Granger Causality and Regime Inference in Bayesian Markov-Switching VARs

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

[a]*Department of Economics, European University Institute*
[b]*Directorate General Research, European Central Bank*
[c]*Department of Economics, University of Melbourne*

Abstract

We derive restrictions for Granger noncausality in Markov-switching vector autoregressive models and also show under which conditions a variable does not affect the forecast of the hidden Markov process. Based on Bayesian approach to evaluating the hypotheses, the computational tools for posterior inference include a novel block Metropolis-Hastings sampling algorithm for the estimation of the restricted models. We analyze a system of monthly US data on money and income. The results of testing in MS-VARs contradict those in linear VARs: the money aggregate M1 is useful for forecasting income and for predicting the next period's state.

Keywords: Granger causality, hidden Markov process, Markov-switching models, mixture 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 is based on the idea that variable x which causes another variable y should precede it. This idea has been formalized such that x is said not to Granger-cause y if past and current information about x does not improve the forecast of y in a mean square error sense; see also Sims (1972). Knowledge of Granger causal relations may allow a researcher to formulate an appropriate model and obtain a better forecast of variables of interest. Note that this concept refers to the forecasting of variables, in contrast to, e.g., the causality concept attributed to Rubin (1974), based on *ceteris paribus* effects (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).

Granger-causality has primarily been studied empirically in vector autoregressions (VARs) with a focus on one-step-ahead forecasts; see, e.g., Lütkepohl (1993). In such a setting, x

*Corresponding author. *Address:* Department of Economics, University of Melbourne, 111 Barry Street, FBE Building, Level 4, 3053 Carlton, Victoria, Australia, *Phone:* +61 3 8344 5310, *Fax:* +61 3 8344 6899, *Email address:* tomasz.wozniak@unimelb.edu.au.
URL: `http://www.texlips.net/awarne/` (Anders Warne), `http://bit.ly/tomaszw` (Tomasz Woźniak)

does not Granger-cause y if the coefficients on lags of x in the equation for y are jointly zero. Among other 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, 2012b; Woźniak, 2012a), as well as dynamic discrete-time bivariate probit model (Mosconi & Seri, 2006). Note that all these works analyze *one-step-ahead* Granger noncausality (see Lütkepohl, 1993; Lütkepohl & Burda, 1997; Dufour, Pelletier & Renault, 2006, for analyses based on h-step-ahead forecasts in VAR models).

Psaradakis, Ravn & Sola (2005) use a Markov-switching (MS) VAR model to analyze *temporary Granger causality* within the money-income system, i.e., causality which holds in some periods but not in others. Technically, this means that they condition the causality analysis on realizations of the hidden Markov process and therefore focus only on linear relations between variables. That is, x does not Granger cause y temporarily if the coefficients on lags of x in the equation for y are zero in some of the states. Since *all* parameters of an MS-VAR model may, in principle, vary with the hidden Markov process, their analysis neglects the possibility that x may be useful for predicting the states that affect the coefficients in the y equation.

The approach to Granger causality that we consider in this paper takes into account the two sources of predictive relations between the variables of interest: first, the linear relations in the VAR model conditional on the states, and second, the fact that all of the variables are used to forecast the future probabilities of the states. The analysis of Granger causality is consequently not conditioned on the realizations of the hidden Markov process, but only on observed variables. Both of these properties make it difficult to conduct classical inference, where multiple sets of restrictions complicates the determination of the overall test level and nonlinearities may affect the asymptotic properties of test statistics.

As a second contribution, we suggest an approach for performing Bayesian inference that allows us to test all of the restrictions of Granger noncausality jointly. The proposed framework consists of Bayesian estimation of the unrestricted model, allowing for Granger causality, and of the restricted models representing hypotheses of noncausality. For this purpose, we construct a novel block Metropolis-Hastings sampling algorithm that allows for the estimation of the restricted models. The hypotheses of Granger causality and noncausality can thereafter be evaluated 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 jointly. Another feature of the posterior odds analysis is that all the hypotheses are treated symmetrically. As a consequence, this method gives arguments *in favor of* a hypothesis. Finally, since a mixture model is a special case of a Markov switching model, our analytical results apply also to such models.

The remainder of the paper is organized as follows. In Section 2 we present the model and the definitions for Granger noncausality and regime independence, while Section 3 provides the restrictions for the considered relations between variables. Section 4 first discusses the use of classical inference when testing for Granger noncausality in MS-VARs, before it considers the pros and cons of instead using Bayesian inference. The block Metropolis-Hastings algorithm, required for estimating the models consistent with Granger noncausality, is described in Section 5. Section 6 gives an empirical illustration of the methodology, using the bivariate money-income system for

monthly US data. Section 7 concludes. All the proofs and technical details are presented in the mathematical and statistical appendices.

2. Theoretical Framework for Granger Causality and Regime Inference

2.1. A Markov Switching VAR Model

Let $\mathbf{y}_T = (y_1, \ldots, y_T)'$ denote a time series of T observations, where each y_t is a N-variate real-valued vector for $t \in \{1, \ldots, T\}$. We consider a class of parametric Markov mixture distribution models in which the stochastic process Y_t depends on the realizations 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 time series up to time $t-1$, $\mathbf{y}_{t-1}$, y_t follows an independent identical normal distribution. The conditional mean process is a VAR model in which an intercept, μ_{s_t}, as well as lag polynomial matrices, $A^{(i)}_{s_t}$, for $i = 1, \ldots, p$, and covariance matrices, Σ_{s_t}, depend on the state $s_t = 1, \ldots, M$:

$$y_t = \mu_{s_t} + A^{(1)}_{s_t} y_{t-1} + \cdots + A^{(p)}_{s_t} y_{t-p} + \epsilon_t, \tag{1}$$

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

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.

The variable s_t is assumed to be an irreducible aperiodic Markov chain with $\Pr(S_0 = i|\mathbf{P}) = \pi_i$, where $\pi = (\pi_1, \ldots, \pi_M)$ is the ergodic distribution of the Markov Switching (MS) process. Its properties are sufficiently described by the $(M \times M)$ transition probabilities matrix:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1M} \\ p_{21} & p_{22} & \cdots & p_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M1} & p_{M2} & \cdots & p_{MM} \end{bmatrix}, \tag{3}$$

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, $\sum_{j=1}^{M} p_{ij} = 1$. Equations (1)–(3) represent a MS-VAR model with M states and p lags.

To establish the notation, 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}: $\mu_{s_t}, A^{(i)}_{s_t}, \Sigma_{s_t}$, for $s_t = 1, \ldots, M$ and $i = 1, \ldots, p$.

2.2. Some Useful Definitions for Analyses of Granger Causality and Regime Inference

Write $y_t = (y'_{1t}, y'_{2t}, y'_{3t}, y'_{4t})'$ for $t = 1, \ldots, T$, where y_{it} is a $N_i \times 1$ vector with $N_1, N_4 = 1, N_2, N_3 \geq 0$ and $\sum_{i=1}^{4} N_i = N$. The variables of interest are given by y_1 and y_4, between which we want to study causal relations[1]. 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:

[1]The proposed analysis holds for $N_1, N_4 \geq 1$ with slight adjustments of the notation.

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

$$u_{t+1} = y_{1.t+1} - E[y_{1.t+1}|\mathbf{y}_t; \theta]. \tag{4}$$

By construction, u_{t+1} has conditional mean zero and positive-definite conditional covariance matrix. And let $\tilde{u}_{t+1} = y_{1.t+1} - E[y_{1.t+1}|\mathbf{v}_{1t}, \mathbf{y}_{3t}; \theta]$ be the one-step-ahead forecast error for $y_{1.t+1}$, 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 y_4 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:

$$E\left[u^2_{t+1}; \theta\right] = E\left[\tilde{u}^2_{t+1}; \theta\right] < \infty \quad \forall t = 1, \ldots, T. \tag{5}$$

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. For this reason, Granger causality under a Bayesian approach concerns the validity of (5) 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:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \\ y_{4t} \end{bmatrix} = \begin{bmatrix} m_{1.s_t} \\ m_{2.s_t} \\ m_{3.s_t} \\ m_{4.s_t} \end{bmatrix} + \sum_{k=1}^{p} \begin{bmatrix} a^{(k)}_{11.s_t} & a^{(k)}_{12.s_t} & a^{(k)}_{13.s_t} & a^{(k)}_{14.s_t} \\ a^{(k)}_{21.s_t} & a^{(k)}_{22.s_t} & a^{(k)}_{23.s_t} & a^{(k)}_{24.s_t} \\ a^{(k)}_{31.s_t} & a^{(k)}_{32.s_t} & a^{(k)}_{33.s_t} & a^{(k)}_{34.s_t} \\ a^{(k)}_{41.s_t} & a^{(k)}_{42.s_t} & a^{(k)}_{43.s_t} & a^{(k)}_{44.s_t} \end{bmatrix} \begin{bmatrix} y_{1t-i} \\ y_{2t-i} \\ y_{3t-i} \\ y_{4t-i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \\ \varepsilon_{4t} \end{bmatrix}. \tag{6}$$

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

$$\Sigma_{s_t} = \text{Var}\left(\begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \\ \varepsilon_{4t} \end{bmatrix}\right) = \begin{bmatrix} \Omega_{11.s_t} & \Omega'_{21.s_t} & \Omega'_{31.s_t} & \Omega_{41.s_t} \\ \Omega_{21.s_t} & \Omega_{22.s_t} & \Omega'_{32.s_t} & \Omega'_{42.s_t} \\ \Omega_{31.s_t} & \Omega_{32.s_t} & \Omega_{33.s_t} & \Omega'_{43.s_t} \\ \Omega_{41.s_t} & \Omega_{42.s_t} & \Omega_{43.s_t} & \Omega_{44.s_t} \end{bmatrix}. \tag{7}$$

For expositional purposes, 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:

$$E\left[y_{1.t+1}|s_{t+1}, \mathbf{y}_t, \theta\right] = y_{1.t+1} - \epsilon_{1.t+1}. \tag{8}$$

Accordingly, the forecast error is given by $\epsilon_{1.t+1}$ and the conditional forecast error variance by $\Omega_{11.s_{t+1}}$. The necessary and sufficient condition for y_4 not to Granger-cause y_1 is that $a^{(k)}_{14.s_t}$ in equation (6) 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:

$$E\left[y_{1.t+1}|\mathbf{y}_t, \theta\right] = \sum_{s_{t+1}} E\left[y_{1.t+1}|s_{t+1}, \mathbf{y}_t, \theta\right] \Pr\left[s_{t+1}|\mathbf{y}_t, \theta\right]. \tag{9}$$

The role for y_4 is different in (9) relative to (8) 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:

$$\Pr\left[s_{t+1}|\mathbf{y}_t, \theta\right] = \sum_{s_t} \Pr\left[s_{t+1}|s_t, \theta\right] \Pr\left[s_t|\mathbf{y}_t, \theta\right]. \tag{10}$$

From this relationship we may conjecture 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 the hidden Markov process $s_{1.t+1}$ but not with the process $s_{2.t+1}$, where $s_{1.t+1}$ and $s_{2.t+1}$ form the joint process s_{t+1}. Similarly, there may be information in $\mathbf{y}_{4t}$ for predicting $s_{2.t+1}$, but not for predicting $s_{1.t+1}$. In such situations, it may still be the case that the prediction of y_1 in (9) 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 for $i = 1, 2$, such that $M = M_1 \cdot M_2$. Such a decomposition can always be performed without imposing any restrictions on the transitions matrix P. For example, we may let

$$s_t = s_{1t} + M_1 \left(s_{2t} - 1\right),$$

determines s_t uniquely from s_{1t} and s_{2t} without imposing any constraints on how these Markov processes evolve over time. In case M is a prime number it follows that M_1 or M_2 is always equal to unity. For non-prime integer values of M it is possible to consider sub-processes s_{it} such that M_1 and M_2 are both greater than unity. In fact, for the purpose of hypotheses testing one should consider all the possible combinations of M_1 and M_2 given M (see also Sections 4.2 and 6).

