Graduate School of Decision Sciences

Universität Konstanz

GSDS
Working Paper
No. 2017-15

Directed Graphs and Variable Selection in Large Vector Autoregressive Models

Ralf Brüggemann
Christian Kascha

August 2017

Graduate School of Decision Sciences

All processes within our society are based on decisions – whether they are individual or collective decisions. Understanding how these decisions are made will provide the tools with which we can address the root causes of social science issues.

The GSDS offers an open and communicative academic environment for doctoral researchers who deal with issues of decision making and their application to important social science problems. It combines the perspectives of the various social science disciplines for a comprehensive understanding of human decision behavior and its economic and political consequences.

The GSDS primarily focuses on economics, political science and psychology, but also encompasses the complementary disciplines computer science, sociology and statistics. The GSDS is structured around four interdisciplinary research areas: (A) Behavioural Decision Making, (B) Intertemporal Choice and Markets, (C) Political Decisions and Institutions and (D) Information Processing and Statistical Analysis.

GSDS – Graduate School of Decision Sciences
University of Konstanz
Box 146
78457 Konstanz

Phone: +49 (0)7531 88 3633
Fax: +49 (0)7531 88 5193
E-mail: gsds.office@uni-konstanz.de

-gsds.uni-konstanz.de

ISSN: 2365-4120
August 2017

Directed Graphs and Variable Selection in Large Vector Autoregressive Models*

Ralf Brüggemann[a] Christian Kascha[b]

first draft: July 17, 2017
this version: August 8, 2017

Abstract

We represent the dynamic relation among variables in vector autoregressive (VAR) models as directed graphs. Based on these graphs, we identify so-called strongly connected components (SCCs). Using this graphical representation, we consider the problem of variable selection. We use the relations among the strongly connected components to select variables that need to be included in a VAR if interest is in forecasting or impulse response analysis of a given set of variables. We show that the set of selected variables from the graphical method coincides with the set of variables that is multi-step causal for the variables of interest by relating the paths in the graph to the coefficients of the 'direct' VAR representation. Empirical applications illustrate the usefulness of the suggested approach: Including the selected variables into a small US monetary VAR is useful for impulse response analysis as it avoids the well-known 'price-puzzle'. We also find that including the selected variables into VARs typically improves forecasting accuracy at short horizons.

Keywords: Vector autoregression, Variable selection, Directed graphs, Multi-step causality, Forecasting, Impulse response analysis

JEL classification: C32, C51, C55, E52

*We thank participants of the SFB 649 final colloquium at Humboldt-University Berlin and University of Kiel Econometrics Seminar for useful comments on earlier versions of this paper. Part of this research has been conducted while the second author was a postdoctoral research at the Department of Economics at the University of Zurich, Switzerland.

[a]Corresponding author: Ralf Brüggemann: University of Konstanz, Department of Economics, Box 129, 78457 Konstanz, Germany, email: ralf.brueggemann@uni-konstanz.de

[b]christian.kascha@posteo.de

1 Introduction

Vector autoregressive (VAR) models are popular tools in the analysis of multiple time series. Typical applications include forecasting or impulse response analysis. The popularity of the VAR model is at least partly due to the fact that it typically does not require strong economic theory assumptions. Often VAR models without any restrictions on the parameters are used to describe the joint dynamics of a set of economic time series.

While the general VAR lag structure allows to uncover dynamic relations between the variables included in the system, the use of unrestricted VARs comes at a cost: The number of parameters to be estimated from the data increases with the square of the number of variables in the system. Even in moderately large VARs the degrees of freedom exhaust quickly. Thus applied researchers have to choose the number of variables to be included in the VAR wisely. On the one hand, a researcher would like to include all relevant variables to avoid omitted variable bias and to get a complete picture of the underlying dynamics. On the other hand, including too many variables makes parameter estimates unreliable and estimation uncertainty may lead to rather uninformative results such as estimated impulse responses with very wide confidence intervals.

Given variables of interest, our paper suggests to use a graphical modeling approach in order to select a 'minimal' VAR containing only variables that are relevant for predicting the variables of interest. This approach is helpful in selecting the relevant variables for VAR analysis in a data-driven way. We argue that this is a useful addition to the toolbox of time series econometricians as on the one hand, it exploits the information from large dimensional data sets but on the other hand eventually uses smaller VAR models for forecasting and structural analysis.

To fix ideas, suppose a researcher is interested in a set of variables denoted by y^{I}, including say GDP growth, the consumer price (CPI) inflation, and a key interest rate. She either wants to forecasts the variables in y^{I} or to conduct an impulse response analysis for the variables in y^{I}. For this purpose, typically a large cross-section of time series on e.g. output, income, consumption, the labor market, orders and inventories, money and credit, interest and exchange rates, financial market variables and various price measures, is available.[1]

In recent years, suggestions have been made on how to include the information from a large dimensional data set into VARs. Factor-augmented VARs (FAVARs) (see e.g. Bernanke, Boivin & Eliasz (2005) and Stock & Watson (2016)) condense the information from a large time series data set into a few factor time series, which are then included in a VAR model. Factor-augmented models have been used for forecasting and structural analysis.[2] Clearly, these models are only suitable if the underlying data has a factor structure, i.e. if the large number of time series are really driven by a small number of common factors (see e.g. Uhlig (2009) on this point). Other potential problems in empirical work on FAVARs include the

[1] Large data sets of this type have been used in various studies. See e.g. Stock & Watson (2003) or McCracken & Ng (2015) with references therein. They typically contain up to 130 variables.

[2] See e.g. Stock & Watson (2002), Ludvigson & Ng (2007), Eickmeier & Ziegler (2008), Ludvigson & Ng (2009), Stock & Watson (2012), Clements (2016) and Cheng & Hansen (2015) for applications of factor-augmented regressions and FAVARs.

identification of factors, determination of the number of factors and cross-correlation among the idiosyncratic disturbances. An alternative are large Bayesian VARs (BVARs) as suggested by Banbura, Giannone & Reichlin (2010).[3] In large dimensional settings, however, these models require to use a very tight prior. Consequently, using a large BVAR might impose more structure on the model than typical VAR users feel comfortable with. Other shrinkage methods[4] have also been used for large VAR models, including the least absolute shrinkage and selection operator (LASSO) (see e.g. Kascha & Trenkler (2015)). The LASSO approach can handle large dimensional VAR models by setting some VAR coefficients to zero and at the same time shrinking the remaining coefficients. While already frequently used in applied work, the theoretical underpinning is still developing and it is not entirely clear how to conduct inference based on LASSO-VAR models.

Consequently, using FAVARs or large BVARs may not be ideal in some situations faced by applied time series econometricians. Researchers may actually prefer to use smaller VARs also because they might be easier to interpret and resemble more closely small scale dynamic stochastic general equilibrium (DSGE) models used in macroeconomics. At the same time, the researcher would like to include additional 'relevant' variables that affect predictions and/or impulse responses for the variables of interest. Against the background of the large number of time series available today, this entails a variable selection procedure.

Our paper is concerned with the question of how to choose variables for the smallest ('minimal') VAR that contains all variables that are 'relevant' in forecasting the variables of interest y^{I}. For structural analysis, our paper addresses the question: What is the smallest VAR containing all variables 'relevant' for impulse responses of the variables of interest y^{I}? To address our research questions, we develop a variable selection strategy based on a graphical modeling approach. The first contribution of our paper is to use so-called strongly connected components (SSCs) and the relation among these components for variable selection. We first represent a sparse VAR structure as a directed graph with vertices and edges. From this graph we identify all SCCs in our VAR model by using a simple graphical modeling algorithm. The concept of SSCs is well-established in the graphical modeling literature but to the best of our knowledge, this concept has not been used in econometrics. We show how the SCCs and their connections to other SCCs are helpful in identifying the set of 'relevant' variables. Effectively, the set of relevant variables can be found from the graphical representation of the SCCs, known as a component graph. In a second theoretical contribution, we show the relation between the SCCs and the concept of multi-step causality as in Dufour & Renault (1998). In particular, given the variables of interest y^{I}, we show that a minimal VAR chosen by SCCs is identical to the VAR that contains y^{I} and all variables that are multi-step causal for y^{I}.

Methodologically, our paper is related to the literature on using graphical models in econometrics. Following the work on causal analysis of multivariate data (see e.g. Lauritzen (1996), Pearl (2000) and Edwards (2000)), graphical models have also been introduced for time series models. Brillinger (1996) and Dahlhaus (2000) are the first papers mentioning the use

[3] See e.g. Carriero, Kapetanios & Marcellino (2009, 2012, 2015), Giannone, Lenza, Momferatou & Onorante (2014), and Koop (2013) for applications of this method.

[4] See also Stock & Watson (2012).

of graphical modeling for time series data and present concepts based on the partial correlation and partial spectral coherence.[5] Dahlhaus & Eichler (2003) introduce causality graphs based on the autoregressive representation. Our work is most closely related to the work of Eichler (2006, 2007, 2012), who shows the close relation of different causality concepts (Granger-causality and multi-step causality) to graphical representations in vector autoregressive models. We add to this work the link from causality structures to variable selection using the concept of strongly connected components.[6]

Our paper is also closely related to the work of Jarociński & Maćkowiak (2017), who investigate the same research question of variable choice for VAR analysis, albeit with a different econometric approach. Based on the concept of Granger-causal priority (see Sims (1982, 2015), Doan & Todd (2010)), their paper evaluates in a Bayesian setup the posterior probability of Granger-causal priority. As pointed out by Jarociński & Maćkowiak (2017), Granger-causal priority is a sufficient condition for Granger-noncausality at all horizons and thus also related to Dufour & Renault (1998). For a given set of variables of interest y^{I}, Jarociński & Maćkowiak (2017) would drop a variable y_j, say, if the variables in y^{I} are likely to be Granger-causal prior to y_j. Thus, their method may also be used to choose variables for VAR analysis.

We illustrate the usefulness of our suggested variable selection in two applications to US macroeconomic data. The first application is similar to the one in Jarociński & Maćkowiak (2017). The variables of interest in y^{I} are US output, CPI inflation and the federal funds rate, three variables often used in stylized three-variable VARs for the US. Given y^{I}, we use our variable selection method based on SCCs to select a minimal VAR from 41 US time series for a period between 1979 and 2014. Starting point is a sparse VAR structure obtained from applying the LASSO to the large VAR. A number of interesting results emerge: First, regardless of the considered estimation period, 10 out of the 41 variables are always selected into the model and the selection is fairly stable over different samples before the financial crisis in 2008. Second, the set of selected variables is remarkably similar to the one obtained by Jarociński & Maćkowiak (2017) from their Bayesian analysis. Third, additional variables are selected into the model in a number of periods. Consequently, the 'minimal VAR' is still relatively large, indicating that the underlying relations are typically quite complex and may not be captured adequately in a three-variable VAR. We also find that including the selected variables into the VAR leads to more reasonable responses to a monetary policy shock, indicating that the selection is useful. Finally, we also conduct pseudo-out-of-sample forecasting experiments. Our results indicate that VARs with only the selected variables outperform both, the baseline VAR and a large VAR with all variables included. Similar results are obtained in the second empirical application, where selecting 'minimal VARs' leads to better forecasts in US output growth and unemployment.

The remainder of the paper is structured as follows. Section 2 shows how VARs can

[5] See also Flamm, Kalliauer, Deistler, Waser & Graef (2012) for an overview of different approaches.

