2006). The complete-data posterior distribution is proportional to the product of the complete-data likelihood function (6) and the prior distribution:

$$p(\theta|\mathbf{y}, \mathbf{S}) \propto p(\mathbf{y}, \mathbf{S}|\theta)p(\theta). \tag{7}$$

Prior distribution. The convenient factorization of the likelihood function (6) is maintained by the choice of the prior distribution in the following form:

$$p(\theta) = \prod_{i=1}^{M} p(\theta_i)p(\mathbf{P}). \tag{8}$$

The independence of the prior distribution of the state-specific parameters for each state and the transition probabilities matrix is assumed. This allows the possibility to incorporate prior knowledge of the researcher about the state-specific parameters of the model, θ_{s_t}, separately for each state.

For the unrestricted MSIAH-VAR(p) model, we assume the following prior specification. Each row of the transition probabilities matrix, $\mathbf{P}$, *a priori* follows an M variate Dirichlet distribution, with parameters set to 1 for all the transition probabilities except the diagonal elements $\mathbf{P}_{ii}$, for $i = 1, \ldots, M$, for which it is set to 10. Therefore, we assume that the states of an economy are persistent over time (see e.g. Kim & Nelson, 1999a). The duration of the states implied by such prior assumptions depends on the number of states. For instance, for the models with two states, $M = 2$, the prior distribution implies the duration of the states equal to eleven periods, whereas for the model with three states, $M = 3$, the duration of the states is equal to six periods.

Further, the state-dependent parameters of the VAR process are collected in vectors:

$$\beta_{s_t} = (\mu'_{s_t}, \text{vec}(A^{(1)}_{s_t})', \ldots, \text{vec}(A^{(p)}_{s_t})')',$$

for $s_t = 1, \ldots, M$. For this group of the parameters we assume the Litterman prior introduced by Doan, Litterman & Sims (1983) and Litterman (1986). *A priori* these parameters follow a $(N + pN^2)$-variate Normal distribution, with mean equal to a vector of zeros and a diagonal covariance matrix. Elements on the diagonal of the covariance matrix are determined by a set of of hyper-parameters, $(\lambda_1, \lambda_2, \lambda_3, c)'$ and are as follows:

$$(\varsigma_i \lambda_3)^2 \qquad \text{for } \mu_{i.s_t}, \tag{9a}$$

$$\left(\frac{\lambda_1}{\exp(ck - c)}\right)^2 \qquad \text{for } A^{(k)}_{ii.s_t}, \tag{9b}$$

$$\left(\frac{\lambda_1 \lambda_2}{\exp(ck - c)} \frac{\varsigma_i}{\varsigma_j}\right)^2 \qquad \text{for } A^{(k)}_{ij.s_t}, \tag{9c}$$

for $i, j = 1, \ldots, N$ and $i \neq j$, and $k = 1, \ldots, p$. We scale the variances of the prior distribution using the variances of the residuals of the autoregressions of order 17 for each of the variables ς_i, for $i = 1, \ldots, N$.

Several remarks concerning the construction of the covariance matrix of the prior distribution for autoregressive parameters are required. First, these prior distributions are the same irrespectively of the state, s_t. We set the value of the hyper-parameter responsible for shrinking of the constant terms, λ_3, to 0.033(3), as in Robertson & Tallman (1999). The overall shrinking hyper-parameter for the autoregressive parameters, λ_1, is set to 0.3 as in Adolfson, Lindé & Villani (2007). The values of the variances of the prior distributions decrease with the indicator for lag, k, according to the exponential pattern proposed by Robertson & Tallman (1999) in the denominator of equations (9b) and (9c). We also set the value of the hyper-parameter c to -0.13412, following the pattern of Robertson & Tallman for monthly data. Finally, we set the value of λ_2 to 1, since we do not want to shrink the off-diagonal parameters of the autoregressive matrices more than the diagonal parameters.

Note that the means of the prior distribution for the off-diagonal elements of matrices A_{s_t} are set to zero. If we condition our analysis on the states, this would mean that we assume *a priori* the Granger noncausality hypothesis. However, in Section 4 we show that, when the states are unknown, the inference about Granger noncausality involves many other parameters of the model. Moreover, we do not shrink these parameters more than the diagonal elements, and consequently, we do not favor the noncausality hypotheses *a priori*.

We model the state-dependent covariance matrices of the MSIAH-VAR process, decomposing each to a $N \times 1$ vector of standard deviations, σ_{s_t}, and a $N \times N$ correlation matrix, $\mathbf{R}_{s_t}$, according to the decomposition:

$$\Sigma_{s_t} = \text{diag}(\sigma_{s_t}) \mathbf{R}_{s_t} \text{diag}(\sigma_{s_t}).$$

Modeling covariance matrices using such a decomposition was proposed in Bayesian inference by Barnard, McCulloch & Meng (2000). We adapt this approach to Markov-switching models, since the algorithm easily enables the imposing of restrictions on the covariance matrix (see the details of the MCMC sampling

algorithm for the unrestricted and the restricted models in Section 6). We model the unrestricted model in the same manner, because we want to keep the prior distributions for the unrestricted and the restricted models comparable. Thus, each standard deviation $\sigma_{s_t.j}$ for $s_t = 1, \ldots, M$ and $j = 1, \ldots, N$, follows a log-Normal distribution, with a mean parameter equal to 0 and the standard deviation parameter set to 2. Finally, we assume that the prior distribution for each of the element of the correlation matrix $\mathbf{R}_{s_t}$ is a uniform distribution on the interval (a, b). For each of the correlation parameter, the values of a and b depend on all the remaining elements of the correlation matrix. a and b are chosen such that while a single correlation parameter is sampled the resulting correlation matrix is positive-definite (the implications of such a prior specification and the algorithm of choosing a and b are discussed in the original paper of Barnard et al., 2000).

To summarize, the prior specification (8) now takes the detailed form of:

$$p(\theta) = \prod_{i=1}^{M} p(\mathbf{P}_i)p(\beta_i)p(\mathbf{R}_i)\left(\prod_{j=1}^{N} p(\sigma_{i.j})\right), \tag{10}$$

where each of the prior distributions is as assumed:

$$\begin{aligned} \mathbf{P}_{i\cdot} &\sim \mathcal{D}_M(\iota_M' + 9I_{M.i\cdot}) \\ \beta_i &\sim \mathcal{N}(\mathbf{0}, \underline{V_\beta}) \\ \sigma_{i.j} &\sim log\mathcal{N}(0, 2) \\ \mathbf{R}_{i.jk} &\sim \mathcal{U}(a, b) \end{aligned}$$

for $i = 1, \ldots, M$ and $j, k = 1, \ldots, N$, where ι_M is a $M \times 1$ vector of ones and $I_{M.i\cdot}$ is i^{th} row of an identity matrix I_M. a and b are as in Algorithm 3. $\underline{V_\beta}$ is a diagonal matrix with the diagonal elements as in (9).

Posterior distribution. The structure of the likelihood function (6) and the prior distribution (10) have an effect on the form of the posterior distribution that is proportional to the product of the two densities. The form of the posterior distribution (7), resulting from the assumed specification, is as follows:

$$p(\theta|\mathbf{y}, \mathbf{S}) \propto \prod_{i=1}^{M} p(\theta_i|\mathbf{y}, \mathbf{S})p(\mathbf{P}|\mathbf{y}, \mathbf{S}). \tag{11}$$

It is now easily decomposed into a posterior density of the transition probabilities matrix:

$$p(\mathbf{P}|\mathbf{S}) \propto p(s_0|\mathbf{P}) \prod_{i=1}^{M}\prod_{j=1}^{M} p_{ij}^{N_{ij}(\mathbf{S})} p(\mathbf{P}), \tag{12}$$

and the posterior density of the state-dependent parameters:

$$p(\theta_i|\mathbf{y}, \mathbf{S}) \propto \prod_{t:S_t=i} p(y_t|\theta_i, \mathbf{y}_{t-1},)p(\theta_i). \tag{13}$$

Since the form of the posterior density for all the parameters is not standard, the commonly used strategy is to simulate the posterior distribution with numerical methods. A Monte Carlo Markov Chain (MCMC) algorithm, the Gibbs sampler (see Casella & George, 1992, and references therein), enables us to simulate the joint posterior distribution of all the parameters of the model by sampling from the full conditional distributions. Such an algorithm has also been adapted to Markov-switching models by Albert & Chib (1993) and McCulloch & Tsay (1994).

In the Block Metropolis-Hastings algorithm, parameters of the model are split into sub-vectors, the full conditional densities of which are of convenient form. Firstly, however, we draw a vector of the states of the

economy, **S**. We initialize the algorithm, conditioning on the starting values for the parameters, $\theta^{(0)}$. Then, using the BLHK filter and smoother (see Frühwirth-Schnatter, 2006, Chapter 11 and references therein), we obtain the probabilities $\Pr(s_t = i|\mathbf{y}, \theta^{(l-1)})$, for $t = 1, \ldots, T$ and $i = 1, \ldots, M$, and then draw $\mathbf{S}^{(l)}$, for l^{th} iteration of the algorithm.

Secondly, we draw from the posterior distribution of the transition probabilities matrix (12), conditioning on the states drawn in the previous step of the current iteration, $\mathbf{P}^{(l)} \sim p(\mathbf{P}|\mathbf{S}^{(l)})$. Assuming the Dirichlet prior distribution and that the hidden Markov process starts from its ergodic distribution, π, makes the posterior distribution not of standard form. In this step of the MCMC sampler, we use the Metropolis-Hastings algorithm as described in (Frühwirth-Schnatter, 2006, Section 11.5.5).

Thirdly, we draw the state-dependent parameters of the VAR process collected in one vector, $\beta = (\beta_1', \ldots, \beta_M')'$. Due to the form of the likelihood function and normal prior distribution, the full conditional distribution is also normal $f(\beta|\mathbf{y}, \mathbf{S}^{(l)}, \mathbf{P}^{(l)}, \sigma^{(l-1)}, \mathbf{R}^{(l-1)}) = \mathcal{N}\left(\bar{\beta}^*, \bar{V}_{\beta^*}\right)$, from which we draw $\beta^{(l)}$. $\bar{\beta}^*$ and $\bar{V}_{\beta^*}$ are the parameters of the full conditional distribution specified in Section 6 (see also Frühwirth-Schnatter, 2006, Section 8.4.3).

Finally, we collect all the standard deviations in one vector, $\sigma = (\sigma_1', \ldots, \sigma_M')'$, and all the unknown correlation coefficients into a vector, $\mathbf{R} = (\text{vecl}(\mathbf{R}_1)', \ldots, \text{vecl}(\mathbf{R}_M)')'$, where function, vecl, stacks all the lower-diagonal elements of the correlation matrix into a vector. In order to draw from the full conditional densities of these two vectors, $f(\sigma|\mathbf{y}, \mathbf{S}^{(l)}, \mathbf{P}^{(l)}, \beta^{(l)}, \mathbf{R}^{(l-1)})$ and $f(\mathbf{R}|\mathbf{y}, \mathbf{S}^{(l)}, \mathbf{P}^{(l)}, \beta^{(l)}, \sigma^{(l)})$, we employ the Griddy-Gibbs sampling algorithm of Ritter & Tanner (1992), as described by Barnard et al. (2000).

The algorithm for the restricted models is presented in detail in Section 6.

3. Definitions and MS-VAR Model for Granger Causality Analysis and Regime Inference

Let $\{y_t : t \in \mathbb{Z}\}$ be a $N \times 1$ multivariate square integrable stochastic process on the integers $\mathbb{Z}$. Write:

$$y_t = (y_{1t}', y_{2t}', y_{3t}', y_{4t}')', \tag{14}$$

for $t = 1, \ldots, T$, where y_{it} is a $N_i \times 1$ vector with $N_1, N_4 \geq 1, N_2, N_3 \geq 0$ and $\sum_{i=1}^{4} N_i = N$. Variables of interest are contained in vectors y_1 and y_4, between which we want to study causal relations. Vectors y_2 and y_3 (that for $N_2 = N_3 = 0$ are empty) may contain auxiliary variables that are also used for forecasting and modeling purposes. Moreover, define two vectors: the first is $(N_1 + N_2)$-dimensional, $v_{1t} = (y_{1t}', y_{2t}')'$, while the second is $(N_3 + N_4)$-dimensional, $v_{2t} = (y_{3t}', y_{4t}')'$, such that:

$$y_t = \begin{bmatrix} v_{1t} \\ v_{2t} \end{bmatrix},$$

with matrix $\mathbf{v}_{it}$ collecting observations of v_{it} up to period t for $i = 1, 2$.

Suppose that the conditional mean $E[y_{t+1}|\mathbf{y}_t; \theta]$ is finite and that the conditional covariance matrix $E\left[(y_{t+1} - E[y_{t+1}|\mathbf{y}_t])(y_{t+1} - E[y_{t+1}|\mathbf{y}_t; \theta])'|\mathbf{y}_t; \theta\right]$ positive definite for all finite t. Further, let u_{t+1} denote the one-step-ahead forecast error for y_{t+1}, conditional on $\mathbf{y}_t$ (and the parameters) when the predictor is given by the conditional expectations, i.e.:

$$u_{t+1} = y_{t+1} - E[y_{t+1}|\mathbf{y}_t; \theta]. \tag{15}$$

By construction, u_{t+1} has conditional mean zero and positive-definite conditional covariance matrix. And let $\tilde{u}_{t+1} = y_{t+1} - E[y_{t+1}|\mathbf{v}_{1t}, \mathbf{y}_{3t}; \theta]$ be the one-step-ahead forecast error for y_{t+1}, conditional on $\mathbf{v}_{1t}$ and $\mathbf{y}_{3t}$ with analogous properties.

We focus on the Granger-causal relations between variables y_1 and y_4. The definition of *Granger causality*, originally given by Granger (1969), states simply that y_4 is not causal for y_1 when the past and current information about, $\mathbf{y}_{4.t}$ cannot improve mean square forecast error of $y_{1.t+1}$.

Definition 1. y_4 does not Granger-cause y_1, denoted by $y_4 \overset{G}{\nrightarrow} y_1$, if and only if:

$$E\left[u_{t+1}^2; \theta\right] = E\left[\tilde{u}_{t+1}^2; \theta\right] < \infty \quad \forall t = 1, \ldots, T. \tag{16}$$

This definition refers to the conditional mean process for a univariate y_{1t}, and holds if and only if the two means conditioned on the full set of variables, $\mathbf{y}_t$, and on the restricted set, $(\mathbf{v}_{1t}, \mathbf{y}_{3t})$, are the same (see Boudjellaba, Dufour & Roy, 1992). It is important to note that the definition involves conditioning on the parameters and under a classical treatment the parameters would be set to their "true" values. Since the latent state variables are not parameters under that approach, we choose to treat them equivalently in our Bayesian setting. Hence, Granger causality in MS-VARs under a Bayesian approach concerns the validity of (16) for any $\theta \in \Theta$.

To model Granger noncausality, we make use of the decomposition of y_t into y_{it} for $i = 1, \ldots, 4$. The system in equation (1) is expressed as:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \\ y_{4t} \end{bmatrix} = \begin{bmatrix} m_{1.s_t} \\ m_{2.s_t} \\ m_{3.s_t} \\ m_{4.s_t} \end{bmatrix} + \sum_{k=1}^{p} \begin{bmatrix} a^{(k)}_{11.s_t} & a^{(k)}_{12.s_t} & a^{(k)}_{13.s_t} & a^{(k)}_{14.s_t} \\ a^{(k)}_{21.s_t} & a^{(k)}_{22.s_t} & a^{(k)}_{23.s_t} & a^{(k)}_{24.s_t} \\ a^{(k)}_{31.s_t} & a^{(k)}_{32.s_t} & a^{(k)}_{33.s_t} & a^{(k)}_{34.s_t} \\ a^{(k)}_{41.s_t} & a^{(k)}_{42.s_t} & a^{(k)}_{43.s_t} & a^{(k)}_{44.s_t} \end{bmatrix} \begin{bmatrix} y_{1t-i} \\ y_{2t-i} \\ y_{3t-i} \\ y_{4t-i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \\ \varepsilon_{4t} \end{bmatrix}. \tag{17}$$

The covariance matrix of the residuals conditional on the regime is given by:

$$\Sigma_{s_t} = \text{Var}\left(\begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \\ \varepsilon_{4t} \end{bmatrix} \right) = \begin{bmatrix} \Omega_{11.s_t} & \Omega'_{21.s_t} & \Omega'_{31.s_t} & \Omega'_{41.s_t} \\ \Omega_{21.s_t} & \Omega_{22.s_t} & \Omega'_{32.s_t} & \Omega'_{42.s_t} \\ \Omega_{31.s_t} & \Omega_{32.s_t} & \Omega_{33.s_t} & \Omega'_{43.s_t} \\ \Omega_{41.s_t} & \Omega_{42.s_t} & \Omega_{43.s_t} & \Omega_{44.s_t} \end{bmatrix}. \tag{18}$$

With expositional purpose in mind, let us first assume that all regimes are known. Next period's prediction of y_1 conditional on s_{t+1} and $\mathbf{y}_t$ is then:

$$E\left[y_{1t+1} | s_{t+1}, \mathbf{y}_t, \theta\right] = y_{1t+1} - \epsilon_{1t+1}. \tag{19}$$

Accordingly, the forecast error is given by ϵ_{1t+1} and the conditional forecast error variance by $\omega_{1s_{t+1}}$. The necessary and sufficient condition for y_4 not to Granger-cause y_1 is that $a^{(k)}_{14,s_t}$ in equation (17) is equal to zero, for all k and t.

Let us now drop the assumption that the regimes are known. While the regime variable s_{t+1} conditional on s_t is independent of $\mathbf{y}_t$, it can be predicted using only past observations of y. Let $\Pr[s_{t+1} | \mathbf{y}_t, \theta]$ denote the probability of a particular state occurring at $t+1$ conditional on the information available at t. The prediction of next period's value of y_1 is then given by:

$$E\left[y_{1t+1} | \mathbf{y}_t, \theta\right] = \sum_{s_{t+1}} E\left[y_{1t+1} | s_{t+1}, \mathbf{y}_t, \theta\right] \Pr\left[s_{t+1} | \mathbf{y}_t, \theta\right]. \tag{20}$$

The role for y_4 is different in (20) relative to (19) in that the history of y_4 can now predict y_1 by containing information which helps predict next period's state.