The definitions of predictive state independence and predictive redundancy are now useful for the discussion:

Definition 2. A system for y_{t+1} which depends on the hidden Markov process $s_{t+1} = (s_{1.t+1}, s_{2.t+1})$

is said to be *predictively state independent* when

$$\Pr\Big[(s_{1.t+1}, s_{2.t+1}) = (j_1, j_2)\Big|\mathbf{y}_t, \theta\Big] = \Pr\Big[s_{1.t+1} = j_1\Big|\mathbf{y}_t, \theta\Big] \cdot \Pr\Big[s_{2.t+1} = j_2\Big|\mathbf{y}_t, \theta\Big], \tag{11}$$

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

Definition 3. The vector $\mathbf{v}_{2t}$ is said to be *predictively redundant* for $s_{1.t+1}$ when

$$\Pr\Big[s_{1.t+1} = j_1\Big|\mathbf{y}_t, \theta\Big] = \Pr\Big[s_{1.t+1} = j_1\Big|\mathbf{v}_{1t}, \theta\Big]. \tag{12}$$

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

Predictive state independence therefore means that the regime predictions of $s_{1.t+1}$ and $s_{2.t+1}$ conditional of $\mathbf{y}_t$ are independent. Predictive redundancy, on the other hand, concerns the possibility that there is no unique information in $\mathbf{v}_{2t}$ for predicting $s_{1.t+1}$ beyond the information contained in $\mathbf{v}_{1t}$. These aspects can be seen from the decomposition of the joint probability into the product of the the conditional probability and the marginal probability. That is,

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

Predictive redundancy can here be interpreted as a property of the first term on the right hand side, i.e., of the marginal probability of $s_{1.t+1}$, while predictive state independence is a feature of the joint probability of $(s_{1.t+1}, s_{2.t+1})$ and can therefore be translated into the behavior of the conditional probability of $s_{2.t+1}$ in the second term on the right hand side. Predictive redundancy does not imply predictive state indepepence or vice versa, and the two concepts therefore concern different properties of a model subject to a hidden Markov process.

If a system for y_{t+1} satisfies the conditions for predictive state independence and $\mathbf{v}_{2t}$ being predictively redundant for $s_{1.t+1}$, it follows that beyond the information in $\mathbf{v}_{1t}$ there is no additional information in $\mathbf{v}_{2t}$ and $s_{2.t+1}$ that can affect the probability of $s_{1.t+1} = j_1$ for any j_1. This opens up for the possibility that v_2 (y_4) does not Granger cause v_1 (y_1) if the parameters of the v_1 sub-system only depend on s_1. A restricted version of the system in equation (1) is therefore given by:

$$\begin{bmatrix} v_{1t} \\ v_{2t} \end{bmatrix} = \begin{bmatrix} \mu_{1.s_{1t}} \\ \mu_{2.s_{2t}} \end{bmatrix} + \sum_{k=1}^{p} \begin{bmatrix} A^{(k)}_{11.s_{1t}} & A^{(k)}_{12.s_{1t}} \\ A^{(k)}_{21.s_{2t}} & A^{(k)}_{22.s_{2t}} \end{bmatrix} \begin{bmatrix} v_{1t-i} \\ v_{2t-i} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}. \tag{13}$$

where the following linear restrictions have been imposed:

$$\mu_{i.s_t} = \mu_{i.s_{it}}, \;\; A^{(k)}_{ij.s_t} = A^{(k)}_{ij.s_{it}}, \quad i, j = 1, 2, \text{ and } k = 1, \ldots, p. \tag{14}$$

If the ϵ_{it} residuals are independent of the regime, equation (14) 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} ($i \neq j$).

In a restricted version, we may also consider the possibility that the marginal distribution of the $\epsilon_{it}|s_t$ is subject to linear restrictions given by:

$$\Sigma_{ii.s_t} = \Sigma_{ii.s_{it}}, \quad i = 1, 2. \tag{15}$$

The restrictions in (15) are necessary but not sufficient for $p(\epsilon_t|s_t) = p(\epsilon_{1t}|s_{1t})p(\epsilon_{2t}|s_{2t})$, i.e., for $\epsilon_{1t}|s_{1t}$ and $\epsilon_{2t}|s_{2t}$ to be independent. The additional requirement is simply that $\Sigma_{12.s_t} = 0$ for all regimes such that the covariance matrix is block diagonal.

In the event that the restrictions in (14) and (15) are satisfied and the covariance matrix is block diagonal 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}. In the next section we shall first consider which restrictions are needed for the conditions in Definitions 2 and 3 to be satisfied by an MS-VAR system. Second, we shall examine the situations when y_4 does not Granger cause y_1 in this setup.

3. Regime Inference and Granger Causality Analysis

3.1. Regime Inference

The first result in this paper concerns the restrictions that the MS-VAR system needs 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 MS-VAR system for y_{t+1} in (1)–(3) with $s_{t+1} = (s_{1.t+1}, s_{2.t+1})$ is predictively state independent and $\mathbf{v}_{2t}$ is predictively redundant for $s_{1.t+1}$ if and only if either:*

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

(ii) *equations (14) and (15) are satisfied,*

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

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

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

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

is satisfied for all $\theta \in \Theta$, where $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$ are transition probabilities matrices associated with the Markov processes s_{1t} and s_{2t} respectively (and of dimensions $M_1 \times M_1$ and $M_2 \times M_2$ respectively).

First, note that conditions (A1) and (A2) imply linear restrictions on parameters of the model. Second, 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). Condition (A2) states that the first out of the two decomposed hidden Markov processes is serally uncorrelated and the marginal distribution of ϵ_{1t} is therefore a mixed normal. As a consequence, the system for y_{t+1} is predictively state independent also under (A2). Predictive redundancy of $\mathbf{v}_{2t}$ for $s_{1.t+1}$ is obtained for the (A2) case since $\Pr[s_{1.t+1} = j_1|\mathbf{y}_t, \theta] = \pi^{(1)}_{j_1}$, while condition (A1), excluding the block diagonality restrictions in (iii), is also sufficient for this property to hold.

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:

$$v_{2t|s_t=2} - v_{2t|s_t=1} = (\mu_{2,2} - \mu_{2,1}) + (A_{21,2} - A_{21,1})\, v_{1t-1} + (A_{22,2} - A_{22,1})\, v_{2t-1}. \tag{16}$$

The right hand side of (16) 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_{2t-1} in that equation is zero for both states, then $\mathbf{v}_{2t-1}$ play 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}.

3.2. Granger Noncausality Analysis

The restrictions for predictive state independence and predictive redundancy presented in the previous section are either sufficient for Granger noncausality (A1) or insufficient (A2), but the analysis has nevertheless established that there is an interesting connection between these concepts and Granger noncausality. Furthermore, the discussion reveals that Granger noncausality in the MS-VAR setting does generally not give rise to a single set of parameter restrictions that the system should satisfy. In this section we shall therefore focus on the necessary and sufficient conditions for this type of noncausal relations.

Additional notation is first required. Specifically, let:

$$\bar{m}_{1t} \equiv E\left[m_{1s_{t+1}}|\mathbf{y}_t, \theta\right], \tag{17a}$$

$$\bar{a}^{(k)}_{1r.t} \equiv E\left[a^{(k)}_{1r.s_{t+1}}|\mathbf{y}_t; \theta\right], \tag{17b}$$

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 $u_{t+1} = z_{t+1} + \epsilon_{1.t+1}$, where:

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

is conditionally on $\mathbf{y}_t$ uncorrelated with $\epsilon_{1.t+1}$.[2] A sufficient, but not necessary, condition for z_{t+1} to be mean zero stationary is that y_t 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, i.e. that its variance exists.

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:*

[2]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.

(A1) *or*

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

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

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

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 the expected value of each random coefficient in the equation for $y_{1.t+1}$ is constant for all regimes $s_t = i$. 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 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.

When the Markov processes $s_{1.t+1}$ and $s_{2.t+1}$ are independent the restrictions of Proposition 2 may be simplified.

Corollary 1. *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 1 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. 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 total number of restrictions is equal to $p + 1$, corresponding to the p restrictions on the lags and one restriction on the Markov transition matrix ($p_{11} + p_{22} = 1$).

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. The total number of restrictions is now equal to $p(3 + N_2 + N_3) + 1$, where there is one restriction on each lag of y_1 ($a_{11,1}^{(k)} = a_{11,2}^{(k)}$), N_2 and N_3 restrictions on each lag of y_2 and y_3, respectively, ($a_{12,1}^{(k)} = a_{12,2}^{(k)}$ and $a_{13,1}^{(k)} = a_{13,2}^{(k)}$), two restrictions on each lag of y_4 ($a_{14.1}^{(k)} = a_{14.2}^{(k)} = 0$), and one restriction on the constant term ($m_{1.1} = m_{1.2}$).

Another case when Corollary 1 is of special interest is for the mixture VAR model, i.e., when $\mathbf{P} = \iota_M \pi'$. For such models it is straightforward to show that y_4 does not Granger cause y_1 if

$$\sum_{j=1}^{M} a_{14,j}^{(k)} \pi_j = 0, \tag{18}$$

for all $k = 1, \ldots, p$. In fact, the Granger noncausality restrictions are unique for the mixture VAR model and the reason is that (A2) with $M_2 = 1$ holds by assumption and, as a consequence, the restrictions in (A1) imply that (A4) is true but the reverse is not true, i.e., (A4) is *minimal* in the sense of Gabriel (1969), while (A1) is not and can therefore be discarded.[3] This result is also quite intuitive since for mixture models the optimal prediction of the regime in the next period is the ergodic probability (π_j) for each regime j, with the implication that y_4 can only improve the one-step-ahead forecasts of y_1 when the ergodic mean of the coefficient on y_4 is nonzero for some lag in the y_1 equation.

4. Bayesian Testing

Restrictions (A1)–(A4) can be tested using either classical or Bayesian inference. Below, we briefly discuss classical tests and point out some important obstacles in the current setting and then present the Bayesian testing procedure.