[6] Graphical modelling has also been used for identifying the instantaneous relations. The first work in this area is the paper by Swanson & Granger (1997), followed by a number of studies that use graphical modeling for identifying structural VAR models (see e.g. Demiralp & Hoover (2003), Hoover, Demiralp & Perez (2009) and Heinlein & Krolzig (2012)).

be represented as directed graphs. We also introduce the concept of strongly connected components and explain how this can be used for variable selection and for finding a 'minimal VAR'. Section 3 relates the graph-theoretical concepts to multi-step causality and shows how variable selection based on both concepts leads to the same set of relevant variables. In Section 4, we illustrate the usefulness of our method in empirical applications. Section 5 concludes.

2 Vector Autoregressive Models, Directed Graphs and Strongly Connected Components

In this section, we explain how VAR models can be represented by directed graphs. We then review the concept of strongly connected components (SCCs) in directed graphs and explain how SCCs can be used for selecting relevant variables.

We denote the VAR model of order p, a VAR(p) for the K-dimensional time series vector $y_t = (y_{1,t}, y_{2,t}, \ldots, y_{K,t})'$ by

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t, \tag{2.1}$$

where $A_1, \ldots, A_p$ are $K \times K$ coefficient matrices and u_t is a zero mean white noise disturbance vector with non-singular covariance matrix Σ_u. We have not included deterministic terms (e.g. intercepts) into the VAR in (2.1) to simplify the notation. Adding deterministic terms would not change the results discussed below and they can be included in empirical work. The VAR(p) can be written in VAR(1) companion form as

$$Y_t = \mathbf{A} Y_{t-1} + U_t, \tag{2.2}$$

where $Y_t = (y_t', y_{t-1}', \ldots, y_{t-p+1}')'$ and $U_t = (u_t', 0, \ldots, 0)'$ are $Kp \times 1$ vectors and

$$\mathbf{A} := \begin{pmatrix} A_1 & A_2 & A_3 & \ldots & A_p \\ I_K & 0 & 0 & \ldots & 0 \\ 0 & I_K & 0 & \ldots & 0 \\ 0 & 0 & \ddots & & \vdots \\ 0 & 0 & \ldots & I_K & 0 \end{pmatrix}. \tag{2.3}$$

To make use of graphical modeling concepts, we represent the VAR in (2.2) as a directed graph. Following the standards in graphical modeling, a directed graph G is described by a set of vertices V and a set of edges E that are ordered pairs of vertices. In our application the vertices correspond to the Kp elements in the vector Y_t and the edges are determined by the elements of the companion form matrix $\mathbf{A}$ as in the following definition:

DEFINITION 1 (Directed VAR Graph) *Given a VAR model as in* (2.1)*, the associated directed graph is the graph* $G = (V, E)$ *with* $V = \{1, 2, 3, \ldots, Kp\}$ *and*

$$(i, j) \in E \quad \Leftrightarrow \quad \mathbf{a}_{ij} \neq 0 \text{ and } i \neq j, \tag{2.4}$$

where $\mathbf{a}_{ij}$ *is the* (ij)*th element of the companion matrix* $\mathbf{A}$*.*

Remark 1 In this graph, a directed edge (i, j) leads from vertex i to vertex j. This is standard in the graphical modeling literature. In our context, using this definition $(i, j) \in E$ implies that the ith variable *depends on* the jth component in Y_t, however, the arrow would point from vertex i to vertex j. Thus, the direction of the arrows is reversed compared to the type of arrows sometimes used to denote (Granger-)causality.

Remark 2 We exclude self-loops since we focus on the relationship between different variables in the VAR as in e.g. Granger-causality analysis. Introducing self-loops would only clutter the resulting graph without any additional insights. Note that arrows from a variable to itself such as $(1, 1)$ are not included. That is, the graph only describes relations between different variables.

Remark 3 Different definitions of directed graphs for VAR models are possible. E.g. Eichler (2007) uses a different definition for VAR(p) models. In particular, Eicher works directly on the matrices $A_i, i = 1, \ldots, p$ and defines

$$(i, j) \notin E \quad \text{if} \quad A_{ij,s} = 0, \forall s = 1, \ldots p,$$

where $A_{ij,s}$ denotes the ijth element of A_s. Obviously, for $p > 1$ this definition would yield a different graph compared to the obtained in Definition 1.

Given a sparse VAR structure, i.e. a VAR with zero restrictions on the VAR coefficients, we may use the associated directed graph to learn about the set of relevant variables. We do so by first defining variables that are included in the set of strongly connected components. To define a strongly connected component, we make use of the concept of a path or a pathway. A path is defined to be a sequence of vertices to go from one vertex to another. More formally, a path P of length k leading from vertex u to u' in graph $G = (V, E)$ is a sequence $P = (v_0, v_1, \ldots, v_k)$ of vertices such that $u = v_0$ and $u' = v_k$ and $(v_{i-1}, v_i) \in E$ for $i = 1, 2, \ldots, k$. If there is a path from u to u', we say that u' is reachable from u, denoted as $u \overset{P}{\rightsquigarrow} u'$. We may now define the strongly connected components of a directed graph:

DEFINITION 2 (Strongly Connected Components (SCCs)) *A strongly connected component (SCC) of a directed graph $G = (V, E)$ is the maximal set of vertices $C \subset V$ such that for every pair of vertices u and v in C, we have $u \overset{P}{\rightsquigarrow} v$ and $v \overset{P}{\rightsquigarrow} u$, i.e. u and v are mutually reachable (see e.g. Tarjan (1972)).*

Remark 4 Note that each vertex of the graph G belongs to exactly one strongly connected component. Consequently, the set of all strongly connected components $C_1, \ldots, C_k$ forms a partition of the graph G such that $V = C_1 \cup \ldots \cup C_k$ (see e.g. Duff & Reid (1978)).

Remark 5 Following Tarjan (1972), a depth-first search algorithm may be used to compute the strongly connected components efficiently. We have implemented a variant of Tarjan's algorithm using Matlab.

Remark 6 Duff & Reid (1978) suggest to order the SCCs such that there is no path from one strongly connected component to another later in the sequence, i.e. the SCCs $C_1, \ldots, C_k$ may be ordered such that there is no path from C_i to any C_j for $j > i$. The associated reordered matrix of the graph is then lower block-triangular. Each block corresponds to one of the SCCs. In the context of our economic applications, the structure of the SCCs may give additional insights on the relevance of different variables.

In the next step, we condense the information of the graph G by moving from graph G to a graph of the SCCs. The resulting graph is called a component graph, which is defined next.

DEFINITION 3 (Component Graph) *A component graph is defined as* $G^{SCC} = (V^{SCC}, E^{SCC})$, *where* $V^{SCC} = \{v_1, \ldots, v_k\}$ *contains a vertex* v_i *for each strongly connected component* $C_i, i = 1, \ldots, k$ *of* G. *There is an edge* $(v_i, v_j) \in E^{SCC}$ *if* G *contains a directed edge* (x, y) *for* $x \in C_i$ *and* $y \in C_j$.

Remark 7 The definition implies that there is only an edge (v_i, v_j) between two strongly connected components C_i and C_j if the original graph G has a directed edge from one member of the SCC C_i to a member of the SCC C_j.

Remark 8 The component graph may be viewed as a condensed view of the original graph. Essentially the component graph collapses all edges of the original graph whose incident are with the same SCC.

Finally, for a given set of variables of interest, say $y^{\mathrm{I}} \subset y$, we would like to identify a 'minimal' set of variables which have to be taken into account when modeling y^{I}. We denote this set of relevant variables as $R(y^{\mathrm{I}})$. Furthermore, we define the set $R_{G^{SCC}}(C_i)$ as the set of all SCCs that are *reachable* from C_i in G^{SCC} (including C_i). That is, $R_{G^{SCC}}(C_i)$ is the set of SCCs to which there is a path from C_i and it can be interpreted as the set of SCCs on which C_i 'depends' and which have to be taken into consideration when forecasting variables in C_i. We now define the 'minimal' VAR system in the following definition.

DEFINITION 4 *Given a subset of interest* $y^I \subset y$, *the minimal VAR is a VAR composed of the series that are contained in the relevant SCCs given by*

$$R(y^I) := \bigcup_{\{C_i \,:\, y^I \cap C_i \neq \emptyset\}} R_{G^{SCC}}(C_i).$$

Remark 9 Definition 4 states that the minimal VAR is the one composed of the series in the SCCs which contain elements of y^I and all SCCs that may be reached from these SCCs.

We illustrate the graph theoretical concepts using a simple example, starting with a four-dimensional VAR(1) with coefficient matrix as shown on the left side in Figure 1. The associated directed graph indicates that this system has two strongly connected components. The first one C_1 consists of variables (vertices) 1 and 3, $C_1 = \{1, 3\}$, as vertex 1 may be

$$\begin{pmatrix} a_{11} & 0 & a_{13} & 0 \\ a_{21} & 0 & 0 & a_{24} \\ a_{31} & 0 & a_{33} & 0 \\ 0 & a_{42} & a_{43} & a_{44} \end{pmatrix}$$

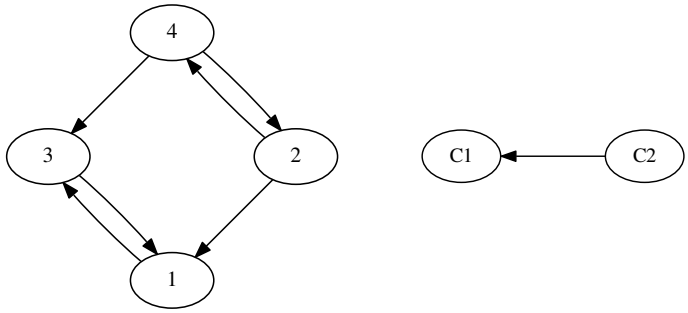


Figure 1: VAR matrix, associated directed graph, and component graph

reached from vertex 3 and vice versa. The second SCC C_2 consists of vertices 2 and 4, $C_2 = \{2, 4\}$. Note that the vertices within each strongly connected components are mutually reachable (compare Definition 2) and each variable (vertex) is in exactly one SCC (see Remark 4).

We may also illustrate Remark 6 as it is easy to see that we may reorder the variables such that a lower block-triangular matrix results. For this purpose, we order the SCC first that has no leaving edges. In our example, the SCC $C_1 = \{1, 3\}$ has no leaving edges and hence is ordered first. The new ordering of variables is then (1, 3, 2, 4), which results in the following reordered VAR matrix

$$A^* = \left(\begin{array}{cc|cc} a_{11} & a_{13} & 0 & 0 \\ a_{31} & a_{33} & 0 & 0 \\ \hline a_{21} & 0 & 0 & a_{24} \\ 0 & a_{43} & a_{42} & a_{44} \end{array}\right) = \left(\begin{array}{c|c} A_{11} & 0 \\ \hline A_{21} & A_{22} \end{array}\right),$$

where the coefficient within the matrix A^* still have their original names. Obviously, this matrix has the desired block-triangular form. After having grouped the variables according to the strongly connected components, we may now draw a corresponding component graph according to the partition of the matrix A^*. The resulting component graph is shown in the right panel of Figure 1. This component graph indicates that component C_1 may be reached from component C_2 but not vice versa. Consequently, if the variable(s) of interest are included in component C_1, then the minimal VAR only includes the variables contained in C_1 but not those contained in C_2. In contrast, if the variable(s) of interest are contained in C_2, a corresponding VAR needs to include the variables from C_2 and in addition, the variables from C_1 as C_1 may be reached from C_2. For instance, if the variable of interest is e.g. variable 3, then the VAR needed in forecasting or structural analysis needs to contain all variables that are in the corresponding component C_1. In our example, these are the variables 1 and 3. In contrast, if the variable of interest is variable 2, then we need to include all variables in C_2 and C_1, i.e. all four variables, in the VAR model. Similarly, we may find the minimal VAR from the component graph even if we have more than one variable of interest. In our example, if variable 1 and 3 are of interest, a VAR for just these two variables suffices as both variables form a strongly connected component (C_1) and no other strongly connected

component can be reached from C_1. In contrast, if variables 2 and 4 are of interest, we need a VAR with all variables from C_1 and C_2 as C_1 may be reached from C_2. In other words, we need all four variables. Now assume that variables 1 and 2 are of interest. Then again, we need to consider all variables from the strong components that include variable 1 and variable 2, which in our example boils down to again using all variables since there are only the two components C_1 and C_2.

