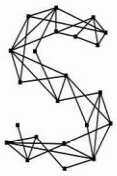


SYRTO

SYstemic Risk TOmography
Signals, Measurements, Transmission Channels, and Policy Interventions

Sparse Graphical Vector Autoregression: A Bayesian Approach

Daniel Felix Ahelegbey, Monica Billio, Roberto Casarin

SYRTO WORKING PAPER SERIES
Working paper n. 7 | 2015

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 320270.

Sparse Graphical Vector Autoregression: A Bayesian Approach

Daniel Felix Ahelegbey*, Monica Billio, Roberto Casarin

Department of Economics, Ca'Foscari University of Venice, Italy

Abstract

In high-dimensional vector autoregressive (VAR) models, it is natural to have large number of predictors relative to the number of observations, and a lack of efficiency in estimation and forecasting. In this context, model selection is a difficult issue and standard procedures may often be inefficient. In this paper we aim to provide a solution to these problems. We introduce sparsity on the structure of temporal dependence of a graphical VAR and develop an efficient model selection approach. We follow a Bayesian approach and introduce prior restrictions to control the maximal number of explanatory variables for VAR models. We discuss the joint inference of the temporal dependence, the maximum lag order and the parameters of the model, and provide an efficient Markov chain Monte Carlo procedure. The efficiency of the proposed approach is showed on simulated experiments and real data to model and forecast selected US macroeconomic variables with many predictors.

Keywords: High-dimensional Models, Large Vector Autoregression, Model Selection, Prior Distribution, Sparse Graphical Models, Bayesian Vector Autoregressive Models

JEL: C11, C15, C52, C55, E17

1. Introduction

High dimensional modeling and large dataset handling have recently gain attention in several fields, particularly in economics and finance. This has become necessary since useful information is often scattered among large number of variables. Building models that allow for extraction of these information from large dataset enhances a better understanding of the modern economic and financial system. Many researchers have shown that combining financial and macroeconomic variables to estimate large vector autoregressive (VAR) models produces better forecasts than standard approaches (see,

*Address: Department of Economics, Ca' Foscari University of Venice, Fondamenta San Giobbe 873/b, 30121, Venice, Italy. Corresponding author: Daniel Felix Ahelegbey, dfkahey@unive.it. Other contacts: billio@unive.it (Monica Billio); r.casarin@unive.it (Roberto Casarin).

March 18, 2015

Banbura et al., 2010; Carriero et al., 2013; Giannone et al., 2005; Koop, 2013; Stock and Watson, 2012). Many others, using datasets of a large number of financial institutions, have shown that the financial system has become highly interconnected and thus, can be represented as a network where linkages among agents sharing common structures play a fundamental role in contagion and the spread of systemic risk (see, Billio et al., 2012; DasGupta and Kaligounder, 2014; Diebold and Yilmaz, 2014; Hautsch et al., 2014; Huang et al., 2012).

In this paper we propose a new Bayesian model for multivariate time series of large dimension by combining graph-based notion of causality (see Lauritzen and Wermuth, 1989; Pearl, 1988; Whittaker, 1990), with the concept of sparsity (see, e.g. Box and Meyer, 1986). Graphical models have been applied in time series analysis for estimating causal structures in VAR models (see Corander and Villani, 2006; Demiralp and Hoover, 2003; Moneta, 2008; Swanson and Granger, 1997) and identification restrictions in structural VAR (Ahelegbey et al., 2015). They have received increasing attention as tools to represent interconnectedness and sources of contagion among financial institutions (see Ahelegbey and Giudici, 2014; Barigozzi and Brownlees, 2014; Billio et al., 2012; Diebold and Yilmaz, 2014; Merton et al., 2013). As described in the following, we contribute to the literature in many ways.

One of the key challenges of high-dimensional models is the complex interactions among variables and the inferential difficulty associated with handling large datasets. For instance, in large VAR models, econometricians encounter the curse of dimensionality problem due to high number of variables relative to the number of data points. The standard Bayesian VAR approach to this problem is to apply Minnesota prior by Doan et al. (1984), as a solution to overfitting. This approach is however inefficient to deal with the problem of indeterminacy (see Donoho, 2006), i.e. when the number of parameters in a system of equations exceeds the number of observations. Two common approaches to the indeterminacy issue discussed in the literature are based alternatively on dimension reduction or variable selection methodologies. For dimension reduction, dynamic factor models, factor augmented VAR and Bayesian model averaging have been extensively discussed and widely considered to extract useful information from a large number of predictors (see Bai and Ng, 2008; Bernanke et al., 2005; Geweke, 1977; Giannone et al., 2005; Koop and Potter, 2004; Stock and Watson, 2006). For variable selection, standard techniques have been applied to reduce the number of predictors, e.g., the Least Absolute Shrinkage and Selection Operator (LASSO) of Tibshirani (1996), and its variants, (see, e.g. Efron et al., 2004; Kyung et al., 2010; Park and Casella, 2008; Zou and Hastie, 2005). The method considered in this paper is related to the latter, thus to variable selection.

Variable selection is a fundamental problem in high-dimensional models, and this

is closely related to the possibility to describe the model with sparsity (Zhang et al., 2012). The idea of sparsity is associated with the notion that a large variation in the dependent variables is explained by a small proportion of predictors (Box and Meyer, 1986). Modeling sparsity has received attention in recent years in many fields, including econometrics, (see Elliott et al., 2013; Gefang, 2014; Korobilis, 2013). See also Belloni and Chernozhukov (2011) for an introduction to high-dimensional sparse econometric models.

This paper introduces and models sparsity in graphical VAR models of large dimension by dealing also with uncertainty in the lag order. It thus substantially extends the graphical VAR model, the inference approach and posterior approximation algorithm given in Ahelegbey et al. (2015). In most empirical analyses, researchers often overlook dependence among series when dealing with multi-equation regression models and large number of predictors, (see, e.g. Korobilis, 2013; Stock and Watson, 2014), since model selection is a difficult issue and such approach is often necessary to avoid the indeterminacy problem. However, this can be unsatisfactory in terms of interpretability and forecasting performance, since temporal dependence in the series is ignored. The graphical approach presented in this paper enables us to deal with this indeterminacy problem by exploiting sparsity to estimate the dynamic causal structure in large VAR models.

Many studies have considered several approaches to model sparse graphs (see, e.g. Carvalho et al., 2008; Jones et al., 2005; Scott and Carvalho, 2008; Shojaie and Michailidis, 2010). Also, there is an increasing interest in sparsity estimation for large VAR models (see, e.g. Davis et al., 2012; De Mol et al., 2008; Gefang, 2014; Kock and Callot, 2012; Medeiros and Mendes, 2012; Song and Bickel, 2011). We contribute to this literature by focusing on graphical VAR models from a Bayesian perspective with suitable prior specifications to deal with sparsity on the temporal dependence. More precisely, we build on the fan-in method of Friedman and Koller (2003) and propose a new approach to sparsity modeling. The idea of the fan-in is based on imposing a maximal number of predictors to ensure sparsity on the graph. Setting an a-priori hard fan-in might be too restrictive for large VAR applications. We therefore propose a prior distribution on the fan-in to allow for different prior information level about the maximal number of predictors for each equation of the VAR model. Thus, we allow for a random fan-in and adapt this prior distribution to the prior probability in variable selection problems. We show that this new prior distribution encourages sparsity on the graph taking into account the lag order. Since there is duality between prior and the penalty in the information criterion, our prior leads to a modified BIC for graphical model selection.

We also contribute to the literature on dynamic relationship identification. Here, we propose an efficient Markov Chain Monte Carlo (MCMC) algorithm to sample jointly,

the graph structure, the lag order and the parameters of the VAR model. Due to the uncertainty on the lag order, we propose an efficient MCMC algorithm that takes advantage of computational power through parallel simulation of the graph and lag order. Inference of the graph and lag order allows us to estimate only the parameters of the relevant explanatory variables.

We show the efficiency and study the performance of our approach through simulation examples and an application to forecast macroeconomic times series with large number of predictors. We consider the standard Lasso-type methods (i.e. LASSO and Elastic-Net) as benchmarks for comparing the identified causal structure and the forecast ability. We find evidence that our sparse graphical VAR model is more parsimonious than the LASSO and Elastic-Net. Furthermore, we find evidence of a gain in the predictive accuracy of our approach over the Lasso-type methods.

The paper is organized as follows: Section 2 presents the graphical concept for VAR models; Section 3 discusses prior distributions and focuses on the interaction between lag order and sparse graph prior distribution; Section 4 discusses the Bayesian inference scheme; Section 5 illustrates the simulation experiments; and Section 6 presents the empirical application.

2. Graphical VAR Models

Graphical models are statistical models that summarize the marginal and conditional independences among random variables by means of graphs (see Brillinger, 1996). The core of such models is representing relationships among variables in the form of graphs using nodes and edges, where nodes denote variables and edges show interactions. They can be represented in a more compact form by $(G, \theta) \in (\mathcal{G} \times \Theta)$, where G is a graph of relationships among variables, θ is the graphical model parameters, $\mathcal{G}$ is the space of the graphs and Θ is the parameter space.

Let X_t be $n \times 1$ vector of observations at time t and assume $X_t = (Y_t', Z_t')$, where Y_t, the $n_y \times 1$ vector of endogenous variables, and Z_t, a $n_z \times 1$, $n_z = (n - n_y)$ vector of exogenous predictors. In a VAR model with exogenous variables, the variables of interest Y_t, is determined by the equation

$$Y_t = \sum_{i=1}^{p} B_i X_{t-i} + \varepsilon_t \tag{1}$$

$t = 1, \ldots, T$, where ε_t is $n_y \times 1$ vector of errors, independent and identically normal, with mean zero and covariance matrix Σ_ε; p is the maximum lag order; B_i, $1 \leq i \leq p$ is $n_y \times n$ matrix of coefficients.

By interpreting (1) as a model with temporal dependence between explanatory and dependent variables, The VAR model can be expressed in a graphical framework (referred to as graphical VAR model), with a one-to-one correspondence between the coefficient matrices and a directed acyclic graph; if $B_{s,ij} \neq 0$ then there is a causal effect of X^j_{t-s} on Y^i_t, $1 \leq s \leq p$. Here we read X^i_t as realization of the i-th element of X at time t.

More formally, we define the relation $B_s = (G_s \circ \Phi_s)$, where G_s is a $n_y \times n$ binary connectivity matrix (also called adjacency matrix), Φ_s is a $n_y \times n$ coefficients matrix, and the operator $(\circ)$ is the element-by-element Hadamard's product (i.e., $B_{s,ij} = G_{s,ij}\ \Phi_{s,ij}$). Based on this definition, we identify a one-to-one correspondence between B_s and Φ_s conditional on G_s, such that $B_{s,ij} = \Phi_{s,ij}$, if $G_{s,ij} = 1$; and $B_{s,ij} = 0$, if $G_{s,ij} = 0$. The above relationship can be presented in a stacked matrix form as, $B = (G \circ \Phi)$, where $B = (B_1, \ldots, B_p)$, $G = (G_1, \ldots, G_p)$ and $\Phi = (\Phi_1, \ldots, \Phi_p)$, where each matrix is of dimension $n_y \times np$.

Let W_t be stacked lags of X_t, where $W_t = (X'_{t-1}, \ldots, X'_{t-p})'$ is of dimension $np \times 1$, with p as the lag order, and $V_t = (Y'_t, W'_t)'$ of dimension $(n_y + np) \times 1$. Suppose the joint, V_t, follows the distribution, $V_t \sim \mathcal{N}(0, \Omega^{-1})$, where $\Sigma = \Omega^{-1}$ is $(n_y + np) \times (n_y + np)$ is the covariance matrix. The joint distribution of the variables in V_t can be summarized with a graphical model, (G, θ), where $G \in \mathcal{G}$ consists of directed edges. In this paper, we focus on modeling directed edges from elements in W_t to elements in Y_t, capturing the temporal dependence among the observed variables. More specifically, $G_{ij} = 0$, means the i-th element of Y_t and j-th element of W_t are conditionally independent given the remaining variables in V_t, and $G_{ij} = 1$ corresponds to a conditional dependence between the i-th and j-th elements of Y_t and W_t respectively given the remaining variables in V_t. The graphical model parameter, $\theta \in \Theta$, consist the coefficients, capturing the strength and sign of the temporal dependence relationship described by G. Based on the above assumption, estimating the model parameters associated with G is equivalent to estimating the precision matrix, Ω, i.e $\theta = \Omega$. The relationship between the parameters of the VAR, $\{B, \Sigma_\varepsilon\}$, and that of the graphical model, Ω, is as follows.

Proposition 1. *Suppose the marginal distribution of* $W_t \sim \mathcal{N}(0, \Sigma_{ww})$ *and the conditional distribution of* $Y_t|W_t \sim \mathcal{N}(BW_t, \Sigma_\varepsilon)$. *There is a correspondence between* $\{B, \Sigma_\varepsilon\}$ *and* Ω, *such that given* Ω, *we obtain* $\Sigma = \Omega^{-1}$ *and* $\{B, \Sigma_\varepsilon\}$ *can be estimated by*

$$B = \Sigma_{yw}\Sigma^{-1}_{ww}, \qquad \Sigma_\varepsilon = \Sigma_{yy} - \Sigma_{yw}\Sigma^{-1}_{ww}\Sigma_{wy} \tag{2}$$

Also given $\{B, \Sigma_\varepsilon\}$ *and* Σ_{ww}, *the precision matrix* $\Omega = \Sigma^{-1}$ *of* (Y_t, W_t) *is estimated by*

$$\Omega = \begin{pmatrix} \Sigma^{-1}_\varepsilon & -\Sigma^{-1}_\varepsilon B \\ -B'\Sigma^{-1}_\varepsilon & \Sigma^{-1}_{ww} + B'\Sigma^{-1}_\varepsilon B \end{pmatrix} \tag{3}$$

Proof. See Appendix A.1. □

Following our definition, $B = (G \circ \Phi)$ and the results of Proposition 1, we identify the relationship between Ω and the dependence structure G, through the sub-matrix $(\Sigma_\varepsilon^{-1}B)$ of Ω. We denote $\Omega^G = \Sigma_\varepsilon^{-1}B$, defined on the space $\mathcal{M}(G)$, i.e. the set of precision matrices where elements of Ω^G corresponds to elements in G. Clearly, if the errors are assumed to be independent and normally distributed, Σ_ε is a diagonal matrix, which when normalized to identity leads to a one-to-one correspondence between B, Ω^G and G such that $B_{ij} = \Omega^G_{ij} = 0$ if $G_{ij} = 0$ and $B_{ij} = \Omega^G_{ij} \neq 0$ if $G_{ij} = 1$. In large VAR models estimation, most empirical papers follow the above assumption on the errors to estimate the model, (see, e.g. Kock and Callot, 2012; Stock and Watson, 2014).

In this paper we assume Σ_ε is a full matrix, i.e, the errors are correlated among the equations of the VAR. Estimating our graphical VAR model therefore involves: the temporal dependence graph, G, the maximum lag order, p, and the set of parameters in Ω which related to $\{B, \Sigma_\varepsilon\}$. Estimating all these jointly is challenging. However, following the Bayesian paradigm allows us to take into account uncertainties on the quantities of interest and inference on these through model averaging, (Giudici and Green, 1999; Madigan and York, 1995). The objective of this paper is to estimate jointly the lag order and graph from the observed time series, and to incorporate the inferred quantities to select the relevant variables to estimate the parameters of the model.

3. Sparse Bayesian Graphical VAR Models

In a system of linear equations where the number of parameters exceeds the number of observations, for instance in large VAR models, we face another problem referred to as indeterminacy, (see Donoho, 2006). Such systems can be modeled by exploiting sparsity. The description of our graphical VAR for high dimensional multivariate time series is completed with the elicitation of the prior distributions for the lag order p, a sparsity prior on the graph, and the prior on G and Ω.