4.1. Classical Inference

Apart from more general problems related to classical inference, such as those related to size and power, two specific obstacles that we need to take into account when attempting to draw inferences from classical tests in an MS-VAR setting are:

- Granger noncausality results in multiple sets of restrictions on the parameters. Consequently, one hypothesis may be represented by several restricted models;
- Some of the restrictions are in the form of nonlinear functions of the parameters.

These problems may potentially be difficult to handle in a classical setting, especially when taken together. Issues related to multiple testing—a subfield of *multiple inference* or *simultaneous inference*—are well-known in statistics; see, e.g., Schaffer (1995) for a review on this topic, and Holm (1979) for details on the so called Holm-Bonferroni method which may be applied to the Granger noncausality restrictions in Proposition 2. The particular procedure suggested by Holm may be used as long as the asymptotic distribution of each individual test statistic is known, and improves upon the so called Bonferroni correction (at least in large samples), but nevertheless remains conservative.

Standard classical tests of nonlinear restrictions on parameters typically rely on computing the matrix of partial derivatives of the restrictions with respect to the parameters. If this matrix has rank less than the number of restrictions, the asymptotic distribution of the test statistic is generally not known, but depends on the rank of this matrix. If this is the case, then the Holm-Bonferroni

[3] A hypothesis is said to be minimal if it does not imply the truth of any other hypothesis in a set containing multiple hypotheses.

method cannot be applied as it relies on the distribution of the tests for each individual hypothesis being known.

However, this problem does not plague the nonlinear Granger noncausality restrictions in (A3) or (A4). The intuitive reason for this result is that the VAR parameter themselves never appear nonlinearly, but only as products with the Markov transition probabilities. As a results the transition probabilities appear in the matrix with partial derivatives individually and not multiplied by any VAR parameter.

To see why this observation is important, let us consider the Granger noncausality restrictions for the mixture VAR model in equation (18) and assume for simplicity that $p = 2$. The matrix with partial derivatives of the two restrictions with respect to only the parameters involved in the restrictions is then

$$\frac{\partial f\left(\theta_{(r)}\right)}{\partial \theta'_{(r)}} = \begin{bmatrix} a_{14,1}^{(1)} & \cdots & a_{14,M}^{(1)} & \pi_1 & \cdots & \pi_M & 0 & \cdots & 0 \\ a_{14,1}^{(2)} & \cdots & a_{14,M}^{(2)} & 0 & \cdots & 0 & \pi_1 & \cdots & \pi_M \end{bmatrix},$$

where

$$\theta_{(r)} = \begin{bmatrix} \pi_1 & \cdots & \pi_M & a_{14,1}^{(1)} & \cdots & a_{14,M}^{(1)} & a_{14,1}^{(2)} & \cdots & a_{14,M}^{(2)} \end{bmatrix}'.$$

The rank of the matrix of partial derivatives of the restrictions is always two since the π_j probabilities are positive.

It should be emphasized that for each possible rank of $\mathbf{P}$, the exact form of the Granger noncausality restrictions in MS-VARs is affected, also in terms of the number of restrictions, and the restrictions become linear when $\mathbf{P}$ has full rank M. For each possible rank of $\mathbf{P}$, the matrix of partial derivatives of the restrictions in (A3) or (A4) has full row rank, due to the assumptions on the behavior of the transition probabilities p_{ij}. However, this also suggests that for nonlinear restrictions where the VAR parameters appear in terms of, say, products or have exponents different from one, then the matrix with partial derivatives can have reduced row rank.[4]

4.2. Bayesian Inference

In this study we make use of Bayesian inference when testing the parameter restrictions. The approach we suggest can deal with both multiple sets of restrictions and nonlinearities; see also Woźniak (2012b, 2012a), where Granger noncausality is studied within the Extended CCC-GARCH model of Jeantheau (1998).[5]

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, $\mathcal{H}_i = \{j : \mathcal{M}_j \text{ represents } i^{th} \text{ hypothesis}\}$. The models that are included in this set are assumed to be minimal. Furthermore, suppose that we are interested in comparing the posterior probability of this hypothesis to hypothesis $\mathcal{H}_0$, represented by the unrestricted model $\mathcal{M}_0$. The credibility of the hypothesis $\mathcal{H}_i$ compared to the hypothesis

[4]One such case is if we are interested in restrictions on the conditional prediction variances of the MS-VAR model; see, e.g., Warne (2000) for Granger noncausality in mean-variance. Another case is if we are concerned with restrictions on the h-step-ahead forecasts for $h \geq 2$.

[5]Two other works use the Bayesian approach to make inference about concepts somehow related to Granger noncausality. Jarociński & Maćkowiak (2013) sample from the space of models in order to determine Granger-causal-priority in the VAR model, while Pajor (2011) uses Bayes factors to assess conditional exogeneity conditions for models with latent variables, and in particular in multivariate stochastic volatility models.

$\mathcal{H}_0$ may then be assessed with the posterior odds ratio given by:

$$\text{POR} = \frac{\Pr\left(\mathcal{H}_i|\mathbf{y}_T\right)}{\Pr\left(\mathcal{H}_0|\mathbf{y}_T\right)} = \frac{\sum_{j\in\mathcal{H}_i}\Pr\left(\mathcal{M}_j|\mathbf{y}_T\right)}{\Pr\left(\mathcal{M}_0|\mathbf{y}_T\right)} = \frac{\Pr\left(\mathcal{H}_i\right)}{\Pr\left(\mathcal{H}_0\right)}\cdot\frac{p\left(\mathbf{y}_T|\mathcal{H}_i\right)}{p\left(\mathbf{y}_T|\mathcal{H}_0\right)}. \tag{19}$$

If we set equal prior probabilities for all the hypotheses, then the posterior odds ratio is equal to the Bayes factor and is given by the ratio of marginal data densities (MDDs) when conditioning on $\mathcal{H}_i$ and $\mathcal{H}_0$, respectively.

The MDD is typically calculated for a given model $\mathcal{M}_j$ rather than for a hypothesis $\mathcal{H}_i$. To determine the MDD for the hypothesis $\mathcal{H}_i$ using the MDDs for the models that are consistent with it, we can utilize the following:

$$p\left(\mathbf{y}_T|\mathcal{H}_i\right) = \sum_{j\in\mathcal{H}_i} p\left(\mathbf{y}_T|\mathcal{M}_j\right)\Pr\left(\mathcal{M}_j|\mathcal{H}_i\right).$$

If we assume that all models $\mathcal{M}_j$ are equally likely apriori given that the hypothesis $\mathcal{H}_i$ is true, then the MDD given the hypothesis is equal to the average of the MDDs for the models.

The restrictions in (A1) and (A3) depend on auxiliary values and can therefore result in multiple models that are consistent with each one of these two conditions, respectively. Condition (A1) depends on the number of states of the hidden Markov process, M, through the decomposition of process s_t into (s_{1t}, s_{2t}) such that $M_1 \cdot M_2 = M$. In our empirical example we find support for $M = 3$ in the bivariate money-income model for monthly US data. For this case, two decompositions are possible: $M_1 = 1$ and $M_3 = 3$, or $M_1 = 3$ and $M_3 = 1$. In order to compute the MDD of condition (A1) we integrate out the possible decompositions by applying:

$$\begin{aligned} p\left(\mathbf{y}_T\middle|(\text{A1})\right) = p\left(\mathbf{y}_T\middle|(\text{A1})\wedge M_1 = 1, M_2 = 3\right)\Pr\left(M_1 = 1, M_2 = 3\middle|(\text{A1})\right)+ \\ + p\left(\mathbf{y}_T\middle|(\text{A1})\wedge M_1 = 3, M_2 = 1\right)\Pr\left(M_1 = 3, M_2 = 1\middle|(\text{A1})\right), \end{aligned} \tag{20}$$

where $p(\mathbf{y}_T|(\text{A1})\wedge M_1 = 1, M_2 = 3)$ and $p(\mathbf{y}_T|(\text{A1})\wedge M_1 = 3, M_2 = 1)$ are estimated by an available MDD estimator (see below), and the conditional prior probabilities may be selected as $\Pr(M_1 = 1, M_2 = 3|(\text{A1})) = \Pr(M_1 = 3, M_2 = 1|(\text{A1})) = 1/2$.

Similarly, condition (A3) depends on the rank of $\mathbf{P}$ and this value is not of interest from the point of view of the Granger causality testing, but nevertheless affects the restrictions. In this case, we can integrate out the rank of $\mathbf{P}$ from the testing problem by computing:

$$p\left(\mathbf{y}_T\middle|(\text{A3})\right) = \sum_{i=1}^{M} p\left(\mathbf{y}_T\middle|(\text{A3})\wedge \text{rank}(\mathbf{P}) = i\right)\Pr\left(\text{rank}(\mathbf{P}) = i\middle|(\text{A3})\right), \tag{21}$$

where $p(\mathbf{y}_T|(\text{A3})\wedge \text{rank}(\mathbf{P}) = i)$ are estimated by a MDD estimator, whereas $\Pr(\text{rank}(\mathbf{P}) = i|(\text{A3})) = 1/M$, for $i = 1, \ldots, M$, is a possible choice for the conditional prior probabilities.

4.3. Testing Noncausality Restrictions in MS-VARs

The crucial element of using the posterior odds ratio in (19) to assess the hypotheses of interest is the computation of MDDs, $p(\mathbf{y}_T|\mathcal{M}_j)$, for the unrestricted and the restricted models. There are several available methods of computing this value. In this paper we use the modified harmonic

mean (MHM) method of Geweke (1999, 2005). Amongst other methods of computing the MDD is the one suggested by Sims et al. (2008) based on an elliptical truncation rather than a normal, and which also belongs to a class of the modified harmonic mean estimators. The difficulty in employing other estimators, such as the bridge sampling estimator by Frühwirth-Schnatter (2004) or the one by Chib & Jeliazkov (2001) would require further studies and adjustments and is left for future research.[6]

Using the posterior odds ratio when testing a noncausality hypothesis represented by a couple of restricted models makes the testing possible. Moreover, as emphasized in Hoogerheide, van Dijk & van Oest (2009), the Bayesian posterior odds ratio procedure gives arguments *in favour of* hypotheses. Therefore, the procedure gives positive arguments supporting particular solutions.

However, the approach also has its costs. First of all, in order to specify the complete model and thereby avoid Bartlett's paradox, prior distributions for the parameters of the model and the prior probabilities of models need to be specified.[7] Moreover, the time required for simulating all the models can be costly, first in the model selection procedure, and second in testing the restrictions of the parameters.

5. Block Metropolis-Hastings Sampler for MS-VAR Models

This section describes the likelihood function, prior distributions and the algorithm for the estimation of the unrestricted and restricted models; the details of the algorithm are given in the Statistical Appendix. Our parametrization allows for the estimation of the restricted models, where the restrictions on the parameters are given by the regime inference restrictions in Proposition 1 or Granger noncausality restrictions in Proposition 2.