3 Econometric Causality Concepts and Graphs

The graph theoretical concepts discussed in Section 2 have a close relation to multi-step causality concepts in time series econometrics. In this section we explain how the two concepts are related and show that the set of variables selected for a minimal VAR by the SCC method as in Definition 4 of Section 2 coincides with the set of variables that are multi-step causal for at least one of the variables of interest.

The simple notion of Granger causality (see Granger (1969)) is known to neglect any indirect effects and influences of 'auxiliary' variables as it is based on 1-step ahead predictability. Consequently, the original definition of Granger non-causality is not helpful in the context of variable selection. A more general causality concept that takes into account all indirect effects of auxiliary variables is known as multi-step causality and has been formally introduced into the literature by Dufour & Renault (1998).[7] Informally, a subset y^A of the variables causes another subset y^B at a specific horizon h if the best linear forecast for y^B at horizon h can be improved by including the variables in y^A in the information set. Dufour & Renault (1998) discuss necessary and sufficient conditions for non-causality at different forecast horizons h. Dufour, Pelletier & Renault (2006) focus on developing corresponding multi-step non-causality tests in the context of VAR models. For our purpose, it is convenient to note that multi-step non-causality at different horizons may be formulated as linear exclusion restrictions on the so-called direct VAR model. For $h \geq 1$, we write this direct VAR model as

$$y_{t+h} = \Pi_1^{(h)} y_t + \cdots + \Pi_p^{(h)} y_{t-p+1} + u_{t+h}^{(h)}, \tag{3.1}$$

where this representation is obtained by successive substitution from the VAR in (2.1). Dufour & Renault (1998) show that $\Pi_1^{(0)} = I_K$, $\Pi_s^{(1)} = A_s$, $\Pi_s^{(h+1)} = A_{s+h} + \sum_{l=1}^{h} A_{h-l+1}\Pi_s^{(l)} = \Pi_{s+1}^{(h)} + \Pi_1^{(h)} A_s$ and the MA$(h-1)$ innovation term $u_{t+h}^{(h)} = \sum_{j=0}^{h-1} \Pi_1^{(h)} u_{t-j}$.

Given sets of indices A, B, let

$$\Pi_{AB,s}^{(h)} \tag{3.2}$$

denote the submatrix of $\Pi_s^{(h)}$ consisting of the intersection of rows with indices in A and columns with indices in B. If A and B are singletons, say $A = \{k\}$, $B = \{l\}$, we simply write $\Pi_{kl,s}^{(h)}$. We reproduce Theorem 3.1 of Dufour & Renault (1998) tailored to the regular, finite VAR case.

[7] The effect of intermediate variables have also been pointed out earlier by e.g. Sims (1980), Penm & Terrell (1986) and Lütkepohl (1993) but Dufour & Renault (1998) were the first who formalized the concept of multi-step causality in a general framework.

THEOREM 1 (Dufour-Renault (1998)) *Given $y^A, y^B \subset y$ and y is generated by a regular, finite-order VAR as in (2.1)*

$$y^B \nrightarrow_h y^A \Leftrightarrow \Pi^{(h)}_{AB,s} = 0, \quad s = 1, \ldots, p,$$

where 0 indicates a zero matrix of appropriate dimension. That is, y^B does not cause y^A at horizon h if and only if all the relevant coefficients in the direct VAR model for horizon h are zero.

By definition y^B *does cause at lag h* y^A if at least one of the parameter matrices in the above theorem is not zero. For some indices A, we denote by $C(y^A)$ the set consisting of the variables in y^A itself and all variables that cause y^A in the above sense. When A is a singleton, say $A = \{i\}$, we write $C(y_i)$ instead.

We first investigate the case of a VAR with $p = 1$. For this case, we show that the coefficients of the direct VAR representation $\Pi^{(h)}_1$ are related to the set of paths in the directed graph representing the VAR model. To show this, note that the direct VAR representation for $p = 1$ is

$$y_{t+h} = \Pi^{(h)}_1 y_t + u^{(h)}_{t+h}, \quad \text{with} \quad \Pi^{(1)}_1 = A_1, \quad \Pi^{(h)}_1 = A_1 \Pi^{(h-1)}_1 \ (h \geq 2) \tag{3.3}$$

By induction, every off-diagonal element $\Pi^{(h)}_{ij,1}$ in $\Pi^{(h)}_1$ can be linked to the set of all simple[8] paths that lead from the associated vertices i to j in h steps. We state this result formally in Theorem 2.

THEOREM 2 *Given two variables y_i and y_j of y, $i \neq j$, and y follows a regular, finite-order VAR(1) as in (2.1), the entry at position (i, j), $\Pi^{(h)}_{ij,1}$, corresponds to the set of simple paths*

$$\mathbb{P}^{(h)}_{ij} = \{P : P = (e_1, \ldots, e_h), e_k \in E, e_1 = (v_0, v_1), v_0 = i, \\ e_h = (v_{h-1}, v_h), v_h = j, P \text{ is simple}\} \tag{3.4}$$

leading from vertex i to vertex j for all $h \in \mathbb{N}$ in that

$$\Pi^{(h)}_{ij,1} = \sum_{P \in \mathbb{P}^{(h)}_{ij}} \prod_{(l,m) \in P} a_{l,m}$$

The proof of Theorem 2 is given in Appendix A.1. $\mathbb{P}^{(h)}_{ij}$ is the set of all simple paths of length h leading from i to j. The theorem essentially states, that the coefficients of the direct VAR $\Pi^{(h)}_{ij,1}$ can be written in terms of sums of products of autoregressive coefficients in A_1, where the indices correspond to edges on different paths from i to j. To illustrate this, consider again our simple VAR(1) from Figure 1. In this example, there are two simple path of length $h = 2$ from variable 2 to variable 3, thus the set of simple paths is

$$\mathbb{P}^{(2)}_{23} = \{\langle (2,1), (1,3) \rangle, \langle (2,4), (4,3) \rangle\}.$$

[8] A path is simple if all its vertices are distinct except possibly the first and the last vertex.

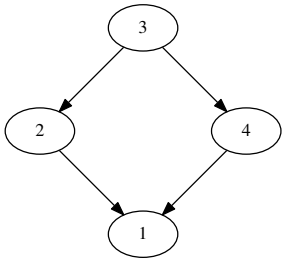
Using the result from Theorem 2, we find for $h = 2$, $i = 2$, and $j = 3$,

$$\Pi_{23,1}^{(2)} = a_{21}a_{13} + a_{24}a_{43}.$$

Note that the indices of the VAR coefficients match with the edges of the path set $\mathbb{P}_{23}^{(2)}$. Obviously, the same $\Pi_{23,1}^{(2)}$ would be obtained from the direct VAR coefficient definition.

At first sight, the result of Theorem 2 seems to imply that variable j would be multi-step causal for variable i, whenever there exists at least one path from i to j. However, $\mathbb{P}_{ij}^{(h)} \neq \emptyset$ does *not* imply that variable y_j causes y_i as the following example illustrates. Consider a VAR(1) with the associated directed graph:

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.5 & 1 & 0 & 0 \\ 0 & 0.5 & 1 & 0.5 \\ -0.5 & 0 & 0 & 1 \end{pmatrix}$$



In this example, we have $\mathbb{P}_{31}^{(1)} = \emptyset$ and $\mathbb{P}_{31}^{(2)} = \{(\,(3,2),(2,1)\,),(\,(3,4),(4,1)\,)\}$ but variable y_3 is not caused by y_1 at neither horizon one nor horizon two since

$$\Pi_{31}^{(2)} = a_{32}a_{21} + a_{34}a_{41} = 1/4 - 1/4 = 0.$$

This 'cancelling-out' effect of course happens very rarely with any real data set and most reasonable estimation methods. Therefore, one might exclude it by assumption.

ASSUMPTION 1 *Given a VAR(1) system, for all $h > 1$ and $i \neq j$*

$$\Pi_{ij}^{(h)} = 0 \Leftrightarrow a_{ik}\Pi_{kj}^{(h-1)} = 0,\ k = 1, \ldots, K.$$

We basically assume that if variable j is multi-step non-causal for variable i, then this is because there is no path form i to j and not because there is a path from i to j with VAR coefficients such that there is 'cancelling-out'. Assumption 1 thus ensures the correspondence between paths and causality as in Corollary 1.

COROLLARY 1 *Given a VAR(1) system and Assumption 1, for all $h \geq 1$ and $i \neq j$:*

$$\mathbb{P}_{ij}^{(h)} \neq \emptyset \Leftrightarrow \Pi_{ij,1}^{(h)} \neq 0.$$

This results states that variable j is multi-step causal for variable i if there is at least one path from variable i to variable j. Under Assumption 1, the strongly connected components can be now be interpreted very easily.

COROLLARY 2 *Given a VAR(1) system and Assumption 1, the strongly connected components are sets of variables that are mutual causal.*

This follows immediately from the definition of a SCC as for each pair i and j, there is a path from i to j and from j to i. The last step is to analyze multi-step causality across SCCs. Consider a strongly connected component C_i and, as in Section 2, denote the set of all SCCs that are reachable from C_i by $R_G^{SCC}(C_i)$. Then all variables in $R_G^{SCC}(C_i)$ are multi-step causal for variables in C_i as there is a path from any variable in C_i to the variables in $R_G^{SCC}(C_i)$.

Finally, based on the foregoing discussion and for a given set of variables of interest y^{I}, we note that the varaibles, which are multi-step causal for y^{I} are all the variables in the SCCs that contain elements of y^{I} and all variables in all SCCs that may reached from these SCCs. Note that under Assumption 1, this coincides with the definition of the minimal set of relevant variables $R(y^{\mathrm{I}})$ from Section 2. We denote the set consisting of y^{I} and all variables multi-step causal for y^{I} by $C(y^{\mathrm{I}})$. In case of 'cancelling-out', the set of relevant variables $R(y^{\mathrm{I}})$ will be larger than the set of causal variables for y^{I}, $C(y^{\mathrm{I}})$. We summarize this result more formally in Corollary 3.

COROLLARY 3 *Given a VAR(1) system, $C(y^I) \subset R(y^I)$, that is, all variables that cause y^I are contained in the set of relevant variables. If Assumption 1 is true, then the set of relevant variables is identical to the set of causal variables, $C(y^I) = R(y^I)$.*

This establishes the relation between the set of relevant variables found from the graph of strongly connected components and the variables that are multi-step causal for the variables of interest. The results can be generalized to VAR(p) models, which can always be written as a VAR(1) in companion form. The corresponding analysis would follow similar arguments.

4 Empirical Applications

To illustrate the usefulness of the graph theoretical approach for selecting the relevant information set, we apply the method on a large set of US economic time series. We focus on the selection of variables, on impulse response analysis and forecasting properties of the selected models.