Since s_{t+1} conditional on s_t is independent of $\mathbf{y}_t$ it follows that:

$$\Pr\left[s_{t+1} | \mathbf{y}_t, \theta\right] = \sum_{s_t} \Pr\left[s_{t+1} | s_t, \theta\right] \Pr\left[s_t | \mathbf{y}_t, \theta\right]. \tag{21}$$

From this relationship we may deduce that there are only two instances when there is no additional information in the history of y_4 for predicting next period's state. The first is when $\Pr[s_{t+1} | s_t, \theta] = \Pr[s_{t+1}; \theta]$, i.e. the Markov process is serially uncorrelated. The second case occurs when $\Pr[s_t | \mathbf{y}_t, \theta] = \Pr[s_t | \mathbf{v}_{1t}, \mathbf{y}_{3t}, \theta]$.

This discussion presumes that the coefficients in the equation for y_1 vary freely with the regime. It is possible, however, that these coefficients vary with s_{1t+1} but not with s_{2t+1}. Similarly, there may be information in $\mathbf{y}_{4t}$ for predicting s_{2t+1}, but not for predicting s_{1t+1}. In such situations, it may still be the case that the prediction of y_1 in (20) does not depend on the history of y_4.

The regime inference question is in fact better addressed in terms of the sub-vectors v_1 and v_2. Apart from decomposing the observed variables into the v_{it} sub-vectors, the parameter vectors and matrices are decomposed analogously. Furthermore, the hidden Markov process is decomposed into two sub-processes, $s_t = (s_{1t}, s_{2t})$, where s_{it} has M_i states that are characterized by transition probability matrices, $\mathbf{P}^{(i)}$ (and ergodic probabilities, $\pi^{(i)}$) for $i = 1, 2$, such that $M = M_1 \cdot M_2$. The construction of the joint transition probabilities matrix, $\mathbf{P}$, is not specified for the moment and will be the subject of further analysis.

Specifically, a restricted version of the system in equation (1) is given by:

$$\begin{bmatrix} v_{1t} \\ v_{2t} \end{bmatrix} = \begin{bmatrix} \mu_{1.s_{1t}} \\ \mu_{2.s_{2t}} \end{bmatrix} + \sum_{k=1}^{p} \begin{bmatrix} A^{(k)}_{11.s_{1t}} & A^{(k)}_{12.s_{1t}} \\ A^{(k)}_{21.s_{2t}} & A^{(k)}_{22.s_{2t}} \end{bmatrix} \begin{bmatrix} v_{1t-i} \\ v_{2t-i} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}. \tag{22}$$

where the following linear restrictions have been imposed:

$$\mu_{i.s_t} = \mu_{i.s_{it}}, \quad A^{(k)}_{ij.s_t} = A^{(k)}_{ij.s_{it}}, \quad i, j = 1, 2, \text{ and } k = 1, \ldots, p. \tag{23}$$

If the ϵ_{it} residuals are independent of the regime, equation (23) states that v_{it} is only directly affected by s_{it}. Indirectly, it may be affected by (lags of) the other regime process s_{jt} through lags of v_{jt} ($i \neq j$).

The marginal distribution of the $\epsilon_{it}|s_t$ may also be subject to linear restrictions given by:

$$\Sigma_{ii.s_t} = \Sigma_{ii.s_{it}}, \quad i = 1, 2. \tag{24}$$

The restrictions in (24) are necessary, but not sufficient for $p(\epsilon_t|s_t) = p(\epsilon_{1t}|s_{1t})p(\epsilon_{2t}|s_{2t})$, meaning that $\epsilon_{1t}|s_{1t}$ and $\epsilon_{2t}|s_{2t}$ are independent. The additional requirement is simply that $\Sigma_{12.s_t} = 0$ for all regimes.

In the event that the restrictions in (23) and (24) are satisfied and the covariance matrix $\Sigma_{12.s_t} = 0$ for all regimes, then v_{it} is only directly influenced by the s_{it} regime process, i.e., through the regime dependent $\mu_{i.s_{it}}$ and $A_{ij.s_{it}}$ matrices. Nevertheless, v_{it} may still be indirectly influenced by lags of the s_{jt} process through lags of v_{jt}.

We are now in a position to give a definition of predictive conditional state independence:

Definition 2. The regime predictions of $s_{1.t+1}$ and $s_{2.t+1}$ conditional on $\mathbf{y}_t$ are independent if and only if:

$$\Pr\left[(s_{1.t+1}, s_{2.t+1}) = (j_1, j_2)|\mathbf{y}_t, \theta\right] = \Pr\left[s_{1.t+1} = j_1|\mathbf{y}_t, \theta\right] \cdot \Pr\left[s_{2.t+1} = j_2|\mathbf{y}_t, \theta\right], \tag{25}$$

for all $\theta \in \Theta$, $j_1 = 1, \ldots, M_1$ with $M_1 \geq 2$, $j_2 = 1, \ldots, M_2$ and $t = 1, \ldots, T$.

In addition, we shall say that there is no unique information in $\mathbf{v_{2t}}$ for predicting $s_{1.t+1}$ when the following holds for all cases considered in this definition:

$$\Pr\left[s_{1.t+1} = j_1|\mathbf{y}_t, \theta\right] = \Pr\left[s_{1.t+1} = j_1|\mathbf{v}_{1t}, \theta\right]. \tag{26}$$

An analogous relation may be specified for the predictions of $s_{2.t+1}$. In the next section we shall first consider under which restrictions on the MS-VAR process that the two regime processes are independent in the sense of Definition 2 as well as when equation (26) holds. Second, we analyse under which conditions y_4 does not Granger cause y_1 in this setup.

4. Regime Inference and Granger Causality Analysis

4.1. Regime Inference

The first result in this paper concerns the restrictions that MS-VAR system in (1) need to satisfy to guarantee that we can make optimal inference from the v_{1t} sub-system about the regimes that affect these variables.

Proposition 1. *The regime predictions of $s_{1t+1}|\mathbf{v}_{1t}$ and $s_{2t+1}|\mathbf{v}_{2t}$ are independent and there is no unique information in $\mathbf{v}_{2t}$ for predicting s_{1t+1} if and only if either:*

(A1): (i) $\mathbf{P} = (\mathbf{P}^{(1)} \otimes \mathbf{P}^{(2)})$,

(ii) *equations (23) and (24) are satisfied,*

(iii) $\Sigma_{12.s_t} = 0$, *and*

(iv) $A^{(k)}_{12.s_{1t}} = 0$,

for all $k = 1, \ldots, p$ *and* $s_{it} = 1, \ldots, M_i$ *with* $M_1 \geq 2$, *and* $i, j = 1, 2$; *or:*

(A2): $\mathbf{P} = (\iota_{M_1} \pi^{(1)\prime} \otimes \mathbf{P}^{(2)})$,

is satisfied for all $\theta \in \Theta$.

First, condition (A1)(i) is a result of forming the full transition probabilities matrix out of the transition probabilities matrices of two independent hidden Markov processes (see Sims, Waggoner & Zha, 2008). Restrictions (A1)(ii) state that the parameters of the equation for v_{1t} change over time with the first Markov process, whereas the parameters of the equation for v_{2t} follow the second Markov process as in equation (22). Furthermore, restrictions (A1)(iii) guarantee independence in the sense discussed below equation (24). Condition (A2) states simply that the first out of the two decomposed hidden Markov processes is serally uncorrelated and we can thereby model the residuals ϵ_{1t} with a mixture of Gaussian distributions.

Second, note that conditions (A1) and (A2) imply linear restrictions on parameters of the model. Furthermore, if we change the restrictions in (A1)(iv) to $A^{(k)}_{21s_{2t}} = 0$, then there is no unique information in $\mathbf{v}_{1t}$ for predicting s_{2t+1}. Moreover, in the Appendix it is shown that:

Corollary 1. *If and only if restrictions in* (A1)(i)-(iii) *are satisfied for all* $\theta \in \Theta$, *then*

$$\Pr\left[(s_{1,t}, s_{2,t}) = (i_1, i_2)|\mathbf{y}_\tau, \theta\right] = \Pr\left[s_{1,t} = i_1|\mathbf{y}_\tau, \theta_1, \mathbf{P}^{(1)}\right] \Pr\left[s_{2,t} = i_2|\mathbf{y}_\tau, \theta_2, \mathbf{P}^{(2)}\right],$$

for all $i_1 = 1, \ldots, M_1$ *with* $M_1 \geq 2$, $i_2 = 1, \ldots, M_2$ *with* $M_2 \geq 2$, *and* $t, \tau = 1, \ldots, T$, *with* $\theta = (\theta_1, \theta_2, \mathbf{P})$.

Hence, for the predictions of s_{1t} and s_{2t} to be independent, it is not sufficient that the Markov processes are independent. In fact, the joint distribution for y_t conditional on s_t (and $\mathbf{y}_{t-1}$) being equal to the product between the marginal distributions for v_{lt} conditional on s_{lt} (and $\mathbf{v}_{t-1}$) for $l = 1, 2$ must also be satisfied. Under these additional restrictions forecasting, filtering and smoothing inference about the two regime variables can be conducted independently. Additionally:

Corollary 2. *If and only if condition* (A1) *is satisfied for all* $\theta \in \Theta$, *then*

$$\Pr\left[(s_{1,t}, s_{2,t}) = (i_1, i_2)|\mathbf{y}_\tau, \theta\right] = \Pr\left[s_{1,t} = i_1|\mathbf{v}_{1,\tau}, \theta_1, \mathbf{P}^{(1)}\right] \Pr\left[s_{2,t} = i_2|\mathbf{v}_\tau; \theta_2, \mathbf{P}^{(2)}\right],$$

for all $i_1 = 1, \ldots, M_1$ *with* $M_1 \geq 2$, $i_2 = 1, \ldots, M_2$, *and* $t, \tau = 1, \ldots, T$, *with* $\theta = (\theta_1, \theta_2, \mathbf{P})$.

In Appendix A (see Lemma 2 and Lemma 3) we present necessary and sufficient conditions for conducting optimal inference on $s_{1,t}$ and $s_{2,t}$ independently in Markov-switching models when the density function for $\epsilon_t|s_t$ meets the criteria for conducting optimal inference on s_t using the algorithm in Hamilton (1994) and Kim (1994).

The intuition behind condition (A1) is, in fact, straightforward. Suppose $p = 1$, $N = M = M_1 = 2$, while ϵ_{2t} is i.i.d.. The restrictions on Σ_{s_t} in (A1) are sufficient for the residual of the equation for v_2 to be i.i.d.. Now consider the experiment of drawing two v_{2t}'s, one for each regime, when v_{1t-1} and v_{2t-1} are fixed. The difference between these two draws is:

$$v_{2t|s_t=2} - v_{2t|s_t=1} = (\mu_{2,2} - \mu_{2,1}) + (A_{21,2} - A_{21,1})\, v_{1t-1} + (A_{22,2} - A_{22,1})\, v_{2t-1}. \tag{27}$$

The right hand side of (27) is zero for all vectors (v_{1t-1}, v_{2t-1}) when the coefficients in the v_2 equation are constant across states. Accordingly, if these restrictions are satisfied, then $\Pr[s_t|\mathbf{v}_{1t}, \mathbf{v}_{2t}] = \Pr[s_t|\mathbf{v}_{1t}, \mathbf{v}_{2t-1}]$ and all information about s_t is found in the equation for v_1. If the coefficient on $v_2 t - 1$ in that equation is zero for both states, then lags of $\mathbf{v}_{2t-1}$ plays no role for predicting regime switches either.

To sum up, condition (A1) tells us exactly under which conditions we can disregard the information in v_{2t} when we are either only interested in the behavior of the variables in the v_{1t} vector or in the s_{1t} regime process. Alternatively, if we are primarily interested in v_{2t} (or in s_{2t}) and would like to treat v_{1t} as being "exogenous", then (A1) provides the set of restrictions that we implicitly impose on the system describing both v_{2t} and v_{1t}.

4.2. Granger Noncausality Analysis

The restrictions for the regime independence presented in the previous section are stronger than what is required for Granger noncausality. In this section we focus on the necessary and sufficient conditions for this type of noncausal relations. Additional notation is first required. Specifically, define:

$$\bar{m}_{1t} \equiv E\left[m_{1s_{t+1}}|\mathbf{y}_t, \theta\right], \tag{28a}$$

$$\bar{a}^{(k)}_{1r.t} \equiv E\left[a^{(k)}_{1r.s_{t+1}}|\mathbf{y}_t; \theta\right], \tag{28b}$$

for all $r \in \{1,\ldots,4\}$ and $k \in \{1,\ldots,p\}$. The one-step-ahead forecast error for y_1 is then given by $u_{t+1} = z_{t+1} + \epsilon_{1t+1}$, where:

$$\begin{aligned} z_{t+1} \equiv (m_{1s_{t+1}} - \bar{m}_{1t}) &+ \sum_{k=1}^{p}\left(a^{(k)}_{11.s_{t+1}} - \bar{a}^{(k)}_{11.t}\right) y_{1.t+1-k} \\ &+ \sum_{k=1}^{p}\left(a^{(k)}_{12.s_{t+1}} - \bar{a}^{(k)}_{12.t}\right) y_{2.t+1-k} + \sum_{k=1}^{p}\left(a^{(k)}_{13.s_{t+1}} - \bar{a}^{(k)}_{13.t}\right) y_{3.t+1-k} \\ &+ \sum_{k=1}^{p}\left(a^{(k)}_{14.s_{t+1}} - \bar{a}^{(k)}_{14.t}\right) y_{4.t+1-k}, \end{aligned}$$

is (conditionaly on $\mathbf{y}_t$) uncorrelated with ϵ_{1t+1}.[1] A sufficient, but not necessary, condition for z_{t+1} to be mean zero stationary is that y_1 is covariance stationary. Another possibility is that y_t is co-integrated. For the remainder of this section, we shall assume that u_{t+1} is mean zero stationary.

This assumption brings us to the main result about Granger noncausality.

Proposition 2. *y_4 does not Granger-cause y_1 if and only if either:*

(A1) *or*

(A3): **(i)** $\sum_{j=1}^{M} m_{1.j}p_{ij} = \bar{m}_1$,

(ii) $\sum_{j=1}^{M} a^{(k)}_{1r.j}p_{ij} = \bar{a}^{(k)}_{1r}$,

(iii) $\bar{a}^{(k)}_{14} = 0$,

for all $i = 1,\ldots,M$, $r = 1,\ldots,4$, and $k = 1,\ldots,p$,

is satisfied for all $\theta \in \Theta$.

The nonlinear restrictions in condition (A3)(i) and (A3)(ii) state that given any regime j, the expected value of each random coefficient in the equation for y_1 is constant. Condition (A3)(iii) sets each expected value of the coefficients on lags of y_4 to zero.

Note that the restrictions of (A3) do not rely on a decomposition of the hidden Markov process. This comes from the fact that these conditions refer solely to the expected value of the parameters of the equation for y_1. At the same time, they do not rule out that the transition matrix $\mathbf{P}$ has reduced rank or that the

[1] See (Krolzig, 1997, Chapter 4) . Blix (1997) derives a general formula for the expectation of $y_{t+\tau}$, $\tau \geq 1$, conditional on $\mathbf{y}_t$ and applies it to rational expectations hypotheses.

Markov process can be decomposed into multiple processes. Hence, the restrictions in (A3) are very general and it is not possible to determine the number of restrictions without specifically referring to the properties of the transition matrix.

From the perspective of the full MS-VAR system, the restrictions in (A3) are weaker than those in (A1). First, each restriction in (A1)(ii) results in a constraint on an individual parameter, whereas each restriction in (A3)(i) and (A3)(ii) involves a function on several parameters.[2] In addition, (A3) does not include any restriction on the covariance matrices, while (A1) has several linear restrictions on Σ_{s_t}. Restriction (A3)(iii) compared to restrictions (A1)(vi) does not constrain any of the parameter to zero, but instead imposes a zero condition on the expected value of the forecasted parameter. Thus, (A3)(iii) is a weaker assumption than (A1)(iv).

However, contrarily to conditions (A1) and (A2), the condition (A3) is not linear in parameters. Still, conditions (A3)(i) and (A3)(ii) have equivalent form, $\sum_{j=1}^{M} m_{1,j}(p_{ij} - p_{kj}) = 0$ for $i,k = 1,\ldots,M$ and $i \neq k$, which for some special cases give rise to linear restrictions. In Section 5 we discuss consequences of the nonlinearity of the restrictions when testing them.

Under specific assumptions, the restrictions of Proposition 2 may be simplified. We present such a possibility in Corollary 3:

Corollary 3. *Suppose that condition (A2) is satisfied for all $\theta \in \Theta$, then condition (A3) is equivalent to:*

(A4): (i) $\sum_{j_1=1}^{M_1} m_{1.(j_1,j_2)}\pi_{j_1}^{(1)} = \bar{m}_1$,

(ii) $\sum_{j_1=1}^{M_1} a_{1r.(j_1,j_2)}^{(k)}\pi_{j_1}^{(1)} = \bar{a}_{1r}^{(k)}$,

(iii) $\bar{a}_{14}^{(k)} = 0$,

for all $j_2 = 1,\ldots,M_2$, $r = 1,\ldots,4$, and $k = 1,\ldots,p$.

Corollary 3 reintroduces the decomposition of the hidden Markov process. One benefit is that the number of the restrictions to be imposed on the model is typically reduced.

This Corollary is of particular interest when $M = 2$. For such Markov processes, the rank of $\mathbf{P}$ can be either one or two. In both cases, condition (A2) is satisfied and, hence, condition (A4) gives two sets of parameter constraints that are equivalent to the set of restrictions in (A3). If the rank of $\mathbf{P}$ is unity ($M_1 = 2, M_2 = 1$), then the two-state Markov process is serially uncorrelated. For this case, (A4) reduces to A4(iii), where $\sum_{j_1=1}^{2} a_{14,j_1}^{(k)}\pi_{j_1}^{(1)} = 0$ for all k, while (A4)(i)-(ii) are satisfied by construction. Notice that all restrictions are nonlinear and that the number of restrictions is equal to $p + 1$.

On the other hand, if the rank of $\mathbf{P}$ is two ($M_1 = 1, M_2 = 2$), then the Markov process is serially correlated with $\mathbf{P} = \mathbf{P}^{(2)}$. Now, condition (A4) states that all coefficients in the equation for y_1 are constant across the regimes, and the coefficients on lags of y_4 are zero, i.e., all restrictions are linear and the total number of equal to $3p + 1$.