3.1. Lag Order Prior Distribution

We allow for different lag orders for the different equations of the VAR model. We denote with p_i the lag order of the i-th equation. We assume for each p_i, $i = 1, \ldots, n_y$, a discrete uniform prior on the set $\{\underline{p}, \ldots, \bar{p}\}$

$$P(p_i) = \frac{1}{(\bar{p} - \underline{p} + 1)} \mathbb{I}_{\{\underline{p}, \ldots, \bar{p}\}}(p_i) \tag{4}$$

where $\mathbb{I}_A(x)$ is the indicator function, that is unity if $x \in A$ and zero otherwise. This is a standard choice in AR model selection problems (e.g., see Casarin et al. (2012)). Alternatively, the lag order can be assumed to follow a truncated Poisson distribution with mean λ and maximum $\bar{p}$ (see Vermaak et al. (2004)), or a discretized Laplace distribution (see Ehlers and Brooks (2004)). Our choice of discrete uniform distribution is fairly informative since $\underline{p}$ and $\bar{p}$ are defined a-priori following standard applications and assigns equal weights to all possible values of p_i. For instance, in estimating VAR models with monthly (quarterly) time series, the standard lag selection procedure often consider $\underline{p} = 1$ and $\bar{p} = 12$ ($\bar{p} = 4$). The alternative lag order prior distributions are more informative and assigns different weights to the possible values of p_i.

3.2. Standard Graph Prior Distribution

Most of the literature on graphical models takes the prior for a graph G with n variables as uniform over all the relevant graphs, i.e., $P(G) = |\mathcal{G}|^{-1}$, where $|\mathcal{G}|$ is the cardinality of $\mathcal{G}$, (see Geiger and Heckerman, 2002; Giudici and Castelo, 2003). This can be attributed to the fact that the number of possible graphs increases super-exponentially with the number of variables, and there is difficulty in calculating the number of possible graphs. Assuming uniform probabilities for all graphs, however, does not ensure sparsity. Thus, many authors have discussed several approaches to penalize globally or locally "dense" graphs (see, e.g. Carvalho et al., 2008; Jones et al., 2005; Scott and Carvalho, 2008; Wang, 2010). See also Telesca et al. (2012) and Shojaie and Michailidis (2009) for the use of explicit information prior to improve the estimation of the graph structure.

Friedman and Koller (2003) proposed a factorization of the graph prior into equation specific terms for DAG models. As argued by the authors, setting an upper bound on the number of explanatory variables for each dependent variable encourages sparsity on the graph. This bound is referred to as the fan-in restriction in the graphical model literature. Let m be the maximum number of explanatory variables for each equation. Restricting the graph model selection to at most f explanatory variables instead of m, $f < m$, reduces the number of possible sets from $\mathcal{O}(2^m)$ to $\binom{m}{f}$, where $\binom{n}{k}$ is the binomial coefficient. A uniform choice on the latter set yields a graph prior

$$P(G) = \prod_{i=1}^{n} P(\pi_i) \propto \prod_{i=1}^{n} \binom{n-1}{|\pi_i|}^{-1} \tag{5}$$

where $\pi_i = \{j = 1, \ldots, n : G_{ij} = 1\}$ is the set of explanatory variables of the i-th equation, $|\pi_i|$ is the number of elements in π_i, and $P(\pi_i)$ is the local graph prior of the i-th equation.

Jones et al. (2005) discussed a prior distribution for penalizing the inclusion of additional edges in dense graphs given by

$$P(G|\gamma) = \gamma^k (1-\gamma)^{m-k} \tag{6}$$

where m is the maximum number of edges and k represents the number of edges in the graph. In their application, the authors use a Bernoulli prior on each edge inclusion with parameter $\gamma = 2/(n-1)$ and set $m = \binom{n}{2}$.

For choices of the prior distribution on γ in the beta family, Scott and Carvalho (2008) showed that γ can be explicitly marginalized out. The uniform prior on γ gives a marginal prior inclusion probability of $1/2$ for all edges and yields model probabilities

$$P(G) = \frac{1}{(m+1)} \binom{m}{k}^{-1} \tag{7}$$

3.3. Sparsity Prior Distribution

We build on the fan-in approach of Friedman and Koller (2003) by introducing a prior distribution on the fan-in to allow for different prior information level about the maximal number of predictors for each equation of the VAR model.

In a multivariate dynamic models with n variables and a lag order p, the number of possible predictors is np. Given that each series has T number of observations, then the number of observations of the model is $T-p$. In large VAR models, it is often natural that the number of predictors is greater than the number of observations, i.e. $np > T-p$. When this happens, we expect that each equation has at most $T-p$ predictors to efficiently estimate the model. In cases where $T-p > np$, we expect that each equation has at most np predictors. Thus the maximal number of explanatory variables required to efficiently estimate a high dimensional model is given by $m_p = \min\{np, T-p\}$. Setting an a-priori hard fan-in (see Friedman and Koller, 2003) might be too restrictive for large VAR applications.

We denote with $\bar{\eta}$, $0 \leq \bar{\eta} \leq 1$, the measure of the maximal density, i.e. the fraction of the predictors that explains the dependent variables. Thus the level of sparsity is given by $(1-\bar{\eta})$. We set the fan-in to $f = \lfloor \bar{\eta} m_p \rfloor$, where f is the largest integer less than $\bar{\eta} m_p$. To allow for different levels of sparsity for the equations in the VAR model, we assume a prior distribution on the maximal density for the different equations. We denote $\bar{\eta}_i$ the maximal density of the i-th equation and assume the prior on $\bar{\eta}_i$, given lag order p_i is beta distributed with parameters $a, b > 0$, $\bar{\eta}_i | p_i \sim \mathcal{B}e(a,b)$, on the interval $[0,1]$

$$P(\bar{\eta}_i | p_i) = \frac{1}{B(a,b)} \bar{\eta}_i^{a-1} (1-\bar{\eta}_i)^{b-1} \tag{8}$$

This leads to random fan-in's for the different equations in the VAR model. Note that the fan-in, f_i, must be directly related to the number of selected predictors in each equation and indirectly related to the number of observations of the model.

3.4. Our Graph Prior Distribution

We define the graph prior for each equation in the VAR model conditional on the sparsity prior. We refer to the prior on the graph of each equation as the local graph prior, denoted by $P(\pi_i|p_i, \gamma, \bar{\eta}_i)$. Following (Scott and Berger, 2010), we consider the inclusion of predictors in each equation as exchangeable Bernoulli trials with prior probability

$$P(\pi_i|p_i, \gamma, \bar{\eta}_i) = \gamma^{|\pi_i|}(1-\gamma)^{np_i-|\pi_i|}\ \mathbb{I}_{\{0,\dots,f_i\}}(|\pi_i|) \tag{9}$$

where $\gamma \in (0,1)$ is the Bernoulli parameter, $|\pi_i|$ is the number of selected predictors out of np_i and $f_i = \lfloor \bar{\eta}_i m_p \rfloor$ is the fan-in restriction for the i-th equation. We assign to each variable inclusion a prior probability, $\gamma = 1/2$, which is equivalent to assign the same prior probability to all models with predictors less than the fan-in f_i, that is

$$P(\pi_i|p_i, \bar{\eta}_i) = \frac{1}{2^{np_i}}\ \mathbb{I}_{\{0,\dots,f_i\}}(|\pi_i|) \tag{10}$$

Alternatively, a uniform prior on γ gives to each variable a marginal prior inclusion probability of $1/2$ and a local graph prior (Foygel and Drton, 2011) given by

$$P(\pi_i|p_i, \bar{\eta}_i) = \binom{np_i}{|\pi_i|}^{-1}\ \mathbb{I}_{\{0,\dots,f_i\}}(|\pi_i|) \tag{11}$$

3.5. Parameter Prior Distribution

There are two main approaches to define parameter priors for graphical models, however a common feature to these approaches is that both are graph conditional parameter priors. On one hand is a vast work on Gaussian DAG models discussing a list of conditions that permits an unconstrained precision matrix Ω (see, e.g. Consonni and Rocca, 2012; Geiger and Heckerman, 2002; Heckerman, 1998; Heckerman and Chickering, 1995; Heckerman and Geiger, 1994). On the other hand is a vast publication on Gaussian decomposable undirected graphical (UG) models with constraints on the precision matrix Ω (see, e.g. Carvalho and Scott, 2009; Lenkoski and Dobra, 2011; Roverato, 2002; Wang and Li, 2012). Note that, an unconstrained Ω characterizes a complete Gaussian DAG or UG model, i.e. a graph with no missing edges. The standard parameter prior for Gaussian DAG models with zero expectations is a Wishart distribution, whereas that of UG models is a G-Wishart distribution. A consequence of the DAG conditional parameter prior is that, once we specify the parameter prior for one complete DAG model, all other

priors can be generated automatically (see Consonni and Rocca, 2012). In this paper, we follow the DAG model framework since it allows us to marginalize out Ω analytically thereby focusing on the structure inference, and the estimation of the model parameters given the structure of dependence (see Section 4 for details).

Following Geiger and Heckerman (2002), we assume a prior distribution on the unconstrained precision matrix, Ω, conditional on any complete DAG, G, for a given lag order p, is Wishart distributed, with probability density function

$$P(\Omega|p,G) = \frac{1}{K_d(\nu,S_0)}|\Omega|^{\frac{(\nu-d-1)}{2}}\text{etr}\Big(-\frac{1}{2}\Omega S_0\Big) \tag{12}$$

where $\text{etr}(A) = \exp\{\text{tr}(A)\}$ and $\text{tr}(A)$ is the trace of a square matrix A, ν is the degree of freedom parameter, S_0 is a $d \times d$ symmetric positive definite matrix, with $d = n_y + np$, the size of the vector of stacked dependent and explanatory variables of the model. The normalizing constant is:

$$K_d(\nu,S_0) = 2^{\frac{\nu d}{2}}|S_0|^{-\frac{\nu}{2}}\Gamma_d\Big(\frac{\nu}{2}\Big) \tag{13}$$

where $\Gamma_d(a) = \pi^{\frac{d(d-1)}{4}}\prod_{i=1}^{d}\Gamma\Big(a+\frac{1-i}{2}\Big)$ is the multivariate gamma function and $\Gamma(\cdot)$ denotes the gamma function.

Based on the assumption that the conditional distribution of the dependent variables given the set of predictors, is described by equation (1), with parameters $\{B,\Sigma_\varepsilon\}$, we assume the prior distribution on $(B,\Sigma_\varepsilon|p,G)$ is an independent normal-Wishart (see, e.g. Geiger and Heckerman, 2002; Heckerman and Geiger, 1994). This is one of the prior distributions extensively applied in the Bayesian VAR literature. Specifically, we assumed the coefficients matrix B is independent and normally distributed, $B|p,G \sim \mathcal{N}(\underline{B},\underline{V})$, and Σ_ε^{-1} is Wishart distributed, $\Sigma_\varepsilon^{-1} \sim \mathcal{W}(\underline{\nu},\underline{S}/\underline{\nu})$. The prior expectation, $\underline{B} = \underline{0}_{n_y \times np}$, is a zero matrix, and the prior variance of the coefficient matrix, $\underline{V} = I_{np\times np}$, is an identity matrix. Also, the prior expectation of $\Sigma_\varepsilon = \frac{1}{\underline{\nu}}\underline{S}$ where is $\underline{S}$ is $n_y \times n_y$ positive semi-definite matrix and $\underline{\nu} > n_y - 1$ is the degrees of freedom.

4. Bayesian Inference

We define G_s as $n_y \times n$ binary connectivity matrix that captures the temporal relationship of variables at time $t-s$ with the variables at time t. We denote with $\vec{G}_p = (G_1,\ldots,G_p)$ as stacked $G_1,\ldots,G_p$, such that $\vec{G}_p$ is of dimension $n_y \times np$. We then define $\vec{G}_{p,i},\ i = 1,\ldots,n_y$ as the local graph associated with the i-th equation which is the i-th row of $\vec{G}_p$. We proceed under the assumption that the series of dependent and explanatory variables is jointly Gaussian, $\mathcal{N}(0,\Omega^{-1})$. Moreover, conditional on the lag

order p, any complete DAG, $\vec{G}_p$, and an unconstrained precision matrix Ω, the likelihood function is given by

$$P(\mathcal{X}|p, \vec{G}_p, \Omega) = (2\pi)^{-\frac{dT_0}{2}} |\Omega|^{\frac{T_0}{2}} \text{etr}\Big(-\frac{1}{2}\Omega\hat{S}\Big) \tag{14}$$

where $\hat{S} = \sum_{i=1}^{T_0} V_t V_t'$, sum of squares matrix of dimension $d \times d$. A nice feature of the unconstrained parameter prior in the DAG mode framework is that it allows for integrating out analytically, the precision matrix, Ω, with respect to its prior to obtain a marginal likelihood function for any DAG, $\vec{G}_p$ with lag p given by

$$P(\mathcal{X}|p, \vec{G}_p) = \int P(\mathcal{X}|p, \vec{G}_p, \Omega)\ P(\Omega|p, \vec{G}_p)\ d\Omega = \frac{K_d(\nu + T_0, S_0 + \hat{S})}{(2\pi)^{\frac{dT_0}{2}} K_d(\nu, S_0)} \tag{15}$$

where $T_0 = T - p$, S_0 and $S_0 + \hat{S}$ are the prior and posterior sum of square matrices, which when normalized are $\underline{\Sigma} = \frac{1}{\nu} S_0$ and $\bar{\Sigma} = \frac{1}{\nu + T_0}(S_0 + \hat{S})$ respectively. Geiger and Heckerman (2002) outlined conditions for the integral in equation (15) to be analytically tractable and to have a close form expression that can be factorized into local marginal likelihoods. A key assumption is that the parameters must be independent within and across equations. In VAR models, the errors are correlated across equations which makes the factorization of (15) quite problematic. Following the idea of large-sample approximation by Kass et al. (1988) and Chickering and Heckerman (1997), we consider the errors of a large VAR model as unobserved variables which can be ignored when dealing with large datasets (see, e.g. Stock and Watson, 2014). Based on this consideration and the assumption that the coefficients in B are independent a-priori within and across equations, we approximate (15) with a pseudo-marginal likelihood given by the product of local densities

$$P(\mathcal{X}|p, \vec{G}_p) \approx \prod_{i=1}^{n_y} P(\mathcal{X}|p_i, \vec{G}_p(y_i, \pi_i)) = \prod_{i=1}^{n_y} \frac{P(\mathcal{X}^{(y_i, \pi_i)}|p_i, \vec{G}_p)}{P(\mathcal{X}^{(\pi_i)}|p_i, \vec{G}_p)} \tag{16}$$

where $\vec{G}_p(y_i, \pi_i)$ is the local graph of the i-th equation with y_i as dependent variable and π_i as the set of predictors; $\mathcal{X}^{(y_i, \pi_i)}$ is the sub-matrix of $\mathcal{X}$ consisting of y_i and π_i; and $\mathcal{X}^{(\pi_i)}$ is the sub-matrix of π_i. This approximation allows us to develop search algorithms to focus on local graph estimation. More specifically, a Markov chain Monte Carlo (MCMC) algorithm using the global score would be less efficient in exploration since the global score would be insensitive to the proposal of edge deletion or addition. Thus, the approximation allows the chain to explore the graph locally at equation level. The pseudo-likelihood has been used in MCMC by Zhou and Schmidler (2009) to circumvent

the intractable normalizing constant problem in random fields. See also Andrieu and Roberts (2009); Maclaurin and Adams (2014) for one of the approximated likelihood in MCMC. The closed form of (16) is given by

$$P(\mathcal{X}^{d_i}|p_i,\vec{G}_p) = \pi^{\frac{-T_i|d_i|}{2}}\ \frac{|\bar{\Sigma}_{d_i}|^{-\frac{(\nu+T_i)}{2}}}{|\underline{\Sigma}_{d_i}|^{-\frac{\nu}{2}}}\ \prod_{i=1}^{|d_i|}\frac{\Gamma\Big(\frac{\nu+T_i+1-i}{2}\Big)}{\Gamma\Big(\frac{\nu+1-i}{2}\Big)} \tag{17}$$

where $d_i \in \big\{(y_i,\pi_i),\ \pi_i\big\}$, and $\mathcal{X}^{d_i}$ is a sub-matrix of $\mathcal{X}$ consisting of $|d_i|\times T_i$ realizations, where $|d_i|$ is the dimension of d_i, $T_i = T - p_i$, $|\underline{\Sigma}_{d_i}|$ and $|\bar{\Sigma}_{d_i}|$ are the determinants of the prior and posterior covariance matrices associated with d_i.