We start from a set of 41 quarterly economic time series that includes a large variety of macroeconomic and financial series. The collection of variables is similar to related studies as e.g. Jarociński & Maćkowiak (2017) and Kascha & Trenkler (2015) and includes real GDP and its components, business cycle indicators, various price measures and interest rates, monetary aggregates and a number of labor market variables. In addition, the data includes exchange rate data together with three key variables for the Euro area (Euro area GDP, Euro area CPI and a Euro area interest rate). The US data is taken from the FRED data base, while the European series are obtained from the Area Wide Model (AWM) data base maintained at the European Central Bank (ECB). A detailed list with variables and data sources is shown in Table A.1.

For some variables, we only have data starting in the mid 1970s. Consequently, our baseline sample starts in 1975. The end of the baseline sample is the last quarter of 2014.

To apply the approach discussed in Section 2, we first transform the data to stationarity. This involves taking logarithms and/or differences depending on the property of the respective

variable.[9] We describe the details of data preparation in Appendix A.2 and document the transformations by reporting the transformation codes listed in Table A.1.

In what follows, we apply the graph theoretical methods to a sparse VAR, i.e. a VAR with a number of zero coefficients in the autoregressive matrices. In our application, these sparse VARs are selected by applying the least absolute selection and shrinkage operator (LASSO) in the context of the VAR model. There is ample evidence in the literature that LASSO is a useful device and often leads to forecasts that are more precise than standard (unrestricted or subset) VARs (see e.g. Kascha & Trenkler (2015) and references therein). While in principle other methods for subset selection may be employed, we only use LASSO and point out that the subset selection in not the main focus of our paper. Instead, we start from a given subset structure and explore how this structure can be used to detect the smallest possible VAR system.

4.1 Variable Selection

To illustrate the variable selection, we need to define a set of 'variables of interest' y^{I}. In the first example, we choose real US GDP, the US consumer prices (CPI), and the federal funds rate as variables of interest. This includes three key economic variables often analyzed with VARs and also corresponds to the variables chosen by Jarociński & Maćkowiak (2017). Then we estimate a LASSO-VAR with $p = 1$ lags in all 41 variables from Table A.1. The shrinkage parameter in the LASSO approach is chosen by the Bayesian information criterion (BIC).[10] We use a balanced data panel by removing all missing observations (after transforming the variables) at the beginning and the end of the sample. In this case, the first available observation corresponds to 1979Q3, the last to $T = 2014Q4$.

To investigate which variables are selected into a VAR and how the selection changes over time, we have used the following expanding window setup. The initial estimation period ends in T_1 and we report the variables selected based on the VAR structure estimated on this sample. We then add recursively observations to the sample and re-estimate the LASSO-VAR with $T_1 + 1, T_1 + 2, \ldots, T - h^* - 1, T - h^*$ observations, where $h^* = 4$ is the maximal forecasting horizon used in the forecasting exercise below. In our baseline specification, we choose $T_1 = 1998Q4$ and thus report recursive selection results for samples ending between 1998Q4 and 2013Q4. Consequently, we report selection results for 61 periods and we show them in graphical form in Figure 2. The rows in the checkerboard graph correspond to the different economic variables, whereas the columns refer to different estimation samples. The filled green squares correspond to the variables of interest (here: GDP, CPI and the Federal Funds Rate), a filled blue square in a specific row indicates that the variable in that row is selected into the minimal VAR in the period corresponding to the column. Accordingly, a white square indicates that the variable has not been selected into the minimal VAR in a particular period.

A number of interesting results emerge: First, there are ten variables that are always selected into the minimal VAR in each of the periods considered. This includes the change of real inventories, employment, the corporate bond spread, the 1-year T-Bill rate, a three-

[9] Transforming the variables to stationarity cancels possible common trends and cointegration relations between the variables. Extending the graphical methods to models with common trends such as cointegrated

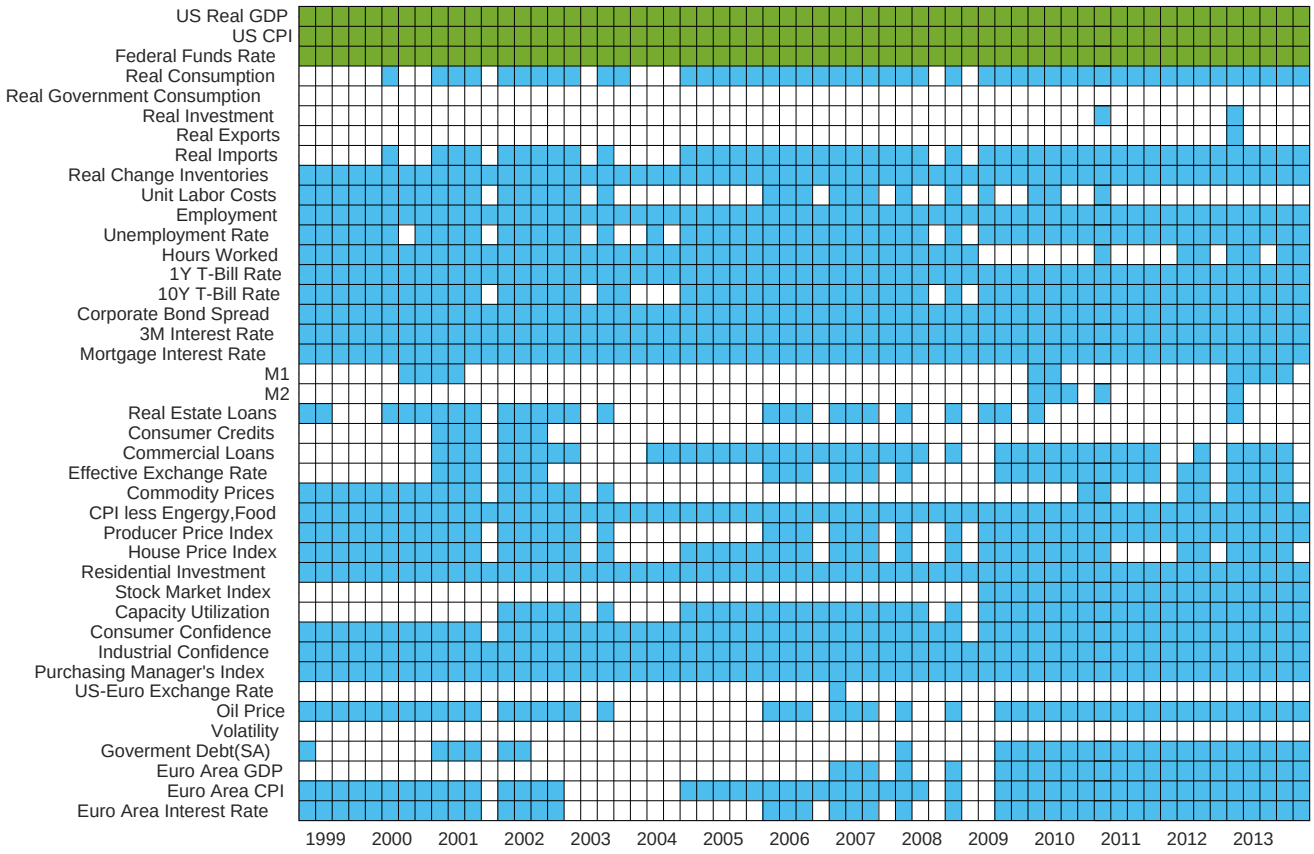


Figure 2: Variable selection results: Variable of interest y^{I}: US Real GDP, CPI and FFR (green). Relevant variables as selected by graphical method (blue) and variables not selected (white). Sample period: 1975Q3-2014Q4.

.

month money market interest rate, the mortgage interest rate, the CPI less food and energy price index, residential investment, an industrial confidence index, and the purchasing manager's index. A tentative interpretation of this result may be that these variables form the minimal set of additional variables that should be considered if a VAR in output, inflation and interest rate is of interest. We also note that some of these variables have typically not been included in related empirical studies on the effects of monetary policy shocks.

We compare our set of selected variables with those from Jarociński & Maćkowiak (2017). As discussed earlier Jarociński & Maćkowiak (2017) use a Bayesian model and compute the posterior probability that the variables of interest are Granger-causal prior to variables in the larger data set. If this probability is low for a particular variable, then this variable is likely to enter the model of the variables of interest. We note that the comparison is somewhat limited by the fact, that we use slightly different data, a different sample period and different variable transformations.[11] Nevertheless, we find some similarity of the sets of

VARs and vector error correction models is left for future research.

[10] As for the implementation of the VAR-LASSO we follow the paper by Kascha & Trenkler (2015) and refer to their paper for details.

[11] Within their Bayesian approach Jarociński & Maćkowiak (2017) use (log) levels for most variables, while we use stationary transforms.

relevant variables identified from the suggested graphical approach and the set of important variables identified from the Bayesian approach in Jarociński & Maćkowiak (2017). We note that (except for residential investment) the variables that have been always selected by our graphical methods show a very low posterior probability (< 0.1) of being Granger causally prior to the variables in y^{I} and consequently are also marked as relevant in the Bayesian analysis of Jarociński & Maćkowiak (2017) (see their Table 1). From the results in Table 1 of Jarociński & Maćkowiak (2017), one would select 22 variables, which are relevant for y^{I} if one selects all variables with a posterior probability less than 0.1. Interestingly, most of these variables are also often (but not always) selected by our methods. With exception of real investment, real exports and M1, our selection rates for the other variables and over the recursive sample range between 68% and 87%. It is also noteworthy that these variables are always selected by our method if the sample includes the post 08/09 crisis period and ends in 2013Q4, which matches the sample end used in Jarociński & Maćkowiak (2017). In addition, we note that real government consumption, stock market volatility, and the US/Euro exchange rate are never selected in our approach, which is also in line high posterior probabilities (ranging from 0.36 to 1) in Jarociński & Maćkowiak (2017). There are only three variables that show a somewhat different pattern in both studies: Real investments, real exports and M1 are almost never selected in our approach, while they are important in the analysis of Jarociński & Maćkowiak (2017). A possible explanation is that in our setup these variables have little additional information not already included in other selected variables (as e.g. the change in real inventories and residential investment, interest rates and real imports). Overall, our selection results seem to match those from Jarociński & Maćkowiak (2017) reasonably well.

In a number of periods additional variables are selected into the minimal VAR. The number of selected variables is quite large (on average 24.4 out of the 41 are selected).[12] This provides evidence that the dynamic relationship between economic variables is more complex than simple, small scale VARs tend to suggest. While there are some changes as we increase the estimation sample, the overall selection of variables is relatively stable for the period before the 2008/2009 economic crisis. Interestingly, when the estimation sample extends beyond the crisis period, we observe that our method tends to select more variables, possibly suggesting that the linkages between variables have become more pronounced.

4.2 Impulse Response Analysis

We illustrate the effect of including the selected variables into a small VAR system on estimated impulse responses. As in Section 4.1, the small VAR consists of the variables of interest with real GDP, the CPI and the federal funds rates (FFR). These type of systems have also been used by Jarociński & Maćkowiak (2017) and Banbura et al. (2010). Following standard specifications from the literature (see e.g. Christiano, Eichenbaum & Evans (1999)), we have used VAR(4) models in the (log) levels of the variables for the comparison of impulse responses. For the VAR with selected variables, we have added the 10 variables (again in (log)

[12]Note that this is again in line with results in Jarociński & Maćkowiak (2017) who find that the number relevant variables ranges between 22-36 variables (see their Table G.2).