The complete-data likelihood function is equal to the joint sampling distribution $p(\mathbf{S}_T, \mathbf{y}_T|\theta)$ for the complete data $(\mathbf{S}_T, \mathbf{y}_T)$ given $\theta = (\theta_1, \ldots, \theta_M, \mathbf{P})$, where $\mathbf{S}_T = (s_0, s_1, \ldots, s_T)'$; see, e.g., Frühwirth-Schnatter (2006). This distribution is further decomposed into a product of a conditional distribution of $\mathbf{y}_T$ given $\mathbf{S}_T$ and θ, and a conditional distribution of $\mathbf{S}_T$ given θ, and by taking into account a convenient partitioning of the vector of parameters into state-specific vectors of the VAR, θ_i, and the matrix with transition probabilities, $\mathbf{P}$:

$$p\big(\mathbf{S}_T, \mathbf{y}_T\big|\theta\big) = p\big(\mathbf{y}_T\big|\mathbf{S}_T, \theta\big)\Pr\big(\mathbf{S}_T\big|\mathbf{P}\big). \tag{22}$$

The two components on the right hand side of equation (22) are the same as in Frühwirth-Schnatter (2006, Section 11.3.1).

We assume that the prior distribution of the state-specific parameters for each state and the

[6] Frühwirth-Schnatter (2004) raises the problem that the MDD estimator can be biased due to the invariance of the likelihood function and the prior distribution of the parameters, with respect to permutations of the regimes' labels. The identification of the model can be insured by ordering restrictions on parameters, and can also be implemented within the Gibbs sampler. In essence, 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). The MS-VAR models considered for causality inference below are globally identified via the ordering imposed on one of the state-specific parameters. In our empirical example we did not encounter any difficulties in finding such restrictions that would effectively not constrain posterior distributions.

[7] See, e.g., Strachan & Van Dijk (2014) for analyses about using improper priors without being exposed to Bartlett's paradox.

transition probabilities matrix are independent:

$$p\left(\theta\right) = \prod_{i=1}^{M} p\left(\theta_i\right) p\left(\mathbf{P}\right). \tag{23}$$

This introduces 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 MS-VAR 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 p_{ii}, for $i = 1, \ldots, M$, for which it is set to the hyper-parameter $\lambda_{\mathbf{P}}$. If $\lambda_{\mathbf{P}} > 1$, the regimes are persistent over time (see e.g. Kim & Nelson, 1999).

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

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

for $s_t = 1, \ldots, M$. These parameters follow a $(N + pN^2)$-variate normal distribution, with mean equal to a vector of zeros and a diagonal covariance matrix, $\underline{V_\beta}$. Note that the means of the prior distribution for the off-diagonal elements of matrices A_{s_t} are set to zero. In more general terms, the means of the prior distributions assumed in this work imply Granger noncausality.

Our prior distribution nests many popular in empirical macroeconomics research prior specifications, including the class of shrinkage prior specifications, and can be easily extended to hierarchical prior structures. Therefore, prior distributions proposed e.g. by Doan, Litterman & Sims (1983), Ni & Sun (2003) or Bańbura, Giannone & Reichlin (2010) could also be used. Furthermore, the mean vector of the normal prior distribution of parameters β_{s_t} is set to a vector of zeros, since in the empirical example in Section 6 we use logarithmic rates of returns of the original variables. If logarithms of the levels of the variables are modeled, then the mean vector of this prior distribution could be set such that it contained ones for the diagonal elements of matrices $A^{(1)}_{s_t}$, for $s_t \in \{1, \ldots, M\}$ (see Doan et al., 1983).

We model the state-dependent covariance matrices of the error term, decomposing each one 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:

$$\Sigma_{s_t} = \text{diag}\left(\sigma_{s_t}\right) \mathbf{R}_{s_t} \text{diag}\left(\sigma_{s_t}\right). \tag{24}$$

Modeling covariance matrices using a decomposition into standard deviations and a correlation matrix, as in equation (24), 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. 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 $\lambda_\sigma > 0$. 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.[8] 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 the operator, vecl, stacks all the lower-diagonal elements of the correlation matrix into a vector.

To summarize, the prior specification (23) now takes the detailed form of:

$$p(\theta) = \prod_{i=1}^{M} p(\mathbf{P}_i)p(\beta_i)p(\mathbf{R}_i)\left(\prod_{j=1}^{N} p(\sigma_{i.j})\right), \tag{25}$$

where each of the prior distributions is as assumed:

$$\begin{aligned}
\mathbf{P}_{i\cdot} &\sim \mathcal{D}_M(\iota_M' + (\lambda_p - 1)I_{M.i\cdot}) \\
\beta_i &\sim \mathcal{N}(\mathbf{0}, \underline{V_\beta}) \\
\sigma_{i.j} &\sim log\mathcal{N}(0, \lambda_\sigma) \\
\mathbf{R}_{i.jk} &\sim \mathcal{U}(a, b)
\end{aligned}$$

for $i = 1, \ldots, M$ and $j, k = 1, \ldots, N$, and $j \neq k$, 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, while a and b are as in Algorithm 3 in the Statistical Appendix.

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. Symbols, l and $l-1$, refer to the iteration of the MCMC sampler. For the first iteration of a MCMC sampler, $l = 1$, initial parameter values come from an EM algorithm, and are denoted by $\theta^{(0)}$.

1. Draw a vector of the states of the economy, $\mathbf{S}_T$. Using the filter and smoother (see, e.g., Frühwirth-Schnatter, 2006, and references therein), we obtain the probabilities $\Pr(s_t = i|\mathbf{y}_T, \theta^{(l-1)})$, for $t = 1, \ldots, T$ and $i = 1, \ldots, M$, and then draw $\mathbf{S}_T^{(l)}$, for l^{th} iteration of the algorithm.
2. Draw from the posterior distribution of the transition probabilities matrix conditioning on the states drawn in the previous step of the current iteration, $\mathbf{P}^{(l)} \sim p(\mathbf{P}|\mathbf{S}_T^{(l)})$. 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 the Statistical Appendix.
3. Draw from the full conditional densities of σ and $\mathbf{R}$, denoted by $p(\sigma|\mathbf{y}_T, \mathbf{S}_T^{(l)}, \mathbf{P}^{(l)}, \beta^{(l-1)}, \mathbf{R}^{(l-1)})$ and $p(\mathbf{R}|\mathbf{y}_T, \mathbf{S}_T^{(l)}, \mathbf{P}^{(l)}, \beta^{(l-1)}, \sigma^{(l)})$, respectively, with the Griddy-Gibbs sampling algorithm of Ritter & Tanner (1992), and described by Barnard et al. (2000).
4. 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 $f(\beta|\mathbf{y}_T, \mathbf{S}_T^{(l)}, \mathbf{P}^{(l)}, \sigma^{(l)}, \mathbf{R}^{(l)}) = \mathcal{N}\left(\bar{\beta}^*, \bar{V}_{\beta^*}\right)$, from which we draw $\beta^{(l)}$. $\bar{\beta}^*$ and $\bar{V}_{\beta^*}$ are the parameters of the full conditional distribution specified in the Statistical Appendix.

[8] Barnard et al. (2000) discusses the implications of such a prior specification and the algorithm of choosing a and b. In addition, this paper mentions alternative prior distributions that could be used as well.

6. Granger Noncausality Analysis of US Money and Income

In the studies on Granger causality using MS-VAR models, Warne (2000) and Psaradakis et al. (2005),[9] the focus 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 lead 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 relationship between money and income is, however, of particular interest to the debate, since economists have not reached a consensus on this topic.

This historical debate 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, 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' – unrestricted – 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.

6.1. Data

The data are similar to those estimated by Christiano & Ljungqvist (1988) and Warne (2000), but the sample is longer and spans a period of 53 years. Two monthly series are included, the M1 money stock and the industrial production index for the US, both containing 646 observations covering the period from 1959:1 to 2012:11 and taken from the Citibase database. The data are seasonally adjusted, transformed into log-returns, and multiplied by 1200.

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. The period of the global financial crisis is also characterised by increased volatility in both series, especially after August 2008. 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.

The summary statistics of both series are presented in Table 1. Income grows yearly by around 2.7% on average, with a standard deviation of approximately 10%. Money has a stronger growth rate of nearly 5.3% on average, with a slightly lower standard deviation than income, around 8.3%.

[9] Warne (2000) uses monthly industrial production data as a proxy for income, whereas Psaradakis et al. (2005) use quarterly real GDP data.

6.2. Granger Noncausality Analysis with VAR Models

To study if a nonlinear approach brings added value to the Granger causality analysis we begin by examing a standard VAR model, i.e., the case of $M = 1$. The block Metropolis-Hastings sampler of Section 5 can be simplied to the single regime case and, thus, be used for standard Bayesian VAR models. This makes it possible to compare the VAR models to more complex MS-VAR ones with MDDs and also to examine if the Granger causal analysis suggests similar conclusions in VARs and MS-VARs.

The prior distributions are as follow:

$$\begin{aligned} \beta &\sim \mathcal{N}(\mathbf{0}, \lambda_\beta I_{N+pN^2}) \\ \sigma_j &\sim log\mathcal{N}(0, 2) \\ \mathbf{R}_{jk} &\sim \mathcal{U}(a, b) \end{aligned}$$

for $j, k = 1, \ldots, N$ and a and b as in Algorithm 3. The value of hyper-parameter λ_β has been determined by a grid search and is set to 0.3.

To estimate VAR models for different lag lengths ($p = 0, \ldots, 17$) using the Metropolis-Hastings algorithm the parameters are initialized by the OLS estimates of the VAR coefficients. Then follows a 10,000-iteration burn-in and, after convergence of the sampler, 50,000 final draws from the posterior.

Table 2 displays the MDDs for each model, computed with the modified harmonic mean (MHM) estimator by Geweke (1999, 2005). As in Christiano & Ljungqvist (1988) and Warne (2000), models with long lags are preferred, and the VAR(14) model yields the highest MDD, denoted by $\ln p_{MHM}(\mathbf{y}_T|p)$, and equal to -4544.68, and is therefore the model we choose for the Granger causality analysis.