5. Bayesian Testing

Restrictions **(A1)–(A4)** can be tested. We first consider classical tests and their limitations and then present the Bayesian testing procedure as a solution. The obstacles in using classical tests are threefold:

- The asymptotic distribution of the parameters of the MS-VAR is unknown;
- The conditions for noncausality may result in several sets of restrictions on parameters. Consequently, one hypothesis may be represented by several restricted models;

[2] This is also true when $M_1 = M$ since now all restrictions in (A1)(ii) concern the constancy of the coefficient matrices and the residual covariance matrices of the v_{2t} sub-system, while the corresponding matrices of the v_{1t} sub-system are allowed to vary freely with the regime.

- Some of the restrictions are in the form of nonlinear functions of parameters of the model.

The proposed solution consists of a new Block Metropolis-Hastings sampling algorithm for the estimation of the restricted models, and of the application of a standard Bayesian test to compare the restricted models to the unrestricted one.

Classical testing. In the general case, all the mentioned problems with classical testing are difficult to cope with. While, the lack of the asymptotic distribution of the parameters could be solved using simulation methods, the problem of testing a hypothesis represented by several restricted models seems unsolvable with existing classical methods.

The problem of the nonlinearity of the restrictions, however, is well known in the studies on testing parameter conditions for Granger noncausality in multivariate models. In the general case, nonlinear restrictions on parameters of the model may result in the matrix of partial derivatives of the restrictions with respect to the parameters not having a full rank. Consequently, the asymptotic distribution of test statistic is not known.

This problem was met in several studies on Granger noncausality testing in time series models. Boudjellaba et al. (1992) derive conditions for Granger noncausality for VARMA models that result in multiple nonlinear restrictions on original parameters of the model. As a solution to the problem of testing the restrictions, they propose a sequential testing procedure. There are two main drawbacks of this method. First, despite properly performed procedure, the test may still appear inconclusive, and second, the confidence level is given in the form of inequalities. The problem of testing non-linear restrictions was examined for h-periods ahead Granger causality for VAR models. Dufour et al. (2006) propose the solution based on formulating a new model for each h, and obtain linear restrictions on the parameters on the model. These restrictions can be easily tested with standard tests. In another work by Dufour (1989) the approach is based on the linear regression theory; its solutions would require separate proofs in order to apply it to Markov-switching VARs. Finally, Lütkepohl & Burda (1997) propose a solution for testing nonlinear hypotheses based on a modification of the Wald test statistic. Given the asymptotic normality of the estimator of the parameters, the method uses a modification that, together with standard asymptotic derivations, overcomes the singularity problem.

Finally, the problem of testing the nonlinear restrictions was faced by Warne (2000), who derives the restrictions for Granger noncausality, noncausality in variance and noncausality in distribution for Markov-switching VAR models. Among the solutions reviewed in this Section, only that proposed by Lütkepohl & Burda (1997) seems applicable to this particular problem. This finding should, however, be followed with further studies proving its applicability.

Bayesian testing. In this study we propose a method of solving the problems of testing the parameter restrictions based on Bayesian inference. This approach to testing the noncausality conditions was used by Woźniak (2011, 2012). Both of the papers work on the Extended CCC-GARCH model of Jeantheau (1998). Two other works use the Bayesian approach to make inference about a concept somehow related to Granger noncausality, namely exogeneity. Jarociński & Maćkowak (2011) use Savage-Dickey's Ratio to test block-exonegeneity in the VAR model, while Pajor (2011) uses Bayes factors to assess exogeneity conditions for models with latent variables, and in particular in multivariate Stochastic Volatility models.

In order to compare the unrestricted model, denoted by $\mathcal{M}_i$, and the restricted model, $\mathcal{M}_j$ and $= j \neq i$, we use the Posterior Odds Ratio (POR), which is a ratio of the posterior probabilities, $\Pr(\mathcal{M}|\mathbf{y})$, attached to each of these models representing the hypotheses:

$$\text{POR} = \frac{\Pr(\mathcal{M}_i|\mathbf{y})}{\Pr(\mathcal{M}_j|\mathbf{y})} = \frac{p(\mathbf{y}|\mathcal{M}_i)}{p(\mathbf{y}|\mathcal{M}_j)} \frac{\Pr(\mathcal{M}_i)}{\Pr(\mathcal{M}_j)}, \tag{29}$$

where $p(\mathbf{y}|\mathcal{M})$ is the marginal density of data and $\Pr(\mathcal{M})$ is the prior probability of a model. In order to compare two competing models, one might also consider using Bayes factors, defined by:

$$\mathcal{B}_{ij} = \frac{p(\mathbf{y}|\mathcal{M}_i)}{p(\mathbf{y}|\mathcal{M}_j)}. \tag{30}$$

Note that if one chooses not to discriminate any of the models *a priori*, setting equal prior probabilities for both of the models ($\Pr(\mathcal{M}_i)/\Pr(\mathcal{M}_j) = 1$), the Posterior Odds Ratio is then equal to a Bayes factor. This method of testing does not have any of the drawbacks of the Likelihood Ratio test, once samples of draws from the posterior distributions of parameters for both the models are available (see Geweke, 1995; Kass & Raftery, 1995).

In this work, in order to asses the credibility of the hypotheses, each of which is represented by several sets of restrictions – and thus several models – we compute Posterior Odds Ratios. The results of this analysis are reported in Table 6 in Section 7. Suppose that a hypothesis is represented by several models. Let $\mathcal{H}_i$ denote the set of indicators of the models that represent this hypothesis, $\mathcal{H}_i = \{j : \mathcal{M}_j \text{ represents } i^{th} \text{ hypothesis}\}$. For instance, in our example, the hypothesis of Granger noncausality in mean is represented by four models, such that $\mathcal{H}_2 = \{1, 2, 4, 5\}$. Further, suppose that one is interested in comparing the posterior probability of this hypothesis to the hypothesis $\mathcal{H}_0$, represented by the unrestricted model $\mathcal{M}_0$. Then the credibility of the hypothesis $\mathcal{H}_i$ compared to the hypothesis $\mathcal{H}_0$ may be assessed with the Posterior Odds Ratio given by:

$$\text{POR} = \frac{\Pr(\mathcal{H}_i|\mathbf{y})}{\Pr(\mathcal{H}_0|\mathbf{y})} = \frac{\sum_{j\in\mathcal{H}_i}\Pr(\mathcal{M}_j|\mathbf{y})}{\Pr(\mathcal{M}_0|\mathbf{y})}. \tag{31}$$

We set equal prior probabilities for all the models, which has the effect that none of the models is preferred *a priori*.

Testing the noncausality restrictions in MS-VARs. Taking into account the complicated structure of the restrictions, we choose Posterior Odds Ratio (29) to assess the hypotheses. The crucial element of this method the is computation of marginal data densities, $p(\mathbf{y}|\mathcal{M})$, for the unrestricted and the restricted models. There are several available methods of computing this value. In this study we choose the Modified Harmonic Mean (MHM) method of Geweke (1999). For a chosen model, given the sample of draws, $\{\theta^{(i)}\}_{i=1}^{S}$, from the posterior distribution of the parameters, $p(\theta|\mathbf{y}, \mathcal{M})$, the marginal density of data is computed using:

$$p(\mathbf{y}|\mathcal{M}) = \left(S^{-1} \sum_{i=1}^{S} \frac{h(\theta^{(i)})}{L(\mathbf{y}; \theta^{(i)}, \mathcal{M}) p(\theta^{(i)}|\mathcal{M})} \right)^{-1}, \tag{32}$$

where $L(\mathbf{y}; \theta^{(i)}, \mathcal{M})$ is a likelihood function od model $\mathcal{M}$. $h(\theta^{(i)})$, as specified in Geweke (1999), is a k-variate truncated normal distribution with mean parameter equal to the posterior mean and covariance matrix set to the posterior covariance matrix of θ. The truncation must be such that $h(\theta)$ had thinner tails than the posterior distribution.

Other methods of computing the marginal density of data may also be employed. Several estimators were derived, taking into account the characteristics of Markov-switching models. The reader is referred to the original papers by Frühwirth-Schnatter (2004), Sims et al. (2008) and Chib & Jeliazkov (2001). Moreover, Frühwirth-Schnatter (2004) rises the problem of the bias of the estimators when the label permutation mechanism is missing in the algorithm sampling from the posterior distribution of the parameters. The bias appears to be due to the invariance of the likelihood function and the prior distribution of the parameters, with respect to permutations of the regimes' labels. Then the model is not globally identified. The identification can be insured by the ordering restrictions on parameters, and can also be implemented within the Gibbs sampler. Simply, it is sufficient that the values taken by one of the parameters of the model in different regimes can be ordered, and that the ordering holds for all the draws from the Gibbs algorithm to assure global identification (see Frühwirth-Schnatter, 2004). We assure that this is the case, i.e. that the MS-VAR models considered for causality inference are globally identified by the ordering imposed on some parameter.

Another element of the testing procedure is the estimation of the unrestricted model and the restricted models representing hypotheses of interest. We present a new Block Metropolis-Hastings sampling algorithm specially constructed for the purpose of testing noncausality hypotheses in the MS-VAR models in Section 6. It enables the imposing of restrictions on parameters resulting from conditions (A1) - (A7), and in effect testing different hypotheses of Granger noncausality between variables. In the algorithm, the

restrictions are imposed on different groups of the parameters of the model. First, linear restrictions on the parameters of the VAR process, β, are implemented according to Frühwirth-Schnatter (2006). Next, parameters of the covariance matrices are decomposed into standard deviations, σ, and correlation parameters, $\mathbf{R}$. To these parameter groups we apply the Griddy-Gibbs sampler of Ritter & Tanner (1992), as in Barnard et al. (2000). Such a form of the sampling algorithm easily allows to restrict any of the parameters. Note that the algorithm of Barnard et al. (2000) has not yet been applied to Markov-switching models. Finally, we restrict the matrix of transition probabilities, $\mathbf{P}$, joining the approach of Sims et al. (2008) with the Metropolis-Hastings algorithm of Frühwirth-Schnatter (2006). The Metropolis-Hastings step needs to be implemented, as we require the hidden Markov process to be irreducible. Moreover, additional parts of the algorithm are constructed in order to impose nonlinear restrictions on the parameters of the VAR process and the decomposed covariance matrix.

To summarize, we propose the following procedure in order to test different Granger noncausality hypotheses in Markov-switching VAR models.

Step 1: Specify the MS-VAR model. Choose the order of VAR process, $p \in \{0, 1, \ldots, p_{\max}\}$, and the number of states, $M \in \{1, \ldots, M_{\max}\}$, using marginal densities of data (estimation of all the models is required).

Step 2: Set the restrictions. For the chosen model, derive restrictions on parameters.

Step 3: Test restrictions (A1) and (A2). Estimate the restricted models and compute for them marginal densities of data. Compare the restricted models to the the unrestricted one using the Posterior Odds Ratio, e.g. according to the scale proposed by Kass & Raftery (1995).

Step 4: Test hypotheses of noncausality. If the model restricted according to (A1) is preferred to the unrestricted model, then noncausality of all kinds is established. In the other case, if the model restricted according to (A2) is preferred to the unrestricted model, in order to test different noncausality hypotheses use conditions (A6)–(A7). In the opposite case use conditions (A3)–(A5). For testing, use the Posterior Odds Ratio as in Step 3.

Advantages and costs of the proposed approach. We start by naming the main advantages of the proposed Bayesian approach to testing the restrictions for Granger noncausality. First, using the Posterior Odds Ratio testing principle, we avoid all the problems of testing nonlinear restrictions on the parameters of the model that appear in classical tests. Secondly, in the context of the controversies concerning the choice of number of states for Markov-switching models in the classical approach (see Psaradakis & Spagnolo, 2003), the Bayesian model selection proposed in Step 1 is a proper method free of such problems. Next, as emphasized in Hoogerheide, van Dijk & van Oest (2009), the Bayesian Posterior Odds Ratio procedure gives arguments *in favour of* hypotheses. Accordingly, the hypothesis preferred by the data is not only *rejected* or *not rejected*, but is actually *accepted* with some probability. Finally, Bayesian estimation is a basic estimation procedure proposed for the MS-VAR models and is broadly described and used in many applied publications.

However, this approach has also its costs. First of all, in order to specify the complete model, prior distributions for the parameters of the model and the prior probabilities of models need to be specified. This necessity gives way to subjective interpretation of the inference, on the one hand, but on the others it may ensure economic interpretation of the model. The other cost of the implementation of the Bayesian approach is the time required for simulation of all the models, first in the model selection procedure, and second in testing the restrictions of the parameters.

6. The Block Metropolis-Hastings sampler for restricted MS-VAR models

This section scrutinizes the MCMC sampler set up for sampling from the full conditional distributions. Each step describes the full conditional distribution of one element of the partitioned parameter vector. The parameter vector is broken up into five blocks: the vector of the latent states of the economy $\mathbf{S}$, the transition probabilities $\mathbf{P}$, the regime-dependent covariance matrices (themselves decomposed into standard

deviations σ and correlations $\mathbf{R}$), and finally the regime-dependent vector of constants plus autoregressive parameters β. For each block of parameters – conditionally on the parameter draws from the four other blocks – we describe how we sample from the posterior distribution. The symbols, l and $l-1$, refer to the iteration of the MCMC sampler. For the first iteration of a MCMC run, $l = 1$, initial parameter values come from an EM algorithm. The rest of this section describes all the constituting blocks that form the MCMC sampler.

6.1. Sampling the vector of the states of the economy

The first drawn parameter is the vector representing the states of the economy, $\mathbf{S}$. Being a latent variable, there are no priors nor restrictions on $\mathbf{S}$. We first use a filter (see Section 11.2 of Frühwirth-Schnatter, 2006, and references therein) and obtain the probabilities $\Pr(s_t = i|\mathbf{y}, \theta^{(l-1)})$, for $t = 1, \ldots, T$ and $i = 1, \ldots, M$, and then draw $\mathbf{S}^{(l)}$, for l^{th} iteration of the algorithm. For the full description of the algorithm used in this work the reader is referred to Droumaguet & Woźniak (2012).

6.2. Sampling the transition probabilities

In this step of the MCMC sampler, we draw from the posterior distribution of the transition probabilities matrix, conditioning on the states drawn in the previous step of the current iteration, $\mathbf{P}^{(l)} \sim p(\mathbf{P}|\mathbf{S}^{(l)})$. For the purpose of testing, we impose restrictions of identical rows of $\mathbf{P}$. Sims et al. (2008) provide a flexible analytical framework for working with restricted transition probabilities, and the reader is invited to consult Section 3 of that work for an exhaustive description of the possibilities provided by the framework. We however limit the latitude given by the reparametrization in order to ensure the stationarity of Markov chain $\mathbf{S}$.

Reparametrization. The transitions probabilities matrix $\mathbf{P}$ is modeled with Q vectors w_j, $j = 1, \cdots, Q$ and each of size d_j. Let all the elements of w_j belong to the $(0, 1)$ interval and sum up to one, and stack all of them into the column vector $\mathbf{w} = (w_1', \ldots, w_Q')'$ of dimension $d = \sum_{j=1}^{Q} d_j$. Writing $p = vec(\mathbf{P}')$ as a M^2 dimensional column vector, and introducing the $(M^2 \times d)$ matrix $\mathbf{M}$, the transition matrix is decomposed as:

$$p = \mathbf{M}\mathbf{w}, \tag{33}$$

where the $\mathbf{M}$ matrix is composed of the M_{ij} sub-matrices of dimension $(M \times d_j)$, where $i = 1, \ldots, M$, and $j = 1, \ldots, Q$:

$$\mathbf{M} = \begin{bmatrix} M_{11} & \ldots & M_{1Q} \\ \vdots & \ddots & \\ M_{M1} & & M_{MQ} \end{bmatrix},$$

where each M_{ij} satisfies the following conditions:

1. For each (i, j), all elements of M_{ij} are non-negative.
2. $\iota_M' M_{ij} = \Lambda_{ij} \iota_{d_j}'$, where Λ_{ij} is the sum of the elements in any column of M_{ij}.
3. Each row of $\mathbf{M}$ has, at most, one non-zero element.
4. M is such that $\mathbf{P}$ is irreducible: for all $j, d_j \geq 2$.

The first three conditions are inherited from Sims et al. (2008), whereas the last condition assures that $\mathbf{P}$ is irreducible, forbidding the presence of an absorbing state that would render the Markov chain $\mathbf{S}$ non-stationary. The non-independence of the rows of $\mathbf{P}$ is described in Frühwirth-Schnatter (2006, Section 11.5.5). Once the initial state s_0 is drawn from the ergodic distribution π of $\mathbf{P}$, direct MCMC sampling from the conditional posterior distribution becomes impossible. However, a Metropolis-Hastings algorithm can be set up to circumvent this issue, since a kernel of joint posterior density of all rows is known: $p(\mathbf{P}|\mathbf{S}) \propto \prod_{j=1}^{Q} \mathcal{D}_{d_j}(w_j)\pi$. Hence, the proposal for transition probabilities is obtained by sampling each w_j from the convenient Dirichlet distribution. The priors for w_j follow a Dirichlet distribution, $w_j \sim \mathcal{D}_{d_j}(b_{1,j}, \ldots, b_{d_j,j})$. We then transform the column vector $\mathbf{w}$ into our candidate matrix of transitions probabilities using equation (33). Finally, we compute the acceptance rate before retaining or discarding the draw.

Algorithm 1. *Metropolis-Hastings for the restricted transition matrix.*

1. $s_0 \sim \pi$. The initial state is drawn from the ergodic distribution of $\mathbf{P}$.
2. $w_j \sim \mathcal{D}_{d_j}(n_{1,j} + b_{1,j}, \ldots, n_{d_j,j} + b_{d_j,j})$ for $j = 1, \ldots, Q$. $n_{i,j}$ corresponds to the number of transitions from state i to state j, counted from $\mathbf{S}$. The candidate transition probabilities matrix – in the transformed notation – are sampled from a Dirichlet distribution.
3. $\mathbf{P}^{new} = \mathbf{Mw}$. The proposal for the transitions probabilities matrix is reconstructed.
4. Accept $\mathbf{P}^{new}$ if $u \leq \frac{\pi^{new}}{\pi^{l-1}}$, where $u \sim \mathcal{U}[0,1]$. π^{new} and π^{l-1} are the vectors of the ergodic probabilities resulting from the draws of the transition probabilities matrix $\mathbf{P}^{new}$ and $\mathbf{P}^{l-1}$ respectively.