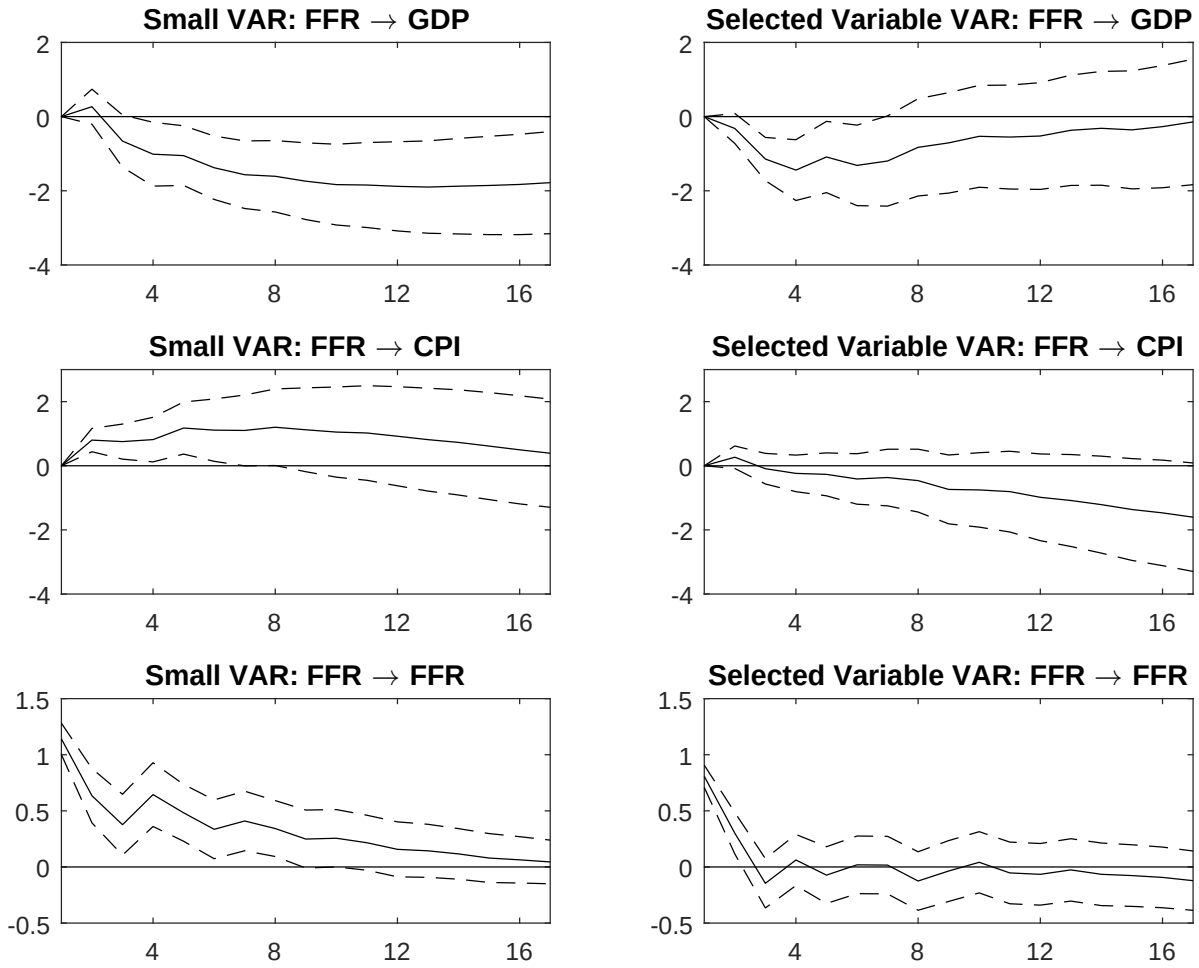


Figure 3: Impulse responses in small VAR and selected variable VAR. Left: Responses to a shock in FFR in 3-variable (small) VAR(4) including GDP, FFR, and CPI. Right: Responses to a shock in FFR in 13-variable VAR(4) with 10 selected additional variables. Sample period: 1975Q3-2007Q4.

levels) that have been selected for all considered sample periods in Section 4.1. We use the standard ordering of variables and thus include the change of real inventories, employment, the CPI less food and energy price index, residential investment, an industrial confidence index, and the purchasing manager's index in the group of 'slow moving' variable, i.e. there are ordered above the federal funds rate variable. In contrast, the corporate bond spread, the 1-year T-Bill rate, a three-month money market interest rate, and the mortgage interest rate are in the group of 'fast moving' variables and are consequently ordered below the federal funds rate. Using a Cholesky decomposition, this ordering implies that a shock in the federal funds rate (typically labeled a monetary policy shock) may have immediate impact on the 'fast moving' variables, while the 'slow moving' variables may only react with a lag of one quarter. All VAR models are estimated by unrestricted multivariate LS (i.e. no shrinkage is applied) and the reported (pointwise) confidence intervals are asymptotic 95% intervals obtained using the 'delta method' (see e.g. Lütkepohl (2005, Section 3.7)). In Figure 3, we report results for a sample that ends in 2007 to exclude the effects of the 2008/09 financial

crisis. The left panel of the figure shows the responses of GDP, CPI and the federal funds rate to a contractionary shock in the federal funds rate within the small, 3 variable VAR. In line with results from the literature, we find the typical pattern in these type of systems with a significant and persistent drop in output. We also find a significant increase in CPI, which is known as the 'price-puzzle' because economic theory suggests a decrease rather than an increase in the price level after tightening monetary policy. In other words, the response of the price level in the small VAR is counterintuitive. Adding the 10 selected variables changes the response patterns substantially. First, the drop in output is now much less persistent. In fact, two years after the shock the response of output is no longer significantly different from zero. Moreover, the price puzzle disappears. While the point estimate of the response shows the expected sign, this effect is not significantly different from zero in the considered sample period. Thus, including the variables selected by our method leads to much more reasonable impulse response patterns and change the interpretation of the results substantially. Using the right information set is of obvious importance for structural analysis.

We report additional results for different sample periods in Figures A.1–A.2 in Appendix A.3. We just note here that similar results have been obtained for the other sample periods. Using a sample that starts in 1972 (earliest possible starting date given our selected variables), the response of the CPI is even significantly negative (see Figure A.1). In the longer sample that extends to the end of 2014, we find that the output response in the selected variable VAR is less strong and only borderline significant while again the positive significant response of CPI disappears by adding our 10 variables (see Figure A.2). It is interesting to note that the changes in the response pattern obtained by just adding 10 variables are to some extent similar to the changes in obtained by Banbura et al. (2010) in their medium (20 variable) and large (131 variables) for monthly data. Also, the working paper version of Jarociński & Maćkowiak (2017), see Jarociński & Maćkowiak (2013), finds that adding selected variables changes the response pattern of output, price level, and short-term interest rate in a similar way as in our paper (however using data for the Euro area). In other words, it seems that our methodology provides a perspective that complements the 'large VAR' idea of Banbura et al. (2010) and the Bayesian approach of Jarociński & Maćkowiak (2017).

4.3 Forecasting Performance

We also investigate the usefulness of our method in forecasting. We first use the same data and variables of interest y^{I} as in Sections 4.1 to 4.2 and report forecasting results for the GDP growth, CPI inflation and the federal funds rate. We compare forecasts of four different VAR models. Our benchmark model is an unrestricted VAR in the three variables of interest (GDP, CPI, federal funds rate). We compare this against predictions obtained from (i) a LASSO-VAR applied to all 41 variables (LVAR), (ii) a LASSO-VAR applied only to the variables of interest and the variables selected in a first step by our method (selLVAR), and (iii) a LASSO-VAR applied to the variables of interest and the variables not selected (irrelevant variables) by our method on the first step (irrLVAR). The selected variable VARs correspond to the VARs discussed in Section 4.1. All VAR models are specified for the transformed variables (according to Table A.1). We compare forecast accuracy in terms of mean squared

Table 1: MSFEs Relative to VAR in Real GDP, CPI and the Federal Funds Rate

		1999Q1-2014Q4			1999Q1-2007Q4		
		$h=1$	$h=2$	$h=4$	$h=1$	$h=2$	$h=4$
US Real GDP	LVAR	0.87	0.96	0.98	0.83	1.02	0.99
	selLVAR	0.81	0.84	0.92	0.85	1.01	1.01
	irrLVAR	0.95	1.04	1.01	0.89	1.03	1.00
US CPI	LVAR	1.07	0.90	1.00	1.35	0.97	1.01
	selLVAR	0.95	0.94	1.00	1.07	0.99	1.01
	irrLVAR	1.11	0.92	1.00	1.24	0.98	1.01
Federal Funds Rate	LVAR	0.36	0.55	0.95	0.36	0.57	1.00
	selLVAR	0.43	0.64	0.95	0.35	0.65	0.98
	irrLVAR	0.47	0.56	0.96	0.47	0.58	1.00

Note: The table shows relative mean squared forecast errors (MSFEs) for different variables, forecasting horizons, and evaluation periods. All results are relative to the forecasting results of an unrestricted VAR in the three variables of interest (real GDP, CPI and the federal funds rate). LVAR denotes results based on a LASSO-VAR applied on all 41 variables, selLVAR denotes results from a LASSO-VAR applied to the minimal VAR with selected variables, irrLVAR denotes results from a LASSO-VAR applied to a VAR including the variables of interest and all variables not selected by the graphical method.

forecast errors (MSFEs) for (iterative) forecasts at horizons $h = 1, 2, 4$. We use the same expanding window setup as in Section 4.1 and report results for different forecast evaluation periods. We discuss results for a baseline evaluation period from 1999Q1-2014Q1. Moreover, we also report results for a shorter evaluation period that ends in 2007Q4 to exclude the 2008/2009 financial crisis period.

From results reported in Table 1, we find gains from variable selection in the forecasting accuracy of GDP growth in the evaluation period ending in 2014. There are also gains for CPI forecasts for short-term predictions ($h = 1$). For other horizons and for the federal funds rate, selecting the variables is not necessarily beneficial in terms of forecasting precision. This might reflect that we model the change in the CPI inflation rate and the change in the federal funds rate, two variables that are quite noisy, not very persistent and inherently difficult to forecast. In this case, applying a LASSO-VAR on all 41 variables is useful because almost no variables remain in the model after shrinkage. Thus, in this situation, the LASSO is quite successful in 'kicking' out variables from the model and consequently, the potential gains from additional (pre-)selection is limited. We also note that the gains from the selection are more pronounced in the period including the financial crisis and in this example disappear almost completely in the shorter evaluation period ending in 2007Q4. Thus, a tentative conclusion is that the variable selection is particularly useful in periods that also include changing economic conditions like the 2008/2009 financial crisis.