Table 3 summarizes the results for the unrestricted and restricted VAR(14) models. Estimation of the restricted VAR model, where the coefficients on lags of money in the income equation are equal to zero, yields an MDD of -4518.43. Expressed in logarithms of base 10 rather than natural logarithms, the posterior odds ratio of the null hypothesis of Granger noncausality from money to income is therefore equal to 11.4. Hence, Bayesian testing provides strong evidence in favor of the hypothesis that money does not Granger cause income within the VAR framework for log-differences.[10]

6.3. Granger Noncausality Analysis with MS-VAR Models

To estimate the number of regimes, M, and the number of lags, p, we consider MS-VAR models with a maximum of four regimes and seven lags. The prior distributions are as defined in Section 5 with a diagonal prior covariance matrix of β_i given by $\underline{V}_\beta = \lambda_\beta I_{N+pN^2}$, with $\lambda_\beta = 0.3$ as in the VAR model analysed before, and $\lambda_\sigma = 2$. The value of the hyper-parameter $\lambda_{\mathbf{P}} = 10$ implies that the states are persistent over time. The expected duration of the states implied by such prior assumptions depends on the number of states, M. For instance, for the models with two states, $M = 2$, the prior distribution implies the duration of the states of around eleven periods, whereas

[10]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. Christiano & Ljungqvist (1988), however, contest this result and argue for a specification error for models with first differences. Warne (2000) also finds that money does not Granger cause income in the bivariate VAR model for log-differences, but that it does in the log-levels specification.

for the model with three states, $M = 3$, the duration of the states is around six periods. The block Metropolis-Hastings algorithm for each model is initialized with the estimates from the EM algorithm of the corresponding model. Then follows a 10,000-iteration burn-in period and, after convergence of the sampler, we sample 100,000 final draws from the posteriors[11].

Table 4 reports the estimated MDDs for the MS-VAR models with 2 and 3 regimes. The case of $M = 4$ is not provided since the computations suggest that MS-VAR models with more than 3 regimes are not supported by the data.[12] The number of lags for the autoregressive coefficients is limited to 7—less than the 17 lags for VAR models—also due to insufficient state occurrences when the number of lags increases. The model preferred by the data is the MS-VAR with 3 regimes and a lag order equal to 3.

Figure 2 plots the regime probabilities from the selected model. State 1 has the highest probabilities of occurrence in the period before 1978, and is characterised by moderate average growth (represented by the μ parameter) and volatility (represented by σ) of the series; see Table 5 for the posterior estimates for the unrestricted MS-VAR model with 3 states and 3 lags. The second state has the highest probabilities of occurrence in the period starting from 1984, and this state is the one with the highest average growth of industrial production and its lowest standard deviation. The second state is also a state of the highest average growth rate of M1 with a moderate level of volatility. The third state has probabilities close to one for the whole year starting in August 2008. This state is also highly likely after July 2011, as well as in the early 1980's and in year 1959. This state has the largest standard deviations, 2.5 times higher than in any other state for income, and 3 times higher than in any other state for money. Moreover, this is the only state in which the average growth of income is negative as measured by the posterior mean of the intercept term of the VAR equation.

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 base 10 of the posterior odds ratio of over 69 in favour of the MS-VAR model when the models are given equal prior probability, thereby strongly supporting the specification of the model where parameters change over time based on a hidden Markov process.

We proceed with the analysis of Granger noncausality for the selected MS(3)-VAR(3) model. The Bayesian testing strategy we employ renders the process straightforward: 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 6 gives the restrictions in (A1)–(A3) for MS-VAR models with three regimes and provides an accounting of the number of restrictions imposed on the parameters.

It can be seen from Table 6 that condition (A1) imposes the largest number of restrictions, and condition (A2) the smallest. The (A1) condition is divided into two models, $\mathcal{M}_1$ and $\mathcal{M}_2$, where the former model mainly covers restrictions on the parameters of the income equation, and the latter mainly on the money equation. The fact that the number of restrictions is greater for $\mathcal{M}_2$ than for

[11] The number of Gibbs algorithm iterations is increased for models that require the simulation of the latent Markov process due to a slightly lower efficiency of simulations for these models.

[12] The computations encountered difficulties for MS-VAR models with 4 regimes that are due to insufficient occurrence of one regime. We assume that the hidden Markov process is stationary which implies nonzero ergodic state probabilities. A situation in which at some Gibbs iteration one of the states has zero occurrences violates this assumption and is not allowed in our algorithms. This restriction made sampling from the posterior distribution of parameters of many of the considered models with 4 states practically impossible. This indicates that the data does not support MS-VAR models with 4 or more regimes, and explains why we only present results with at most 3 regimes.

$\mathcal{M}_1$ is explained by the fact that the former models also includes zero restrictions on parameters in the income equation. The restrictions satisfied by these two models allow the regime process to be serially correlated, while condition (A2) with model $\mathcal{M}_3$ implies that it is not. As can be seen from the Table, these three models are minimal and from Proposition 1 it follows that if one of them is true, then there is not any information in the history of money for improving the predictions of next periods state of the parameters which can affect income.[13]

Models $\mathcal{M}_4$–$\mathcal{M}_6$ jointly imply that condition (A3) holds and are based on the different values for the rank of the matrix with Markov transition probabilities. The first two of these (A3) models have nonlinear restrictions, while the last model has only linear restrictions. It is interesting to note that the number of restrictions for these models is increasing with the rank of the $\mathbf{P}$ matrix.

Table 7 reports natural logarithms of the MDDs given the model and logarithms of the Bayes factors, $\log_{10} \mathcal{B}_{j0}$ for $j = 0, \ldots, 6$. A positive logarithm of the Bayes factor is to be interpreted as evidence in favour of the restricted model. In a symmetric way, negative logarithm of the Bayes factor indicates that the unrestricted model is preferred by the data.

The results in Table 7 show that model $\mathcal{M}_6$ has the highest MDD among the six restricted models and is comparable to the MDD of the unrestricted model, $\mathcal{M}_0$. The other models ($\mathcal{M}_1$–$\mathcal{M}_5$), however, are much less probable than the unrestricted model, as represented by the large negative values of the logarithms of the Bayes factors. Moreover, the MDDs and Bayes factors of conditions (A1)–(A3) are reported. Due to the inclusion of model $\mathcal{M}_6$, only condition (A3) is given some posterior support.

Table 8 presents a summary of the assessment of the considered hypotheses. The hypothesis that the history of money does not imporve the forecast of the regime in the next period (see Proposition 1) is covered by the three minimal individual hypotheses represented by models $\mathcal{M}_1$–$\mathcal{M}_3$. The logarithm of the Bayes factor is here close to -17 when compared with the unrestricted MS-VAR model and, hence, the empirical evidence for US money and income is strongly in favor of the model where the history of money is useful to improving the predictions of the regimes of the parameters which can affect income.

Turning to the Granger noncausality hypothesis, it should be noted that we here represent it by the four models $\mathcal{M}_2, \mathcal{M}_4, \mathcal{M}_5, \mathcal{M}_6$. Model $\mathcal{M}_1$ also implies that money is Granger noncausal for income, but has been excluded from the joint hypothesis. The reason is that $\mathcal{M}_1$ is not minimal since, when true, it implies that the hypothesis $\mathcal{M}_6$ is also true. Classical inference on a multiple hypothesis using, e.g., the Holm-Bonferroni method (see, e.g., Holm, 1979), is based on the condition that all of the individual hypotheses in a multiple hypothesis are minimal. For this reason we also opt to exclude $\mathcal{M}_1$ from the multiple hypothesis $\mathcal{H}_2$ when using Bayesian inference. Since the MDD of $\mathcal{M}_1$ is low compared with some of the models included in $\mathcal{H}_2$, the results in Table 8 are barely affected by this requirement.

From Table 8 it can be seen that the MDD of the joint noncausality hypothesis is lower than the MDD of the unrestricted model by roughly 8.7 natural logarithm units. Translated to logarithms of the base 10 scale, this yields a Bayes factor of roughly -3.8, suggesting that the empirical evidence of Granger noncausality from money to income on monthly US data is, at best, very weak when we condition on MS-VAR models. In other words, the Bayes factor of these two hypotheses is equal to $10^{3.8}$ in favor of Granger causality from money to income.

[13] Model $\mathcal{M}_1$ is trivial in the sense that the regime process s_{1t}, which is allowed to affect the parameters in the income equation, has a unit dimension and is therefore observed. Consequently, the history of money cannot improve the predictions of $s_{1.t+1}$.

A byproduct of the analysis for Markov-switching model is Granger noncausality for micture VARs. As already discussed in Section 3, the mixture-VARs are nested within the MS-VARs by setting the rank of the transition probabilities matrix, **P**, to unity. Table 9 reports the results of testing for Granger noncausality in mixture VARs with the number of mixture components set to 3 and the number of lags of VAR equal to 3.[14] The results are qualitatively similar to those for MS-VARs, with a Bayes factor equal to $10^{3.6}$ in favor of Granger causality over noncausality.

7. Conclusions

In this paper we derive sets of restrictions on the parameters of MS-VAR models that can be used to test for Granger noncausality and for examining which observed variables have information relevant for improving the predictions of the underlying and unobserved Markov process that determines the regimes.

It is shown that both the Granger noncausality and the regime inference hypotheses imply multiple sets of restrictions on the parameters of the MS-VAR. The number of such sets depends not only on the lag order of the VAR but also on the dimension of the observable variable vector and on the number of regimes. Granger noncausality results in some of the 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.

In this paper we have proposed a method of testing the 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 to multiple testing. It requires, however, an algorithm for the estimation of the unrestricted model and of the restricted models, representing the hypotheses of interest. The algorithm we have suggested allows for restrictions on all groups of parameters of the model, i.e., parameters on the constant term, lagged variables, variances and covariances of the innovations, and the transition probabilities of the hidden Markov process. It combines several existing algorithms in order to maintain the desired properties of the model and the efficiency of estimation.

In the empirical investigation we found that Granger noncausality from monthly US money to income established for linear (single regime) VARs is contradicted by the evidence from nonlinear models. The causality analysis of MS-VARs suggest that money is essential for the forecasting of the probabilities of the states which influence the behavior of income. Although Granger noncausality is given a non-zero posterior probability, the posterior probability of the Granger causality hypothesis is several 1000s times larger for MS-VARs.

Since mixture VAR models are nested in MS-VAR models, our analytical results on Granger noncausality can also be applied to such models. In the empirical example for US money and income, we also find strong support in favor of the hypothesis that money Granger causes income in a mixture VAR. Moreover, we find that MS-VARs dominate mixture VARs, with a Bayes factor of about $10^{16.4}$, while mixture VARs strongly dominate the linear VAR specification, with a Bayes factor of about $10^{41.8}$.

One limitation of the analysis on Granger causality in MS-VAR models is that we only consider *one-step-ahead* forecasts. The conditions for Granger noncausality using *multi-steps-ahead* forecast

[14] Notice that three regimes and three lags may not be the preferred choice of these parameters if we were to allow only for mixture VARs when estimating them.

could be further explored. It is notable that the conditions we have provided on regime inference applies to multi-steps-ahead forecast of the regime process and can therefore be made use of for such noncausality analysis. Still, establishing conditions for the noncausality h-steps-ahead for the autoregressive parameters, including covariances, would potentially require tedious algebra.

The Granger noncausality analysis that we have presented in this paper focuses on the properties of the mean squared errors of the forecasts. It is possible that, e.g., money does not Granger cause income from this perspective, but may nevertheless incorporate important information which is valuable for determining higher moments than the mean of the predictive distribution of income. Warne (2000) provides additional noncausality concepts, 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.