4.1. Posterior Approximation

Inferring jointly the lag and the graph allows for selecting the relevant predictors to estimate the model parameters (B,Σ_ε). In order to approximate the posterior distributions of the equations of interest, the standard approach is to consider a collapsed Gibbs sampling. At the j-th iteration, the sampler consists of the following steps:

1. Sample jointly, $p^{(j)}$, $\bar{\eta}^{(j)}$ and $\vec{G}_p^{(j)}$ from $P(p,\bar{\eta},\vec{G}_p|\mathcal{X})$.
2. Estimate $B^{(j)}$ and $\Sigma_\varepsilon^{(j)}$ directly from $P(B,\Sigma_\varepsilon|p^{(j)},\vec{G}_p^{(j)},\mathcal{X})$.

As regards to the first step, standard MCMC algorithms (Madigan and York, 1995) such as Gibbs sampling and Metropolis-Hastings (MH) apply only to model selection with fixed dimensions. In model selection problems with unknown lag order, the dimension of the model varies with the lag order. The algorithm extensively applied for this problem is the reversible jump (RJ) MCMC (Green, 1995). In graphical models especially, the space of possible graphs increases super-exponentially with the number of variables (Chickering et al., 2004). Therefore, sampling from a distribution on a union of varying graph dimension using the RJ algorithm will require a higher number of iterations to thoroughly search the space of all possible graphs. In our graphical VAR, the inferential difficulty increases due to the random fan-in restriction.

We propose an alternative algorithm for sampling the graph taking into consideration the random fan-in and estimating the lag order. At the j-th iteration of the Gibbs, we consider for each equation $i = 1,\dots,n_y$ and each lag order $p_i = \underline{p},\dots,\bar{p}$, a sample of $\bar{\eta}_i^{(j)}$ and $\vec{G}_{p,i}^{(j)}$ from $P(\bar{\eta}_i,\vec{G}_{p,i}|p_i,\mathcal{X}) \propto P(\bar{\eta}_i|p_i)P(\pi_i|p_i,\bar{\eta}_i)P(\mathcal{X}|p_i,\vec{G}_{p,i})$. By conditioning on each possible lag order, we are able to apply standard MCMC algorithm thereby avoiding movement between models of different dimensions since the dimension is fixed for each lag. After J iterations, we average the draws, $\vec{G}_{p,i}^{(j)}$, over J and estimate $\hat{\vec{G}}_{p,i}$, for each $p_i = \underline{p},\dots,\bar{p}$, using the criterion discussed in Appendix B.2. This procedure estimates the local graph for the possible lags of $p_i \in \{\underline{p},\dots,\bar{p}\}$. Next, we find $(\hat{p}_i,\hat{G}_{\hat{p},i})$ which

minimizes a penalized local log-likelihood (BIC) score given in (22). Given the estimated graph and lag order, $(\hat{p}_i, \hat{G}_{\hat{p},i})$, we select the relevant predictors for each equation to estimate the model parameters (B, Σ_ε).

4.2. Graphical Model Selection

Graphical model selection is a challenge since the dimension of the graph space to explore increases super-exponentially with the number of variables. In this paper we apply MCMC and build on the MCMC algorithm described in Grzegorczyk and Husmeier (2011); Madigan and York (1995). Our algorithm differs from that of the above mentioned papers in two aspects: the initialization and the inclusion of the random fan-in restriction.

As regards to the initialization, we propose a strategy which improves the mixing of the chain. In MCMC search algorithms the space exploration crucially depends on the choice of the starting point of the chain. A set of burn-in chain iterations is often used to have a good starting point. However, Brooks et al. (2011) pointed out that any sample that is believed to be representative of the equilibrium distribution is an equally good starting point. In view of this, we propose an initialization which extracts variables (and their lags) with reliable information to improve predictions of the dependent variables. Let $\vec{G}_{p,i}$ denote the local graph of the i-th equation, $\mathbf{V}^i_{p,x}$, the vector of all possible explanatory variables with lags up to p for each equation, with $p \in [\underline{p}, \ldots, \bar{p}]$, and $\mathbf{V}_y$, the vector of dependent variables. We run the following steps:

1. Initialize the graph $\vec{G}_p$ as $n_y \times np$ zero matrix, i.e, $\vec{G}_{p,i}$ is $1 \times np$ zero vector.
2. For each equation, $i = 1, \ldots, n_y$:

   2a. Test whether or not predictions of $y_i \in \mathbf{V}_y$ is improved by incorporating information from each $x_k \in \mathbf{V}^i_{p,x}$, i.e, $P(y_i|x_k) > P(y_i)$. Following a Minnesota type of prior, we assume recent lags (specifically lag 1) of dependent variables are more reliable to influence current realizations. Based on this idea, we set $\vec{G}_p(y_i, x_k) = 1$ if x_k is equal to lag 1 of y_i, and retain x_k in $\mathbf{V}^i_{p,x}$.

   2b. For x_k not equal to lag 1 of y_i, we compare the probability of the null hypothesis, $H_0 = P(\mathcal{X}|p_i, \vec{G}_p(y_i, \emptyset))$, where $\emptyset$ denote the empty set, against the probability of the alternative, $H_1 = P(\mathcal{X}|p_i, \vec{G}_p(y_i, \{x_k\}))$. If $H_1 > H_0$, we reject the null, set $\vec{G}_p(y_i, x_k) = 1$ and retain x_k in $\mathbf{V}^i_{p,x}$. If $H_1 \leq H_0$, we set $\vec{G}_p(y_i, x_k) = 0$ and remove x_k from $\mathbf{V}^i_{p,x}$.
3. We then denote $N_p(\pi_i)$ as the set of variables, $x'_k s$, retained in $\mathbf{V}^i_{p,x}$.

In our experience, the above initialization provides a good starting point for graphical model selection. See Figure B.5 for a comparison of the convergence diagnostics of a random initialization MCMC and our initialization for the graph simulation. Using the

set $N_p(\pi_i)$ of candidate predictors of the dependent variable of the i-th equation, we start our MCMC search algorithm. We proceed with the local causal search by investigating the combination of variables in $N_p(\pi_i)$ that produces the highest scoring local graph(s).

As regards the inclusion of the random fan-in restriction, we denote with $m_p = \min\{np, T-p\}$, the maximal number of predictors required to efficiently estimate the model, for $p \in [\underline{p}, \ldots, \bar{p}]$. At the j-th iteration, let $\vec{G}_{p,i}^{(j-1)}$ be the current local graph and $\pi_i^{(j-1)}$, the current set of predictors in $\vec{G}_{p,i}^{(j-1)}$, then for each equation, $i = 1, \ldots, n_y$, the Gibbs iterates the following steps:

1. Draw the sparsity parameter for the forward proposal probability, $\bar{\eta}_i^{(*)}$ from a $\mathcal{B}e(a,b)$ and set the fan-in $f_i^{(*)} = \lfloor m_p \bar{\eta}_i^{(*)} \rfloor$.
2. If the number of elements in $\pi_i^{(j-1)}$ is less than the fan-in, i.e. $|\pi_i^{(j-1)}| < f_i^{(*)}$, then randomly draw a x_k from the set of candidate predictors, $N_p(\pi_i)$, and add or remove the edge between y_i and x_k, i.e. $\vec{G}_p^{(*)}(y_i, x_k) = 1 - \vec{G}_p^{(j-1)}(y_i, x_k)$. Here we set the forward proposal probability to $Q(\vec{G}_{p,i}^{(*)} | \vec{G}_{p,i}^{(j-1)}, \bar{\eta}_i^{(*)}) = 1/|N_p(\pi_i)|$. If $|\pi_i^{(j-1)}| \geq f_i^{(*)}$, then randomly draw a variable, x_k, from the current set of predictors, $\pi_i^{(j-1)}$, and remove the edge between y_i and x_k, i.e. $\vec{G}_p^{(*)}(y_i, x_k) = 0$. In this case, the forward proposal probability is $Q(\vec{G}_{p,i}^{(*)} | \vec{G}_{p,i}^{(j-1)}, \bar{\eta}_i^{(*)}) = 1/|\pi_i^{(j-1)}|$.
3. To obtain the reverse proposal probability, we denote $\pi_i^{(*)}$, the set of predictors in $\vec{G}_{p,i}^{(*)}$ taking into consideration the changes made in step 2.
4. Next, we draw the sparsity parameter for the reverse proposal probability, $\bar{\eta}_i^{(**)}$ from a $\mathcal{B}e(a,b)$ and set $f_i^{(**)} = \lfloor m_p \bar{\eta}_i^{(**)} \rfloor$.
5. If $|\pi_i^{(*)}| < f_i^{(**)}$, the reverse move will involve a random draw of a variable from $N_p(\pi_i)$ to add or delete from $\vec{G}_{p,i}^{(*)}$. Thus, the reverse proposal probability is given by $Q(\vec{G}_{p,i}^{(j-1)} | \vec{G}_{p,i}^{(*)}, \bar{\eta}_i^{(**)}) = 1/|N_p(\pi_i)|$. If $|\pi_i^{(*)}| \geq f_i^{(**)}$, the reverse will randomly draw a variable from $\pi_i^{(*)}$ to delete from $\vec{G}_{p,i}^{(*)}$. The reverse proposal probability in this case is given by $Q(\vec{G}_{p,i}^{(j-1)} | \vec{G}_{p,i}^{(*)}, \bar{\eta}_i^{(**)}) = 1/|\pi_i^{(*)}|$.
6. From equation (10), the ratio of the local graph priors simplifies to 1 and the acceptance probability is given by $A(\vec{G}_{p,i}^{(*)}, \bar{\eta}_i^{(*)} | \vec{G}_{p,i}^{(j-1)}, \bar{\eta}_i^{(**)}) = \min\{1,\ R_A\}$ where

$$R_A = \left\{ \frac{P(\mathcal{X}|p_i, \vec{G}_{p,i}^{(*)})}{P(\mathcal{X}|p_i, \vec{G}_{p,i}^{(j-1)})} \ \frac{Q(\vec{G}_{p,i}^{(j-1)} | \vec{G}_{p,i}^{(*)}, \bar{\eta}_i^{(**)})}{Q(\vec{G}_{p,i}^{(*)} | \vec{G}_{p,i}^{(j-1)}, \bar{\eta}_i^{(*)})} \right\} \tag{18}$$

   where $P(\mathcal{X}|p_i, \vec{G}_{p,i}) = P(\mathcal{X}|p_i, \vec{G}_p(y_i, \pi_i))$, and can be computed from equations (16) and (17). Note that without the fan-in restriction, the proposal distribution is symmetric, thus, the prior and inverse proposal ratio in (18) simplifies to 1.
7. Sample $u \sim \mathcal{U}_{[0,1]}$ and if $u < \min\{1, R_A\}$, then accept changes made in the local graph and set $\vec{G}_{p,i}^{(j)} = \vec{G}_{p,i}^{(*)}$, otherwise reject changes and set $\vec{G}_{p,i}^{(j)} = \vec{G}_{p,i}^{(j-1)}$.

A description of the pseudo-code for the graph selection is given in Appendix D.

4.3. Duality between Priors and Penalties

Thanks to the duality between prior distributions and the penalization of likelihood functions, it is possible to define an information criterion for choosing the optimal lag order and estimating the graph for graphical VAR models. This criterion has been used in the first step of our Gibbs sampler (see Section 4.1) and is defined as:

$$(\hat{p}, \hat{G}) \;=\; \arg\max_{p, \vec{G}_p} P(p) P(\vec{G}_p|p) P(\mathcal{X}|p, \vec{G}_p). \tag{19}$$

Several authors have considered extensions of the BIC to sparse model selection problems (see Bogdan et al., 2004; Chen and Chen, 2008; Foygel and Drton, 2011). Our extension allows for a more stringent penalty to address the tendency of the BIC to select large size models when dealing with high-dimensional data. To define our lag and graph selection criteria, we proceed by integrating out the hyper-parameter, $\bar{\eta}_i$, analytically as follows.

Proposition 2. *For choices of the prior distribution $P(\bar{\eta}_i|p_i)$ as beta distributed in (8), and $P(\pi_i|p_i, \bar{\eta}_i)$ according to (10), with $|\pi_i| = k$, $\bar{\eta}_i$ can be explicitly marginalized out as*

$$P(\pi_i|p_i) = \frac{1}{2^{np_i}} \sum_{j=0}^{m_p - 1} \mathbb{I}_{\{0,\ldots,j\}}(|\pi_i|) \Big(I_{\frac{j+1}{m_p}}(a, b) - I_{\frac{j}{m_p}}(a, b) \Big) \tag{20}$$

where $I_z(a, b) = \int_0^z (B(a, b))^{-1} (\bar{\eta}_i)^{a-1} (1 - \bar{\eta}_i)^{b-1} \, d\bar{\eta}_i$, is the incomplete beta function (Abramowitz and Stegun, 1964, p. 263).

Proof. See Appendix A.2. □

Corollary 4.1. A uniform prior on $\bar{\eta}_i$, means $a = b = 1$ and yields

$$P(\pi_i|p_i) = \frac{1}{2^{np_i}} \left(1 - \frac{|\pi_i|}{m_p}\right) \tag{21}$$

Proof. See Appendix A.3. □

Proposition 3. *Let $P(\pi_i|p_i)$ be the local graph prior given in (21) evaluated at the values of π_i such that $|\pi_i| = k$. If $\varphi(k) = -\log P(\pi_i|p_i)$ is considered a function of π_i, with $|\pi_i| = k$, $k = 0, \ldots, np_i$, then $\varphi(k)$ is a convex function given $p_i > 0$ and $n > 0$.*

Proof. See Appendix A.4. □

From (21), it follows that $P(\pi_i|p_i) \leq \frac{1}{2^{np_i}}, \forall |\pi_i| \leq m_p$. We define a criteria for graph and lag order selection for each equation the following

$$BIC(p_i, \vec{G}_{p,i}) = -2 \log P(\mathcal{X}|p_i, \vec{G}_{p,i}) + |\pi_i| \log T + 2np_i \log 2 \tag{22}$$

where $\vec{G}_{p,i}$ is the i-th equation local graph, π_i the set of selected predictors and $|\pi_i|$ the number of variables in π_i. Following a similar approach proposed by Chib and Greenberg (1995), we use the estimated graph $\hat{\vec{G}}_{p,i}$ to evaluate the score and to select the lag order $\hat{p}_i$. Selecting the local graph and lag order for each equation may automatically produce asymmetric lags for the different equations.

4.4. Model Estimation

The posterior of Ω conditional on the lag order $\hat{p}$ and a given graph $\hat{G}_p$ is Wishart distributed (see Geiger and Heckerman, 2002). Since Ω is directly related to B and Σ_ε, (see Proposition 1), we proceed with the posterior estimation of the model parameters focusing on B and Σ_ε. Thus we estimate $\hat{B}$ and $\hat{\Sigma}_\varepsilon$ from $P(B, \Sigma_\varepsilon|\hat{p}, \hat{G}_p, \mathcal{X})$ with an independent normal-Wishart. By conditioning on $\hat{G}_p$, we estimate the parameters $\{\bar{B}_{G,i}, \bar{V}_{G,i}\}$ that corresponds to the non-zero elements of the $i-$th equation graph $\hat{G}_{p,i}$. We define the selection matrix $E_i = (e_{j_1}, \ldots, e_{j_{|\pi_i|}})$, where E_i is of dimension $np \times |\pi_i|$, $j_k \in \pi_i$ is an element of the set of predictors of the i-th equation, and e_k is the standard orthonormal basis of the set of real np-dimensional vectors. The posterior mean and variance of $\{\bar{B}_{G,i}, \bar{V}_{G,i}\}$ is given by

$$\bar{B}_{G,i} = \bar{V}_{G,i}(\underline{V}_{G,i}^{-1}\underline{B}_{G,i} + \bar{\sigma}_i^{-2} W_{G,i}' Y_i) \tag{23}$$

$$\bar{V}_{G,i} = (\underline{V}_{G,i}^{-1} + \bar{\sigma}_i^{-2} W_{G,i}' W_{G,i})^{-1} \tag{24}$$

with

$$W_{G,i} = W E_i, \qquad \underline{B}_{G,i} = \underline{B}_i E_i, \qquad V_{G,i} = E_i' \underline{V}_i E_i \tag{25}$$

where $W_{G,i} \in W'$, is the set of selected predictors of the i-th equation; W' is stacked $W_1', \ldots, W_{T_0}'$, such that W' is of dimension $T_0 \times np$; Y is stacked $Y_1', \ldots, Y_{T_0}'$, such that Y is of dimension $T_0 \times n_y$; Y_i is the i-th column of Y; $\underline{B}_{G,i}$ and $\underline{V}_{G,i}$, are the prior mean and variance of $W_{G,i}$ respectively; $\bar{\sigma}_i^2, i = 1, \ldots, n_y$, is the variance of residuals from the posterior of Σ_ε, where the posterior of Σ_ε^{-1} is Wishart distributed with scale matrix

$$\bar{S} = \underline{S} + (Y' - \bar{B}W')'(Y - W\bar{B}') \tag{26}$$

and degrees of freedom $\bar{\nu} = \underline{\nu} + T_0$, where $\bar{B} = (\bar{B}_{G,1}, \ldots, \bar{B}_{G,n_y})$, the stacked posterior mean of the coefficients, such that $\bar{B}$ is of dimension $n_y \times np$ with positions of non-zero

elements corresponding to non-zero elements in $\hat{G}_p$.