To provide some additional evidence on the forecasting performance, we also report results from a forecasting exercise that focusses on two key indicators of real economic activity, US real GDP and the unemployment rate. We start from the same 41 time series as before but y^{I} now contains the output growth and the unemployment rate. We apply the variable selection based on the strongly connected components as explained in Section 2 and Figure

Table 2: MSFEs Relative to VAR in Real GDP and the Unemployment Rate

		1999Q1-2014Q4			1999Q1-2007Q4		
		$h=1$	$h=2$	$h=4$	$h=1$	$h=2$	$h=4$
US Real GDP	LVAR	0.82	0.99	0.99	0.70	1.01	1.00
	selLVAR	0.79	0.91	0.92	0.72	1.00	1.00
	irrLVAR	0.95	1.07	1.01	0.82	1.04	1.01
Unemployment Rate	LVAR	1.13	1.06	0.99	0.98	0.99	0.98
	selLVAR	0.98	0.98	0.92	1.01	0.98	0.92
	irrLVAR	1.17	1.13	1.01	1.07	1.05	1.01

Note: The table shows relative mean squared forecast errors (MSFEs) for different variables, forecasting horizons, and evaluation periods. All results are relative to the forecasting results of an unrestricted VAR in the two variables of interest (real GDP and the unemployment rate). LVAR denotes results based on a LASSO-VAR applied on all 41 variables, selLVAR denotes results from a LASSO-VAR applied to the minimal VAR with selected variables, irrLVAR denotes results from a LASSO-VAR applied to a VAR including the variables of interest and all variables not selected by the graphical method.

A.3 in Appendix A.3 reports the variables that have been selected in different periods. The models and setup of the forecasting exercise is very similar to the previous exercise. Table 2 shows MSFEs of the different VARs relative to an unrestricted (small VAR) with just two variables.

We again find gains in forecasting accuracy when using the graph-theoretic approach to select the variables of the VAR. In our baseline sample, for both considered variables, the MSFE of the selected variable VAR is smallest for all considered forecasting horizons. The corresponding MSFEs from a VAR with all 41 variables (estimated by LASSO) is larger. In other words, selecting the variables first and then applying LASSO leads to more precise forecasts than solely relying on LASSO. To investigate the usefulness of the selection further, we also report results for VAR that include the two variables of interest, together with the set of non-selected variables. As expected, including the 'irrelevant variables' deteriorates the predictive accuracy. MSFEs from this model are higher than both, the LVAR and the selected variable VAR. In addition, the irrLVAR results are also often worse than those from a standard VAR (indicated by relative MSFEs larger 1). Again, this highlights the role for variable selection since applying only LASSO does not succeed to 'kick out' all noise due to irrelevant variables. In the shorter evaluation period, the gains from selecting the variables disappear and predictions from the LVAR are as good as prediction of the selLVAR. Inspection of the recursive forecasts (not shown to conserve space) reveals that the selected variable VAR is especially useful in predicting the sharp drop in output during the 08/09 crises.

Overall, the results indicate that the set of selected variables is typically helpful in getting more precise forecast of key economic variables. Consequently, the results support our hypothesis that the graph-theoretic approach is a useful tool for selection variables for VAR analysis.

5 Conclusion

This paper uses concepts from graph theory for variable selection in VAR models. To this end, we identify strongly connected components from the directed graph representing the dynamic relationships among the variables in a sparse VAR. We suggest to use relations among the strongly connected components in a so-called component graph to identify a minimal set of variables that we need to include in a VAR analysis for a small set of variables if forecasts or impulse responses are of interest. The paper adds to the existing literature by introducing a graphical method, which to the best of our knowledge has not been used for variable selection in econometrics.

We also show that there is a simple relation between the graph theoretical concept and multi-step causality and relate the paths in the graph to coefficients of a direct VAR system. It follows from the results in the paper that the set of relevant variables selected from the graphical approach coincides with the set of variables that are multi-step causal for the variables of interest.

We illustrate the usefulness of the variable selection method in forecasting and structural analysis in empirical applications on US macroeconomic data. We use a small monetary system (US real GDP, CPI inflation and the federal funds interest rates) as the variables of interest. Given this set, we apply the graphical approach to select additional variables out of a large set of US macroeconomic variables. The selected VAR typically includes some variables from the real sector (changes of inventories, employment, residential investment), forward looking indicators (industrial confidence and the purchasing manager's index), different interest rates (corporate bond spread, money market rates, 1 year T-Bill rate) and a CPI related measure (CPI less food and energy). The selection of variable seems sensible from an economic point of view. Interestingly, we find that this list includes some variables that up to now, researchers typically have not included in small monetary systems. We also find that the variable selection suggested by our method is comparable to what related papers have found on similar data but with completely different (Bayesian) methods. Moreover, we find that including the selected variables for impulse response analysis is useful: In the small monetary system in output, inflation and interest rate, we find that including the selected variables avoids the so-called 'prize puzzle'. In addition, short-term forecasts for real GDP and inflation improve by including the selected variables. In a second application, we take real output and unemployment as the variables of interest and find from pseudo-out-of-sample forecasting experiments that including the selected variables improves forecasting accuracy of both variables.

Overall our empirical results suggest that using graphical modeling for variable selection is a useful addition to the VAR econometricians' toolbox. The method complements existing methods for large data sets and is particularly useful if a researcher prefers to work with smaller scale models e.g. for maintaining consistency with small scale theoretical models. Moreover, compared to alternative methods for large data sets, a graphical representation of the strongly connected components may give useful insights on the (causal) relationships among the VAR variables.

Extensions of the current paper could use graphical models for variable selection taking

also the contemporaneous relationships among variables into account. Moreover, extending the approach to models with integrated and cointegrated variables would be of interest. We leave this for future research.

References

Banbura, M., Giannone, D. & Reichlin, L. (2010). Large Bayesian vector autoregressions, *Journal of Applied Econometrics* **25**: 71–92.

Bernanke, B. S., Boivin, J. & Eliasz, P. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach, *Quarterly Journal of Economics* **120**(1): 387–422.

Brillinger, D. R. (1996). Remarks concerning graphical models for time series and point processes, *Revista de Econometria* **16**: 1–23.

Carriero, A., Clark, T. E. & Marcellino, M. (2015). Bayesian VARs: Specification choices and forecast accuracy, *Journal of Applied Econometrics* **30**(1): 46–73.

Carriero, A., Kapetanios, G. & Marcellino, M. (2009). Forecasting exchange rates with a large Bayesian VAR, *International Journal of Forecasting* **25**(2): 400–417.

Carriero, A., Kapetanios, G. & Marcellino, M. (2012). Forecasting government bond yields with large Bayesian vector autoregressions, *Journal of Banking and Finance* **36**(7): 2026–2047.

Cheng, X. & Hansen, B. E. (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach, *Journal of Econometrics* **186**(2): 280–293.

Christiano, L. J., Eichenbaum, M. & Evans, C. (1999). Monetary policy shocks: What have we learned and to what end?, *in* J. Taylor & M. Woodford (eds), *The Handbook of Macroeconomics*, Amsterdam: Elsevier Science Publication.

Clements, M. P. (2016). Real-time factor model forecasting and the effects of instability, *Computational Statistics and Data Analysis* **100**: 661–675.

Dahlhaus, R. (2000). Graphical interaction models for multivariate time series, *Metrika* **51**(2): 157–172.

Dahlhaus, R. & Eichler, M. (2003). Causality and graphical models for time series, *in* P. Green, N. Hjort & S. Richardson (eds), *Highly structured stochastic systems*, Oxford University Press, pp. 115–137.

Demiralp, S. & Hoover, K. D. (2003). Searching for the causal structure of a vector autoregression, *Oxford Bulletin of Economics and Statistics* **65**: 745–767.

Doan, T. A. & Todd, R. (2010). Causal ordering for multivariate linear systems. URL: https://estima.com/forum/download/file.php?id=333&sid=253771ed992e2cfb19632374231e9946

Duff, I. S. & Reid, J. (1978). An implementation of Tarjan's algorithm for the block triangularization of a matrix, *ACM Transactions on Mathematical Software* **4**(2): 137–147.

Dufour, J. M., Pelletier, D. & Renault, E. (2006). Short run and long run causality in time series: Inference, *Journal of Econometrics* **132**(2): 337–362.

Dufour, J. M. & Renault, E. (1998). Short run and long run causality in time series: Theory, *Econometrica* **66**(5): 1099–1125.

Edwards, D. (2000). *Introduction to Graphical Modelling*, Springer Texts in Statistics, 2 edn, Springer-Verlag, New York.

Eichler, M. (2006). Graphical modeling of dynamic relationships in multivariate time series, *in* B. Schelter, M. Winterhalder & J. Timmer (eds), *Handbook of Time Series Analysis: Recent Developments and Applications*, Wiley.

Eichler, M. (2007). Granger causality and path diagrams for multivariate time series, *Journal of Econometrics* **137**(2): 334–353.

Eichler, M. (2012). Graphical modelling of multivariate time series, *Probability Theory and Related Fields* **153**(1-2): 233–268.

Eickmeier, S. & Ziegler, C. (2008). How successful are dynamic factor models at forecasting output and inflation? A meta-analytic approach, *Journal of Forecasting* **27**(3): 237–265.

Flamm, C., Kalliauer, U., Deistler, M., Waser, M. & Graef, A. (2012). Graphs for dependence and causality in multivariate time series, *in* L. Wang & H. Garnier (eds), *System Identification, Environmental Modelling, and Control System Design*, Springer London, London, pp. 133–151.

Giannone, D., Lenza, M., Momferatou, D. & Onorante, L. (2014). Short-term inflation projections: A Bayesian vector autoregressive approach, *International Journal of Forecasting* **30**(3): 635–644.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods, *Econometrica* **37**: 424–438.

Heinlein, R. & Krolzig, H. M. (2012). Effects of monetary policy on the US dollar/UK pound exchange rate. Is there a 'delayed overshooting puzzle'?, *Review of International Economics* **20**(3): 443–467.

Hoover, K., Demiralp, S. & Perez, S. J. (2009). Empirical identification of the vector autoregression: The causes and effects of U.S. M2, *in* J. Castle & N. Shepard (eds), *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, Oxford University Press, Oxford, pp. 37–58.

Jarociński, M. & Maćkowiak, B. (2013). Granger causal priority and choice of variables in vector autoregressions, *Working Paper Series 1600*, Euroepan Central Bank.

Jarociński, M. & Maćkowiak, B. (2017). Granger causal priority and choice of variables in vector autoregressions, *Review of Economics and Statistics* **99**(2): 319–329.

Kascha, C. & Trenkler, C. (2015). Forecasting VARs, model selection, and shrinkage, *Working Paper ECON 15-07*, Department of Economics, University of Mannheim.

Koop, G. M. (2013). Forecasting with medium and large Bayesian VARs, *Journal of Applied Econometrics* **28**(2): 177–203.

Lauritzen, S. L. (1996). *Graphical Models*, Oxford Statistical Science Series, Oxford University Press, Oxford.

Ludvigson, S. C. & Ng, S. (2007). The empirical risk-return relation: A factor analysis approach, *Journal of Financial Economics* **83**(1): 171–222.

Ludvigson, S. C. & Ng, S. (2009). Macro factors in bond risk premia, *Review of Financial Studies* **22**(12): 5027–5067.

Lütkepohl, H. (1993). Testing for causation between two variables in higher dimensional VAR models, *in* H. Schneeweiß & K. F. Zimmermann (eds), *Studies in Applied Econometrics*, Springer-Verlag, Heidelberg, pp. 75–91.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*, Berlin: Springer-Verlag.

McCracken, M. W. & Ng, S. (2015). FRED-MD: A monthly database for macroeconomic research, *Working Paper 2015-012B*, Federal Reserve Bank of St. Louis.

Pearl, J. (2000). *Causality*, Cambridge University Press, New York.

Penm, J. & Terrell, R. (1986). *The 'Derived' Moving-Average Model and its Role in Causality*, Applied Probability Trust, Sheffield, pp. 99–111.

Sims, C. A. (1980). Macroeconomics and reality, *Econometrica* **48**: 1–48.

Sims, C. A. (1982). Policy analysis with econometric-models, *Brookings Papers on Economic Activity* (1): 107–164.