Acknowledgments

This paper was presented at the Society for Nonlinear Dynamics and Econometrics 22nd Annual Symposium in New York, USA, the Econometric Society Australasian Meeting 2012 in Melbourne, Australia, and at the 22^{nd} EC^2 Conference: *Econometrics for Policy Analysis: after the Crisis and Beyond* in Florence, Italy, as well as during seminars at the Australian National University, National Bank of Poland, Deutsche Bundesbank, Freie Universität Berlin, Queensland University of Technology Business School, Deakin University, Monash University, Università degli Studi di Padova and the University of Melbourne. The authors thank the participant of the seminars and in particular Joshua Chan, Peter R. Hansen, Andrzej Kociecki, Helmut Lütkepohl, Massimiliano Marcellino, Mateusz Pipień, Rodney Strachan for their useful comments on the paper. The opinions expressed in the paper are those of the authors and do not necessarily reflect those of the European Central Bank (ECB).

References

Bańbura, M., Giannone, D., & Reichlin, L. (2010). Large Bayesian Vector Auto Regressions. *Journal of Applied Econometrics*, *92*, 71– 92.

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. (1994). Simplified Conditions for Noncausality Between Vectors in Multivariate ARMA Models. *Journal of Econometrics*, *63*, 271–287.

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 Coefficient. *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.

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.

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.

Gabriel, K. R. (1969). Simultaneous Test Procedures — Some Theory of Multiple Comparisons. *The Annals of Mathematical Statistics*, *40*, 224–250.

Geweke, J. (1999). Using Simulation Methods for Bayesian Econometric Models: Inference, Development, and Communication. *Econometric Reviews*, *18*, 1–73.

Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*. John Wiley & Sons, Inc.

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.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

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

Jarociński, M., & Maćkowiak, B. (2013). Granger-Causal-Priority and Choice of Variables in Vector Autoregressions. European Central Bank Working Paper Series.

Jeantheau, T. (1998). Strong Consistency of Estimators for Multivariate ARCH Models. *Econometric Theory*, *14*, 70–86.

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

Kim, C.-J., & Nelson, C. R. (1999). 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.

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.
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.
Mosconi, R., & Seri, R. (2006). Non-Causality in Bivariate Binary Time Series. *Journal of Econometrics*, *132*, 379–407.
Ni, S., & Sun, D. (2003). Noninformative Priors and Frequentist Risks of Bayesian Estimators of Vector-Autoregressive Models. *Journal of Econometrics*, *115*, 159–197.
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.
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.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688 – 701.
Schaffer, J. P. (1995). Multiple Hypothesis Testing. *Annual Review of Psychology*, *46*, 561–584.
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.
Strachan, R. W., & Van Dijk, H. K. (2014). Divergent Priors and Well Behaved Bayes Factors. *Central European Journal of Economic Modelling and Econometrics*, *6*, 1–31.
Tobin, J. (1970). Money and Income: Post Hoc Ergo Propter Hoc? *The Quarterly Journal of Economics*, *84*, 301–317.
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. (2012a). *Testing Causality Between Two Vectors in Multivariate GARCH Models*. EUI Working Papers ECO 2012/20 European University Institute Florence, Italy.
Woźniak, T. (2012b). *Granger Causal Analysis of VARMA-GARCH models*. EUI Working Papers ECO 2012/19 European University Institute Florence, Italy.

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 $\xi_{t|\tau}(j) = \Pr[s_t = j|y_\tau, \mathcal{W}_\tau]$, for $j = 1, \ldots, M$, 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 $(s_{1,t}, s_{2,t}) = (j_1, j_2)$, for $j_1 = 1, \ldots, M_1$ and $j_2 = 1, \ldots, M_2$, 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

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

Collecting these results we find that

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

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

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

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

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

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 $j = 1, \ldots, M$:

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

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

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

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*

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

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

Proof. Let us first prove this for all $\tau < t$. We have already established in Lemma 2 that $\xi^{(1)}_{\tau|\tau}$ and $\xi^{(2)}_{\tau|\tau}$ 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 $\xi^{(1)}_{\tau|\tau}$ and $\xi^{(2)}_{\tau|\tau}$ we obtain $\xi_{t|\tau} = [(\mathbf{P}^{(1)\prime})^{t-\tau}\xi^{(1)}_{\tau|\tau} \otimes (\mathbf{P}^{(2)\prime})^{t-\tau}\xi^{(2)}_{\tau|\tau}] = (\xi^{(1)}_{t|\tau} \otimes \xi^{(2)}_{t|\tau})$, 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

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

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

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

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

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

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 $j_1 \in \{1, \ldots, M_1\}$, $j_2 \in \{1, \ldots, M_2\}$, $r \in \{1, 2\}$, and $k \in \{1, \ldots, p\}$. 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

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

where $\tilde{\Sigma}_{11,j} \equiv \Sigma_{11,j} - \Sigma_{12,j}\Sigma_{22,j_2}^{-1}\Sigma'_{12,j}$. 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 $j_1 \in \{1, \ldots, M_1\}$, $j_2 \in \{1, \ldots, M_2\}$, $r \in \{1, 2\}$, and $k \in \{1, \ldots, p\}$; and (c) $\Sigma_{12,j} = 0$ for all $j \in \{1, \ldots, M\}$. Under (i) to (iii) we find that $\eta_t = (\eta_t^{(1)} \otimes \eta_t^{(2)})$ for all t, with $\eta_t^{(l)}$ 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 $\eta_t^{(1)}$ and $\eta_t^{(2)}$ 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

$$\Pr\Big[s_t = j|\mathbf{y}_t; \theta\Big] = \Pr\Big[s_{1,t} = j_1\Big|\mathbf{v}_{1,t}, \mathbf{v}_{2,t}; \vartheta_1, \mathbf{P}^{(1)}\Big]\Pr\Big[s_{2,t} = j_2\Big|\mathbf{v}_{1,t-1}, \mathbf{v}_{2,t}; \vartheta_2, \mathbf{P}^{(2)}\Big],$$

where $\theta = (\vartheta_1, \vartheta_2, P)$ and $\vartheta_i = \{\mu_{i.s_{it}}, A_{ij.s_{it}}, \Sigma_{ii,s_{it}}\}$ for $i = 1, 2$ are the parameters for the density of $\epsilon_{it}|s_{it}$. When $M_2 \geq 2$ it also follows that $\Pr[s_{1,t} = j_1|\mathbf{v}_{1,t}, \mathbf{v}_{2,t}; \vartheta_1] = \Pr[s_{1,t} = j_1|\mathbf{v}_{1,t}, \mathbf{v}_{2,t-1}; \vartheta_1]$.

The final stage is now straightforward. $\mathbf{v}_{2,t}$ is assumed to be predictively redundant for $s_{1,t+1}$ and this regime process is not serially uncorrelated when (A2) has already been covered, it follows that $\mathbf{v}_{2,t}$ must not contain any information in addition to $\mathbf{v}_{1,t}$ for predicting $s_{1,t}$. This means that 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 $j_1 \in \{1, \ldots, M_1\}$ and $k \in \{1, \ldots, p\}$ and for $M_2 \geq 1$. Hence, we have shown that

$$\Pr\Big[(s_{1,t}, s_{2,t}) = (j_1, j_2)\Big|\mathbf{y}_t; \theta\Big] = \Pr\Big[s_{1,t} = j_1\Big|\mathbf{v}_{1,t}; \vartheta_1\Big] \Pr\Big[s_{2,t} = j_2\Big|\mathbf{y}_t; \vartheta_2\Big],$$

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,

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

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

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

The necessary and sufficient conditions for this expression to be invariant with respect to $\mathbf{y}_{4t}$ are, for all t, given by

$$\begin{aligned}
&\text{(i)} && \bar{m}_{1,t} = E\Big[m_{1,s_{t+1}}\Big|\mathbf{v}_{1t}, \mathbf{y}_{3t}; \theta\Big], && \\
&\text{(ii)} && \bar{a}^{(k)}_{1r,t} = E\Big[a^{(k)}_{1r,s_{t+1}}\Big|\mathbf{v}_{1t}, \mathbf{y}_{3t}; \theta\Big], && r \in \{1, \ldots, 4\} \text{ and } k \in \{1, \ldots, p\}, \\
&\text{(iii)} && \bar{a}^{(k)}_{14,t} = 0, && k \in \{1, \ldots, p\}.
\end{aligned}$$

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

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

From the definitions of $\bar{m}_{1,t}$ and $\bar{a}^{(k)}_{1r,t}$ in both of the equations (17) we find that these random matrices can be expressed as

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

and

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

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 Granger noncausality implies that

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

must hold for all i_1, i_2, and t, while $(m_{1,(j_1,j_2)}, a^{(k)}_{1r,(j_1 j_2)})$ only depends on j_2. By Lemma 3 and the proof of Proposition 1 (see also Warne, 2000, Corollary 2) we know that equation (A.31) can only be satisfied under (A1). The remaining parameter restrictions, $p_{ij} = p^{(1)}_{i_1 j_1} p^{(2)}_{i_2 j_2}$, 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 (A3) is satisfied.

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

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

Statistical Appendix: Block Metropolis-Hastings Algorithm for MS-VARs with restrictions

This section describes all the constituting blocks that form the MCMC sampler.

B.1. Simulating Hidden Markov Process

The first drawn parameter is the vector representing the states of the economy, $\mathbf{S}_T$. Being a latent variable, there are no prior distributions nor restrictions specified for $\mathbf{S}_T$. We first use a BLHK filter and smoother (see Section 11.2 of Frühwirth-Schnatter, 2006, and references therein) and obtain the probabilities $\Pr(s_t = i|\mathbf{y}_T, \theta^{(l-1)})$, for $t = 1, \ldots, T$ and $i = 1, \ldots, M$, and then draw $\mathbf{S}_T^{(l)}$, 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).

B.2. Sampling 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, $\mathbf{P}^{(l)} \sim p(\mathbf{P}|\mathbf{S}_T^{(l)})$. For the purpose of testing restriction (A2), 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 their paper 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 $\mathbf{S}_T$.

Reparametrization. The transition probabilities matrix $\mathbf{P}$ is modeled with Q vectors w_j, $j = 1, \cdots, Q$ and each of size $d_j \leq M$. 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:

$$p = \mathbf{M}\mathbf{w}, \tag{B.1}$$

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 $\mathbf{S}_T$ non-stationary. The lack of 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 step can be set up to circumvent this issue, since a kernel of joint posterior density of all rows is known: $p(\mathbf{P}|\mathbf{S}_T) \propto \prod_{j=1}^{Q} \mathcal{D}_{d_j}(w_j)\pi$. 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 (B.1). Finally, we compute the acceptance rate before retaining or discarding the draw.

Algorithm 1. *Metropolis-Hastings step 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 $\mathbf{S}_T$. The candidate transition probabilities matrix – in the transformed notation – are sampled from a Dirichlet distribution.
3. $\mathbf{P}^{new} = \mathbf{M}\mathbf{w}$. The proposal for the transitions probabilities matrix is reconstructed.
4. Accept $\mathbf{P}^{new}$ if $u \leq (\pi^{new}/\pi^{l-1})$, where $u \sim \mathcal{U}[0, 1]$. π^{new} and π^{l-1} are the ergodic probabilities of $s_0^{(l)}$ that are computed from $\mathbf{P}^{new}$ and $\mathbf{P}^{l-1}$ respectively.