5. Simulation Study

5.1. Metrics for Performance Evaluation

We investigate the effectiveness of our graphical approach with the sparsity prior against one without sparsity prior together with the LASSO of (Tibshirani, 1996) and the Elastic-net (ENET) of Zou and Hastie (2005). We evaluate the efficiency of the algorithms in terms of the estimated graph, the predictive performance of the estimated models on out-of-sample observations and computational cost in terms of run time.

Given the graph of the data generating process (DGP), we extract from the estimated graph the number of true links correctly predicted as TP; FP as number of true zero edges predicted as positives; TN as number of true zero edges correctly predicted; and FN as number of true links unidentified. We evaluate the graph estimation performance based on the number of predicted positive links ($PP = TP + FP$), the graph accuracy (ACC) and precision (PRC) given as

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \qquad PRC = \frac{TP}{TP + FP} \tag{27}$$

Furthermore, we evaluate the graph estimation performance in terms of log-likelihood and BIC scores. Following the expression in (22), the graph BIC is obtained as

$$BIC_G = \sum_{i=1}^{n_y} BIC(p_i, \vec{G}_{p,i}) = -2L_G + \sum_{i=1}^{n_y} (|\hat{\pi}_i| \log T + 2n\hat{p}_i \log 2) \tag{28}$$

where $L_G = \sum_{i=1}^{n_y} L_i$, with L_G is the log-likelihood of the estimated graph and $L_i = \log P(\mathcal{X}|p_i, \vec{G}_p(y_i, \pi_i))$ is the log-likelihood of the local graph of the i-th equation.

We evaluate the model estimation performance based on the out-of-sample joint density and point forecasts. The log-predictive score (LPS) is the most common measure of the joint predictive density discussed in the literature. Since the competing models might have different number of variables and lags across the equations, the predictive AIC presents a meaningful comparison for purposes of parsimony and is given by

$$AIC_M \;=\; -\,2\log P(Y_{\tau_1}|X_{\tau_0}; \hat{B}, \hat{\Sigma}_\varepsilon) \;+\; 2|\hat{B}| \tag{29}$$

for $\tau_1 = \tau_0 + 1, \ldots, T$, where τ_0 is the number of observations for the training sample, X_{τ_0} is the training sample dataset; Y_{τ_1} is the out-of-sample observations of the dependent variables; $|\hat{B}|$ is the number of non-zero coefficients in $\hat{B}$; $\hat{\Sigma}_\varepsilon$ is the estimated error covariance matrix; and $\log P(Y_{\tau_1}|X_{\tau_0}; \hat{B}, \hat{\Sigma}_\varepsilon)$ is the log predictive score.

For point forecast, the mean squared forecast error (MSFE) is the most common measure discussed in the literature. To compare the joint point forecasts, we compute the mean MSFE (MMSFE) following

$$MMSFE = \frac{1}{T-\tau_0-1}\sum_{\tau_1=\tau_0+1}^{T}\Big(\frac{1}{n_y}\sum_{i=1}^{n_y}(Y^i_{\tau_1}-\hat{Y}^i_{\tau_1})^2\Big) \tag{30}$$

where $Y^i_{\tau_1}$ and $\hat{Y}^i_{\tau_1}$ are the out-of-sample observed and predicted values of the i-th dependent variable respectively.

5.2. Simulation Study Set-up and Results

The set-up of the data generating process (DGP) of the simulated study is as follows

$$Y_t = BX_{t-1} + \varepsilon_t, \qquad \varepsilon_t \overset{iid}{\sim} \mathcal{N}(0, I_{n_y}) \tag{31}$$

$t = 1, \ldots, T$, where I_{n_y} is n_y dimensional identity matrix, B is $n_y \times n$ coefficient matrix, Y_t and X_t are is a $n_y \times 1$ and $n \times 1$ respectively. To analyze different sparsity levels, we generate the coefficients matrix B such that, the number of non-zero coefficients for each equation is drawn from a uniform on $\{0, \ldots, 40\}$. We considered a large dimensional model by setting $n_y = 10, n = 100$. We replicate the simulation and estimation exercises 100 times. The 100 replicatons have been conducted on a cluster multiprocessor system which consists of 4 nodes; each comprises four Xeon E5-4610 v2 2.3GHz CPUs, with 8 cores, 256GB ECC PC3-12800R RAM, Ethernet 10Gbit, 20TB hard disk system with Linux. The simulation study in Table 1 takes about 14 minutes of CPU time. For each replication, we generate a sample size, $T = 60$ and use $T_0 = 50$ for model estimation and 10 for out-sample forecast analysis.

We run 20,000 Gibbs iterations for the graph estimation and 2000 iterations for parameter estimations. We applied the standard approach of (Tibshirani, 1996) and Zou and Hastie (2005) for the LASSO and ENET estimation respectively. We set $\underline{p} = 1$ and $\bar{p} = 4$ and implement a parallel estimation for the LASSO and ENET. For each $p \in [\underline{p}, \bar{p}]$, we sequentially use one variable as the dependent variable and the remaining as the predictors. We apply a five-fold cross validation to select the regularization parameter λ with minimal plus one standard error point (index1SE).

Figure B.6 shows the convergence diagnostics of the graph simulation and the local graph BIC for the lags. The figure of the PSRF indicates convergence of the chain. We also notice from Figure B.6d that the posterior distribution on the lag order for each equation of the simulation experiment using our modified BIC favors lag order $p = 1$. We report in Table 1, the performance of the LASSO, ENET, BGVAR and SBGVAR for

the inference of the DGP in (31).

	LASSO	ENET	BGVAR	SBGVAR
DGP Average number of links = 201.5				
PP	108.40	131.30	249.42	69.88
TP	62.04	67.92	**97.76**	48.62
ACC	95.35	95.08	93.61	**95.65**
PRC	58.06	51.85	39.31	**69.58**
L_G	145.28	169.83	**242.35**	160.71
BIC_G	1457.40	1509.23	1837.14	**1307.80**
LPS	-243.07	-304.32	-236.79	**-166.49**
AIC_M	702.94	871.24	972.42	**472.74**
MMSFE	0.67	0.69	0.62	**0.59**
Time (in seconds)	55.76	50.18	162.52	**42.47**

Table 1: Average graph and model estimation performance of algorithms over 100 replications. PP - number of predicted positive links; TP - number of true positive links; ACC - graph accuracy; PRC - graph precision; L_G - graph log-likelihood; BIC_G - graph BIC; LPS - log predictive score; AIC_M - predictive AIC; and $MMSFE$ - mean of MSFE. Bold values indicate the best choice for each metric.

We proceed by comparing the effectiveness of the algorithms in estimating the graph of the true DGP, when the DGP average links number is 201.5. Table 1 shows that without the sparsity prior distribution, the BGVAR overestimates the number of links compared to the other algorithms. The Lasso-type methods (LASSO and Elastic-Net) fall in the middle with a lower number of links compared to that of the DGP. The SBGVAR on the other hand recorded the least number of edges. This is quite expected since the idea is to select the subset of the explanatory variables that explains a large variation in the dependent variables.

By including more edges than the true DGP, the graphical search algorithm without sparsity prior (BGVAR) records the highest true positive links but relatively low accuracy and precision compared to the other algorithms. Again the Lasso-type methods fall in the middle, recording a lower number of true positive links but with a higher accuracy and precision than the BGVAR. The sparsity prior graphical approach instead had the least number of true positive edges but tends to be more accurate and much precise than the other algorithms. The log-likelihood score of the graph favored the BGVAR but the graph BIC score favored the SBGVAR. Thus the BIC score confirms the outcome of the graph accuracy metric which shows that though the SBGVAR records the least edges, it produced a better representation of the temporal dependence in the simulated dataset than the Lasso-type methods and the BGVAR.

The log predictive score, predictive AIC and the MMSFE all favor the SBGVAR over the other competitors. One would expect the Lasso-type methods to perform better than the graphical VAR, however this is not the case according to the above simulation results. This is attributable to the fact that the Lasso-type techniques perform both

model selection and parameter estimation simultaneously. This may seem to be an advantage but on the other hand it affects the estimated parameters, since it shrinks all coefficients at the same rate (see Gefang, 2014). In addition, the Lasso-type methods only focus on estimating the coefficients in each equation neglecting the interaction among the errors across the different equations. The graphical approach instead focus on selecting and estimating only the coefficients of relevant variables taking into consideration the interaction among the errors across the different equations. Thus the latter achieves better parameter estimation efficiency than the Lasso-type models. The result shows that the sparsity prior on the graph enables us to identify the small set of the most influential explanatory variables that explains a large variation in the dependent variables. Also, the SBGVAR produce a more parsimonious model with better out-of-sample forecasts than the Lasso-type methods.

On the computational intensity, the SBGVAR spends less time than the other algorithms. Interestingly, it records about one-fourth of the run time of the BGVAR. Thus, the sparsity prior of the fan-in restriction helps to reduce the run time by considering a relatively lower search space in terms of the number of combinations of explanatory variables. The higher run time of the Lasso-type methods is due to the cost of cross-validation to select the regularization parameter.

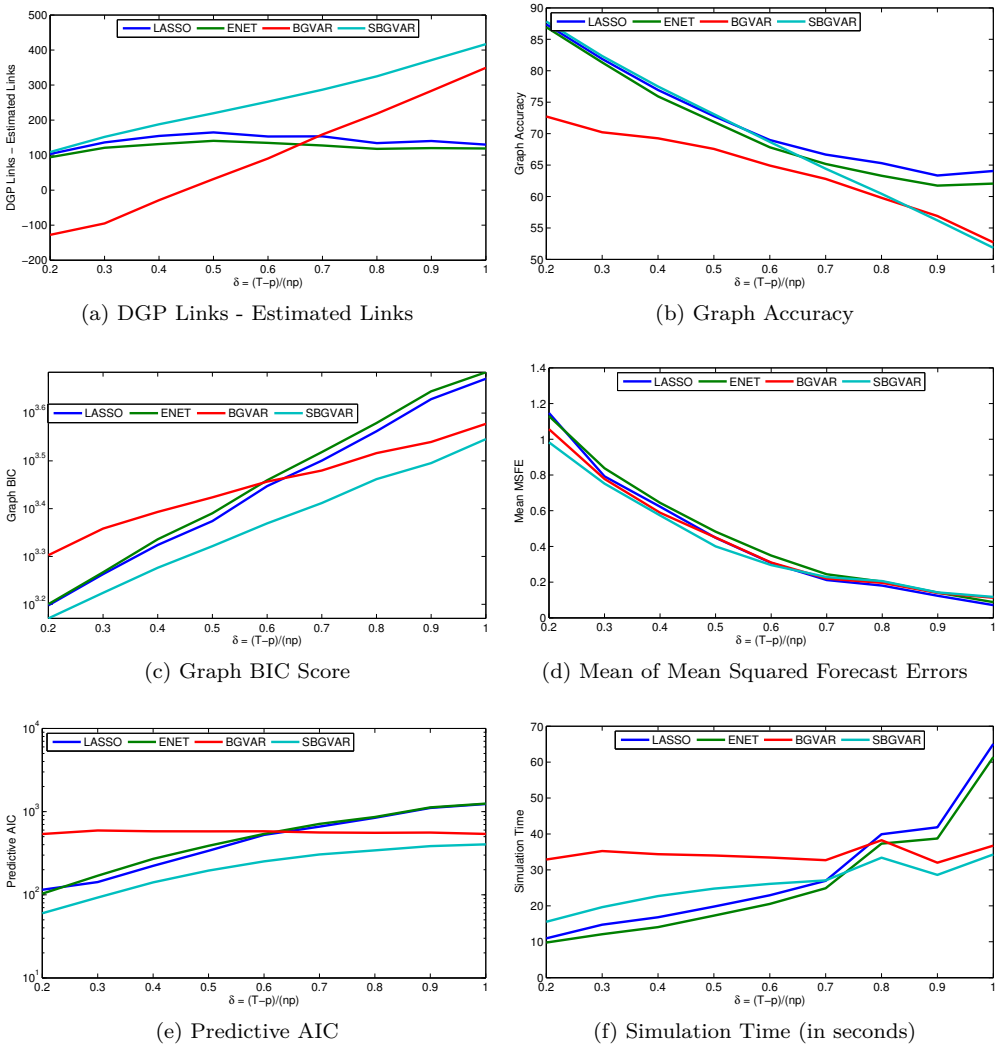
5.3. Sparsity and Indeterminacy Evaluation

A system of linear equations is said to be under-determined (or indeterminate) when the number of parameters to estimate exceeds the number of observations (see Donoho, 2006). Such systems can be modeled by exploiting sparsity. Here, we investigate the performance of the graphical model approaches against the standard Lasso-type methods for different level of indeterminacy and sparsity of the DGP.

For a VAR model with n_y dependent variables and n explanatory variables for each equation, with a maximum lag order p, we have a total of $n_y np$ number of coefficients to estimate. Given a multiple time series with T observations, the total number of observations of the dependent variables is given by $(T-p)n_y$. Following Donoho (2006), we measure the level of indeterminacy by $\delta = (T-p)n_y/n_y np = (T-p)/np$, and the level of sparsity by $\rho = kn_y/(T-p)n_y = k/(T-p)$, where k is the number of non-zero coefficients in each equation of the DGP.

Following Donoho and Stodden (2006), we formulate our experiment by setting the DGP to generate a VAR model with $n_y = 10$, $n = 100$ and lag order $p = 1$. For different level of indeterminacy, we set $T-p$ to take values $\{20, \ldots, 100\}$. For each $T-p$, we generate for each equation, $k = \lceil \rho(T-p) \rceil$, where ρ takes values $\{0.2, 0.3, \ldots, 1\}$. This is to allow for different sparsity levels for each level of indeterminacy.

We proceed by comparing the effectiveness of the LASSO, ENET, BGVAR and SBGVAR in estimating the true DGP by setting $\underline{p} = \bar{p} = 1$. For each T and k, we replicate the simulation and estimation exercise 10 times with the magnitude of the coefficients drawn from a uniform on $[-1, 1]$. In each replication, we estimate the model and perform a 1-step ahead forecast. Figure 1 shows the estimation performance of the algorithms for the different levels of indeterminacy averaged over the different levels of sparsity.



(a) DGP Links - Estimated Links

(b) Graph Accuracy

(c) Graph BIC Score

(d) Mean of Mean Squared Forecast Errors

(e) Predictive AIC

(f) Simulation Time (in seconds)

Figure 1: Estimation performance of the algorithms for different level of indeterminacy averaged over different level of sparsity. The LASSO is in green, ENET in blue, BGVAR in red and SBGVAR in cyan.

Figure 1a shows the difference between the average DGP number of links and the

estimated links of the algorithms. Except for the BGVAR, all the other algorithms estimated a lower number of links compared to that of the DGP. More specifically, the BGVAR seems to overestimate the number of DGP links for lower under-determined models, whereas the SBGVAR underestimates the number of DGP links regardless of the level of indeterminacy. We also see that, the difference between the DGP and the estimated links of the BGVAR and SBGVAR increases overtime while the Lasso-type methods are relatively stable regardless of the level of indeterminacy.