Sims, C. A. (2015). Causal orderings and exogeneity, Lecture Notes.
URL: `http://sims.princeton.edu/yftp/Times15F/GCP15.pdf`

Stock, J. H. & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors, *Journal of the American Statistical Association* **97**(460): 1167–1179.

Stock, J. H. & Watson, M. W. (2003). Forecasting output and inflation: The role of asset prices, *Journal of Economic Literature* **41**(3): 788–829.

Stock, J. H. & Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors, *Journal of Business and Economic Statistics* **30**(4): 481–493.

Stock, J. H. & Watson, M. W. (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics, *in* J. B. Taylor & H. Uhlig (eds), *Handbook of Macroeconomics*, Vol. 2, Elsevier, chapter 8, pp. 415–525.

Swanson, N. R. & Granger, C. W. J. (1997). Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions, *Journal of the American Statistical Association* **92**(437): 357–367.

Tarjan, R. E. (1972). Depth-first search and linear graph algorithms, *SIAM Journal on Computing* **1**(2): 146–160.

Uhlig, H. (2009). Comment on 'How has the Euro changed the monetary transmission mechanism?', *in* D. Acemoglu, K. Rogoff & M. Woodford (eds), *NBER Macroeconomics Annual 2008*, Vol. 23, University of Chicago Press, pp. 141–152.

A Appendix

A.1 Proofs

A.1.1 Proof of Theorem 2

We prove the result of Theorem 2. Assume that the vector y_t is generated by a VAR(1). We need to show that

$$\Pi_{ij,1}^{(h)} = \sum_{P \in \mathbb{P}_{ij}^{(h)}} \prod_{(l,m)\in P} a_{l,m}. \tag{A.1}$$

Let $i, j \in \mathbb{N}_K : i \neq j$. Consider the case of $h = 1$ first. Here we have

$$\Pi_{ij,1}^{(1)} = a_{ij}, \tag{A.2}$$

which follows directly from the definition of the direct VAR coefficient in $\Pi_{ij,1}^{(1)}$ in (3.3). For $h = 1$, there is only one path of length 1 from i to j. Consequently,

$$\mathbb{P}_{ij}^{(1)} = \{\langle (i,j) \rangle\}. \tag{A.3}$$

The equality follows directly from the definition of $\mathbb{P}_{ij}^{(h)}$ in (3.4). Expression (A.1) is true by (A.2) and (A.3).

Next we consider cases with $h > 1$. For $h - 1$, the results in (A.1) may be written as

$$\Pi_{ij,1}^{(h-1)} = \sum_{P \in \mathbb{P}_{ij}^{(h-1)}} \prod_{(l,m)\in P} a_{l,m}.$$

We use

$$\mathbb{P}_{ij}^{(h)} = \cup_{(i,k)\in E,\, P\in\mathbb{P}_{kj}^{(h-1)}} \{(i,k)\} \cup P,$$

where we express the set of paths from i to j as the union of the path from of length one from i to k and the set of paths of length $h - 1$ from k to j. Using (3.3), we have

$$\begin{aligned}
\Pi_{ij,1}^{(h)} &= \sum_{k=1}^{K} a_{ik} \Pi_{kj,1}^{(h-1)} \\
&= \sum_{k=1}^{K} a_{ik} \left(\sum_{P \in \mathbb{P}_{kj}^{(h-1)}} \prod_{(l,m)\in P} a_{l,m} \right) \\
&= \sum_{k=1}^{K} \sum_{P \in \mathbb{P}_{kj}^{(h-1)}} a_{ik} \prod_{(l,m)\in P} a_{l,m} \\
&= \sum_{k=1}^{K} \sum_{P \in \mathbb{P}_{kj}^{(h-1)}} \prod_{(l,m)\in \{(i,k)\}\cup P} a_{l,m} \\
&= \sum_{P \in \mathbb{P}_{ij}^{(h)}} \prod_{(l,m)\in P} a_{l,m},
\end{aligned}$$

which is the result in (A.1). By simple induction this holds for all $h \in \mathbb{N}$. Because i, j was arbitrary it holds for all i, j with $i \neq j$. □

A.1.2 Proof of Corollary 1

Assume that y_t is generated by a VAR(1) and Assumption 1 holds. We need to show that

$$\mathbb{P}_{ij}^{(h)} \neq \emptyset \Leftrightarrow \Pi_{ij,1}^{(h)} \neq 0. \tag{A.4}$$

Let $i, j \in \mathbb{N}_K : i \neq j$ and $h \in \mathbb{N}$. Then

$$\begin{aligned}
\mathbb{P}_{ij}^{(h)} \neq \emptyset &\Leftrightarrow \exists\ P = (e_1, \ldots, e_h) \\
&\Leftrightarrow \exists\ (i_k, j_k)_{k=1}^{h} :\ i_1 = i,\ j_h = j \\
&\quad a_{i_k, j_k} \neq 0, k = 1, \ldots, h, j_k = i_{k+1} \\
&\Leftrightarrow \exists\ P : \prod_{(l,m)\in P} a_{l,m} \neq 0 \\
&\Leftrightarrow \Pi_{ij,1}^{(h)} = \sum_{P \in \mathbb{P}_{ij}^{(h)}} \prod_{(l,m)\in P} a_{l,m} \neq 0. \qquad \text{(Assumption 1)}
\end{aligned}$$

□

A.1.3 Proof of Corollary 2

Assume that y_t is generated by a VAR(1) and Assumption 1 holds. We need to show for all C_k: If $y_i, y_j \in C_k$, then $y_i \in C(y_j)$ and $y_j \in C(y_i)$, i.e. the strongly connected components C_k are sets of mutual causal variables. We show this by picking a C_k and $y_i, y_j \in C_k$. Then,

$$\mathbb{P}_{ij}^{(h_1)} \neq \emptyset \wedge \mathbb{P}_{ji}^{(h_2)} \neq \emptyset \quad \text{for some } h_1, h_2 \geq 1,$$

i.e. there it at least one path from i to j and from j to i. This follows from Definition 2 of the SCCs. From Corollary 1 we have

$$\Pi_{ij}^{(h_1)} \neq 0 \wedge \Pi_{ji}^{(h_2)} \neq 0 \quad \text{for some } h_1, h_2 \geq 1.$$

The last result implies that $y_j \in C(y_i)$ and $y_i \in C(y_j)$ and completes the proof. □

A.1.4 Proof of Corollary 3

Assume that y_t is generated by a VAR(1) and Assumption 1 holds. Given Assumption 1 we only proof the second part, i.e. we show that

$$R(y_i) = C(y_i). \tag{A.5}$$

Let $y_k \in y$, where y denotes the vector of VAR variables. Then,

$$\begin{aligned}
y_k \in R(y_i) &\Leftrightarrow (y_k \in y_i) \vee (\exists h \in \mathbb{N} : \mathbb{P}_{ik}^{(h)} \neq \emptyset) && \text{(Def. of } R(y_i)) \\
&\Leftrightarrow (y_k \in y_i) \vee (\exists h \in \mathbb{N} : \Pi_{ik}^{(h)} \neq 0) && \text{(Corollary 1)} \\
&\Leftrightarrow y_k \in C(y_i), && \text{(Def. of } C(y_i))
\end{aligned}$$

which completes the proof. □

A.2 Data

We describe the data used in the empirical illustrations. Raw data for most series are obtained from the FRED database and Table A.1 shows the corresponding FRED mnemonics. We construct some variables from splicing two series in order to obtain long time series: As a measure for the exchange rate, we use the US/DM exchange rate (`EXGEUS`) until 1998Q4. From 1999Q1 we use `EXUSEU` and splice both series accordingly. The resulting variable is called `EXCH`. We follow McCracken & Ng (2015) and use `OILPRICE` (Spot Oil Price) until 1985Q4 and `MCOILWTICO` (Crude Oil Price, Cushing) since 1986Q1, since the former series has been discontinued. The resulting series is labeled `POIL` in our data set. To obtain a crude measure of stock market volatility, we simply use the squared stock market returns, since the time series of volatility indices in FRED are rather short. This series is called `VOLA`. Seasonally adjusted series have been taken from FRED where necessary. The time series on Government Debt (`GFDEBTN`) has been seasonally adjusted by the authors using X-ARIMA-13. The resulting series is `GFDEBTNSA`. The Euro area time series have been added using the update 15 to the AWM database maintained at the ECB. The AWM mnemonics for the real GDP, CPI, and a short-term interest rates are `YER`, `HICP`, and `STN`. `HICP` has been seasonally adjusted by the authors using X-ARIMA-13. We use `EMUGDP`, `EMUHICPSA`, and `EMURS` to denote the three Euro area variables.

The last columns in Table A.1 lists the transformation codes 1-6, corresponding to the following transformations of the series y_t: (1) no transformation, y_t, (2) Δy_t, (3) $\Delta^2 y_t$, (4) $400 \times \log(y_t)$, (5) $400 \times \Delta \log(y_t)$, (6) $400 \times \Delta^2 \log(y_t)$.

Table A.1: Variables, Data Sources and Transformations

Name	Mnemonic	Transf.Code
Real GDP	GDPC96	5
CPI	CPIAUCSL	6
Federal Funds Rate	FEDFUNDS	2
Real Consumption	PCECC96	5
Real Government Consumption	GCEC1	5
Real Investment	GPDIC1	5
Real Exports	EXPGSC1	5
Real Imports	IMPGSC1	5
Change in Real Inventories	CBIC96	1
Unit Labor Cost	ULCNFB	5
Employment	PAYEMS	5
Unemployment Rate	UNRATE	2
Hours worked	HOHWMN02USQ065S	1
1-year T-Bill Rate	GS1	2
10-year T-Bill Rate	GS10	2
Corporate Bond Spread	AAAFFM	1
Lending Rate to NFCs	TB3MS	3
Mortgage Rate	MORTG	2
M1	M1SL	6
M2	M2SL	6
Government Debt GFDEBTNSA	GFDEBTN. seas.adj: X-13	5
Real Estate Loans	REALLN	5
Consumer Credits	TOTALSL	5
Commercial Loans	BUSLOANS	5
Dollar/Euro Exchange Rate (EXCH)	spliced from EXUSEU and EXUSEU	5
Effective Exchange Rate	NNUSBIS	5
Oil Price (POIL)	spliced from OILPRICE and MCOILWTICO	5
Commodity Prices	CUSR0000SAC	6
Consumer Prices (excl. food, energy)	CPILFESL	6
Producer Price Index	PPIACO	5
House Prices	USSTHPI	6
Real Housing Investment	PRFI	5
Total Share Prices	SPASTT01USQ661N	5
Volatility Index	VIXCLS	5
Capacity Utilization	CUMFNS	2
Consumer Confidence	CSCICP03USM665S	2
Industrial Confidence	BSCICP03USM665S	2
Purchasing Manager's Index	NAPM	1
Real GDP (Euro Area)	AWM mnemonic: YER	5
CPI (Euro Area)	AWM mnemonic: HICP, seas.adj: X-13	6
Short term interest rate (Euro Area)	AWM mnemonic: STN	2

Note: The table shows FRED and AWM database names together with the transformation codes. See Appendix A.2 for a detailed description of the transformations.