B.3. Sampling Second and Independent Hidden Markov Process

Regime inference from proposition (A1) involves two independent Markov processes. Equation (13) 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 B.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, $\mathbf{P} = (\mathbf{P}^{(1)} \otimes \mathbf{P}^{(2)})$.

B.4. Sampling Correlation Coefficients and Standard Deviations

Adapting the approach proposed by Barnard et al. (2000) to Markov-switching models, we sample from the full conditional distribution of unrestricted and restricted covariance matrices. We thus decompose each covariance matrix of the MS-VAR process into a vector of standard deviations (σ_{s_t}) and a correlation matrix ($\mathbf{R}_{s_t}$) as in equation (24).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 $(\hat{\sigma}_{i.s_t} - 3\hat{\sigma}_{\hat{\sigma}_{i.s_t}}, \hat{\sigma}_{i.s_t} + 3\hat{\sigma}_{\hat{\sigma}_{i.s_t}})$ 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

$$p\Big(\sigma_i \Big| \mathbf{y}_T, \mathbf{S}_T, \mathbf{P}, \beta, \sigma_{-i}, \mathbf{R}\Big).$$

   This is done by evaluating a kernel on a grid of points, using the proportionality relation, with the likelihood function times the prior: $\sigma_i|\mathbf{y}_T, \mathbf{S}_T, \mathbf{P}, \beta, \sigma_{-i}, \mathbf{R} \propto p(\mathbf{y}_T|\mathbf{S}_T, \theta) \cdot p(\sigma_i)$. 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

$$p\Big(\sigma_{i.s_t} \Big| \mathbf{y}_T, \mathbf{S}_T, \mathbf{P}, \beta, \sigma_{-i.s_t}, \mathbf{R}\Big),$$

   evaluating a kernel thanks to the proportionality relation, with the likelihood function times the prior: $\sigma_{i.s_t}|\mathbf{y}_T, \mathbf{S}_T, \mathbf{P}, \beta, \sigma_{-i.s_t}, \mathbf{R} \propto p(\mathbf{y}_T|\mathbf{S}_T, \theta) \cdot p(\sigma_{i.s_t})$.

Algorithm 3. *Griddy-Gibbs for the correlations* The algorithm iterates on all the correlation parameters $\mathbf{R}_{i.s_t}$ for $i = 1, \ldots, (N-1)N/2$ 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 interval (a, b) into G grid points.

1. Depending on the restriction scheme, set correlation parameters to 0.
2. Regime-invariant correlations: Draw from the univariate density

$$p\Big(\mathbf{R}_i \Big| \mathbf{y}_T, \mathbf{S}_T, \mathbf{P}, \beta, \sigma, \mathbf{R}_{-i}\Big),$$

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

$$p\Big(\mathbf{R}_{i.s_t} \Big| \mathbf{y}_T, \mathbf{S}_T, \mathbf{P}, \beta, \sigma, \mathbf{R}_{-i.s_t}\Big),$$

   evaluating a kernel thanks to the proportionality relation, with the likelihood function times the prior: $\mathbf{R}_{i.s_t}|\mathbf{y}_T, \mathbf{S}_T, \mathbf{P}, \beta, \sigma, \mathbf{R}_{-\mathbf{R}_{i.s_t}} \propto p(\mathbf{y}_T|\mathbf{S}_T, \theta) \cdot p(\mathbf{R}_{i.s_t})$.

B.5. Sampling 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:

$$y_t = Z_t\beta_0 + Z_t D_{i.1}\beta_1 + \cdots + Z_t D_{i.M}\beta_M + \epsilon_t, \tag{B.2}$$

$$\epsilon_t \sim i.i.\mathcal{N}(\mathbf{0}, \Sigma_{s_t}). \tag{B.3}$$

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:

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

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. $\beta|\mathbf{y}_T, \mathbf{S}_T, \mathbf{P}, \sigma, \mathbf{R} \sim \mathcal{N}(\overline{\beta}, \overline{V_\beta})$. 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

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

B.6. Simulating Restrictions in the Form of Functions of the Parameters

Some of the restrictions for Granger noncausality presented in Section 3 will be in the form of complicated functions of parameters. Suppose some restriction is in the form:

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

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. In such a case, $\beta|\mathbf{y}_T, \mathbf{S}_T, \mathbf{P}, \mathbf{R}, \sigma$ is 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, $p(\beta|\mathbf{y}_T, \mathbf{S}_T, \mathbf{P}, \sigma, \mathbf{R})$, 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:

$$\alpha(\beta^{l-1}, \beta^{new}) = \min\left[\frac{p(\mathbf{y}_T|\mathbf{S}_T, \mathbf{P}, \beta^{new}, \sigma, \mathbf{R})p(\beta^{new})}{p(\mathbf{y}_T|\mathbf{S}_T, \mathbf{P}, \beta^{l-1}, \sigma, \mathbf{R})p(\beta^{l-1})}, 1\right]. \tag{B.4}$$

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 (B.4) 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 B.5.
2. Compute $\beta_i = g(\beta_{-i}, \sigma, \mathbf{R}, \mathbf{P})$.
3. Return the vector $(\beta_1, \ldots, \beta_{i-1}, \mathbf{g}(\beta_{-\mathbf{i}}, \sigma, \mathbf{R}, \mathbf{P}), \beta_{i+1}, \ldots, \beta_k)'$

Figures

Figure 1: Log-differenced monthly data on US money (M1) and income (industrial production) over the sample period 1959:1–2012:11.

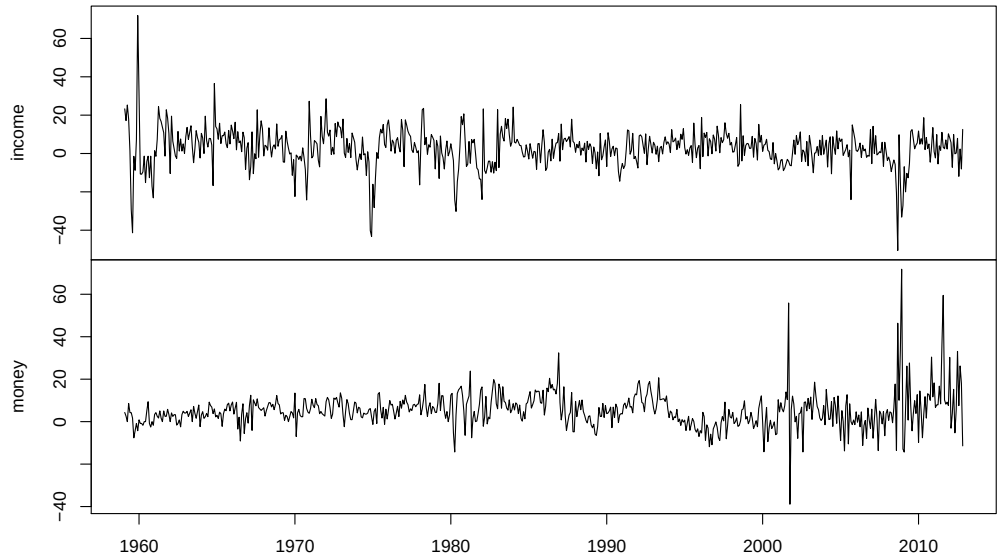
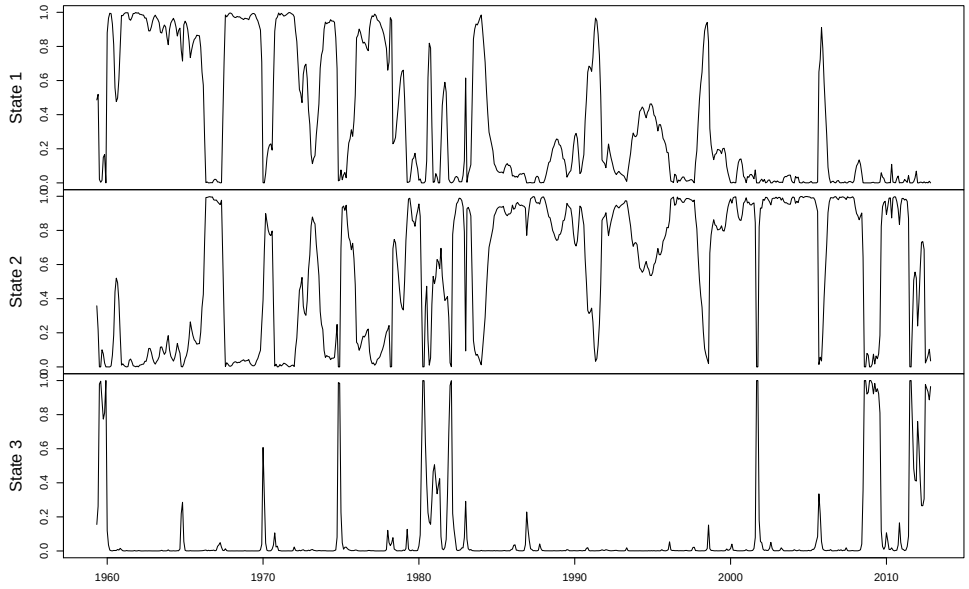


Figure 2: Estimated marginal posterior probabilities of regimes, $\Pr[s_t|\mathbf{y}_T]$, for the unrestricted MS-VAR model with 3 states and 3 lags over the sample period 1959:4–2012:11.



Tables

Table 1: Summary statistics of the data for the sample period 1959:1–2012:11.

Variable	Mean	Median	Standard Deviation	Minimum	Maximum
Δy	2.771	3.312	9.984	-50.553	71.977
Δm	5.29	4.764	8.279	-38.886	71.788

Note: Data Source: Citibase. Δy is the US industrial production index and Δm is the US M1 money stock. Both series are seasonally adjusted, transformed into log-returns and multiplied by 1200.

Table 2: Model selection for VAR(p) models over the sample period 1959:1–2012:11.

Lags (p)	0	1	2	3	4	5	6	7	8
$\ln p_{MHM}(\mathbf{y}_T\|p)$	-4739.81	-4654.11	-4642.09	-4613.66	-4616.61	-4609.95	-4592.88	-4585.46	-4582.27
Lags (p)	9	10	11	12	13	14	15	16	17
$\ln p_{MHM}(\mathbf{y}_T\|p)$	-4581.74	-4584.36	-4556.16	-4553.18	-4551.15	**-4544.68**	-4546.68	-4549.14	-4546.53

Note: $\ln p_{MHM}(\mathbf{y}_T|p)$ denotes the marginal data density using the modified harmonic mean estimator suggested by Geweke (1999, 2005) and computed for VAR models with different lag order, p.

Table 3: Bayesian Granger Noncausality Tests for VAR models with 14 lags over the sample 1959:1–2012:11.