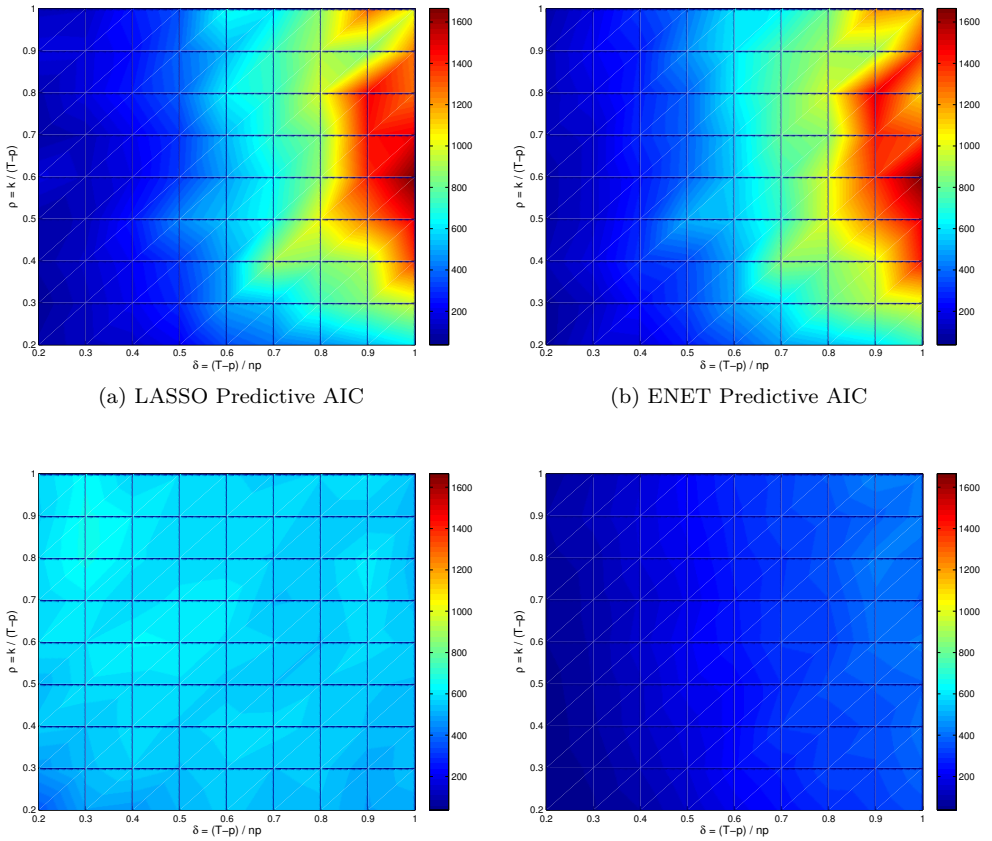
The graph accuracy in Figure 1b shows that all the algorithms experienced a deterioration in the accuracy of the prediction of the graph associated with the DGP. However, on average the SGBVAR performs slightly better at the graph estimation for lower under-determined models than the Lasso-type methods.

In Figure 1c, the graph BIC of the algorithms increases with the level of indeterminacy. This is not surprising since the BIC is a direct function of the number of observations which increases with the level of indeterminacy. Again we observe that the graph BIC score favors the graph estimated by the SBGVAR over the other competing algorithms. This shows that though the SBGVAR recorded the minimum number of links, it produce a better representation of the graph associated with the DGP.

For model estimation performance, Figure 1d shows that all the algorithms perform better at out-of-sample point forecasts for higher under-determined models. The MMSFE of the algorithms are not significantly different though we find that it favors the SBGVAR for lower under-determined models. The predictive AIC (in Figure 1e) on the other strongly favors the SBGVAR for all level of indeterminacy.

On the computational intensity, we notice (from Figure 1f) an increase in run time with the level of indeterminacy for all algorithms except the BGVAR which seems slightly constant over time. Overall, the Lasso-type methods achieve a lower run time for lower under-determined models whiles the SBGVAR achieves lower run time for higher under-determined models.

We focus attention on the model estimation performance of the algorithms for the different levels of indeterminacy and sparsity. Figure 2 shows the heatmap of the predictive AIC of the models of the algorithms estimated over the levels of sparsity and indeterminacy of the DGP. The color bar shows the different range of values of the predictive AIC, where blue represents lower AIC, and red for highest AIC. Clearly, we notice a significant difference between the results of the Lasso-type methods and that of the graphical model approaches. Thus the LASSO and ENET are not significantly different from each other, whiles the BGVAR and SBGVAR are quite different, dominated by cyan and blue respectively. The figure shows that the predictive AIC favor the SBGVAR over all levels of sparsity and indeterminacy. The Lasso-type methods only performs better than the



(a) LASSO Predictive AIC

(b) ENET Predictive AIC

(c) BGVAR Predictive AIC

(d) SBGVAR Predictive AIC

Figure 2: Heatmap of the predictive AIC of the models estimated by the four algorithms over the different levels of indeterminacy and sparsity in the data generating process. The result is an average of 10 replication exercises for each δ and ρ. The color bar shows the different range of values of the predictive AIC, where blue represents lower AIC, and red for highest AIC.

BGVAR for lower under-determined models with different level of sparsity, whiles the BGVAR dominates in higher under-determined models.

The results of this exercise confirm that of our first simulation experiment. Firstly, the sparsity prior on the graph space induces sparsity on the estimated graph of the temporal relationship among the variables. Secondly, the random fan-in restriction helps to reduce the computational complexity by considering a relatively lower search space in terms of the number of combinations of explanatory variables. Thirdly, though the SBGVAR under-estimates the number of links compared to the DGP and other algorithms, it is able to identify the small set of the most influential explanatory variables that explains a large variation in the dependent variables. Thus, the SBGVAR produces a more parsimonious

model with competitive out-of-sample joint point forecasts and better density forecasts than the competing models.

6. Forecasting VAR with Many Predictors

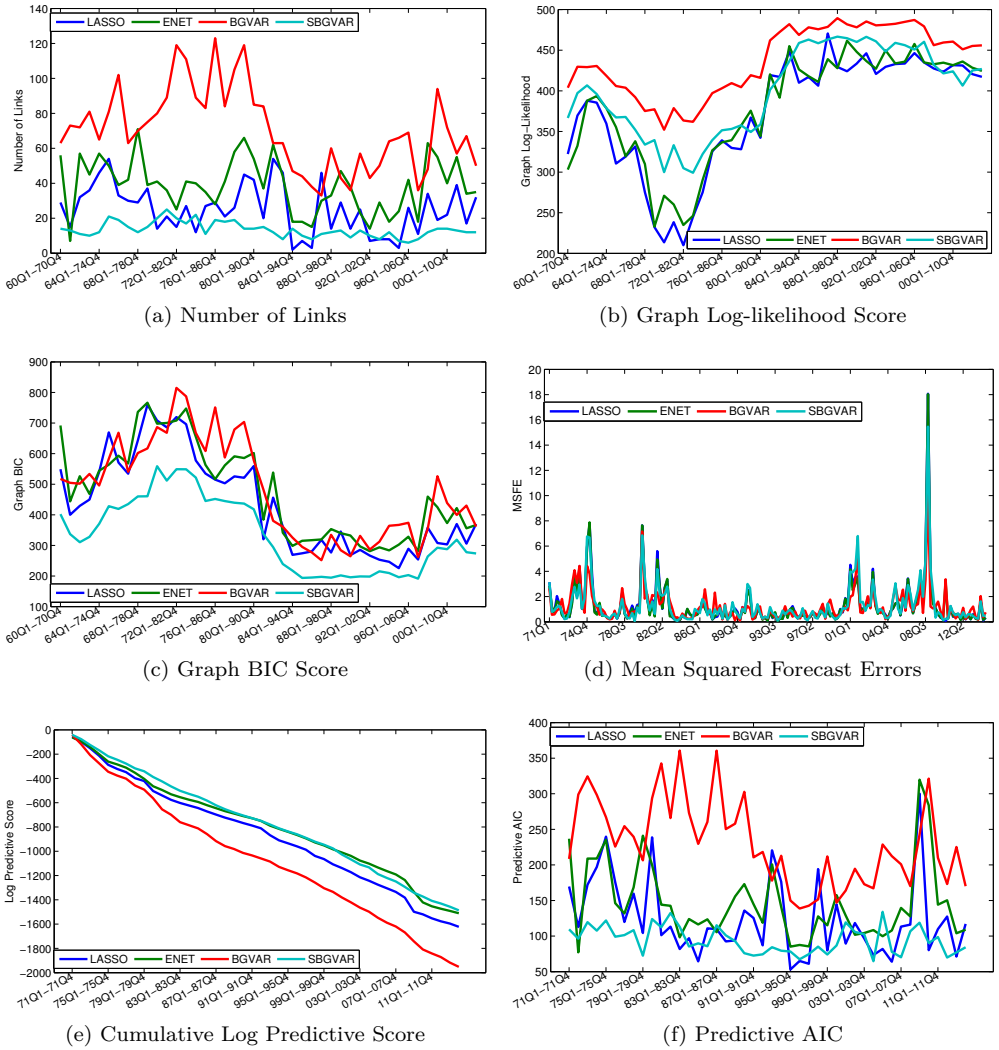
Several studies have shown empirically that applying large VAR models for macroeconomic time series produces better forecasts than standard approaches (see Banbura et al., 2010; Carriero et al., 2013; Giannone et al., 2005; Koop, 2013; Stock and Watson, 2012). In the literature, researchers typically work with a single model with fixed or time varying coefficients (see Koop and Korobilis, 2013). It is therefore important to allow for changes in structure and/or parameters to understand the dynamic evolution of the relationship among variables. As part of our contribution, we apply our graphical scheme to model and forecast selected macroeconomic variables with large number of predictors.

The dataset is quarterly observations of 130 US-macroeconomic variables. All series were downloaded from St. Louis' FRED database and cover the quarters from $1959Q1$ to $2014Q3$. Some series had missing observations which are completed with earlier version of the database used by Korobilis (2013). We follow the adjustment codes of De Mol et al. (2008); Stock and Watson (2012) and Korobilis (2013) to transform all the series into stationarity. See Appendix C for the list of series and adjustment codes. We consider 6 series as dependent variables and the remaining 124 as predictors. The dependent variables are: consumer price index (CPIAUCSL), Federal funds rate (FEDFUNDS), real gross domestic product (GDPC96), real gross private domestic investment (GPDIC96), industrial production index (INDPRO) and real personal consumption expenditure (PCECC96).

We set the minimum and maximum lag order equal to $\underline{p} = 1$ and $\bar{p} = 4$ respectively according to the literature. We consider a moving window with a starting sample from $1960Q1$ to $1970Q4$ to estimate the model and to forecast 1 to 4-quarters ahead. We then move the window forward by 4-quarters. Our last sample covers $2003Q1$ to $2013Q4$, and the final forecast is up to $2014Q3$.

Figures B.7 in Appendix B show the convergence diagnostics of the graph simulation and the local graph BIC for the lags for the macroeconomic application. The figure of the PSRF indicates convergence of the chain. Clearly, the global log score of the graph seems to increase with the lag order in Figure B.7b whereas the total number of links of the different lags seems quite close as displayed in Figure B.7a. However, we notice from Figure B.7d that the posterior distribution on the lag order for each equation of the macroeconomic application using our modified BIC score favors lag order $p = 1$.

We report in Figure 3, the graph and model estimation performance of the Lasso-type methods and the graphical VAR approaches in modeling and forecasting the selected



(a) Number of Links (b) Graph Log-likelihood Score

(c) Graph BIC Score (d) Mean Squared Forecast Errors

(e) Cumulative Log Predictive Score (f) Predictive AIC

Figure 3: Performance of the algorithms in modeling and forecasting selected macroeconomic variables with many predictors over the sample period $1960Q1-2014Q3$. Figures 3a - 3c show the graph estimation performance, whilst 3d - 3f depict the model estimation performance.

macroeconomic variables over the sample period $1960Q1 - 2014Q3$. The graph estimation performance is compared in terms of the number of link (PP - predicted positive edges), the log-likelihood of the graph (L_G) and the BIC score of the graph (BIC_G). The model estimation performance is compared in terms of the log predictive score (LPS), the predictive AIC (AIC_M) and the average of the mean squared forecast errors (MMSFE). Table 2 presents the averages of the graph and model estimation performance of the

algorithms including the computational time over the sample period.

From Figure 3, we observe that the BGVAR estimated more edges than the other algorithms in a greater part of the sample period. This is followed by the Lasso-type methods, (ENET, then the LASSO) and the SBGVAR records the least number of links over the entire sample period. In scoring the estimated graphs, the BGVAR obtained the highest log likelihood over the entire period whiles the SBGVAR records the minimum at the beginning but showed significant improvement over the rest of the sample period. The BIC score of the graph however favored the SBGVAR over the other algorithms. The summary of the averages in Table 2 shows that the by including more edges than the other algorithms the BGVAR records the highest log likelihood of the graph whilst the SBGVAR with the least number of links obtained the minimum BIC score indicating that the SGBVAR graph presents a better representation of the temporal dependence in the macroeconomic application than the Lasso-type methods and the BGVAR.

Figure 3d shows the evolution of the out-of-sample joint point forecasts of the models estimated by the algorithms. We observe from the plot that the MSFE are not very different from each other. However, in terms of the out-of-sample joint density forecasts, the BGVAR model presents the minimum cumulative log predictive score and that of the SBGVAR model dominates the LASSO but is very competitive against the ENET model. When adjusted for the number of selected variables used for the forecasting analysis, the SBGVAR model obtain the minimum predictive AIC whilst the BGVAR model performed worst than the Lasso-type models. From Table 2, we see that on average the Lasso-type models obtain the minimum MMSFE and this indicate that they produce slightly better point forecasts than the graphical approaches. The average log predictive score and AIC on the other hand are in favor of the SGBVAR over the Lasso-type models.

	LASSO	ENET	BGVAR	SBGVAR
PP	25.14	39.09	70.82	13.50
L_G	372.26	379.45	**437.21**	400.05
BIC_G	434.09	473.71	481.01	**333.46**
LPS	-36.87	-34.33	-44.36	**-33.79**
AIC_M	124.01	146.84	230.36	**94.57**
MMSFE	1.30	1.26	**1.24**	1.32
Time (in seconds)	57.93	42.26	65.74	**23.77**

Table 2: Average graph and model estimation performance of algorithms in modeling and forecasting selected macroeconomic series from $1960Q1 - 2014Q3$. PP - number of predicted positive edges; L_G - graph log-likelihood; BIC_G - graph BIC; LPS - log predictive score; AIC_M - predictive AIC; and $MMSFE$ - mean of MSFE. Bold values indicate the best choice for each metric.

On the computational intensity, the result shows that on average the SBGVAR spends less simulation time on graph sampling and parameter estimation than the other algorithms. Interestingly, it records about one-fourth of the run time of the BGVAR.