A.3 Additional Results

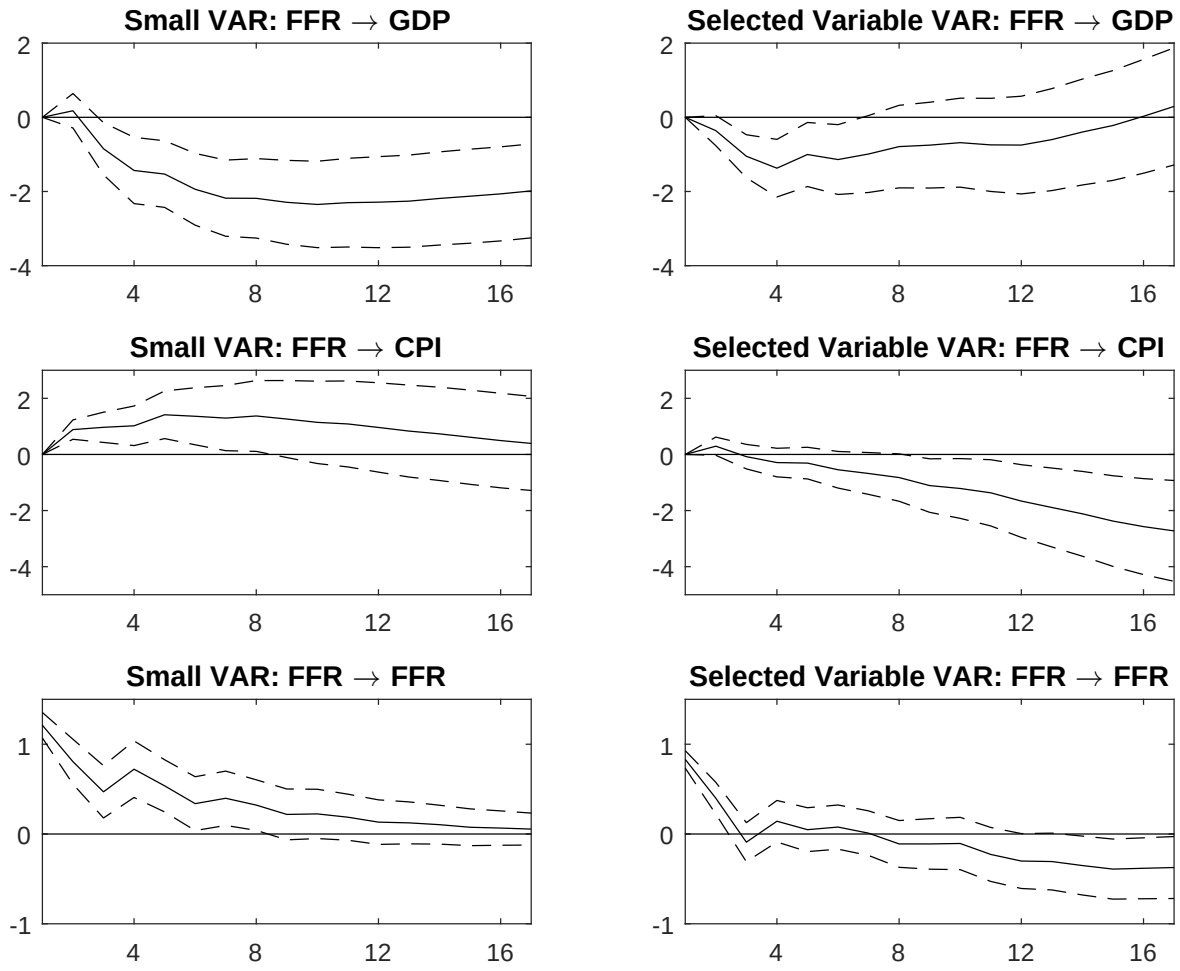


Figure A.1: Impulse responses in small VAR and selected variable VAR. Left: Responses to a shock in FFR in 3-variable (small) VAR(4) including GDP, FFR, and CPI. Right: Responses to a shock in FFR in 13-variable VAR(4) with 10 selected additional variables. Sample period: 1972Q1-2007Q4.

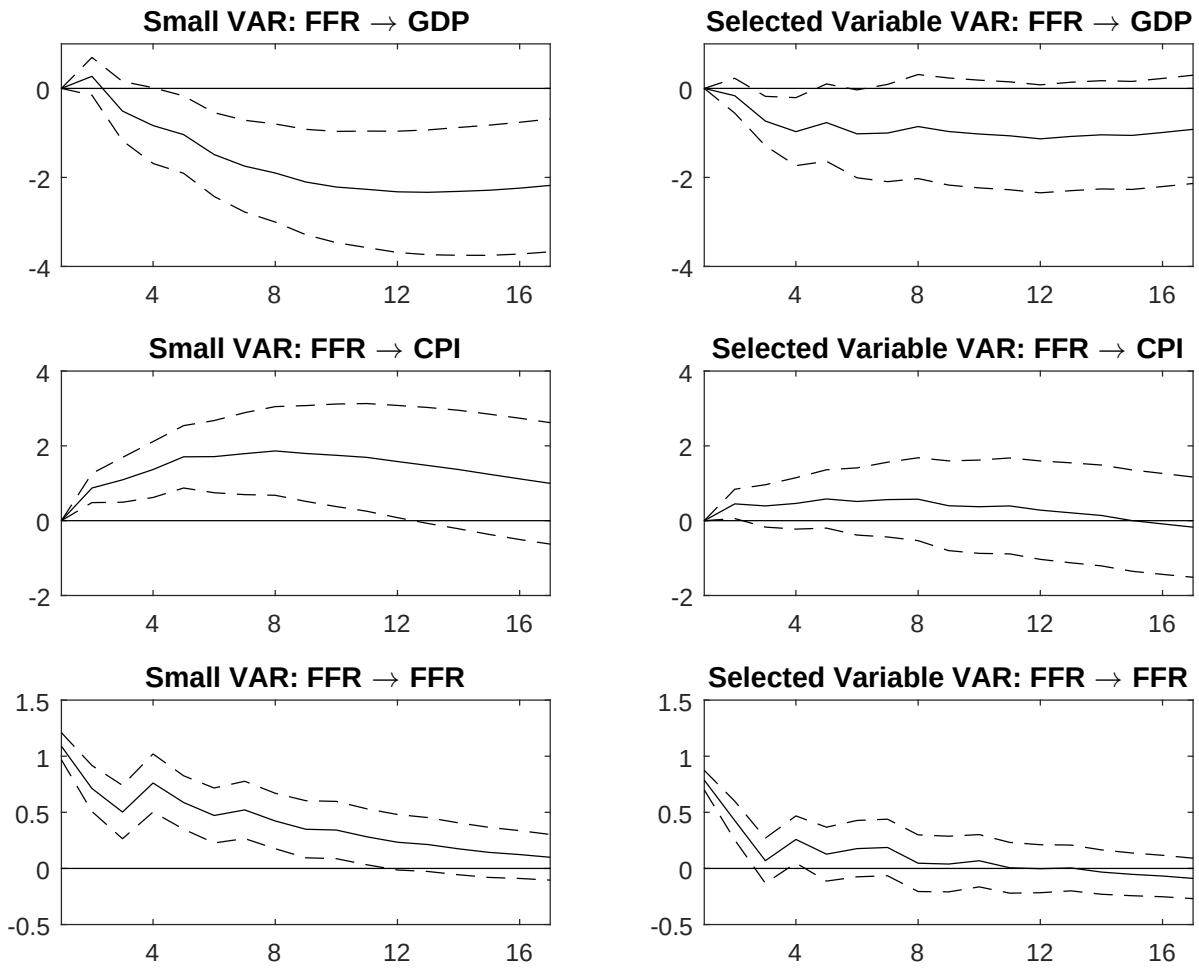


Figure A.2: Impulse responses in small VAR and selected variable VAR. Left: Responses to a shock in FFR in 3-variable (small) VAR(4) including GDP, FFR, and CPI. Right: Responses to a shock in FFR in 13-variable VAR(4) with 10 selected additional variables. Sample period: 1975Q3-2014Q4.

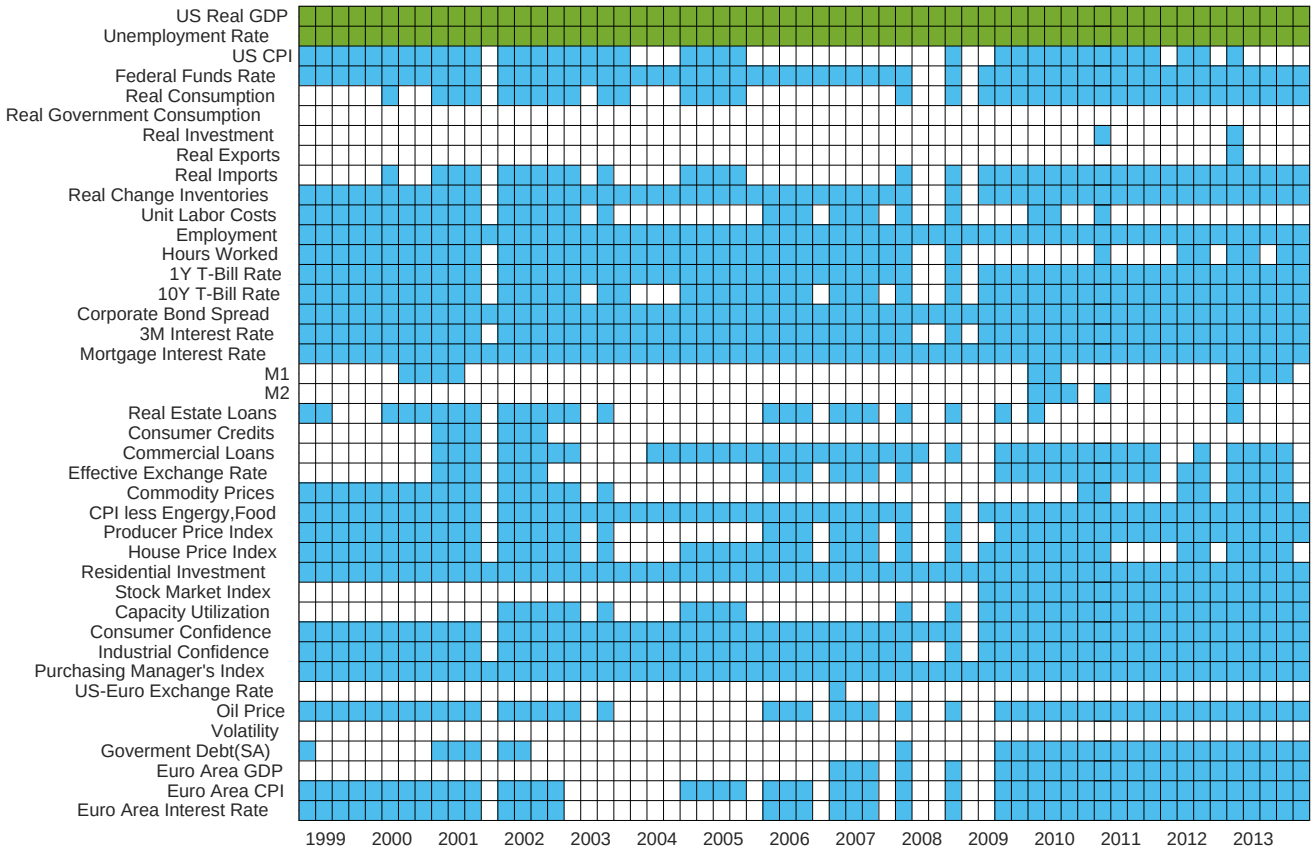


Figure A.3: Variable selection results: Variable of interest y^{I}: GDP growth and unemployment (green). Relevant variables as selected by graphical method (blue) and variables not selected (white). Sample period: 1975Q3-2014Q4.