6.3. Sampling a second and independent hidden Markov process

Regime inference from proposition (A1) involves two independent Markov processes. Equation (22) decomposes the vector of observations into two sub-vectors. Equations contained within each sub-vector are subject to switches from a different and independent Markov process. Sims et al. (2008, section 3.3.3) cover a similar decomposition.

Adding a Markov process is trivial in the sense it involves repeating the steps of Section 6.1 and algorithm 1 subsequently for a second process, yielding two distinct transition probabilities matrices $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$. The transition probabilities matrix for the whole system is formed out of the transition probabilities matrices of two independent hidden Markov processes, $\mathbf{P} = (\mathbf{P}^{(1)} \otimes \mathbf{P}^{(2)})$.

6.4. Sampling the covariance matrices

Adapting the approach proposed by Barnard et al. (2000) to Markov-switching models, we sample from the full conditional distribution of non-restricted and restricted covariance matrices. We thus decompose each covariance matrix of the MSIAH-VAR process into a vector of standard deviations (σ_{s_t}) and a correlation matrix ($\mathbf{R}_{s_t}$) using the equality:

$$\Sigma_{s_t} = \text{diag}(\sigma_{s_t})\mathbf{R}_{s_t}\text{diag}(\sigma_{s_t}).$$

This decomposition – statistically motivated – enables the partition of the covariance matrix parameters into two categories that are well suited for the restrictions we want to impose on the matrices. In a standard covariance matrix, restricting a variance parameter to some value has some impact on the depending covariances, whereas here variances and covariances (correlations) are treated as separate entities. The second and not the least advantage of the approach of Barnard et al. (2000) lies in the employed estimation procedure, the griddy-Gibbs sampler. The method introduced in Ritter & Tanner (1992) is well suited for sampling from an unknown univariate density $p(\mathbf{X}_i|\mathbf{X}_j, i \neq j)$. This is done by approximating the inverse conditional density function, which is done by evaluating $p(\mathbf{X}_i|\mathbf{X}_j, i \neq j)$ thanks to a grid of points. Imposing the desired restrictions on the parameters, and afterwards iterating a sampler for every standard deviation $\sigma_{i.s_t}$ and every correlation $\mathbf{R}_{j.s_t}$, we are able to simulate desired posteriors of the covariance matrices. While adding to the overall computational burden, the griddy-Gibbs sampler gives us full latitude to estimate restricted covariance matrices of the desired form.

Algorithm 2. *Griddy-Gibbs for the standard deviations.* The algorithm iterates on all the standard deviation parameters $\sigma_{i.s_t}$ for $i = 1, \ldots, N$ and $s_t = 1, \ldots, M$. Similarly to Barnard et al. (2000) we assume log-normal priors, $\log(\sigma_{i.s_t}) \sim \mathcal{N}(0, 2)$. The grid is centered on the residuals' sample standard deviation $\hat{\sigma}_{i.s_t}$ and divides the interval $(\hat{\sigma}_{i.s_t} - 2\hat{\sigma}_{\hat{\sigma}_{i.s_t}}, \hat{\sigma}_{i.s_t} + 2\hat{\sigma}_{\hat{\sigma}_{i.s_t}})$ into G grid points. $\hat{\sigma}_{\hat{\sigma}_{i.s_t}}$ is an estimator of the standard error of the estimator of the sample standard deviation.

1. Regime-invariant standard deviations: Draw from the unknown univariate density $p(\sigma_i|\mathbf{y}, \mathbf{S}, \mathbf{P}, \beta, \sigma_{-i}, \mathbf{R})$. This is done by evaluating a kernel on a grid of points, using the proportionality relation, with the likelihood function times the prior: $\sigma_i|\mathbf{y}, \mathbf{S}, \mathbf{P}, \beta, \sigma_{-i}, \mathbf{R} \propto p(\mathbf{y}|\mathbf{S}, \theta) \cdot p(\sigma_i)$. Reconstruct the c.d.f. from the grid through deterministic integration and sample from it.
2. Regime-varying standard deviations: For all regimes $s_t = 1, \ldots, M$, draw from the univariate density $p(\sigma_{i.s_t}|\mathbf{y}, \mathbf{S}, \mathbf{P}, \beta, \sigma_{-i.s_t}, \mathbf{R})$, evaluating a kernel thanks to the proportionality relation, with the likelihood function times the prior: $\sigma_{i.s_t}|\mathbf{y}, \mathbf{S}, \mathbf{P}, \beta, \sigma_{-i.s_t}, \mathbf{R} \propto p(\mathbf{y}|\mathbf{S}, \theta) \cdot p(\sigma_{i.s_t})$.

Algorithm 3. *Griddy-Gibbs for the correlations* The algorithm iterates on all the correlation parameters $\mathbf{R}_{i.s_t}$ for $i = 1, \ldots, \frac{(N-1)N}{2}$ and $s_t = 1, \ldots, M$. Similarly to Barnard et al. (2000), we assume uniform distribution on the feasible set of correlations, $\mathbf{R}_{i.s_t} \sim \mathcal{U}(a, b)$, with a and b being the bounds that keep the implied covariance matrix positive definite; see the aforementioned reference for details of setting a and b. The grid divides (a, b) into G grid points.

1. Depending on the restriction scheme, set correlations parameters to 0.
2. Regime-invariant correlations: Draw from the univariate density $p(\mathbf{R}_i|\mathbf{y}, \mathbf{S}, \mathbf{P}, \beta, \sigma, \mathbf{R}_{-i})$, evaluating a kernel thanks to the proportionality relation, with the likelihood function times the prior: $\mathbf{R}_i|\mathbf{y}, \mathbf{S}, \mathbf{P}, \beta, \sigma, \mathbf{R}_{-i} \propto p(\mathbf{y}|\mathbf{S}, \theta) \cdot p(\mathbf{R}_i)$.
3. Regime-varying correlations: For all regimes $s_t = 1, \ldots, M$, draw from the univariate density $p(\mathbf{R}_{i.s_t}|\mathbf{y}, \mathbf{S}, \mathbf{P}, \beta, \sigma, \mathbf{R}_{-i.s_t})$, evaluating a kernel thanks to the proportionality relation, with the likelihood function times the prior: $\mathbf{R}_{i.s_t}|\mathbf{y}, \mathbf{S}, \mathbf{P}, \beta, \sigma, \mathbf{R}_{-\mathbf{R}_{i.s_t}} \propto p(\mathbf{y}|\mathbf{S}, \theta) \cdot p(\mathbf{R}_{i.s_t})$.

6.5. Sampling the vector autoregressive parameters

Finally, we draw the state-dependent autoregressive parameters, β_{s_t} for $s_t = 1, \ldots, M$. The Bayesian parameter estimation of finite mixtures of regression models when the realizations of states is known has been precisely covered in Frühwirth-Schnatter (2006, Section 8.4.3). The procedure consists of estimating all the regression coefficients simultaneously by stacking them into $\beta = (\beta_0, \beta_1, \ldots, \beta_M)$, where β_0 is a common regression parameter for each regime, and hence is useful for the imposing of restrictions of state invariance for the autoregressive parameters. The regression model becomes:

$$y_t = Z_t\beta_0 + Z_t D_{i.1}\beta_1 + \cdots + Z_t D_{i.M}\beta_M + \epsilon_t, \tag{34}$$

$$\epsilon_t \sim i.i.\mathcal{N}(\mathbf{0}, \Sigma_{s_t}). \tag{35}$$

We have here introduced the $D_{i.s_t}$, which are M dummies taking the value 1 when the regime occurs and set to 0 otherwise. A transformation of the regressors Z_T also has to be performed in order to allow for different coefficients on the dependent variables, for instance to impose zero restrictions on parameters. In the context of VARs, Koop & Korobilis (2010, Section 2.2.3) detail a convenient notation that stacks all the regression coefficients on a diagonal matrix for every equation. We adapt this notation by stacking all the regression coefficients for all the states on diagonal matrix. If $z_{n.s_t.t}$ corresponds to the row vector of $1 + Np$ independent variables for equation n, state s_t (starting at 0 for regime-invariant parameters), and at time t, the stacked regressor Z_t will be of the following form:

$$Z_t = \text{diag}(z_{1.0.t}, \ldots, z_{N.0.t}, z_{1.1.t}, \ldots, z_{N.1.t}, \ldots, z_{1.M.t}, \ldots, z_{N.M.t}).$$

This notation enables the restriction of each parameter, by simply setting $z_{n.s_t.t}$ to 0 where desired.

Algorithm 4. *Sampling the autoregressive parameters.* We assume normal prior for β, i.e. $\beta \sim \mathcal{N}(\mathbf{0}, \underline{V_\beta})$.

1. For all Z_ts, impose restrictions by setting $z_{n,s_t,t}$ to zero accordingly.
2. $\beta|\mathbf{y}, \mathbf{S}, \mathbf{P}, \sigma, \mathbf{R} \sim \mathcal{N}(\overline{\beta}, \overline{V_\beta})$. Sample β from the conditional normal posterior distribution, with the following parameters:

$$\overline{V_\beta} = \left(\underline{V_\beta^{-1}} + \sum_{t=1}^{T} Z_t' \Sigma_{s_t}^{-1} Z_t\right)^{-1}$$

and

$$\overline{\beta} = \overline{V_\beta}\left(\sum_{t=1}^{T} Z_t' \Sigma_{s_t}^{-1} y_t\right).$$

6.6. Simulating restrictions in the form of functions of the parameters.

Some of the restrictions for Granger noncausality presented in Section 4 will be in the form of complicated functions of parameters. Suppose some restriction is in the form:

$$\theta_i = g(\theta_{-i}),$$

where $g(.)$ is a scalar function of all the parameters of the model but θ_i. The restricted parameter, θ_i, in this study may be one of the parameters from the autoregressive parameters, β, or standard deviations, σ. In such a case, the full conditional distributions for β or σ are no longer independent and need to be simulated with a Metropolis-Hastings algorithm.

Restriction on the vector autoregressive parameters β. In this case, the deterministic function restricting parameter β_i will be of the following form:

$$\beta_i = g(\beta_{-i}, \sigma, \mathbf{R}, \mathbf{P}).$$

We draw from the full conditional distribution of the vector autoregressive parameters, $p(\beta|\mathbf{y}, \mathbf{S}, \mathbf{P}, \sigma, \mathbf{R})$, using the Metropolis-Hastings algorithm:

Algorithm 5. *Metropolis-Hastings for the restricted vector autoregressive parameters β.*

1. Form a candidate draw, β^{new}, using Algorithm 6.
2. Compute the probability of acceptance of a draw:

$$\alpha(\beta^{l-1}, \beta^{new}) = \min\left[\frac{p(\mathbf{y}|\mathbf{S}, \mathbf{P}, \beta^{new}, \sigma, \mathbf{R})p(\beta^{new})}{p(\mathbf{y}|\mathbf{S}, \mathbf{P}, \beta^{l-1}, \sigma, \mathbf{R})p(\beta^{l-1})}, 1\right]. \tag{36}$$

3. Accept β^{new} if $u \leq \alpha(\beta^{l-1}, \beta^{new})$, where $u \sim \mathcal{U}[0, 1]$.

The algorithm has its justification in the block Metropolis-Hastings algorithm of Greenberg & Chib (1995). The formula for computing the acceptance probability from equation (36) is a consequence of the choice of the candidate generating distributions. For the parameters β_{-i}, it is a symmetric normal distribution, as in step 2 of Algorithm 4, whereas β_i is determined by a deterministic function.

Algorithm 6. *Generating a candidate draw β.*

1. Restrict parameter β_i to zero. Draw all the parameters $(\beta_1, \ldots, \beta_{i-1}, \mathbf{0}, \beta_{i+1}, \ldots, \beta_k)'$ according to the algorithms described in Section 6.5.
2. Compute $\beta_i = g(\beta_{-i}, \sigma, \mathbf{R}, \mathbf{P})$.
3. Return the vector $(\beta_1, \ldots, \beta_{i-1}, \mathbf{g}(\beta_{-\mathbf{i}}, \sigma, \mathbf{R}, \mathbf{P}), \beta_{i+1}, \ldots, \beta_k)'$

7. Granger causal analysis of US money-income data

In both studies focusing on Granger causality analysis within Markov-switching vector autoregressive models, Warne (2000) and Psaradakis et al. (2005),[3] the focus of study is the causality relationship between U.S. money and income. At the heart of this issue is the empirical analysis conducted in Friedman & Schwartz (1971) asserting that money changes led income changes. The methodology was rejected by Tobin (1970) as a *post hoc ergo propter hoc* fallacy, arguing that the timing implications from money to income could be generated not only by monetarists' macroeconomic models but also by Keynesian models. Sims (1972) initiated the econometric analysis of the causal relationship from the Granger causality perspective. While a Granger causality study concentrates on forecasting outcomes, macroeconomic theoretical modeling tries to remove the question mark over the neutrality of monetary policy for the business cycle. The causal

[3]The total US economic activity is approached from two different perspectives in these papers: Warne (2000) uses monthly income data, whereas Psaradakis et al. (2005) use quarterly output data.

relationship between money and income is, however, of particular interest to the econometric debate, since over the past forty years researchers have not reached a consensus.

This historical debate between econometricians is well narrated by Psaradakis et al. (2005), and the interested reader is advised to consult this paper for a depiction of events. Without detailing the references of the aforementioned paper, there is a problem in the instability of the empirical results found for the causality between money and output. Depending on the samples considered (postwar onwards data, 1970s onwards data, 1980s onwards, 1980s excluded, etc.), the existence and intensity of the causal effect of money on output are subject to different conclusions. Hence, the strategy of Psaradakis et al. (2005): to set up a Markov-switching VAR model that assumes four states of the economy: 1. both variables cause each other; 2. money does not cause output; 3. output does not cause money; 4. none of the variables causes another.

As outlined in the introduction, with the approach of Warne (2000) which we follow, the MS-VAR models are 'standard' ones, and we perform Bayesian model selection through the comparison of their marginal densities of data, to determine the number of states as well as the number of autoregressive lags. Moreover, we perform an analysis with precisely stated definitions of Granger causality for Markov-switching models. In this section, we use the Bayesian testing apparatus to investigate this relationship once again.

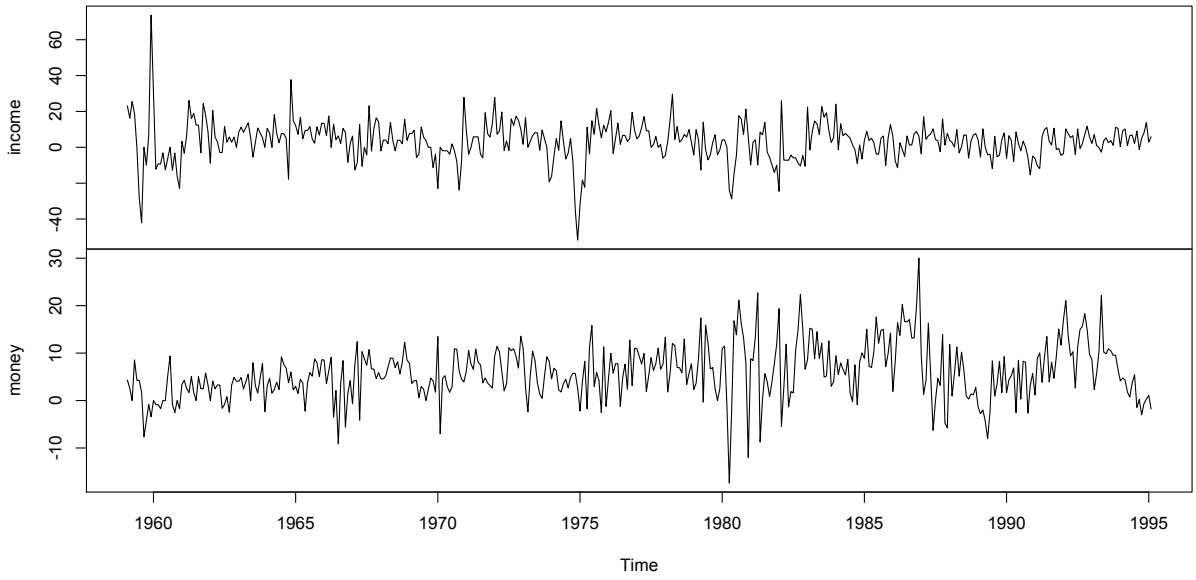

Figure 1: Log-differentiated series of money and income.

Data. The data are identical to those estimated by Warne (2000) and cover the same time period as in the original paper. Two monthly series are included, the US money stock M1 and the industrial production, both containing 434 observations covering the period, from 1959:1 to 1995:2, and both were extracted from the Citibase database. As in the original paper, the data are seasonally adjusted, transformed into log levels, and multiplied by 1200. Warne (2000) performed Johansen tests for cointegration, and – unlike for level series – trace statistics indicated no cointegration for differentiated series. Similarly, we work with the first difference of the series.

The summary statistics of both series are presented in Table 1. Income grows yearly by 3% on average, with a standard deviation of 11%, which seems a lot, but one has to note that we manipulate the monthly series for which the rates are annualized. Money has a stronger growth rate of nearly 6% on average, with a lower standard deviation than the income, below 6%.

Figure 1 plots the transformed series. Observation indicates that at least some heteroskedasticity is present, as can be seen with the money series, where a period of higher volatility starts around 1980.

Table 1: Summary statistics

Variable	Mean	Median	Standard Deviation	Minimum	Maximum
Δy	3.396	4.18	10.99	-51.73	73.72
Δm	5.851	5.24	5.79	-17.39	30.03

Data Source: Citibase.

Table 2: Model selection for VAR(p) – determination of number of lags

Lags	0	1	2	3	4	5	6	7	8
lnMHM	-3149.63	-2991.7	-2983.4	-2966.49	-2970.25	-2954.49	-2948.57	-2944	-2939.52
Lags	9	10	11	12	13	14	15	16	17
lnMHM	-2936.67	-2941.2	-2917.97	**-2916.77**	-2917.87	-2926.21	-2923.23	-2930.82	-2936.96

Summary statistics and series observations all seem to indicate the possibility of different states in the series, in which case MS-VAR models can provide a useful framework for analysis. We, however, start our analysis with Granger causality testing in the context of linear VAR models.