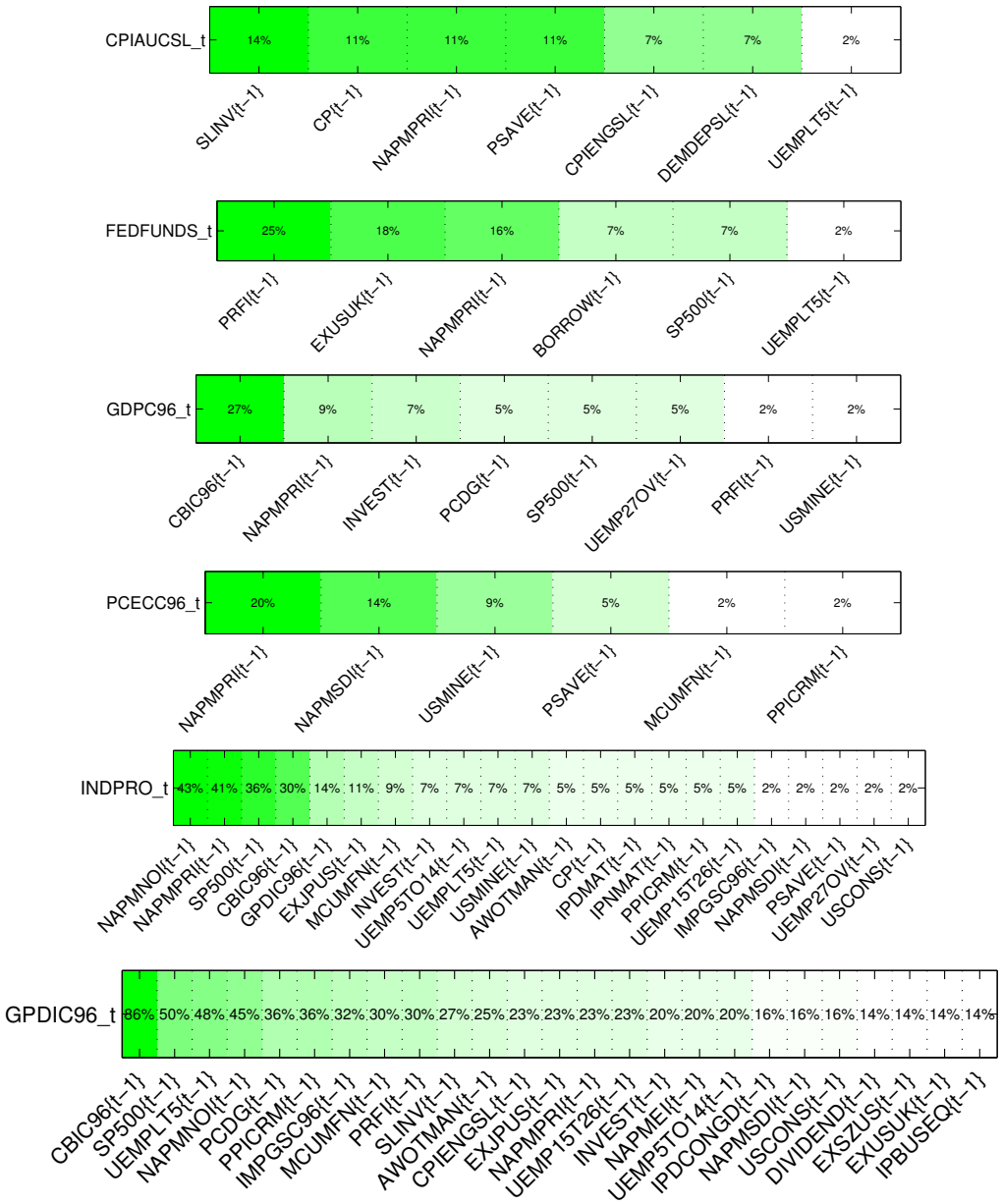


Figure 4: Frequency of inclusion of the most influential variables that explain a large variation in the dependent variables of the macroeconomic application averaged over the sample period $1960Q1-2014Q3$. CPIAUCSL is consumer price index, FEDFUNDS - Federal funds rate, GDPC96 - real gross domestic product, GPDIC96 - real gross private domestic investment, INDPRO - industrial production index and PCECC96 - real personal consumption expenditure.

In summary, we find evidence that the graphical VAR approach with our new graph prior distribution induces sparsity on the graph structure. In modeling and forecasting our selected macroeconomic series, the result shows that the SBGVAR better represents the temporal dependence, since it is more parsimonious than the competitors. Furthermore, we find evidence of a gain in the predictive performance of the SBGVAR approach over the Lasso-type methods. It is also less computationally intensive compared to the graphical approach without sparsity prior and the Lasso-type methods.

In Figure 4, we report the frequency of the most influential variables that explain a large variation in the dependent variables of the macroeconomic application, averaged over the sample period $1960Q1-2014Q3$. For convenience and clarity of presentation, we report only the top explanatory variables of real investment (GPDIC96) with frequency up to 14%. We find strong evidence supporting the effect of financial variables on the real sector of the US economy. More specifically, we find that over the sample period, S&P 500 and exchange rates especially with Japan, Switzerland and UK, have strong effects on real investment (GPDIC96) and industrial production index (INDPRO), and weak effects on real gross domestic product (GDPC96) and on the Federal funds rate (FEDFUNDS). The results are in line with the recommendation by Diebold and Yilmaz (2015) suggesting the importance of monitoring the connectedness between real activity and stock returns (or financial variables). Furthermore, it offers some insight for further evaluation of macro-financial linkages which have long been at the core of the IMF's mandate to oversee the stability of the global financial system. The figure also shows that apart from real investment and industrial production index, that report a higher number of predictors over the sample period, the rest can be predicted by a handful of macroeconomic variables.

7. Conclusion

This paper develops a Bayesian approach to model dependence in high-dimensional multivariate time series and to address over-parametrization in large vector autoregressive (VAR) models. The methodology discussed in the paper is based on combining graphical model notion of causality with a new sparsity prior distribution on the graph space to address model selection problems in multivariate time series of large dimension. In particular, this work builds on the application of restriction on the explanatory variables (fan-in) in the VAR model by allowing for different prior information level about the maximal number of predictors for each equation. This prior distribution proves to be efficient in reducing the number of possible combinations of predictors to explore for each equation when determining the dependence in a large VAR model. Furthermore,

the Bayesian paradigm allows us to take into account uncertainties about the maximum lag order, the dependence structure and coefficients of the VAR through model averaging.

In both simulation study and empirical macroeconomic application to real datasets, we find evidence that our sparsity prior distribution enables us to control the fraction of explanatory variables with temporal causal effect on the dependent variables. The comparison with the standard Lasso-type methods (LASSO and Elastic-Net) as a benchmark shows that our model is more sparse and parsimonious than the benchmark. The results show that, compared to the competing methods, the sparse graphical approach is able to recover the small set of predictors that explains a large variation in the dependent variables of the large VAR model. More specifically, the BIC score of the graph of temporal dependence and the predictive AIC of the estimated model all favor the sparse graphical VAR model over the Lasso-type methods. Furthermore, we fine evidence of a gain in sampling the graph of the temporal dependence among variables which allows to impose zero restrictions supported by the data on the non relevant components of the predictors in order to estimate the coefficients of the selected variables. On the macroeconomic application, we find evidence supporting the effect of financial variables on the real sector of the US economy. Thus, our methodology and result offers insight for further research into empirical evaluation of macro-financial linkages which has long been the core of the IMF's mandate to oversee the stability of the global financial system.

Acknowledgments

Earlier versions of this work have been presented at the 34th International Symposium on Forecasting (ISF 2014), Rotterdam; Workshop on Networks in Economics and Finance, Louvain-la-Neuve, 2014; the 8th International Conference on Computational and Financial Econometrics, (CFE 2014), Pisa; the European Winter Meetings of the Econometric Society, (EWM 2014), Madrid; and the 6th Italian Congress of Econometrics and Empirical Economics (ICEEE 2015), Salerno. We are grateful for useful comments from the participants in these conferences. This research used the SCSCF multiprocessor cluster system at University Ca' Foscari of Venice. Author's research is supported by funding from the European Union, Seventh Framework Programme FP7/2007-2013 under grant agreement SYRTO-SSH-2012-320270, by the Global Risk Institute in Financial Services and the Institut Europlace de Finance, Systemic Risk grant, and by the Italian Ministry of Education, University and Research (MIUR) PRIN 2010-11 grant MISURA.

References

Abramowitz, M., Stegun, I. A., 1964. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Courier Dover Publications.

Ahelegbey, D. F., Billio, M., Casarin, R., 2015. Bayesian Graphical Models for Structural Vector Autoregressive Processes. Journal of Applied Econometrics (forthcoming).

Ahelegbey, D. F., Giudici, P., 2014. Bayesian Selection of Systemic Risk Networks. Advances in Econometrics: Bayesian Model Comparison 34, 117–153.

Andrieu, C., Roberts, G. O., 2009. The Pseudo-Marginal Approach for Efficient Monte Carlo Computations. The Annals of Statistics, 697–725.

Bai, J., Ng, S., 2008. Large Dimensional Factor Analysis. Foundations and Trends in Econometrics 3 (2), 89–163.

Banbura, M., Giannone, D., Reichlin, L., 2010. Large Bayesian Vector Autoregressions. Journal of Applied Econometrics 25, 71 – 92.

Barigozzi, M., Brownlees, C., 2014. NETS: Network Estimation for Time Series. Working Paper, Social Science Research Network.

Belloni, A., Chernozhukov, V., 2011. High Dimensional Sparse Econometric Models: An Introduction. Springer.

Bernanke, B., Boivin, J., Eliasz, P. S., 2005. Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach. The Quarterly Journal of Economics 120 (1), 387–422.

Billio, M., Getmansky, M., Lo, A. W., Pelizzon, L., 2012. Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors. Journal of Financial Economics 104 (3), 535 – 559.

Bogdan, M., Ghosh, J. K., Doerge, R., 2004. Modifying the Schwarz Bayesian Information Criterion to Locate Multiple Interacting Quantitative Trait Loci. Genetics 167 (2), 989–999.

Box, G. E., Meyer, R. D., 1986. An Analysis for Unreplicated Fractional Factorials. Technometrics 28 (1), 11–18.

Brillinger, D. R., 1996. Remarks Concerning Graphical Models for Time Series and Point Processes. Revista de Econometria 16, 1–23.

Brooks, S., Gelman, A., Jones, G., Meng, X.-L., 2011. Handbook of Markov Chain Monte Carlo. CRC press.

Carriero, A., Clark, T. E., Marcellino, M., 2013. Bayesian VARs: Specification Choices and Forecast Accuracy. Journal of Applied Econometrics 25, 400–417.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., West, M., 2008. High-dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. Journal of the American Statistical Association 103 (484).

Carvalho, C. M., Scott, J. G., 2009. Objective Bayesian Model Selection in Gaussian Graphical Models. Biometrika 96 (3), 497–512.

Casarin, R., Dalla Valle, L., Leisen, F., 2012. Bayesian Model Selection for Beta Autoregressive Processes. Bayesian Analysis 7 (2), 385–410.

Casella, G., Robert, C. P., 2004. Monte Carlo Statistical Methods. Springer Verlag, New York.

Chen, J., Chen, Z., 2008. Extended Bayesian Information Criteria for Model Selection with Large Model Spaces. Biometrika 95 (3), 759–771.

Chib, S., Greenberg, E., 1995. Hierarchical Analysis of SUR Models with Extensions to Correlated Serial Errors and Time-varying Parameter Models. Journal of Econometrics 68 (2), 339–360.

Chickering, D. M., Heckerman, D., 1997. Efficient Approximations for the Marginal Likelihood of Bayesian Networks with Hidden Variables. Machine Learning 29 (2-3), 181–212.

Chickering, D. M., Heckerman, D., Meek, C., 2004. Large-Sample Learning of Bayesian Networks is NP-Hard. Journal of Machine Learning Research 5, 1287–1330.

Consonni, G., Rocca, L. L., 2012. Objective Bayes Factors for Gaussian Directed Acyclic Graphical

Models. Scandinavian Journal of Statistics 39 (4), 743–756.

Corander, J., Villani, M., 2006. A Bayesian Approach to Modelling Graphical Vector Autoregressions. Journal of Time Series Analysis 27(1), 141–156.

DasGupta, B., Kaligounder, L., 2014. On Global Stability of Financial Networks. Journal of Complex Networks, 1–59.

Davis, R. A., Zang, P., Zheng, T., 2012. Sparse Vector Autoregressive Modeling. arXiv preprint arXiv:1207.0520.

De Mol, C., Giannone, D., Reichlin, L., 2008. Forecasting Using a Large Number of Predictors: Is Bayesian Shrinkage a Valid Alternative to Principal Components? Journal of Econometrics 146 (2), 318–328.

Demiralp, S., Hoover, K. D., 2003. Searching for the Causal Structure of a Vector Autoregression. Oxford Bulletin of Economics and Statistics 65, 745–767.

Diebold, F., Yilmaz, K., 2014. On the Network Topology of Variance Decompositions: Measuring the Connectedness of Financial Firms. Journal of Econometrics 182(1), 119–134.

Diebold, F., Yilmaz, K., 2015. Financial and Macroeconomic Connectedness: A Network Approach to Measurement and Monitoring. Oxford University Press.

Doan, T., Litterman, R., Sims, C., 1984. Forecasting and Conditional Projection Using Realistic Prior Distributions. Econometric Reviews 3, 1–100.

Donoho, D., Stodden, V., 2006. Breakdown Point of Model Selection when the Number of Variables Exceeds the Number of Observations. In: Neural Networks, 2006. IJCNN'06. International Joint Conference on. IEEE, pp. 1916–1921.

Donoho, D. L., 2006. High-Dimensional Centrally Symmetric Polytopes with Neighborliness Proportional to Dimension. Discrete Computational Geometry 35, 617–652.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least Angle Regression. The Annals of Statistics 32 (2), 407–499.

Ehlers, R. S., Brooks, S. P., 2004. Bayesian Analysis of Order Uncertainty in ARIMA Models. Federal University of Parana, Tech. Rep.

Elliott, G., Gargano, A., Timmermann, A., 2013. Complete Subset Regressions. Journal of Econometrics 177 (2), 357–373.

Foygel, R., Drton, M., 2011. Bayesian Model Choice and Information Criteria in Sparse Generalized Linear Models. arXiv preprint arXiv:1112.5635.

Friedman, N., Koller, D., 2003. Being Bayesian About Network Structure. Journal of Machine Learning 50 (1-2), 95–125.

Gefang, D., 2014. Bayesian Doubly Adaptive Elastic-Net Lasso for VAR Shrinkage. International Journal of Forecasting 30 (1), 1–11.

Geiger, D., Heckerman, D., 2002. Parameter Priors for Directed Acyclic Graphical Models and the Characterization of Several Probability Distributions. Annals of statistics 30 (5), 1412–1440.

Gelman, A., Rubin, D. B., 1992. Inference from Iterative Simulation Using Multiple Sequences, (with discussion). In: Statistical Science. Vol. 7. pp. 457–511.

Geweke, J., 1977. The Dynamic Factor Analysis of Economic Time-Series Models. SSRI workshop series. Social Systems Research Institute, University of Wisconsin-Madison.

Giannone, D., Reichlin, L., Sala, L., 2005. Monetary Policy in Real Time. In: NBER Macroeconomics Annual 2004. Vol. 19. MIT Press, pp. 161–224.

Giudici, P., Castelo, R., 2003. Improving Markov Chain Monte Carlo Model Search for Data Mining. Machine Learning 50 (1-2), 127–158.

Giudici, P., Green, P. J., 1999. Decomposable Graphical Gaussian Model Determination. Biometrika 86 (4), 785–801.

Green, P. J., 1995. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. Biometrika 82 (4), 711–732.

Grzegorczyk, M., Husmeier, D., 2011. Non-homogeneous Dynamic Bayesian Networks for Continuous Data. Machine Learning 83 (3), 355–419.

Hautsch, N., Schaumburg, J., Schienle, M., 2014. Financial Network Systemic Risk Contributions. Review of Finance.

Heckerman, D., 1998. A Tutorial on Learning with Bayesian Networks. Springer.

Heckerman, D., Chickering, D. M., 1995. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. In: Machine Learning. pp. 20–197.

Heckerman, D., Geiger, D., 1994. Learning Gaussian Networks. In: Uncertainty in Artificial Intelligence. pp. 235–243.

Huang, X., Zhou, H., Zhu, H., 2012. Systemic Risk Contributions. Journal of financial services research 42 (1-2), 55–83.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., West, M., 2005. Experiments in Stochastic Computation for High-dimensional Graphical Models. Statistical Science, 388–400.

Kass, R., Tierney, L., Kadane, J., 1988. Asymptotics in Bayesian Computation. Bayesian Statistics 3, 261–278.

Kock, A. B., Callot, L., 2012. Oracle Inequalities for High Dimensional Vector Autoregressions. Aarhus University, CREATES Research Paper 16.

Koop, G., 2013. Forecasting with Medium and Large Bayesian VARs. Journal of Applied Econometrics 28, 177 – 203.

Koop, G., Korobilis, D., 2013. Large Time-Varying Parameter VARs . Journal of Econometrics 177 (2), 185 – 198.