$\mathcal{M}_j$	$\ln p_{MHM}(\mathbf{y}_T \mid \mathcal{M}_j)$	$\log_{10} \mathcal{B}_{j0}$
$\mathcal{H}_0$: *Unrestricted VAR(14) model*		
$\mathcal{M}_0$	-4544.68	0
$\mathcal{H}_1$: *Granger Noncausality from Money to Income*		
$\mathcal{M}_1$	-4518.43	11.4

Note: $\ln p_{MHM}(\mathbf{y}_T \mid \mathcal{M}_j)$ denotes the marginal data density using the modified harmonic mean estimator suggested by Geweke (1999, 2005) and computed for the jth model, $\mathcal{M}_j$, and $\log_{10} \mathcal{B}_{j0}$ denotes a logarithm of base 10 of the Bayes factor of the jth model to model $\mathcal{M}_0$. Model $\mathcal{M}_1$ is the restricted VAR(14) model with the restrictions for Granger noncausality hypothesis.

Table 4: Model selection for MS-VAR models with M states and p lags over the sample 1959:1–2012:11.

	Models with No. of states $M = 2$							
Lags (p)	0	1	2	3	4	5	6	7
$\ln p_{MHM}(\mathbf{y}_T \vert p, M)$	-4578.72	-4456.06	-4440.07	-4412.75	-4418.02	-4420.24	-4411.57	-4411.62
	Models with No. of states $M = 3$							
Lags (p)	0	1	2	3	4	5	6	7
$\ln p_{MHM}(\mathbf{y}_T \vert p, M)$	-4567.22	-4415.06	-4402.98	**-4384.28**	-4390.89	-4392.09	-4387.6	-4390.79

Note: $\ln p_{MHM}(\mathbf{y}_T|p, M)$ denotes the marginal data density using the modified harmonic mean estimator suggested by Geweke (1999, 2005) and computed for MS-VAR models with different number of states, M, and lag order, p.

Table 5: Estimation results for an MS-VAR model with 3 regimes and 3 lags over the sample 1959:1–2012:11.

	State 1											
	μ_1	$A_1^{(1)}$		$A_1^{(2)}$		$A_1^{(3)}$		σ_1	$\mathbf{R}_{1.12}$	p_{11}	p_{12}	p_{13}
Δy	0.365	0.327	-0.322	-0.044	0.535	0.031	0.350	8.571	0.207	0.907	0.072	0.021
	(0.509)	(0.069)	(0.235)	(0.069)	(0.257)	(0.061)	(0.182)	(0.526)	(0.09)	(0.032)	(0.032)	(0.014)
Δm	0.892	0.036	0.483	-0.013	0.127	-0.007	0.152	2.880				
	(0.391)	(0.025)	(0.090)	(0.023)	(0.086)	(0.022)	(0.069)	(0.223)				
	State 2											
	μ_2	$A_2^{(1)}$		$A_2^{(2)}$		$A_2^{(3)}$		σ_2	$\mathbf{R}_{2.12}$	p_{21}	p_{22}	p_{23}
Δy	0.431	0.045	0.086	0.257	0.023	0.233	-0.021	5.849	-0.059	0.048	0.925	0.027
	(0.370)	(0.058)	(0.048)	(0.047)	(0.045)	(0.054)	(0.042)	(0.238)	(0.067)	(0.019)	(0.022)	(0.012)
Δm	1.489	0.028	0.288	-0.093	0.114	-0.071	0.268	6.297				
	(0.393)	(0.059)	(0.057)	(0.052)	(0.052)	(0.049)	(0.051)	(0.349)				
	State 3											
	μ_3	$A_3^{(1)}$		$A_3^{(2)}$		$A_3^{(3)}$		σ_3	$\mathbf{R}_{3.12}$	p_{31}	p_{32}	p_{33}
Δy	-0.028	0.473	0.043	-0.074	-0.088	-0.016	-0.131	20.149	-0.14	0.064	0.164	0.772
	(0.543)	(0.202)	(0.147)	(0.231)	(0.178)	(0.209)	(0.190)	(1.818)	(0.141)	(0.044)	(0.070)	(0.078)
Δm	0.163	0.030	0.107	0.136	0.120	-0.375	0.249	20.753				
	(0.544)	(0.192)	(0.152)	(0.236)	(0.177)	(0.215)	(0.203)	(1.803)				

Note: The Table reports posterior means (standard deviations) of the parameters of the unrestricted MS(3)-VAR(3) model, $\mathcal{M}_0$. μ_m denotes a constant term from state m in the conditional mean equation, $A_m^{(k)}$ is the matrix of autoregressive coefficients at lag $k \in \{1,2,3\}$ of the VAR part in state m, σ_m is the standard deviation in state m of the error term, $\mathbf{R}_{m.12}$ is the correlation coefficient in state m of the error term (for $m \in \{1,2,3\}$), and p_{mn} are the elements of the transition probability matrix $\mathbf{P}$ for $m, n \in \{1,2,3\}$.

Table 6: Summary of regime inference and Granger noncausality restrictions on the parameters of MS-VAR models with $M = 3$.

	Condition (A1)	# restrictions
$\mathcal{M}_1$:	$\mu_{1.s_t} = \mu_1$, $A^{(l)}_{11.s_t} = A^{(l)}_{11}$, $A^{(l)}_{12.s_t} = 0$, for $l = 1, \ldots, p$, and $\Sigma_{11.s_t} = \Sigma_{11}$, $\Sigma_{12.s_t} = 0$	$5p + 7$
$\mathcal{M}_2$:	$\mu_{2.s_t} = \mu_2$, $A^{(l)}_{21.s_t} = A^{(l)}_{21}$, $A^{(l)}_{22.s_t} = A^{(l)}_{22}$, $A^{(l)}_{12.s_t} = 0$, for $l = 1, \ldots, p$, and $\Sigma_{22.s_t} = \Sigma_{22}$, $\Sigma_{12.s_t} = 0$	$7p + 7$
	Condition (A2)	
$\mathcal{M}_3$:	$\mathbf{P} = \iota_3 \pi'$	4
	Condition (A3)	
$\mathcal{M}_4$:	$\mathbf{P} = \iota_3 \pi'$, and $\sum_{j=1}^{3} A^{(l)}_{12.j} \pi_j = 0$, for $l = 1, \ldots, p$	$p + 4$
$\mathcal{M}_5$:	$\mathbf{P} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ c & 1-c \end{bmatrix} \begin{bmatrix} p_{1\cdot} \\ p_{2\cdot} \end{bmatrix}$, $\sum_{j=1}^{3} \mu_{1.j}\left(p_{1j} - p_{2j}\right) = 0$, $\sum_{j=1}^{3} A^{(l)}_{11.j}\left(p_{1j} - p_{2j}\right) = 0$, and $\sum_{j=1}^{3} A^{(l)}_{12.j} p_{ij} = 0$ for $i = 1, 2$ and $l = 1, \ldots, p$	$3p + 2$
$\mathcal{M}_6$:	$\mu_{1.s_t} = \mu_1$, $A^{(l)}_{11.s_t} = A^{(l)}_{11}$, and $A^{(l)}_{12.s_t} = 0$, for $l = 1, \ldots, p$	$5p + 2$

Note: The parameters are as defined in the note of Table 5. Additinally, ι_n is a n-dimensional vector of ones, $p_{i\cdot}$ is the ith row of $\mathbf{P}$, π is a vector of ergodic state probabilities, c is a parameter estimated by maximizing the value of the full conditional posterior density of $\mathbf{P}$.

Table 7: Noncausality and regime inference testing in a MS-VAR models for US monthly data on money and income, 1959:1–2012:11.

$\mathcal{M}_j$	Restrictions	$\ln p_{MHM}\left(\mathbf{y_T}\vert\mathcal{M}_j\right)$	$\log_{10}\mathcal{B}_{j0}$
$\mathcal{H}_0$: *Unrestricted model*			
$\mathcal{M}_0$	MS(3)-VAR(3)	-4384.28	0
$\mathcal{H}_1$: *History of money has no effect on the regime forecast*			
$\mathcal{M}_1$	(A1) with $M_1 = 1, M_2 = 3$	-4441.32	-24.77
$\mathcal{M}_2$	(A1) with $M_1 = 3, M_2 = 1$	-4556.22	-74.68
	(A1)	-4442.01	-25.07
$\mathcal{M}_3$	(A2)	-4422.07	-16.41
$\mathcal{H}_2$: *Granger noncausality*			
$\mathcal{M}_2$	(A1) and $M_1 = 3, M_2 = 1$	-4556.22	-74.68
$\mathcal{M}_4$	(A3) and rank($\mathbf{P}$) = 1	-4430.35	-20.01
$\mathcal{M}_5$	(A3) and rank($\mathbf{P}$) = 2	-4488.47	-45.25
$\mathcal{M}_6$	(A3) and rank($\mathbf{P}$) = 3	-4391.57	-3.17
	(A3)	-4392.67	-3.64

Note: For the definition of $\ln p_{MHM}\left(\mathbf{y_T}\vert\mathcal{M}_j\right)$ and $\log_{10}\mathcal{B}_{j0}$ see the note to Table 3. For the exact restrictions on parameters for the restricted models see Table 6.

Table 8: Summary of Bayesian hypotheses testing on regime inference and Granger noncausality for US monthly data on money and income, 1959:1–2012:11.

$\mathcal{H}_i$	Hypothesis	Represented by models	$\ln p_{MHM}(\mathbf{y}_T \mid \mathcal{H}_i)$	$\log_{10} \mathcal{B}_{j0}$
$\mathcal{H}_0$	Unrestricted model	$\mathcal{M}_0$	-4384.28	0
$\mathcal{H}_1$	History of money does not impact on the regime forecast of income	$\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$	-4423.17	-16.89
$\mathcal{H}_2$	Granger noncausality	$\mathcal{M}_2$, $\mathcal{M}_4$, $\mathcal{M}_5$, $\mathcal{M}_6$	-4392.96	-3.77

Note: For the definition of $\ln p_{MHM}\left(\mathbf{y_T} \mid \mathcal{M}_j\right)$ and $\log_{10} \mathcal{B}_{j0}$ see the note to Table 3.

Table 9: Bayesian hypothesis testing of Granger noncausality in mixture VAR models for US monthly data on money and income, 1959:1–2012:11.

$\mathcal{M}_j$	Restrictions	$\ln p(\mathbf{y}_T \vert \mathcal{M}_j)$	$\log_{10} \mathcal{B}_{j0}$
$\mathcal{H}_0$: *Unrestricted model*			
$\mathcal{M}_3$	mix(3)-VAR(3)	-4422.07	0
$\mathcal{H}_1$: *Granger noncausality from money to income*			
$\mathcal{M}_4$	(A3) rank($\mathbf{P}$) = 1	-4430.35	-3.60

Note: For the definition of $\ln p_{MHM}\left(\mathbf{y_T} \vert \mathcal{M}_j\right)$ and $\log_{10} \mathcal{B}_{j0}$ see the note to Table 3. For the exact restrictions on parameters for the restricted models see Table 6.