Granger causal analysis with the VAR model. The reason why we begin by studying Granger causality with linear models is that we want to relate to the standard methodology, and to illustrate whether a non-linear approach brings added value to the analysis by comparing the results. Also, the Block Metropolis-Hastings sampler of Section 6 can easily be simplified to a Block Metropolis-Hastings sampler for VAR models. By doing so, estimating linear VAR models and comparing marginal densities, we will also compare whether or not these models are preferred by the data to more complex MS-VAR ones.

We estimate the data with the VAR models for different lag lengths, $p = 0, \ldots, 17$. Each of the Metropolis-Hastings algorithms is initiated by the OLS estimates of the VAR coefficients. Then follows a 10,000-iteration burn-in and, after convergence of the sampler, 5000 final draws are to constitute the posteriors. The prior distributions are as follow:

$$\begin{aligned} \beta_i &\sim \mathcal{N}(\mathbf{0}, 100 I_{N+pN^2}) \\ \sigma_{i.j} &\sim log\mathcal{N}(0, 2) \\ \mathbf{R}_{i.jk} &\sim \mathcal{U}(a, b) \end{aligned}$$

for $i = 1, \ldots, M$ and $j, k = 1, \ldots, N$ and a and b as in Algorithm 3.

Table 2 displays the marginal density of data for each model, computed with the modified harmonic mean obtained by applying formula (32) to the posteriors draws. As in Warne (2000), models with long lags are preferred. The VAR(12) model, i.e. with 12 lags for the autoregressive coefficients, yields the highest lnMHM and hence is the model we choose for the Granger causality analysis. Table B.7 in Appendix B displays, for each parameter of the model, the mean, standard deviations, naive standard errors, autocorrelations of the Markov Chain at lag 1 and lag 10. Low autocorrelation at lag 10 indicates that the sampler has good properties.

The set of restrictions to impose on the parameters for vector autoregressive moving average models were covered in Sims (1972) and Boudjellaba et al. (1992). Translated into the VAR representation, and in the case of a bivariate VAR(p) model:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \sum_{i=1}^{p} \begin{bmatrix} A_{11}^{(i)} & A_{12}^{(i)} \\ A_{21}^{(i)} & A_{22}^{(i)} \end{bmatrix} \begin{bmatrix} y_{1,t-i} \\ y_{2,t-i} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix},$$

for $t = 1, \ldots, T$, the restrictions for money, $y_{2,t}$, being Granger noncausal on income, $y_{1,t}$, read:

$$A_{12}^{(i)} = 0 \text{ for } i = 1, \ldots, p.$$

Note that these restrictions, with assumed normal residual terms, are simultaneously encompassing Granger noncausality in mean, variance, and distribution.

Table 3: Noncausality and conditional regime independence in a VAR(12) model. Numerical efficiency results for these models are presented in table C.9 of Appendix C.

$\mathcal{M}_j$	Hypothesis	Restrictions	# restrictions	$\ln p(\mathbf{y}\|\mathcal{M}_j)$	$\ln \mathcal{B}_{j0}$
	$\mathcal{H}_0$: *Unrestricted model*				
$\mathcal{M}_0$	VAR(12)	-	0	-2,916.77	0
	$\mathcal{H}_1$: *Granger noncausality from money to income*				
$\mathcal{M}_1$	(A1)	$A_{12}^{(i)} = 0$	p	-2,901.63	15.13
for $i = 1, \ldots, p$.					

The estimation of the restricted VAR(12) model, with its upper-right autoregressive coefficients $A_{12}^{(i)}$ set to 0 for all lags returns posteriors that yield a lnMHM of -2901.63. Expressed in logarithms, the posterior odds ratio of the null hypothesis of Granger causality from money to income is equal to 15.13. Table 3 summarizes the results for VAR models. This is a very strong acceptance of the restricted model $\mathcal{M}_1$ over the nonrestricted one $\mathcal{M}_0$, hence Bayesian testing provides evidence in favor of Granger noncausality from money to income, within the VAR framework. This result is in line with Christiano & Ljungqvist (1988), where Granger noncausality from money to output is established for the VAR model with log-differences with US data. The authors contest this result and argue for a specification error for models with first differences. We continue our analysis with nonlinear models that allow switches within their parameters.

Granger causal analysis with MS-VARs. We estimate the data MSIAH(m)-VAR(p) models for different number of regimes $M = 2, 3, 4$ and different lag lengths, $p = 0, \ldots, 6$. Each of the Block Metropolis-Hastings algorithm is initiated by the estimates from the EM algorithm of the corresponding model. Then follows a 10,000-iteration burn-in and, after convergence of the sampler, we sample 5000 final draws from the posteriors. The prior distributions are as defined in Section 2.

Table 4 reports the lnMHMs for the estimated models with 2 regimes. Though we also estimated models with 3 or 4 regimes, estimation encountered difficulties of low occurrences of regimes. These phenomena indicate that the data does not support MS-VAR models with 3 or more regimes, and explains why we only present results with 2 regimes. The number of estimated lags for the autoregressive coefficients is limited to 6 lags – less than the 12 lags for VAR models – also due to insufficient state occurrences when the number of AR parameters increases. The model preferred by the data is the MSIAH(2)-VAR(4), i.e. with 2 regimes and VAR process of order 4. Table B.8 in Appendix B displays, for each parameter of the model, the mean, standard deviations, naive standard errors, and autocorrelations of the Markov chains at lag 1 and lag 10. Decaying autocorrelation between draws indicates that the sampler has desirable properties.

Figure 2 plots the regime probabilities from the selected model. In comparison with the second regime, the first regime matches times of higher variance for both variables. As well the constant for income growth, $\mu_{1,1}$, is negative during the occurrences of the first regime. Hence, the first regime can be interpreted as the bad state of the economy.

Note that comparing the best unrestricted MS-VAR model from Table 4 to the best VAR model of Table 3 (that is to the restricted model) yields a logarithm of the posterior odds ratio of 6.41 in favor of the MS-VAR model.

Similarly to Warne (2000), we proceed with the analysis of Granger noncausality for the selected MSIAH(2)-VAR(4) model. The Bayesian testing strategy we employ renders the process straightforward:

Table 4: Model selection for MSIAH(2)-VAR(p) – determination of the lag order

Lags	0	1	2	3	4	5	6
lnMHM	-3,002.64	-2,926.42	-2,903.89	-2,898.21	**-2,895.22**	-2,914.87	-2,913.49

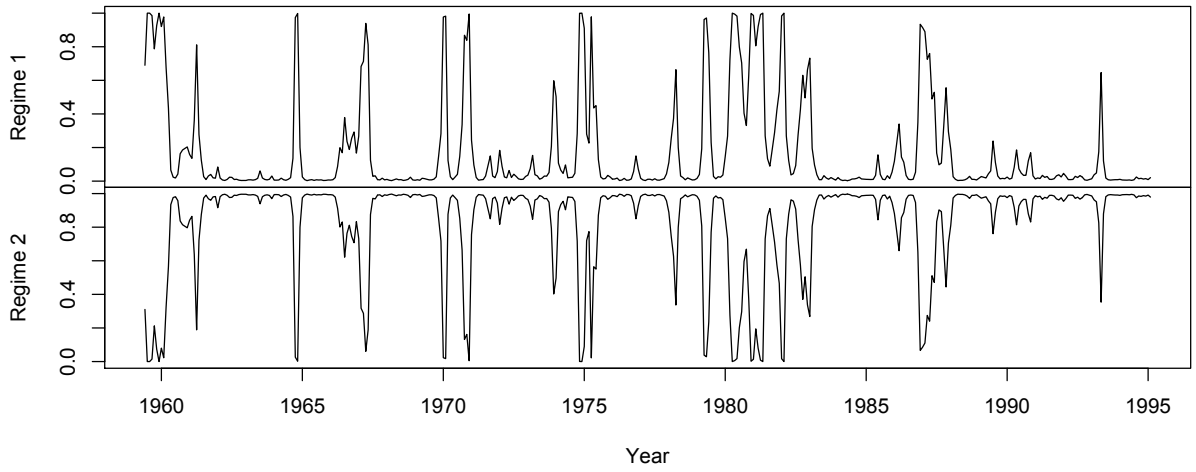

Figure 2: Estimated probabilities of regimes for a MSIAH(2)-VAR(4) model

each type of causality implies different restrictions on the model parameters; we impose them, estimate the models and compute all marginal densities of data. Table 5 summarizes all the sets of restrictions to impose when testing the noncausality from money to income, and also logarithms of the marginal densities of data given the model, $\ln p(\mathbf{y}|\mathcal{M}_j)$, and logarithms of the Bayes factors, $\ln \mathcal{B}_{j0}$ for $j = 1, \ldots, 7$. A positive logarithm of the Bayes factor is to be interpreted as evidence in favor of the restricted model. In a symmetric way, negative logarithm of the Bayes factor indicates that the non-restricted model is preferred by the data.

Analysis of Table 5 shows that only model $\mathcal{M}_5$ is more probable *a posteriori* than the unrestricted model $\mathcal{M}_0$. This model represents one of the sets of restrictions for Granger noncausality in mean. All other models, however, are less probable than the unrestricted model, which is represented with the negative values of the logarithms of the Bayes factors.

Table 6 presents a summary of the assessment of the considered hypotheses. We found strong support for Granger noncausality in mean. This hypothesis has much bigger posterior probability compared to all other hypotheses, including the unrestricted model. Warne (2000) found a similar result, but holding only at the 10% level of significance. However, Bayesian testing establishes this strong result, and the conditional mean of income is invariant to the history of money.

However, when it comes to the regime inference, we find a strong evidence that money play a significant role in forecasting the state of the economy one period ahead. In Table 6, in the second row, we read that the hypothesis that the history of money aggregate M1 does not impact on the regime forecast of income is rejected. It is evidenced by the negative value of Posterior Odds Ratio comparing this hypothesis to the unrestricted model.

Summary. The results of Bayesian testing for Granger causality from money to input on the US monthly series covering the period 1959–1995 are in line with the narration of Psaradakis et al. (2005), in the sense that the strongly established result of noncausality in mean within VAR models (which is equivalent to the noncausality in variance and in distribution) changes within the MS-VAR models. Allowing non-linearity

Table 5: Noncausality and conditional regime independence in a MSIAH(2)-VAR(4) model. Numerical efficiency results for these models are presented in table C.9 of Appendix C.

$\mathcal{M}_j$	Hypothesis	Restrictions	# restrictions	$\ln p(\mathbf{y}\|\mathcal{M}_j)$	$\ln \mathcal{B}_{j0}$
	$\mathcal{H}_0$: *Unrestricted model*				
$\mathcal{M}_0$	MS(2)-VAR(4)	-	0	-2895.22	0
	$\mathcal{H}_1$: *History of money does not impact on the regime forecast of income*				
$\mathcal{M}_1$	(A1) $M_1 = 1, M_2 = 2$	$\mu_{1,s_t} = \mu_1, A^{(i)}_{11,s_t} = A^{(i)}_{11}, A^{(i)}_{12,s_t} = 0$ $\Sigma_{11,s_t} = \Sigma_{11}, \Sigma_{12,s_t} = 0$	$3p$+4	-2964.72	-69.50
$\mathcal{M}_2$	(A1) $M_1 = 2, M_2 = 1$	$\mu_{2,s_t} = \mu_2, A^{(i)}_{21,s_t} = A^{(i)}_{21}, A^{(i)}_{22,s_t} = A^{(k)}_{22}$, $\Sigma_{22,s_t} = \Sigma_{22}, \Sigma_{12,s_t} = 0, A^{(i)}_{12,s_t} = 0$	$4p$+4	-2921.54	-26.32
	(A2) $M_1 = 1, M_2 = 2$	Always holds, no restrictions	-	-	-
$\mathcal{M}_3$	(A2) $M_1 = 2, M_2 = 1$	$p_{11} = p_{21}$	1	-2907.39	-12.17
	$\mathcal{H}_2$: *Granger noncausality*				
	(A1) or	-	-	-	-
$\mathcal{M}_4$	(A4) $M_1 = 1, M_2 = 2$	$\mu_{1,s_t} = \mu_1, A^{(i)}_{11,s_t} = A^{(i)}_{11}, A^{(i)}_{12,s_t} = 0$	$3p$+1	-2880.63	14.59
$\mathcal{M}_5$	(A4) $M_1 = 2, M_2 = 1$	$p_{11} = p_{21}, \sum_{j=1}^{2} A^{(i)}_{12,j}\pi_j = 0$	p+1	-2897.24	-2.02

for $i = 1, \ldots, p$.

Note: The derived restrictions are such that Model $\mathcal{M}_1$ is nested within Model $\mathcal{M}_4$. In order to use the Posterior Odds Ratio for the hypotheses from equation (31), we formally exclude from the parameter space of Model $\mathcal{M}_4$ the point restrictions on the covariance matrices that characterize Model $\mathcal{M}_1$. This way we subtract from the parameter space a region of zero probability mass, which does not impact on the posterior simulations.

Table 6: Summary of the hypotheses testing

$\mathcal{H}_i$	Hypothesis	Represented by models	$\ln \frac{\Pr(\mathcal{H}_i\|\mathbf{y})}{\Pr(\mathcal{H}_0\|\mathbf{y})}$
$\mathcal{H}_0$	Unrestricted model	$\mathcal{M}_0$	0
$\mathcal{H}_1$	History of money does not impact on the regime forecast of income	$\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_4$	-12.17
$\mathcal{H}_2$	Granger noncausality	$\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_4$, $\mathcal{M}_5$	14.59

in the models' coefficients, here by a Markov chain permitting switches between regimes of the economy, and testing for causality from money to income yields a different result and the strong non-causal evidence is decomposed. We found that the history of money helps to predict the regimes of income. However, we did find evidence for Granger noncausality in mean from money to income, as did Warne (2000).

These findings have particular consequences for the forecasting of the income. Despite the fact that past information about money does not change the forecast of the conditional mean of income, it is still crucial for its modeling. Past observations of money improves the forecast of the state of the economy when modeled with a Markov-switching process. Therefore, if one is interested in forecasting regime switches in the income equation, then one should add the money variable into the considered system.

8. Conclusions

The contribution of this paper is the determination of a set of parametric restrictions, both for Granger causality analysis and for conducting regime inference, within framework of Markov-switching Vector Autoregressions models. The starting point for these concepts is the one-period ahead forecast errors for the conditional expectations operator.

It is shown that both Granger noncausality and regime inference are not associated with a unique set of restrictions on the parameters of the MS-VAR. However, the number of such sets is finite and depends on the dimension of the observable variable vector and on the number of Markov regimes. Granger noncausality generally result in some of these sets containing nonlinear restrictions, with the nonlinearity being dependent on the rank of the matrix with Markov transition probabilities. Moreover, the number of restrictions actually being tested depends on the rank of this matrix, thereby making these concepts difficult to deal with in classical testing.

We propose a method of testing the nonlinear restrictions for the hypotheses of Granger noncausality and for conducting regime inference. The employed Bayes factors and Posterior Odds Ratios overcome the limitations of the classical approach. It requires, however, an algorithm of estimation of the unrestricted model and of the restricted models, representing hypotheses of interest. The algorithm we proposed, allows for the restriction of all the groups of parameters of the model in an appropriate way. It combines several existing algorithms and improves them in order to maintain the desired properties of the model and the efficiency of estimation. The estimation method allows us to use all the existing methods of computing of the marginal density of data that are required for both Bayes factors and Posterior Odds Ratios.

In the empirical illustration of the methodology, we have found that in the USA money does not cause income in mean. We have, however, found that the money impacts on the forecast of the future state of the economy. If the empirical analysis is to be something more than just an illustration of the methodology, and in effect be conclusive, robustness checks are required. In particular, considering more relevant variables in the system could impact on the conclusions of the analysis of the Granger causality between money and income.

As the main limitation of the whole analysis of Granger causality for MS-VAR models, we find that only *one-period ahead* Granger causality is considered in this study. The conditions for *h-periods ahead* noncausality should be further explored. We only mention that potentially establishing that one variable does not improve the forecast of the hidden Markov process, taking into account the Markov property, may imply the same for all periods in the future. Still, establishing conditions for the noncausality h-periods ahead for the autoregressive parameters, including covariances, would potentially require tedious algebra. This statement is motivated by the complexity of formulating forecasts with MS-VAR models.

This line of research could be extended to other definitions of Granger causality, namely the second order Granger causality and the Granger causality in distribution. These two forms being more restrictive than the one we consider, a refined analysis on the causal nature between economic variables could be proposed.

Appendix A. Mathematical Appendix: Proofs

Proof of Proposition 1

It is straightforward to show that (A2) implies that there is no information in $\mathbf{v}_{2t}$ for predicting $s_{1,t+1}$ since it implies that $\Pr[s_{1,t+1}|\mathbf{y}_t] = \Pr[s_{1,t+1}]$. Let us therefore focus on the only remaining possibility, i.e. that $\Pr[s_{1,t}|\mathbf{y}_t] = \Pr[s_{1,t}|\mathbf{v}_{1,t}]$. To prove that condition (A1) is necessary and sufficient for this to hold, we shall proceed in two steps. The first step involves finding a general condition for predictions of $s_{1,t}$ (and $s_{2,t}$) to be invariant with respect to alternative information sets. In the second step we show that when $\epsilon_t|s_t$ is Gaussian, then the parameter restrictions in (A1) are necessary and sufficient for the invariance condition in the first step to be satisfied under the two information sets of interest.

Let $\xi_{t|\tau}(j) = \Pr[s_t = j|y_\tau, \mathcal{W}_\tau]$, where y_t is a vector of variables and $\mathcal{W}_\tau$ is the history of an observable vector w_t up to and including period τ. The vector w_t can, for example, be defined such that it contains y_{t-1} and various exogenous variables observable at time t. Furthermore, let $\eta_t(j) = f_{y_j}(y_t|s_t = j, \mathcal{W}_t)$ be the density function for y_t conditional on the state and the history of w_t. We stack these functions into $M \times 1$ vectors $\xi_{t|\tau}$ and η_t, respectively. From e.g. Hamilton (1994) we have that $\xi_{t|t}$, $\xi_{t|t-1}$, and η_t are related according to:

$$\xi_{t|t} = \frac{\left(\xi_{t|t-1} \odot \eta_t\right)}{\iota_q'\left(\xi_{t|t-1} \odot \eta_t\right)}, \qquad t = 1, 2, \ldots, \tag{A.1}$$

while

$$\xi_{t|t-1} = \mathbf{P}'\xi_{t-1|t-1}, \qquad t = 2, 3, \ldots, \tag{A.2}$$

and $\xi_{1|0} = \rho$, a $M \times 1$ vector of positive constants summing to unity. Here, $\odot$ denotes the Hadamard (element-by-element) product and ι_M the $M \times 1$ unit vector.