Koop, G., Potter, S., 2004. Forecasting in Dynamic Factor Models Using Bayesian Model Averaging. The Econometrics Journal 7 (2), 550–565.

Korobilis, D., 2013. Hierarchical Shrinkage Priors for Dynamic Regressions with Many Predictors. International Journal of Forecasting 29 (1), 43–59.

Kyung, M., Gill, J., Ghosh, M., Casella, G., 2010. Penalized Regression, Standard Errors, and Bayesian Lassos. Bayesian Analysis 5 (2), 369–411.

Lauritzen, S. L., Wermuth, N., 1989. Graphical Models for Associations Between Variables, Some of Which Are Qualitative and Some Quantitative. Annals of Statistics 17, 31–57.

Lenkoski, A., Dobra, A., 2011. Computational Aspects Related to Inference in Gaussian Graphical Models with the G-Wishart Prior. Journal of Computational and Graphical Statistics 20 (1).

Maclaurin, D., Adams, R. P., 2014. Firefly Monte Carlo: Exact MCMC with Subsets of Data. arXiv preprint arXiv:1403.5693.

Madigan, D., York, J., 1995. Bayesian Graphical Models for Discrete Data. International Statistical Review 63 (2), 215–232.

Medeiros, M. C., Mendes, E., 2012. Estimating High-dimensional Time Series Models. CREATES Research Paper 37.

Merton, R. C., Billio, M., Getmansky, M., Gray, D., Lo, A. W., Pelizzon, L., 2013. On a New Approach for Analyzing and Managing Macrofinancial Risks. Financial Analysts Journal 69 (2).

Moneta, A., 2008. Graphical Causal Models and VARs: An Empirical Assessment of the Real Business Cycles Hypothesis. Empirical Economics 35 (2), 275–300.

Park, T., Casella, G., 2008. The Bayesian Lasso. Journal of the American Statistical Association 103 (482), 681–686.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann publishers, San Mateo, CA, USA,.

Roverato, A., 2002. Hyper Inverse Wishart Distribution for Non-decomposable Graphs and its Application to Bayesian Inference for Gaussian Graphical Models. Scandinavian Journal of Statistics 29 (3), 391–411.

Scott, J. G., Berger, J. O., 2010. Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem. The Annals of Statistics 38 (5), 2587–2619.

Scott, J. G., Carvalho, C. M., 2008. Feature-inclusion Stochastic Search for Gaussian Graphical Models. Journal of Computational and Graphical Statistics 17 (4).

Shojaie, A., Michailidis, G., 2009. Analysis of Gene Sets Based on the Underlying Regulatory Network. Journal of Computational Biology 16 (3), 407–426.

Shojaie, A., Michailidis, G., 2010. Penalized Likelihood Methods for Estimation of Sparse High Dimensional Directed Acyclic Graphs. Biometrika 97 (3), 519–538.

Song, S., Bickel, P. J., 2011. Large Vector Auto Regressions. Tech. rep., Humboldt University, Collaborative Research Center 649.

Stock, J. H., Watson, M. W., 2006. Forecasting with Many Predictors. Handbook of Economic Forecasting 1, 515–554.

Stock, J. H., Watson, M. W., 2012. Generalized Shrinkage Methods for Forecasting using Many Predictors. Journal of Business & Economic Statistics 30 (4), 481–493.

Stock, J. H., Watson, M. W., 2014. Estimating Turning Points Using Large Data Sets. Journal of Econometrics 178, 368–381.

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.

Telesca, D., Müller, P., Kornblau, S. M., Suchard, M. A., Ji, Y., 2012. Modeling Protein Expression and Protein Signaling Pathways. Journal of the American Statistical Association 107 (500), 1372–1384.

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267–288.

Vermaak, J., Andrieu, C., Doucet, A., Godsill, S., 2004. Reversible jump Markov chain Monte Carlo Strategies for Bayesian Model Selection in Autoregressive Processes. Journal of Time Series Analysis 25 (6), 785–809.

Wang, H., 2010. Sparse Seemingly Unrelated Regression Modelling: Applications in Econometrics and Finance. Computational Statistics & Data Analysis 54 (11), 2866–2877.

Wang, H., Li, S. Z., 2012. Efficient Gaussian Graphical Model Determination Under G-Wishart Prior Distributions. Electronic Journal of Statistics 6, 168–198.

Whittaker, J., 1990. Graphical Models in Applied Multivariate Statistics. John Wiley, Chichester.

Woodbury, M. A., 1950. Inverting Modified Matrices. Memorandum Report, 42, Statistical Research Group, Princeton University.

Zhang, Z., Wang, S., Liu, D., Jordan, M. I., 2012. EP-GIG Priors and Applications in Bayesian Sparse Learning. The Journal of Machine Learning Research 13 (1), 2031–2061.

Zhou, X., Schmidler, S. C., 2009. Bayesian Parameter Estimation in Ising and Potts Models: A Comparative StudyWith Applications to Protein Modeling. Tech. rep., Duke University.

Zou, H., Hastie, T., 2005. Regularization and Variable Selection via the Elastic-Net. Journal of the Royal Statistical Society: Series B, Statistical Methodology 67 (2), 301–320.

Appendix A. Proofs

Appendix A.1. Proof of Proposition 1

Proof. Let X_t be $n \times 1$ vector of observations at time t, $Y_t \subseteq X_t$ a $n_y \times 1$ vector of dependent variables, W_t the stacked lags of X_t, where $W_t = (X'_{t-1}, \dots, X'_{t-p})'$ is of $np \times 1$ dimension, with p as the maximum lag order. Suppose the joint distribution of $(Y'_t, W'_t) \sim \mathcal{N}(\mu, \Omega^{-1})$, where μ is the $((np + n_y) \times 1)$ vector of means and Ω^{-1} is $(np + n_y) \times (np + n_y)$ matrix of covariances. Without loss of generalization, we assume μ is a zero vector.

Suppose the marginal distribution of $W_t \sim \mathcal{N}(0, \Sigma_{ww})$ and the conditional distribution of $Y_t|W_t \sim \mathcal{N}(BW_t, \Sigma_\varepsilon)$, where B is $n_y \times np$ matrix of coefficients and Σ_ε is $n_y \times n_y$ covariance matrix of the errors. Then given Ω as the precision matrix of (Y_t, W_t), we can obtain $\Sigma = \Omega^{-1}$ which can be expressed as

$$\Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yw} \\ \Sigma_{wy} & \Sigma_{ww} \end{pmatrix} \tag{A.1}$$

where Σ_{wy} is $np \times n_y$ the covariances between W_t and Y_t, and Σ_{yy} is $n_y \times n_y$ covariances among Y_t. Then B and Σ_ε can be obtained from Σ by

$$B = \Sigma_{yw}\Sigma_{ww}^{-1}, \qquad \Sigma_\varepsilon = \Sigma_{yy} - \Sigma_{yw}\Sigma_{ww}^{-1}\Sigma_{wy} \tag{A.2}$$

Now given Σ as in (A.1), $\Omega = \Sigma^{-1}$ can be obtained as:

$$\Omega = \begin{pmatrix} (\Sigma_{yy} - \Sigma_{yw}\Sigma_{ww}^{-1}\Sigma_{wy})^{-1} & -(\Sigma_{yy} - \Sigma_{yw}\Sigma_{ww}^{-1}\Sigma_{wy})^{-1}\Sigma_{yw}\Sigma_{ww}^{-1} \\ -(\Sigma_{ww} - \Sigma_{wy}\Sigma_{yy}^{-1}\Sigma_{yw})^{-1}\Sigma_{wy}\Sigma_{yy}^{-1} & (\Sigma_{ww} - \Sigma_{wy}\Sigma_{yy}^{-1}\Sigma_{yw})^{-1} \end{pmatrix} \tag{A.3}$$

To complete the proof, we report the well-known Sherman-Morrison-Woodbury formula (see Woodbury, 1950). Thus, the inverse of a partitioned symmetric matrix is given by

$$\begin{aligned} (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} &= A_{11}^{-1} + A_{11}^{-1}A_{12}\Big(A_{22} - A_{21}A_{11}^{-1}A_{12}\Big)^{-1}A_{21}A_{11}^{-1} \\ -(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}A_{12}A_{22}^{-1} &= -A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \end{aligned} \tag{A.4}$$

Following (A.4) and the expressions in (A.2), (A.3) can be simplified as

$$\Omega = \begin{pmatrix} \Sigma_\varepsilon^{-1} & -\Sigma_\varepsilon^{-1}B \\ -B'\Sigma_\varepsilon^{-1} & \Sigma_{ww}^{-1} + B'\Sigma_\varepsilon^{-1}B \end{pmatrix} \tag{A.5}$$

□

Appendix A.2. Proof of Proposition 2

Proof. From the prior distributions in (8) and (10), $\bar{\eta}_i$ can be marginalized out as

$$P(\pi_i|p_i) = \frac{1}{2^{np_i}} \int_0^1 \mathbb{I}_{\{0,\ldots,f_i\}}(|\pi_i|) \frac{1}{B(a,b)}(\bar{\eta}_i)^{a-1}(1-\bar{\eta}_i)^{b-1}\, d\bar{\eta}_i \tag{A.6}$$

where $f_i = \lfloor \bar{\eta}_i m_p \rfloor$ with $m_p = \min\{np_i, T-p_i\}$, $\mathbb{I}_{\{0,\ldots,f_i\}}(|\pi_i|)$ is the indicator function

$$\mathbb{I}_{\{0,\ldots,f_i\}}(|\pi_i|) = \begin{cases} \mathbb{I}_{\{0\}}(|\pi_i|), & 0 \le \bar{\eta}_i < \frac{1}{m_p} \\ \quad\vdots & \cdots \\ \mathbb{I}_{\{0,\ldots,m_p-1\}}(|\pi_i|), & \frac{m_p-1}{m_p} \le \bar{\eta}_i < 1 \\ \mathbb{I}_{\{0,\ldots,m_p\}}(|\pi_i|), & \bar{\eta}_i = 1 \end{cases} \tag{A.7}$$

Let $f(\bar{\eta}_i) = (B(a,b))^{-1}\ \bar{\eta}_i^{a-1}(1-\bar{\eta}_i)^{b-1}$. From (A.6)

$$\begin{aligned} P(\pi_i|p_i) &= \frac{1}{2^{np_i}} \left[\mathbb{I}_{\{0\}}(|\pi_i|) \int_0^{\frac{1}{m_p}} f(\bar{\eta}_i)\, d\bar{\eta}_i + \ldots + \mathbb{I}_{\{0,\ldots,m_p-1\}}(|\pi_i|) \int_{\frac{m_p-1}{m_p}}^1 f(\bar{\eta}_i)\, d\bar{\eta}_i \right] \\ &= \frac{1}{2^{np_i}} \left[\sum_{j=0}^{m_p-1} \mathbb{I}_{\{0,\ldots,j\}}(|\pi_i|) \left(I_{\frac{j+1}{m_p}}(a,b) - I_{\frac{j}{m_p}}(a,b) \right) \right] \end{aligned} \tag{A.8}$$

where $I_z(a,b) = \int_0^z f(\bar{\eta}_i) d\bar{\eta}_i$ is the incomplete beta function (Abramowitz and Stegun, 1964, p. 263). □

Appendix A.3. Proof of Corollary 4.1

Proof. By assuming a uniform prior on $\bar{\eta}_i$, $f(\bar{\eta}_i) = 1$. Furthermore, the difference between the incomplete beta functions in (A.8) is $I_{\frac{j+1}{m_p}}(a,b) - I_{\frac{j}{m_p}}(a,b) = \frac{1}{m_p}$. Thus

$$P(\pi_i|p_i) = \frac{1}{2^{np_i}} \frac{1}{m_p} \sum_{j=0}^{m_p-1} \mathbb{I}_{\{0,\ldots,j\}}(|\pi_i|) = \frac{1}{2^{np_i}} \left(1 - \frac{|\pi_i|}{m_p}\right) \tag{A.9}$$

□

Appendix A.4. Proof of Proposition 3

Proof. The function $\varphi(k)$ is convex if and only if $\varphi''(k) > 0$, $\forall k$. By defining $\varphi(k) = -\log P(\pi_i|p_i) = np_i \log(2) + \log(m_p) - \log(m_p - k)$, for $|\pi_i| = k$, it can be shown that

$$\varphi''(k) = \frac{1}{(m_p-k)^2} > 0 \tag{A.10}$$

□

Appendix B. Convergence Diagnostics and Posterior Approximation

Appendix B.1. Convergence Diagnostics

For our graphical approach, we monitor the convergence of the MCMC chain using the potential scale reduction factor (PSRF), see Gelman and Rubin (1992). See also Casella and Robert (2004), ch. 12, for a review on methods for convergence monitoring in MCMC. The PSRF monitors the within-chain and between-chain covariances of the global log posterior densities of the sampled structures to test whether The chain is said to have properly converged if $PSRF \leq 1.2$. Figure B.5 display a comparison of the MCMC convergence diagnostics for a random initialization and our initialization procedure of the graph averaged over lags. Figures B.6 and B.7 shows plots of links and graph score at each MCMC iteration, the local graph BIC for the lags for the simulation experiments and the macroeconomic application respectively.

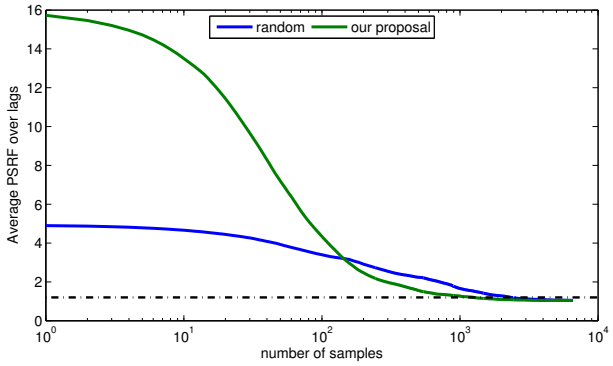
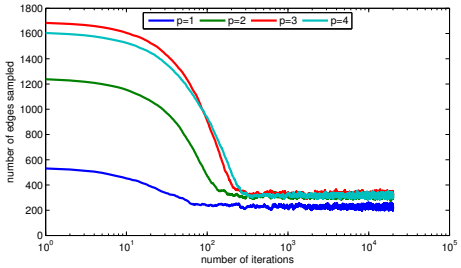


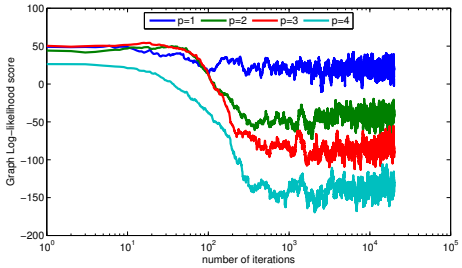
Figure B.5: Comparison of the MCMC convergence diagnostics for a random initialization (in blue) and our initialization (in green) procedure of the graph averaged over lags. The black dashed line is 1.2, and colored lines close to this line indicate convergence of the chain.

Appendix B.2. Edge Posterior Approximation

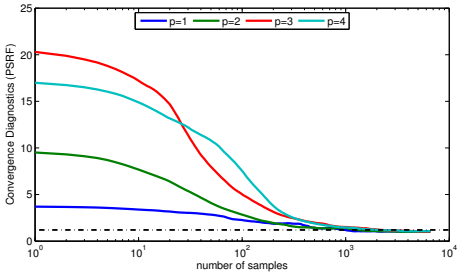
We estimate the posterior probability of the edge by $\hat{e}_{ij}$, which is the average of the MCMC samples from the G_{ij} posterior distribution. For variable selection purposes, we define the estimator G^*_{ij} of the edge from X^j to X^i based on a one sided posterior credibility interval for the edge posterior distribution, and find the interval lower bound $G^*_{ij} = 1$ if $\hat{e}_{ij} - z_{(1-\alpha)}\sqrt{\frac{\hat{e}_{ij}(1-\hat{e}_{ij})}{n_{eff}}} > 0.5$, where n_{eff} is the effective sample size representing the number of independent posterior samples of the graph, and $z_{(1-\alpha)}$ is the z-score of the normal distribution at the $(1-\alpha)$ significance level.