Let s_t be represented by two Markov processes, $s_{1,t}$ and $s_{2,t}$, which are not necessarily independent. Define j such that $j \equiv j_2 + M_2(j_1 - 1)$ when $(s_{1,t}, s_{2,t}) = (j_1, j_2)$, where $M_1, M_2 \geq 1$ and $M = M_1 M_2 \geq 2$. Then $\xi_{t|\tau}(j) = \xi_{t|\tau}(j_1, j_2) = \Pr[s_{1,t} = j_1, s_{2,t} = j_2|y_\tau, \mathcal{W}_\tau]$, while $\xi^{(1)}_{t|\tau}(j_1) = \sum_{j_2=1}^{M_2} \xi_{t|\tau}(j_1, j_2)$ and similarly for $\xi^{(2)}_{t|\tau}(j_2)$. More compactly, this means that $\xi^{(1)}_{t|\tau} = [I_{M_1} \otimes \iota'_{M_2}]\xi_{t|\tau}$ and $\xi^{(2)}_{t|\tau} = [\iota'_{M_1} \otimes I_{M_2}]\xi_{t|\tau}$. The following result about Hadamard and Kronecker products will prove useful below:

Lemma 1. *If and only if* $\eta_t = (\eta^{(1)}_t \otimes \eta^{(2)}_t)$ *with* $\eta^{(l)}_t$ *being* $M_l \times 1$ *for* $l = 1, 2$, *then*

$$\left(I_{M_1} \otimes \iota'_{M_2}\right)\left(\xi_{t|t-1} \odot \eta_t\right) = \left(\left[I_{M_1} \otimes \eta^{(2)\prime}_t\right]\xi_{t|t-1}\right) \odot \eta^{(1)}_t, \tag{A.3}$$

while

$$\left(\iota'_{M_1} \otimes I_{M_2}\right)\left(\xi_{t|t-1} \odot \eta_t\right) = \left(\left[\eta^{(1)\prime}_t \otimes I_{M_2}\right]\xi_{t|t-1}\right) \odot \eta^{(2)}_t. \tag{A.4}$$

Proof. The j:th element of $(\xi_{t|t-1} \odot \eta_t)$ is given by $\xi_{t|t-1}(j_1, j_2)\eta^{(1)}_t(j_1)\eta^{(2)}_t(j_2)$. Premultiplying this $M \times 1$ vector by $[I_{M_1} \otimes \iota'_{M_2}]$ we obtain a $M_1 \times 1$ vector whose j_1:th element is

$$\eta^{(1)}_t(j_1) \sum_{j_2=1}^{M_2} \xi_{t|t-1}(j_1, j_2)\eta^{(2)}_t(j_2).$$

Now define

$$\gamma_{t|t-1}(j_1) \equiv \begin{bmatrix} \xi_{t|t-1}(j_1, 1) \\ \vdots \\ \xi_{t|t-1}(j_1, M_2) \end{bmatrix}, \qquad j_1 = 1, \ldots, M_1. \tag{A.5}$$

Then

$$\gamma_{t|t-1}(j_1)'\eta^{(2)}_t = \sum_{j_2=1}^{M_2} \xi_{t|t-1}(j_1, j_2)\eta^{(2)}_t(j_2).$$

Collecting these results we find that

$$\left[I_{M_1} \otimes \iota'_{M_2}\right]\left[\xi_{t|t-1} \odot \left(\eta^{(1)}_t \otimes \eta^{(2)}_t\right)\right] = \begin{bmatrix} \gamma_{t|t-1}(1)'\eta^{(2)}_t \\ \vdots \\ \gamma_{t|t-1}(M_1)'\eta^{(2)}_t \end{bmatrix} \odot \eta^{(1)}_t. \tag{A.6}$$

Define the $M_2 \times M_1$ matrix $\gamma_{t|t-1}$ according to $\gamma_{t|t-1} \equiv [\gamma_{t|t-1}(1) \ \cdots \ \gamma_{t|t-1}(M_1)]$. It then follows that

$$\gamma'_{t|t-1}\eta_t^{(2)} = \begin{bmatrix} \gamma_{t|t-1}(1)'\eta_t^{(2)} \\ \vdots \\ \gamma_{t|t-1}(M_1)'\eta_t^{(2)} \end{bmatrix}. \tag{A.7}$$

Moreover, $\xi_{t|t-1} = \text{vec}(\gamma_{t|t-1})$, with vec being the column stacking operator. Next,

$$\begin{aligned} \gamma'_{t|t-1}\eta_t^{(2)} &= \left[\eta_t^{(2)\prime} \otimes I_{M_1}\right]\text{vec}\left(\gamma'_{t|t-1}\right) \\ &= \left[\eta_t^{(2)\prime} \otimes I_{M_1}\right]K_{M_2,M_1}\text{vec}\left(\gamma_{t|t-1}\right) \\ &= K_{M_1,1}\left[I_{M_1} \otimes \eta_t^{(2)\prime}\right]\xi_{t|t-1} \\ &= \left[I_{M_1} \otimes \eta_t^{(2)\prime}\right]\xi_{t|t-1}, \end{aligned} \tag{A.8}$$

where $K_{m,n}$ is the $mn \times mn$ commutation matrix, $K_{m,1} = I_m$, and the third equality follows by Theorem 3.9 in Magnus & Neudecker (1988). Collecting these last results we have established (A.3). The result (A.4) follows by similar arguments. □

If $s_{1,t}$ and $s_{2,t}$ are independent, it follows that

$$\begin{aligned} \xi_{t|t-1}^{(1)} &= \left[I_{M_1} \otimes \iota'_{M_2}\right]\left[\mathbf{P}^{(1)\prime} \otimes \mathbf{P}^{(2)\prime}\right]\xi_{t-1|t-1} \\ &= \mathbf{P}^{(1)\prime}\xi_{t-1|t-1}^{(1)}, \end{aligned} \tag{A.9}$$

since $\mathbf{P}^{(2)}\iota_{M_2} = \iota_{M_2}$. Similarly, $\xi_{t|t-1}^{(2)} = \mathbf{P}^{(2)\prime}\xi_{t-1|t-1}^{(2)}$. However, this does not mean that $\xi_{t|t-1}^{(1)}$ and $\xi_{t|t-1}^{(2)}$ are independent since $\xi_{t-1|t-1}^{(1)}$ and $\xi_{t-1|t-1}^{(2)}$ need not be independent.

Lemma 2. *If and only if* (i) $\eta_t = \varphi_t(\eta_t^{(1)} \otimes \eta_t^{(2)})$ *where* φ_t *is a scalar and* $\eta_t^{(l)}$ *a* $M_l \times 1$ *vector,* (ii) $\eta_t^{(1)}$ *and* $\eta_t^{(2)}$ *are vectors of density functions for independent random variables, and* (iii) $s_{1,t}$ *and* $s_{2,t}$ *are independent, then for all* $t = 1, \ldots, T$

$$\xi_{t|t}^{(l)} = \frac{\left(\xi_{t|t-1}^{(l)} \odot \eta_t^{(l)}\right)}{\iota'_{M_l}\left(\xi_{t|t-1}^{(l)} \odot \eta_t^{(l)}\right)}, \qquad l = 1, 2, \tag{A.10}$$

with $\xi_{t|\tau} = (\xi_{t|\tau}^{(1)} \otimes \xi_{t|\tau}^{(2)})$, *where* $\xi_{t|\tau}^{(1)}$ *and* $\xi_{t|\tau}^{(2)}$ *are independent for* $\tau = t, t-1$.

Proof. Note first that $\iota'_M = \iota'_{M_1}(I_{M_1} \otimes \iota'_{M_2}) = \iota'_{M_2}(\iota'_{M_1} \otimes I_{M_2})$. For $l = 1$ we know that $\xi_{t|t}^{(1)} = [I_{M_1} \otimes \iota'_{M_2}]\xi_{t|t}$. From equation (A.1) we thus have that

$$\begin{aligned} \xi_{t|t}^{(1)} &= \left[I_{M_1} \otimes \iota'_{M_2}\right]\left[\xi_{t|t-1} \odot \eta_t\right]\left[\iota'_{M_1}\left(I_{M_1} \otimes \iota'_{M_2}\right)\left(\xi_{t|t-1} \odot \eta_t\right)\right]^{-1} \\ &= \left[\left(\left[I_{M_1} \otimes \eta_t^{(2)\prime}\right]\xi_{t|t-1}\right) \odot \eta_t^{(1)}\right]\left[\iota'_{M_1}\left(\left[\left(I_{M_1} \otimes \eta_t^{(2)\prime}\right)\xi_{t|t-1}\right] \odot \eta_t^{(1)}\right)\right]^{-1}, \end{aligned} \tag{A.11}$$

by Lemma 1 and since the scalar φ_t cancels. A similar expression is obtained for $\xi_{t|t}^{(2)}$. Let $\rho = (\rho^{(1)} \otimes \rho^{(2)})$ where the elements of $\rho^{(l)}$ are positive and sum to unity. Then

$$\begin{aligned} \xi_{1|1}^{(1)} &= \left[\left(\rho^{(1)} \otimes \eta_1^{(2)\prime}\rho^{(2)}\right) \odot \eta_1^{(1)}\right]\left[\iota'_{M_1}\left(\left[\rho^{(1)} \otimes \eta_1^{(2)\prime}\rho^{(2)}\right] \odot \eta_1^{(1)}\right)\right]^{-1} \\ &= \left[\rho^{(1)} \odot \eta_1^{(1)}\right]\left[\iota'_{M_1}\left(\rho^{(1)} \odot \eta_1^{(1)}\right)\right]^{-1}, \end{aligned} \tag{A.12}$$

and similarly for $\xi_{1|1}^{(2)}$. By (ii) it follows that $\xi_{1|1}^{(1)}$ and $\xi_{1|1}^{(2)}$ are independent. Thus, $\xi_{1|1} = (\xi_{1|1}^{(1)} \otimes \xi_{1|1}^{(2)})$. Moreover, by (iii) we have that $\xi_{2|1}^{(l)} = \mathbf{P}^{(l)\prime}\xi_{1|1}^{(l)}$, which are also independent for $l = 1, 2$. Thus, $\xi_{2|1} = (\xi_{2|1}^{(1)} \otimes \xi_{2|1}^{(2)})$ and so on for $t = 2, 3, \ldots, T$, thereby establishing sufficiency.

To prove necessity, suppose (i) is not true, i.e., $M_i \geq 2$ for $i = 1, 2$. Let $\eta_t = (\eta_t^{(1)} \otimes \eta_t^{(2)}) \odot \psi_t$, where $\psi_t \neq (\psi_t^{(1)} \otimes \psi_t^{(2)})$ for $M_l \times 1$ vectors $\psi_t^{(l)}$. Then, for example

$$\begin{aligned}\xi_{t|t}^{(1)} &= \left[\left(I_{M_1} \otimes \eta_t^{(2)\prime}\right)\left(\xi_{t|t-1} \odot \psi_t\right) \odot \eta_t^{(1)}\right]\left[\iota_{M_1}'\left(\left[I_{M_1} \otimes \eta_t^{(2)\prime}\right]\left[\xi_{t|t-1} \odot \psi_t\right] \odot \eta_t^{(1)}\right)\right]^{-1} \\ &\neq \left[\left(\left[I_{M_1} \otimes \eta_t^{(2)\prime}\right]\xi_{t|t-1}\right) \odot \eta_t^{(1)}\right]\left[\iota_{M_1}'\left(\left[\left(I_{M_1} \otimes \eta_t^{(2)\prime}\right)\xi_{t|t-1}\right] \odot \eta_t^{(1)}\right)\right]^{-1}.\end{aligned} \tag{A.13}$$

The only case when the inequality can be replaced with an equality is if $\psi_t = (\psi_t^{(1)} \otimes \psi_t^{(2)})$. Next, if (ii) does not hold, then for instance $\xi_{1|1}^{(1)}$ and $\xi_{1|1}^{(2)}$ cannot be independent. Finally, if (iii) does not hold, then $\xi_{t|t-1}^{(1)} \neq \mathbf{P}^{(1)\prime}\xi_{t-1|t-1}^{(1)}$ and depends on $\xi_{t-1|t-1}^{(2)}$ as well. Thus, $\xi_{2|1}^{(1)}$ and $\xi_{2|1}^{(2)}$ cannot be independent even if $\xi_{1|1}^{(1)}$ and $\xi_{1|1}^{(2)}$ are. □

Assumptions (i) and (ii) in Lemma 2 are useful for the above proof, but can in practise be more conveniently expressed as restrictions on marginal and conditional densities via the decomposition $y_t = (v_{1t}, v_{2t})$. For all $j = 1, \ldots, M$ we may express the joint density for y_t as

$$\eta_t(j) = f_{y_j}\left(y_t | s_t = j, \mathcal{W}_t\right) = f_{v_{1j}}\left(v_{1t} | s_t = j, v_{2t}, \mathcal{W}_t\right) f_{v_{2j}}\left(v_{2t} | s_t = j, \mathcal{W}_t\right).$$

This standard decomposition ensures that the densities of interest concern independent random variables and may therefore be taken as an interpretation of assumption (ii) in Lemma 2 once the conditions that we consider next are met.

To deal with assumption (i) we first of all require that the marginal density for v_{2t} depends only on s_{2t}. That is, for all $j = 1, \ldots, M$:

$$f_{v_{2j}}\left(v_{2t} | s_t = j, \mathcal{W}_t\right) = f_{v_{2j_2}}\left(v_{2t} | s_{2t} = j_2, \mathcal{W}_t\right), \quad j_2 = 1, \ldots, M_2. \tag{A.14}$$

Concerning the conditional density for v_{1t} the restrictions can be written as:

$$f_{v_{1j}}\left(v_{1t} | s_t = j, v_{2t}, \mathcal{W}_t\right) = \begin{cases} f_{v_{1j_1}}\left(v_{1t} | s_{1t} = j_1, \mathcal{W}_t\right) & \text{if } M_2 > 1, \\ f_{v_{1j_1}}\left(v_{1t} | s_{1t} = j_1, v_{2t}, \mathcal{W}_t\right) & \text{otherwise} \end{cases} \tag{A.15}$$

for all $j_1 = 1, \ldots, M_1$ and $j = 1, \ldots, M$. In other words, the conditional density for v_{1t} must be such that it does not depend on s_{2t}. If $M_2 > 1$, then v_{2t} varies with s_{2t} and, hence, the density of v_{1t} must be invariant with respect to v_{2t}. On the other hand, when $M_2 = 1$, then by (A.14) we have that v_{2t} is invariant with respect to s_t and is therefore not otherwise required to be subject to a constraint. The restrictions in (A.14) and (A.15) are more convenient than assumptions (i) and (ii) when we attempt to determine the restrictions that a specific density function for y_t must satisfy.

In fact, the conditions in Lemma 2 have even further implications:

Lemma 3. *If and only if the conditions in Lemma 2 are satisfied, then*

$$\xi_{t|\tau} = \left(\xi_{t|\tau}^{(1)} \otimes \xi_{t|\tau}^{(2)}\right), \tag{A.16}$$

for all $t, \tau = 1, \ldots, T$, with $\xi_{t|\tau}^{(1)}$ and $\xi_{t|\tau}^{(2)}$ being independent.

Proof. Let us first prove this for all $\tau < t$. We have already established in Lemma 2 that $\xi_{\tau|\tau}^{(1)}$ and $\xi_{\tau|\tau}^{(2)}$ are independent for all τ. By equation (22.3.13) in Hamilton (1994) we have that $\xi_{t|\tau} = (\mathbf{P}')^{t-\tau}\xi_{\tau|\tau}$ for $\tau = 1, \ldots, t-1$. By independence of $s_{1,t}$ and $s_{2,t}$ and of $\xi_{\tau|\tau}^{(1)}$ and $\xi_{\tau|\tau}^{(2)}$ we obtain $\xi_{t|\tau} = [(\mathbf{P}^{(1)\prime})^{t-\tau}\xi_{\tau|\tau}^{(1)} \otimes (\mathbf{P}^{(2)\prime})^{t-\tau}\xi_{\tau|\tau}^{(2)}] = (\xi_{t|\tau}^{(1)} \otimes \xi_{t|\tau}^{(2)})$, which are thus independent.

To show (A.16) for $\tau > t$ it is sufficient to consider $\tau = T$ since the algorithm for computing smooth probabilities is valid for any $\tau > t$. From Kim (1994) (see also (Lindgren, 1978; Hamilton, 1994)) we get

$$\xi_{t|T} = \xi_{t|t} \odot \left[\mathbf{P}\left(\xi_{t+1|T} \ominus \xi_{t+1|t}\right)\right], \qquad t = 1, \ldots, T-1, \tag{A.17}$$

where $\ominus$ denotes element-by-element division. To show that $\xi_{t|T} = (\xi_{t|T}^{(1)} \otimes \xi_{t|T}^{(2)})$, with $\xi_{t|T}^{(l)}$ independent for $l = 1, 2$, we begin with $t = T - 1$. By Lemma 2 we have that $\xi_{T|\tau} = (\xi_{T|\tau}^{(1)} \otimes \xi_{T|\tau}^{(2)})$ for $\tau = T, T-1$. Accordingly,

$$\left[\xi_{T|T} \ominus \xi_{T|T-1}\right] = \left[\left(\xi_{T|T}^{(1)} \ominus \xi_{T|T-1}^{(1)}\right) \otimes \left(\xi_{T|T}^{(2)} \ominus \xi_{T|T-1}^{(2)}\right)\right]. \tag{A.18}$$

Let $\psi_T^{(l)} \equiv \mathbf{P}^{(l)}(\xi_{T|T}^{(l)} \ominus \xi_{T|T-1}^{(l)})$ for $l = 1, 2$. We then obtain

$$\mathbf{P}\Big[\xi_{T|T} \ominus \xi_{T|T-1}\Big] = \Big[\psi_T^{(1)} \otimes \psi_T^{(2)}\Big] \equiv \psi_T. \tag{A.19}$$

Hence, $\xi_{T-1|T} = (\xi_{T-1|T-1} \odot \psi_T)$. With $\xi_{t|T}^{(1)} = [I_{M_1} \otimes \imath'_{M_2}]\xi_{t|T}$ it follows by Lemma 1 and Lemma 2 that

$$\begin{aligned} \xi_{T-1|T}^{(1)} &= \Big[\Big(I_{M_1} \otimes \psi_T^{(2)\prime}\Big)\xi_{T-1|T-1}\Big] \odot \psi_T^{(1)} \\ &= \psi_T^{(2)\prime}\xi_{T-1|T-1}^{(2)}\Big(\xi_{T-1|T-1}^{(1)} \odot \psi_T^{(1)}\Big), \end{aligned} \tag{A.20}$$

since $\xi_{T-1|T-1} = (\xi_{T-1|T-1}^{(1)} \otimes \xi_{T-1|T-1}^{(2)})$. From the definition of $\psi_T^{(2)}$ we find that

$$\begin{aligned} \psi_T^{(2)\prime}\xi_{T-1|T-1}^{(2)} &= \Big(\xi_{T|T}^{(2)} \ominus \xi_{T|T-1}^{(2)}\Big)' \mathbf{P}^{(2)\prime}\xi_{T-1|T-1}^{(2)} \\ &= \Big(\xi_{T|T}^{(2)} \ominus \xi_{T|T-1}^{(2)}\Big)' \xi_{T|T-1}^{(2)} \\ &= \sum_{j_2=1}^{M_2} \xi_{T|T}^{(2)}(j_2). \end{aligned} \tag{A.21}$$

This is equal to unity and we thus get

$$\xi_{T-1|T}^{(1)} = \xi_{T-1|T-1}^{(1)} \odot \Big[\mathbf{P}^{(1)}\Big(\xi_{T|T}^{(1)} \ominus \xi_{T|T-1}^{(1)}\Big)\Big]. \tag{A.22}$$

Proceeding with $\xi_{T-1|T}^{(2)}$, the above arguments imply that

$$\xi_{T-1|T}^{(2)} = \xi_{T-1|T-1}^{(2)} \odot \Big[\mathbf{P}^{(2)}\Big(\xi_{T|T}^{(2)} \ominus \xi_{T|T-1}^{(2)}\Big)\Big], \tag{A.23}$$

and, hence, by Lemma 2, $\xi_{T-1|T}^{(l)}$ are independent for $l = 1, 2$ and $\xi_{T-1|T} = (\xi_{T-1|T}^{(1)} \otimes \xi_{T-1|T}^{(2)})$. For the remaining t, backwards recursions, using the above arguments, implies the result. Necessity follows by the arguments in Lemma 2. □

Notice that condition (i) of Lemma 2 is only sufficient in forecast situations. If s_t is serially uncorrelated, then $\mathbf{P}' = \pi\imath'_M$, with π being the vector of ergodic probabilities. Accordingly, for all $\tau < t$, $\xi_{t|\tau} = (\mathbf{P}')^{t-\tau}\xi_{\tau|\tau} = \pi$ since $\imath'_M\pi = \imath'_q\xi_{\tau|\tau} = 1$. Hence, if $s_{1,t}$ and $s_{2,t}$ are independent and serially uncorrelated, then $\xi_{t|\tau} = (\xi_{t|\tau}^{(1)} \otimes \xi_{t|\tau}^{(2)}) = (\pi^{(1)} \otimes \pi^{(2)})$ for all $\tau < t$.

This completes step one in the proof of Proposition 1. We have established necessary and sufficient conditions for how the information used to predict s_t can be split into information valuable for predicting $s_{1,t}$ but not $s_{2,t}$, and vice versa, and when information can be "thrown away" without affecting the regime predictions. Note that the conditions in Lemma 2 are very general in the sense that they apply to any vector of density functions η_t. For example, the functional form can vary over t as well as over states. The crucial underlying assumption is that s_t conditional on s_{t-1} is independent of information available at time $t-1$. If this assumption is violated, then the algorithms for computing regime predictions are no longer valid.

The assumption that $s_{1,t}$ and $s_{2,t}$ are independent, in fact, increases the level of generality of the results. For example, it allows $M_2 = 1$ in which case $\eta_t = \varphi_t\eta_t^{(1)}$ (with the scalar φ_t being a marginal density which is invariant with respect to s_t) is necessary and sufficient for regime predictions based on the vector densities η_t and $\eta_t^{(1)}$ to be equivalent.

When $M_1, M_2 \geq 2$ we allow for the possibility that two subsystems of the model can contain information for predicting one independent regime process each but not the other regime process, while a third subsystem is completely non-informative about regimes. By considering r independent Markov chains, these results can be generalized further. For our purposes, however, the above results are sufficient.

Now let us return to the MS-VAR with conditionally Gaussian residuals. Here we find that for each $j \in \{1, \ldots, M\}$ the joint log density is

$$\ln\Big(\eta_t(j)\Big) = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln\Big(\det\Big[\Sigma_j\Big]\Big) - \frac{1}{2}\epsilon'_{t|j}\Sigma_j^{-1}\epsilon_{t|j}, \tag{A.24}$$

where $\epsilon_{t|j} = y_t - \mu_j - \sum_{k=1}^{p} A_j^{(k)} y_{t-k}$. Let n_1 and n_2 be the number of $v_{1,t}$ and $v_{2,t}$ variables, respectively, with $n_1 + n_2 = N$. The marginal density for $v_{2,t}$, conditional on $s_t = j$ and $\mathbf{y}_{t-1}$, is

$$\ln\Big(\eta_t^{(2)}(j)\Big) = -\frac{n_2}{2}\ln(2\pi) - \frac{1}{2}\ln\Big(\det\Big[\Sigma_{22,j}\Big]\Big) - \frac{1}{2}\epsilon'_{2,t|j}\Sigma_{22,j}^{-1}\epsilon_{2,t|j}. \tag{A.25}$$

If this density is invariant with respect to $s_{1,t}$, then (a) $\Sigma_{22,(j_1,j_2)} = \Sigma_{22,j_2}$, $\mu_{2,(j_1,j_2)} = \mu_{2,j_2}$, and $A^{(k)}_{2r,(j_1,j_2)} = A^{(k)}_{2r,j_2}$ for all $j_1 \in \{1,\dots,M_1\}$, $j_2 \in \{1,\dots,M_2\}$, $r \in \{1,2\}$, and $k \in \{1,\dots,p\}$. For $M_2 = 1$ these restrictions imply that the parameters in the marginal density for $v_{2,t}$ are constant across states.

Under the restrictions in (a), the density for $v_{1,t}$, conditional on $s_t = j = j_2 + M_2(j_1 - 1)$, $v_{2,t}$, and $\mathbf{y}_{t-1}$, is

$$\begin{aligned}\ln\left(\eta^{(1)}_t(j)\right) = &-\frac{n_1}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\det\left[\tilde{\Sigma}_{11,j}\right]\right) + \epsilon'_{2,t|j_2}\Sigma^{-1}_{22,j_2}\Sigma'_{12,j}\tilde{\Sigma}^{-1}_{11,j}\epsilon_{1,t|j} \\ &- \frac{1}{2}\epsilon'_{1,t|j}\tilde{\Sigma}^{-1}_{11,j}\epsilon_{1,t|j} - \frac{1}{2}\epsilon'_{2,t|j_2}\Sigma^{-1}_{22,j_2}\Sigma'_{12,j}\tilde{\Sigma}^{-1}_{11,j}\Sigma_{12,j}\Sigma^{-1}_{22,j_2}\epsilon_{2,t|j_2},\end{aligned} \tag{A.26}$$

where $\tilde{\Sigma}_{11,j} \equiv \Sigma_{11,j} - \Sigma_{12,j}\Sigma^{-1}_{22,j_2}\Sigma'_{12,j}$. If this density function is invariant with respect to $s_{2,t}$ for $M_2 \geq 2$, then (b) $\Sigma_{11,(j_1,j_2)} = \Sigma_{11,j_1}$, $\mu_{1,(j_1,j_2)} = \mu_{1,j_1}$, and $A^{(k)}_{1r,(j_1,j_2)} = A^{(k)}_{1r,j_1}$ for all $j_1 \in \{1,\dots,M_1\}$, $j_2 \in \{1,\dots,M_2\}$, $r \in \{1,2\}$, and $k \in \{1,\dots,p\}$; and (c) $\Sigma_{12,j} = 0$ for all $j \in \{1,\dots,M\}$. Under (i) to (iii) we find that $\eta_t = (\eta^{(1)}_t \otimes \eta^{(2)}_t)$ for all t, with $\eta^{(l)}_t$ being the marginal density of $v_{l,t}$ conditional on $s_{l,t}$ and $\mathbf{y}_{t-1}$. If these linear restrictions are not satisfied, then η_t cannot be decomposed into the (Kronecker) product between a M_1 and a M_2 vector density. If $M_2 = 1$, then condition (c) can, for now, be dispensed with.

To satisfy the remaining condition in Lemma 2 we need to let $s_{1,t}$ and $s_{2,t}$ be independent. For $M_2 \geq 2$ we have that $\eta^{(1)}_t$ and $\eta^{(2)}_t$ are vectors of densities for independent random variables ($\epsilon_{1,t}|s_{1,t}$ and $\epsilon_{2,t}|s_{2,t}$) from, in particular, restrictions (c), and for $M_2 = 1$ this is not needed since φ_t is just a scalar which cancels in (A.1). By Lemma 2 it then follows that

$$\Pr\left[s_t = j|\mathbf{y}_t;\theta\right] = \Pr\left[s_{1,t} = j_1\middle|\mathbf{v}_{1,t},\mathbf{v}_{2,t};\theta_1,\mathbf{P}^{(1)}\right]\Pr\left[s_{2,t} = j_2\middle|\mathbf{v}_{1,t-1},\mathbf{v}_{2,t};\theta_2,\mathbf{P}^{(2)}\right].$$

When $M_2 \geq 2$ it also follows that $\Pr[s_{1,t} = j_1|\mathbf{v}_{1,t},\mathbf{v}_{2,t};\theta_1] = \Pr[s_{1,t} = j_1|\mathbf{v}_{1,t},\mathbf{v}_{2,t-1};\theta_1]$.

The final stage is now straightforward. Since $\mathbf{v}_{2,t}$ is assumed to be non-informative about $s_{1,t}$, the restrictions (c) must also hold for $M_2 = 1$. Furthermore, we may also infer that: (d) $A^{(k)}_{12,j_1} = 0$ for all $j_1 \in \{1,\dots,M_1\}$ and $k \in \{1,\dots,p\}$ and for $M_2 \geq 1$. Hence, we have shown that

$$\Pr\left[(s_{1,t},s_{2,t}) = (j_1,j_2)\middle|\mathbf{y}_t;\theta\right] = \Pr\left[s_{1,t} = j_1\middle|\mathbf{v}_{1,t};\theta_1\right]\Pr\left[s_{2,t} = j_2\middle|\mathbf{y}_t;\theta_2\right],$$

implies that (A1) is satisfied. To prove the reverse is straightforward. Q.E.D.

Proof of Proposition 2

Given that u_{t+1} is mean zero stationary we know that $E[u^2_{t+1};\theta] \leq E[\tilde{u}^2_{t+1};\theta]$ since $(\mathbf{v}_{1t},\mathbf{y}_{3t}) \subset \mathbf{y}_t$ for all t. In particular,

$$E\left[\tilde{u}^2_{t+1};\theta\right] = E\left[u^2_{t+1};\theta\right] + E\left[\left(E\left[y_{1,t+1}\middle|\mathbf{y}_t;\theta\right] - E\left[y_{1,t+1}\middle|\mathbf{v}_{1t},\mathbf{y}_{3t};\theta\right]\right)^2;\theta\right]. \tag{A.27}$$

Accordingly, the variances of u_{t+1} and $\tilde{u}_{t+1}$ are equal if and only if $E[y_{1,t+1}|\mathbf{y}_t;\theta] = E[y_{1,t+1}|\mathbf{v}_{1t},\mathbf{y}_{3t};\theta]$ for all t.

The prediction of $y_{1,t+1}$ conditional on $\mathbf{y}_t$ is given by

$$\begin{aligned}E\left[y_{1,t+1}|\mathbf{y}_t;\theta\right] = \bar{m}_{1,t} + \sum_{k=1}^{p}\Big(&\bar{a}^{(k)}_{11,t}y_{1,t+1-k} + \bar{a}^{(k)}_{12,t}y_{2,t+1-k} \\ &+ \bar{a}^{(k)}_{13,t}y_{3,t+1-k} + \bar{a}^{(k)}_{14,t}y_{4,t+1-k}\Big).\end{aligned} \tag{A.28}$$

The necessary and sufficient conditions for this expression to be invariant with respect to $\mathbf{y}_{3t}$ are, for all t, given by

$$\begin{aligned}&\text{(i)} \quad \bar{m}_{1,t} = E\left[m_{1,s_{t+1}}\middle|\mathbf{v}_{1t},\mathbf{y}_{3t};\theta\right], \\ &\text{(ii)} \quad \bar{a}^{(k)}_{1r,t} = E\left[a^{(k)}_{1r,s_{t+1}}\middle|\mathbf{v}_{1t},\mathbf{y}_{3t};\theta\right], \qquad r \in \{1,\dots,4\} \text{ and } k \in \{1,\dots,p\}, \\ &\text{(iii)} \quad \bar{a}^{(k)}_{13,t} = 0, \qquad\qquad\qquad\qquad\quad k \in \{1,\dots,p\}.\end{aligned}$$

To prove the claim in Proposition 2 we therefore have to show that (i)–(iii) are equivalent to [(A1) or (A3)].

$$\text{Noncausality in mean} \Rightarrow \left[\text{(A1) or (A3)}\right]$$

From the definitions of $\bar{m}_{1,t}$ and $\bar{a}^{(k)}_{1r,t}$ in both of the equations (28) we find that these random matrices can be expressed as

$$\bar{m}_{1,t} = \sum_{i=1}^{M}\sum_{j=1}^{M} m_{1,j}p_{ij}\Pr\left[s_t = i\middle|\mathbf{y}_t;\theta\right], \tag{A.29}$$

and

$$\bar{a}^{(k)}_{1r,t} = \sum_{i=1}^{M}\sum_{j=1}^{M} a^{(k)}_{1r,j}p_{ij}\Pr\left[s_t = i\middle|\mathbf{y}_t;\theta\right]. \tag{A.30}$$

From these two equations it can be seen that $\bar{m}_{1,t}$ and $\bar{a}^{(k)}_{1r,t}$ depend on t, and thus potentially on $\mathbf{y}_{4t}$, only via the filter probabilities $\Pr[s_t = i|\mathbf{y}_t;\theta]$.

Suppose first that $(\bar{m}_{1,t}, \bar{a}^{(k)}_{1r,t})$ indeed varies with t. It now follows that noncausality in mean implies that

$$\Pr\left[(s_{1,t}, s_{2,t}) = (i_1, i_2)\middle|\mathbf{y}_t;\theta\right] = \Pr\left[s_{1,t} = i_1\middle|\mathbf{v}_{1,t};\theta\right]\Pr\left[s_{2,t} = i_2\middle|\mathbf{y}_t;\theta\right], \tag{A.31}$$

must hold for all i_1, i_2, and t, while $(m_{1,(j_1,j_2)}, a^{(k)}_{1r,(j_1j_2)})$ only depends on j_2. By Corollary 2 we know that equation (A.31) can only be satisfied under (A1). The remaining parameter restrictions, $p_{ij} = p^{(1)}_{i_1j_1}p^{(2)}_{i_2j_2}$, are also satisfied under (A1).

Notice that the formulation in (A.31) covers the case when $n_2 = 1$, i.e. when $\mathbf{y}_{3t}$ is empty and all auxiliary variables are located in $\mathbf{y}_{2t}$, as well as the cases when $n_2 \geq 2$. It is therefore more general than one where $\Pr[s_{1,t} = i_1|\mathbf{v}_{1,t};\theta]$ is replaced with $\Pr[s_{1,t} = i_1|\mathbf{v}_{1t}, \mathbf{y}_{3t};\theta]$.

It remains to examine the case when $(\bar{m}_{1,t}, \bar{a}^{(k)}_{1r,t})$ is invariant with respect to t. From equations (A.29)–(A.30) we now have that $\sum_{j=1}^{M} m_{1,j}p_{ij} = \bar{m}_1$, $\sum_{j=1}^{M} a^{(k)}_{1r,j}p_{ij} = \bar{a}^{(k)}_{1r}$, with $\bar{a}^{(k)}_{14} = 0$ for all i, r, and k. Hence, condition (B2) is satisfied.

$$\Big[\text{(A1) or (A3)}\Big] \Rightarrow \textsc{Noncausality in mean}$$

Evaluating equation (A.28) under (A1) and (A3), respectively, gives the result. Q.E.D.