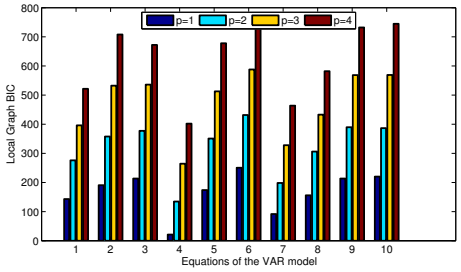
(a) Number of links and iterations



(b) Log score and iterations

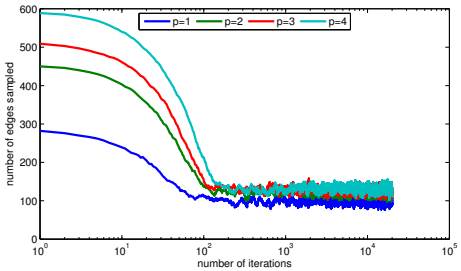


(c) Convergence diagnostics (PSRF)

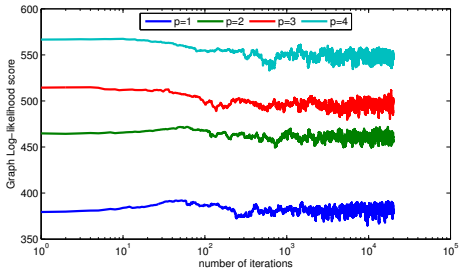


(d) Local graph BIC for the lags

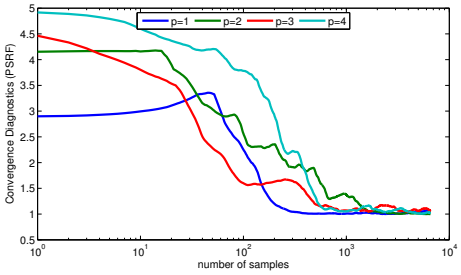
Figure B.6: Plots of (B.6a) links and (B.6b) graph score at each MCMC iteration, with (B.6c) convergence diagnostics and (B.6d) local graph BIC for the lags for each equation of the simulation experiments.



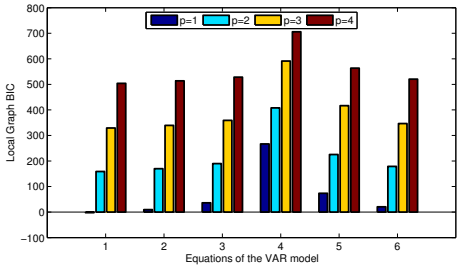
(a) Number of links and iterations



(b) Log score and iterations



(c) Convergence diagnostics (PSRF)



(d) Local graph BIC for the lags

Figure B.7: Plots of (B.7a) links and (B.7b) graph score at each MCMC iteration, with (B.7c) convergence diagnostics and (B.7d) local graph BIC for the lags for each equation of the macroeconomic application.

Appendix C. Data Description For Large Macroeconomic Application

Table C.3 provides a description and stationarity transformation codes used for our large macroeconomic application in Section 6.

Appendix D. Pseudo-Code for Sparse Graph Selection

Algorithm 1 presents a description of the pseudo-code for the sparse graphical model selection.

No.	Mnemonic	Description	Tcode
1	CPIAUCSL*	Consumer Price Index for All Urban Consumers: All Items	6
2	FEDFUNDS*	Effective Federal Funds Rate	2
3	GDPC96*	Real Gross Domestic Product, 3 Decimal	5
4	GPDIC96*	Real Gross Private Domestic Investment, 3 decimal	5
5	INDPRO*	Industrial Production Index	5
6	PCECC96*	Real Personal Consumption Expenditures	5
7	AAA	Moody's Seasoned Aaa Corporate Bond Yield	2
8	AHECONS	Ave Hr Earnings Of Prod & Nonsupervisory Employees: Construction	6
9	AHEMAN	Ave Hr Earnings of Prod & Nonsupervisory Empl: Manufacturing	6
10	AWHMAN	Ave Wkly Hr of Prod & Nonsupervisory Empl: Manufacturing	5
11	AWOTMAN	Ave Wkly Overtime Hrs of Prod & Nonsup. Empl: Manufacturing	5
12	BAA	Moody's Seasoned Baa Corporate Bond Yield	2
13	BORROW	Total Borrowings of Depository Institutions from the Federal Reserve	6
14	BUSLOANS	Commercial and Industrial Loans, All Commercial Banks	6
15	CBIC96	Real Change in Private Inventories	1
16	CCFC	Corporate: Consumption of Fixed Capital	6
17	CIVPART	Civilian Labor Force Participation Rate	5
18	CONSUMER	Consumer Loans at All Commercial Banks	5
19	CP	Corporate Profits After Tax (without IVA and CCAdj)	6
20	CPIAPPSL	Consumer Price Index for All Urban Consumers: Apparel	6
21	CPIENGSL	Consumer Price Index for All Urban Consumers: Energy	6
22	CPILEGSL	Consumer Price Index for All Urban Consumers: All Items Less Energy	6
23	CPIMEDSL	Consumer Price Index for All Urban Consumers: Medical Care	6
24	CPITRNSL	Consumer Price Index for All Urban Consumers: Transportation	6
25	CPIUFDSL	Consumer Price Index for All Urban Consumers: Food	6
26	CPIULFSL	Consumer Price Index for All Urban Consumers: All Items Less Food	6
27	CURRCIR	Currency in Circulation	6
28	CURRSL	Currency Component of M1	6
29	DEMDEPSL	Demand Deposits at Commercial Banks	6
30	DIVIDEND	Corporate Profits after tax with IVA and CCAdj: Net Dividends	6
31	DMANEMP	All Employees: Durable goods	5
32	DPIC96	Real Disposable Personal Income	6
33	EMRATIO	Civilian Employment-Population Ratio	5
34	EXCAUS	Canada / U.S. Foreign Exchange Rate	5
35	EXJPUS	Japan / U.S. Foreign Exchange Rate	5
36	EXPGSC96	Real Exports of Goods & Services, 3 Decimal	5
37	EXSZUS	Switzerland / U.S. Foreign Exchange Rate	5
38	EXUSUK	U.S. / U.K. Foreign Exchange Rate	5
39	FINSLC96	Real Final Sales of Domestic Product	5
40	GCEC96	Real Government Consumption Expenditures & Gross Investment	5
41	GDPDEF	Gross Domestic Product: Implicit Price Deflator	5
42	GPDICTPI	Gross Private Domestic Investment: Chain-type Price Index	6
43	GS1	1-Year Treasury Constant Maturity Rate	2
44	GS10	10-Year Treasury Constant Maturity Rate	2
45	GS3	3-Year Treasury Constant Maturity Rate	2
46	GS5	5-Year Treasury Constant Maturity Rate	2
47	GSAVE	Gross Saving	5
48	HOUST	Housing Starts: Total: New Privately Owned Housing Units Started	4
49	HOUST1F	Privately Owned Housing Starts: 1-Unit Structures	4
50	HOUST5F	Privately Owned Housing Starts: 5-Unit Structures or More	4
51	HOUSTMW	Housing Starts in Midwest Census Region	4
52	HOUSTNE	Housing Starts in Northeast Census Region	4
53	HOUSTS	Housing Starts in South Census Region	4
54	HOUSTW	Housing Starts in West Census Region	4
55	IMPGSC96	Real Imports of Goods & Services, 3 Decimal	5
56	INVEST	Securities in Bank Credit at All Commercial Banks	5
57	IPBUSEQ	Industrial Production: Business Equipment	5
58	IPCONGD	Industrial Production: Consumer Goods	5
59	IPDCONGD	Industrial Production: Durable Consumer Goods	5
60	IPDMAT	Industrial Production: Durable Materials	5
61	IPFINAL	Industrial Production: Final Products (Market Group)	5
62	IPMAT	Industrial Production: Materials	5
63	IPNCONGD	Industrial Production: Nondurable Consumer Goods	5

Table C.3: Data description and transformation codes. 1 = no transformation, 2 = first difference, 4 = log, 5 = 100×(first difference of log), 6 = 100×(second difference of log). *- The dependent variables.

No.	Mnemonic	Description	Tcode
64	IPNMAT	Industrial Production: nondurable Materials	5
65	LOANS	Loans and Leases in Bank Credit, All Commercial Banks	6
66	M1SL	M1 Money Stock	6
67	M1V	Velocity of M1 Money Stock	5
68	M2SL	M2 Money Stock	6
69	M2V	Velocity of M2 Money Stock	5
70	MCUMFN	Capacity Utilization: Manufacturing (NAICS)	1
71	MPRIME	Bank Prime Loan Rate	2
72	MZMSL	MZM Money Stock	6
73	NAPM	ISM Manufacturing: PMI Composite Index	1
74	NAPMEI	ISM Manufacturing: Employment Index	1
75	NAPMII	ISM Manufacturing: Inventories Index	1
76	NAPMNOI	ISM Manufacturing: New Orders Index	1
77	NAPMPI	ISM Manufacturing: Production Index	1
78	NAPMPRI	ISM Manufacturing: Prices Index	1
79	NAPMSDI	ISM Manufacturing: Supplier Deliveries Index	1
80	NDMANEMP	All Employees: Nondurable goods	5
81	NONREVSL	Total Nonrevolving Credit Owned and Securitized, Outstanding	6
82	OPHPBS	Business Sector: Real Output Per Hour of All Persons	5
83	PAYEMS	All Employees: Total nonfarm	5
84	PCDG	Personal Consumption Expenditures: Durable Goods	5
85	PCECTPI	Personal Consumption Expenditures: Chain-type Price Index	5
86	PCESV	Personal Consumption Expenditures: Services	5
87	PCND	Personal Consumption Expenditures: Nondurable Goods	5
88	PFCGEF	Producer Price Index: Finished Consumer Goods Excluding Foods	6
89	PINCOME	Personal Income	6
90	PNFI	Private Nonresidential Fixed Investment	6
91	PPIACO	Producer Price Index: All Commodities	6
92	PPICPE	Producer Price Index: Finished Goods: Capital Equipment	6
93	PPICRM	Producer Price Index: Crude Materials for Further Processing	6
94	PPIFCF	Producer Price Index: Finished Consumer Foods	6
95	PPIFCG	Producer Price Index: Finished Consumer Goods	6
96	PPIFGS	Producer Price Index: Finished Goods	6
97	PPIITM	Producer Price Index: Intermediate Materials: Supplies & Components	6
98	PRFI	Private Residential Fixed Investment	6
99	PSAVE	Personal Saving	5
100	REALLN	Real Estate Loans, All Commercial Banks	6
101	SAVINGSL	Savings Deposits - Total	6
102	SLEXPND	State & Local Government Current Expenditures	6
103	SLINV	State & Local Government Gross Investment	6
104	SP500	S&P 500	5
105	SRVPRD	All Employees: Service-Providing Industries	5
106	TB3MS	3-Month Treasury Bill: Secondary Market Rate	2
107	TB6MS	6-Month Treasury Bill: Secondary Market Rate	2
108	TCDSL	Total Checkable Deposits	6
109	TOTALSL	Total Consumer Credit Owned and Securitized, Outstanding	6
110	TVCKSSL	Travelers Checks Outstanding	6
111	UEMP15T26	Number of Civilians Unemployed for 15 to 26 Weeks	5
112	UEMP27OV	Number of Civilians Unemployed for 27 Weeks and Over	5
113	UEMP5TO14	Number of Civilians Unemployed for 5 to 14 Weeks	5
114	UEMPLT5	Number of Civilians Unemployed - Less Than 5 Weeks	5
115	ULCNFB	Nonfarm Business Sector: Unit Labor Cost	5
116	UNRATE	Civilian Unemployment Rate	2
117	USCONS	All Employees: Construction	5
118	USEHS	All Employees: Education & Health Services	5
119	USFIRE	All Employees: Financial Activities	5
120	USGOOD	All Employees: Goods-Producing Industries	5
121	USGOVT	All Employees: Government	5
122	USINFO	All Employees: Information Services	5
123	USLAH	All Employees: Leisure & Hospitality	5
124	USMINE	All Employees: Mining and logging	5
125	USPBS	All Employees: Professional & Business Services	5
126	USPRIV	All Employees: Total Private Industries	5
127	USTPU	All Employees: Trade, Transportation & Utilities	5
128	USTRADE	All Employees: Retail Trade	5
129	USWTRADE	All Employees: Wholesale Trade	5
130	WASCUR	Compensation of Employees: Wages & Salary Accruals	6

Algorithm 1 Graphical VAR Model Selection Algorithm

for $p \in [\underline{p}, \ldots, \bar{p}]$ **do**

Initialize $\vec{G}_p$ as $(n_y \times np)$ zero matrix, set $m_p = \min\{np, T-p\}$

$\mathbf{V}^i_{p,x}$ the vector of all possible explanatory variables up to lag p

$\mathbf{V}_y$ the vector of dependent variables

for each $y_i \in \mathbf{V}_y$ **do**

for each $x_k \in \mathbf{V}^i_{p,x}$ **do**

if x_k is equal to lag 1 of y_i **then**

Set $\vec{G}_p(y_i, x_k) = 1$

else

Compute $H_0 = P(\mathcal{X}|p_i, \vec{G}_p(y_i, \emptyset))$ and $H_1 = P(\mathcal{X}|p_i, \vec{G}_p(y_i, \{x_k\}))$

if $H_1 > H_0$ **then**

Set $\vec{G}_p(y_i, x_k) = 1$ and retain x_k in $\mathbf{V}^i_{p,x}$

else

Set $\vec{G}_p(y_i, x_k) = 0$ and remove x_k from $\mathbf{V}^i_{p,x}$

Set $N_p(\pi_i)$ the set of variables, $x_k's$, retained in $\mathbf{V}^i_{p,x}$

for $j = 1 \to J$, the total iterations **do**

for each $y_i \in \mathbf{V}_y$ **do**

Set $\pi_i^{(j-1)}$ = the set explanatory variables of y_i in $\vec{G}^{(j-1)}_{p,i}$

Draw $\bar{\eta}_i^{(*)}$ from a $\mathcal{B}e(a,b)$ and set $f_i^{(*)} = \lfloor m_p \bar{\eta}_i^{(*)} \rfloor$

if $|\pi_i^{(j-1)}| < f_i^{(*)}$ **then**

Set $Q(\vec{G}^{(*)}_{p,i}|\vec{G}^{(j-1)}_{p,i}, \bar{\eta}_i^{(*)}) = 1/|N_p(\pi_i)|$,

Draw $x_k \in N_p(\pi_i)$

Add/remove edge; i.e. $\vec{G}_p^{(*)}(y_i, x_k) = 1 - \vec{G}_p^{(j-1)}(y_i, x_k)$

else

Set $Q(\vec{G}^{(*)}_{p,i}|\vec{G}^{(j-1)}_{p,i}, \bar{\eta}_i^{(*)}) = 1/|\pi_i^{(j-1)}|$

Draw $x_k \in \pi_i^{(j-1)}$

Remove edge; i.e. $\vec{G}_p^{(*)}(y_i, x_k) = 0$

Set $\pi_i^{(*)}$ = the set explanatory variables of y_i in $\vec{G}^{(*)}_{p,i}$

Draw $\bar{\eta}_i^{(**)}$ from a $\mathcal{B}e(a,b)$ and set $f_i^{(**)} = \lfloor m_p \bar{\eta}_i^{(**)} \rfloor$

if $|\pi_i^{(*)}| < f_i^{(**)}$ **then**

Set $Q(\vec{G}^{(j-1)}_{p,i}|\vec{G}^{(*)}_{p,i}, \bar{\eta}_i^{(**)}) = 1/|N_p(\pi_i)|$

else

Set $Q(\vec{G}^{(j-1)}_{p,i}|\vec{G}^{(*)}_{p,i}, \bar{\eta}_i^{(**)}) = 1/|\pi_i^{(*)}|$

Sample $u \sim \mathcal{U}_{[0,1]}$

Compute R_A following equation (18)

if $u < \min\{1, R_A\}$ **then**

$\vec{G}^{(j)}_{p,i} = \vec{G}^{(*)}_{p,i}$

else

$\vec{G}^{(j)}_{p,i} = \vec{G}^{(j-1)}_{p,i}$