Appendix B. Summary of the posterior densities simulations

Table B.7: VAR(12): posterior properties

	Mean	Std. dev.	Naive Std. error	Autocorr. lag 1	Autocorr. lag 10
Standard deviations					
σ_1	9.192	0.137	0.002	0.028	0.006
σ_2	4.912	0.095	0.001	0.046	0.002
Correlations					
ρ_1	-0.025	0.058	0.001	0.060	-0.014
Intercepts					
μ_1	-0.004	0.300	0.004	0.001	-0.009
μ_2	0.582	0.266	0.004	-0.011	0.006
Autoregressive coefficients					
$A_{11}^{(1)}$	0.284	0.049	0.001	-0.007	0.005
$A_{12}^{(1)}$	0.138	0.088	0.001	-0.006	-0.028
$A_{21}^{(1)}$	0.027	0.027	0.000	-0.024	-0.016
$A_{22}^{(1)}$	0.361	0.049	0.001	0.020	0.027
$A_{11}^{(2)}$	0.076	0.049	0.001	-0.009	0.014
$A_{12}^{(2)}$	0.108	0.094	0.001	-0.034	-0.014
$A_{21}^{(2)}$	-0.044	0.026	0.000	-0.001	0.012
$A_{22}^{(2)}$	-0.005	0.052	0.001	0.007	-0.001
$A_{11}^{(3)}$	0.068	0.049	0.001	0.002	0.011
$A_{12}^{(3)}$	0.133	0.093	0.001	-0.035	0.009
$A_{21}^{(3)}$	-0.054	0.026	0.000	-0.014	-0.009
$A_{22}^{(3)}$	0.199	0.052	0.001	0.001	-0.001
$A_{11}^{(4)}$	0.085	0.049	0.001	0.004	0.009
$A_{12}^{(4)}$	-0.053	0.092	0.001	-0.014	-0.008
$A_{21}^{(4)}$	-0.024	0.027	0.000	0.012	-0.011
$A_{22}^{(4)}$	-0.106	0.051	0.001	-0.026	0.002
$A_{11}^{(5)}$	-0.054	0.049	0.001	-0.003	-0.010
$A_{12}^{(5)}$	0.032	0.094	0.001	-0.019	-0.010
$A_{21}^{(5)}$	0.007	0.026	0.000	0.008	-0.005
$A_{22}^{(5)}$	0.228	0.051	0.001	0.004	0.008
$A_{11}^{(6)}$	0.004	0.047	0.001	0.000	0.009
$A_{12}^{(6)}$	0.106	0.095	0.001	0.009	0.019
$A_{21}^{(6)}$	0.000	0.026	0.000	0.004	0.011
$A_{22}^{(6)}$	0.067	0.052	0.001	0.008	-0.010
$A_{11}^{(7)}$	0.035	0.048	0.001	-0.002	-0.007
$A_{12}^{(7)}$	-0.100	0.095	0.001	-0.008	0.003
$A_{21}^{(7)}$	0.001	0.025	0.000	0.017	-0.002
$A_{22}^{(7)}$	-0.012	0.053	0.001	-0.025	-0.008
$A_{11}^{(8)}$	0.031	0.048	0.001	0.035	-0.017
$A_{12}^{(8)}$	0.056	0.094	0.001	0.005	-0.005
$A_{21}^{(8)}$	0.052	0.025	0.000	-0.015	0.005
$A_{22}^{(8)}$	0.104	0.051	0.001	0.011	0.010
$A_{11}^{(9)}$	0.015	0.048	0.001	-0.016	0.019
$A_{12}^{(9)}$	-0.054	0.093	0.001	0.006	0.004
$A_{21}^{(9)}$	-0.043	0.025	0.000	0.016	-0.004

	Mean	Std. dev.	Naive Std. error	Autocorr. lag 1	Autocorr. lag 10
$A_{22}^{(9)}$	0.181	0.052	0.001	0.023	-0.012
$A_{11}^{(10)}$	0.020	0.047	0.001	0.023	0.020
$A_{12}^{(10)}$	0.008	0.090	0.001	0.007	-0.022
$A_{21}^{(10)}$	-0.008	0.026	0.000	-0.010	-0.005
$A_{22}^{(10)}$	-0.077	0.052	0.001	0.018	-0.012
$A_{11}^{(11)}$	0.008	0.048	0.001	-0.017	0.021
$A_{12}^{(11)}$	-0.064	0.093	0.001	-0.014	0.001
$A_{21}^{(11)}$	-0.036	0.026	0.000	0.007	-0.006
$A_{22}^{(11)}$	-0.023	0.052	0.001	-0.022	0.001
$A_{11}^{(12)}$	-0.069	0.044	0.001	0.008	0.003
$A_{12}^{(12)}$	-0.042	0.087	0.001	-0.031	0.006
$A_{21}^{(12)}$	0.061	0.024	0.000	0.010	-0.013
$A_{22}^{(12)}$	-0.029	0.049	0.001	-0.004	-0.002

Table B.8: MSIAH(2)-VAR(4): posterior properties

	Mean	Std. dev.	Naive Std. error	Autocorr. lag 1	Autocorr. lag 10
Transition probabilities					
$p_{1,1}$	0.734	0.066	0.001	0.557	-0.005
$p_{2,1}$	0.059	0.018	0.000	0.624	0.088
Standard deviations					
$\sigma_{1,1}$	17.129	1.207	0.017	0.625	0.150
$\sigma_{2,1}$	8.746	0.646	0.009	0.559	0.111
$\sigma_{1,2}$	6.983	0.276	0.004	0.669	0.173
$\sigma_{2,2}$	4.011	0.179	0.003	0.666	0.105
Correlations					
$\rho_{1,1}$	-0.173	0.127	0.002	0.203	0.008
$\rho_{1,2}$	0.078	0.070	0.001	0.284	0.018
Intercepts regime 1					
$\mu_{1,1}$	-0.213	0.949	0.013	0.014	0.032
$\mu_{2,1}$	1.107	0.885	0.013	0.101	0.011
Autoregressive coefficients regime 1					
$A^{(1)}_{11,1}$	0.497	0.147	0.002	0.128	0.016
$A^{(1)}_{12,1}$	0.209	0.287	0.004	0.142	-0.018
$A^{(1)}_{21,1}$	0.069	0.075	0.001	0.156	0.027
$A^{(1)}_{22,1}$	0.419	0.156	0.002	0.222	-0.002
$A^{(2)}_{11,1}$	-0.253	0.191	0.003	0.238	0.020
$A^{(2)}_{12,1}$	-0.134	0.361	0.005	0.191	-0.005
$A^{(2)}_{21,1}$	-0.018	0.094	0.001	0.131	0.025
$A^{(2)}_{22,21}$	-0.092	0.202	0.003	0.237	0.002
$A^{(3)}_{11,1}$	0.172	0.218	0.003	0.173	0.001
$A^{(3)}_{12,1}$	-0.176	0.376	0.005	0.105	0.008
$A^{(3)}_{21,1}$	-0.126	0.122	0.002	0.265	0.006
$A^{(3)}_{22,1}$	0.112	0.217	0.003	0.191	0.004
$A^{(4)}_{11,1}$	-0.490	0.217	0.003	0.325	0.078
$A^{(4)}_{12,1}$	0.409	0.343	0.005	0.164	0.019
$A^{(4)}_{21,1}$	0.088	0.106	0.001	0.252	0.029
$A^{(4)}_{22,1}$	0.098	0.205	0.003	0.281	0.031
Intercepts regime 2					
$\mu_{1,2}$	0.295	0.634	0.009	0.163	-0.005
$\mu_{2,2}$	2.058	0.420	0.006	0.210	-0.012
Autoregressive coefficients regime 2					
$A^{(1)}_{11,2}$	0.237	0.059	0.001	0.391	0.041
$A^{(1)}_{12,2}$	0.028	0.099	0.001	0.333	-0.002
$A^{(1)}_{21,2}$	-0.026	0.031	0.000	0.259	0.025
$A^{(1)}_{22,2}$	0.398	0.058	0.001	0.297	-0.024
$A^{(2)}_{11,2}$	0.130	0.048	0.001	0.210	0.014
$A^{(2)}_{12,2}$	0.165	0.088	0.001	0.195	0.013
$A^{(2)}_{21,2}$	-0.032	0.028	0.000	0.194	0.005
$A^{(2)}_{22,2}$	0.092	0.057	0.001	0.321	0.038

	Mean	Std. dev.	Naive Std. error	Autocorr. lag 1	Autocorr. lag 10
$A^{(3)}_{11,2}$	0.099	0.053	0.001	0.377	0.057
$A^{(3)}_{12,2}$	0.214	0.086	0.001	0.195	0.006
$A^{(3)}_{21,2}$	-0.014	0.026	0.000	0.176	0.023
$A^{(3)}_{22,2}$	0.285	0.053	0.001	0.284	0.007
$A^{(4)}_{11,2}$	0.106	0.052	0.001	0.394	0.039
$A^{(4)}_{12,2}$	-0.174	0.092	0.001	0.272	0.014
$A^{(4)}_{21,2}$	-0.019	0.025	0.000	0.200	0.009
$A^{(4)}_{22,2}$	-0.066	0.055	0.001	0.323	0.031

Appendix C. Characterization of estimation efficiency

Table C.9: Characterization of the efficiency in the models' estimations

	RNE			*Autocorr. lag 1*			*Autocorr. lag 10*			*Geweke z-score*		
$\mathcal{M}_j$	Median	Min	Max	Median	Min	Max	Median	Min	Max	Median	Min	Max
Vector autoregressive models												
$\mathcal{M}_0$	1.00	0.85	1.19	0.00	-0.03	0.06	0.00	-0.03	0.03	-0.10	-2.37	2.38
$\mathcal{M}_1$	1.00	0.76	1.08	0.01	-0.03	0.07	0.00	-0.04	0.02	0.07	-2.57	2.43
Markov switching vector autoregressive models												
$\mathcal{M}_0$	0.48	0.10	1.00	0.24	0.01	0.67	0.02	-0.02	0.17	-0.56	-2.14	3.27
$\mathcal{M}_1$	0.47	0.06	1.00	0.17	-0.02	0.78	0.01	-0.03	0.29	0.22	-1.98	2.58
$\mathcal{M}_2$	0.71	0.13	1.12	0.14	-0.02	0.71	0.01	-0.03	0.08	0.13	-2.10	1.59
$\mathcal{M}_3$	0.30	0.02	0.94	0.27	0.03	0.89	0.04	-0.01	0.56	-0.32	-2.43	1.94
$\mathcal{M}_4$	0.46	0.08	0.83	0.25	0.07	0.78	0.01	-0.03	0.23	-0.20	-1.57	1.56
$\mathcal{M}_5$	0.22	0.02	0.43	0.44	0.12	0.85	0.07	-0.01	0.50	-0.10	-2.39	2.16

Table C.9 reports statistics for assessing the efficiency of each estimated model. Three types of statistics are presented: the relative numerical efficiency of Geweke (1989), autocorrelations at different lags, and the convergence diagnostic of Geweke (1992). Statistics should be presented separately for each parameter of each model, but to save space, we summarize each model with a median, minimum, and maximum.

The relative numerical efficiency represents the ratio of the variance of a hypothetical draw from the posterior density over the variance of the Gibbs sampler. Thus, it can be interpreted as a measure of the computational efficiency of the algorithm. The columns of Table C.9, unsurprisingly, tell us that the algorithm for VAR models is more efficient than that for MS-VAR. The same observation can be made when comparing unrestricted models with restricted ones. What is interesting for us is the magnitude of the RNE statistics between unrestricted and restricted models. Those are comparable, which is a good sign that the algorithm for constrained models are, computationally, reasonable efficient.

The columns displaying the autocorrelations at lag 1 and lag 10 are here to ensure that there is a decay over time. This is the case here, and the Gibbs samplers explore the entire posterior distribution.

Geweke (1992) introduces the z scores test which tests the stationarity of the draws from the posterior distribution simulation comparing the mean of the first 30% of the draws with the last 40% of the draws. We compare the two means with a z-test. Typically, values outside $(-2, 2)$ indicate that the mean of the series is still drifting, and this occurs for some parameters in each models, except $\mathcal{M}_4$ and $\mathcal{M}_6$ for MS-VARs. Increasing the burn in period might improve the scores and stationarity of the MCMC chain.

References

Adolfson, M., Lindé, J., & Villani, M. (2007). Forecasting Performance of an Open Economy DSGE Model. *Econometric Reviews*, *26*, 289–328.

Albert, J. H., & Chib, S. (1993). Bayes Inference via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts. *Journal of Business & Economic Statistics*, *11*, 1–15.

Barnard, J., McCulloch, R., & Meng, X.-l. (2000). Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, with Applicationto Shrinkage. *Statistica Sinica*, *10*, 1281–1311.

Blix, M. (1997). Rational Expectations in a VAR with Markov Switching. In *Rational Expectations and Regime Shifts in Macroeconometrics*. PhD Thesis, Monograph Series No. 31, Institute for International Economic Studies, Stockholm University, Sweden.

Boudjellaba, H., Dufour, J.-M., & Roy, R. (1992). Testing Causality Between Two Vectors in Multivariate Autoregressive Moving Average Models. *Journal of the American Statistical Association*, *87*, 1082– 1090.

Boudjellaba, H., Dufour, J.-M., & Roy, R. (1994). Simplified conditions for noncausality between vectors in multivariate ARMA models. *Journal of Econometrics*, *63*, 271–287.

Casella, G., & George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, *46*, 167–174.

Chib, S., & Jeliazkov, I. (2001). Marginal Likelihood from the Metropolis-Hastings Output. *Journal of the American Statistical Association*, *96*, 270– 281.

Christiano, L. J., & Ljungqvist, L. (1988). Money does granger-cause output in the bivariate money-output relation. *Journal of Monetary Economics*, *22*, 217 – 235.

Christopoulos, D. K., & León-Ledesma, M. A. (2008). Testing for Granger (Non-) causality in a Time-Varying Coeffi cient. *Journal of Forecasting*, (pp. 293–303).

Comte, F., & Lieberman, O. (2000). Second-Order Noncausality in Multivariate GARCH Processes. *Journal of Time Series Analysis*, *21*, 535–557.

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

Doan, T., Litterman, R. B., & Sims, C. A. (1983). Forecasting and Conditional Projection Using Realistic Prior Distributions. *NBER Working Paper*, *1202*, 1–71.

Droumaguet, M., & Woźniak, T. (2012). *Bayesian Testing of Granger Causality in Markov-Switching VARs*. Working Paper Series 2012/06 European University Institute Florence, Italy. Download: http://cadmus.eui.eu/bitstream/handle/1814/20815/ECO_2012_06.pdf?sequence=1.

Dufour, J.-M. (1989). Nonlinear hypotheses, inequality restrictions, and non-nested hypotheses: Exact simultaneous tests in linear regressions. *Econometrica*, *54*, 335–355.

Dufour, J.-M., Pelletier, D., & Renault, E. (2006). Short run and long run causality in time series: inference. *Journal of Econometrics*, *132*, 337–362.

Friedman, M., & Schwartz, A. (1971). *A monetary history of the United States, 1867-1960* volume 12. Princeton Univ Pr.

Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *Econometrics Journal*, *7*, 143–167.

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

Geweke, J. (1989). Bayesian inference in econometric models using monte carlo integration. *Econometrica*, *57*, pp. 1317–1339.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. Bernardo, J. Berger, A. Dawid, & A. Smith (Eds.), *Bayesian Statistics 4* (pp. 169–194). Clarendon Press, Oxford, UK.

Geweke, J. (1995). Bayesian Comparison of Econometric Models.

Geweke, J. (1999). Using simulation methods for bayesian econometric models: inference, development,and communication. *Econometric Reviews*, *18*, 1–73.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, *37*, 424–438.

Greenberg, E., & Chib, S. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, *49*, 327–335.

Hamilton, J. D. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, *57*, 357–384.

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

Hoogerheide, L. F., van Dijk, H. K., & van Oest, R. (2009). Simulation Based Bayesian Econometric Inference: Principles and Some Recent Computational Advances. Handbook of Computational Econometrics chapter 7. (pp. 215–280). Wiley.

Jarociński, M., & Maćkowak, B. (2011). Choice of Variables in Vector Autoregressions.

Jeantheau, T. (1998). Strong consistency of estimators for multivariate arch models. *Econometric Theory*, *14*, 70–86.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kim, C. J. (1994). Dynamic Linear Models with Markov-Switching. *Journal of Econometrics*, *60*, 1–22.

Kim, C.-J., & Nelson, C. R. (1999a). Has the U.S. Economy Become More Stable? A Bayesian Approach Based on a Markov-Switching Model of the Business Cycle. *Review of Economics and Statistics*, *81*, 608–616.

Kim, C.-J., & Nelson, C. R. (1999b). *State-space models with regime switching: classical and Gibbs-sampling approaches with applications*. MIT press.

Koop, G., & Korobilis, D. (2010). Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends in Econometrics*, *3*, 267–358.

Krolzig, H. (1997). *Markov-switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis*. Springer Verlag.

Lechner, M. (2011). The Relation of Different Concepts of Causality Used in Time Series and Microeconometrics. *Econometric Reviews*, *30*, 109–127.

Lindgren, G. (1978). Markov Regime Models for Mixed Distributions and Switching Regressions. *Scandinavian Journal of Statistics*, *5*, 81–91.
Litterman, R. B. (1986). Forecasting with Bayesian Vector Autoregressions: Five Years of Experience. *Journal of Business & Economic Statistics*, *4*, 25–38.
Lütkepohl, H. (1993). *Introduction to Multiple Time Series Analysis*. Springer-Verlag.
Lütkepohl, H., & Burda, M. M. (1997). Modified Wald tests under nonregular conditions. *Journal of Econometrics*, *78*, 315–332.
Magnus, J., & Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Chichester: John Wiley.
McCulloch, R. E., & Tsay, R. S. (1994). Statistical analysis of economic time series via markov switching models. *Journal of Time Series Analysis*, *15*, 523–539.
Pajor, A. (2011). A Bayesian Analysis of Exogeneity in Models with Latent Variables. *Central European Journal of Economic Modelling and Econometrics*, *3*, 49–73.
Psaradakis, Z., Ravn, M. O., & Sola, M. (2005). Markov switching causality and the money-output relationship. *Journal of Applied Econometrics*, *20*, 665–683.
Psaradakis, Z., & Spagnolo, N. (2003). On the determination ot the number of regimes in markov-switching autoregressive models. *Journal of Time Series Analysis*, *24*, 237–252.
Ritter, C., & Tanner, M. A. (1992). Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler. *Journal of the American Statistical Association*, *87*, 861– 868.
Robertson, J. C., & Tallman, E. W. (1999). Vector Autoregressions: Forecasting and Reality. *Federal Reserve Bank of Atlanta Economic Review*, (pp. 4–18).
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688 – 701.
Sims, C. A. (1972). Money, Income, and Causality. *The American Economic Review*, *62*, 540 – 552.
Sims, C. A., Waggoner, D. F., & Zha, T. (2008). Methods for inference in large multiple-equation markov-switching models. *Journal of Econometrics*, *146*, 255–274.
Taylor, J. B., & Williams, J. C. (2009). A Black Swan in the Money Market. *American Economic Journal: Macroeconomics*, *1*, 58–83.
Tobin, J. (1970). Money and Income: Post Hoc Ergo Propter Hoc? *The Quarterly Journal of Economics*, *84*, 301–317. ArticleType: research-article / Full publication date: May, 1970 / Copyright © 1970 Oxford University Press.
Warne, A. (2000). *Causality and Regime Inference in a Markov Switching VAR*. Working Paper Series 118 Sveriges Riksbank (Central Bank of Sweden).
White, H., & Lu, X. (2010). Granger Causality and Dynamic Structural Systems. *Journal of Financial Econometrics*, *8*, 193–243.
Woźniak, T. (2011). *Granger causal analysis of VARMA-GARCH models*. Unpublished manuscript, Department of Economics, European University Institute.
Woźniak, T. (2012). *Testing Causality Between Two Vectors in Multivariate GARCH Models*. Working Paper Series 1139 Department of Economics, The University of Melbourne Melbourne. Download: http://www.economics.unimelb.edu.au/research/1139.pdf.