Ca' Foscari University of Venice

Department of Economics

Working Paper

Daniel Felix Ahelegbey, Monica Billio and Roberto Casarin

Bayesian Graphical Models for Structural Vector Autoregressive Processes

ISSN: 1827/3580
No. 36/WP/2012

Working Papers
Department of Economics
Ca' Foscari University of Venice
No. 36/WP/2012
ISSN 1827-3580

Bayesian Graphical Models for Structural Vector Autoregressive Processes

Daniel Felix Ahelegbey[†] **Monica Billio**[†] **Roberto Casarin**[†]

[†] *University Ca' Foscari of Venice*

December 2012

Abstract
Vector autoregressive models have widely been applied in macroeconomics and macroeconometrics to estimate economic relationships and to empirically assess theoretical hypothesis. To achieve the latter, we propose a Bayesian inference approach to analyze the dynamic interactions among macroeconomics variables in a graphical vector autoregressive model. The method decomposes the structural model into multivariate autoregressive and contemporaneous networks that can be represented in the form of a directed acyclic graph. We then simulated the networks with an independent sampling scheme based on a single-move Markov Chain Monte Carlo (MCMC) approach. We evaluated the efficiency of our inference procedure with a synthetic data and an empirical assessment of the business cycles hypothesis.

Keywords: Bayesian Graphical models, Markov Chain Monte Carlo, Structural Vector Autoregression, Directed Acyclic Graph, Bayesian Inference, Dynamic Bayesian Network.

JEL Codes: C11, C15, C53, G17

Address for correspondence:
Daniel Felix Ahelegbey
Department of Economics
Ca' Foscari University of Venice
Cannaregio 873, Fondamenta S.Giobbe
30121 Venezia - Italy
Phone: (++39) 041 2349149
Fax: (++39) 041 2349176
e-mail: dfkahey@yahoo.com

The Working Paper Series
is available only on line
(http://www.unive.it/nqcontent.cfm?a_id=86302)
For editorial correspondence, please contact:
wp.dse@unive.it

Department of Economics
Ca' Foscari University of Venice
Cannaregio 873, Fondamenta San Giobbe
30121 Venice Italy
Fax: ++39 041 2349210

1. Introduction

A vector autoregressive (VAR) model is an extension of an autoregressive (AR) model to the multivariate case. VAR allows analyzing how realizations of variables in past times influence current realizations. Introduced by Sims (1980), VAR has widely been applied in macroeconomics and macroeconometrics to estimate economic relationships and to empirically assess theoretical hypothesis. VAR models have been proven to be very suitable in evaluating the impact of economic shocks on key macroeconomic variables such as production, investment, consumption, monetary policy, interest rates, the effects of fiscal policy, and the dynamics of financial time series etc.

Despite the advantages of the VAR, the problem of identification in the estimation of the structural model renders some limitation to the use of the model for forecasting. The standard approach to overcome this problems attempts to estimate the VAR in reduced form. However, it is clear that the reduced form is unable to offer economic interpretations. To deal with the economic interpretations of the model, the standard techniques imposes restrictions to recover the structural parameters from the reduced form parameters. The identification problem has been discussed by Cooley and Leroy (1985), Bernanke (1986), King et al. (1991) etc.

In structural analysis, certain assumptions about the causal structure of the data under investigation are often imposed. Imposing such restrictions to the dynamics of the structure leads to a cost in the generalization of the results. Since the VAR are suppose to empirically assess theoretical hypothesis, it is quite hard to provide convincing restrictions without relying on theories which undermines the use of such models to achieve the purpose for which it was designed for. It is therefore clear that inferences drawn from such models overlaid with restrictions that are difficult to defend poses another problem. Furthermore, drawing causal relations from correlations among variables on which we have data is another limitation when using standard estimation techniques. It can also be noted that, inference of contemporaneous interaction of variables using impulse response functions from residuals leads to high standard error problems which affects the accuracy in forecasting.

The method discussed in this paper draws on the use of graphical models for structural VAR (SVAR) analysis. Graphical models presents the idea that interaction among random variables in a system can be represented in the form of graphs where the nodes represents the variables and the edges shows the interactions. (See Pearl (1988), Lauritzen and Wermuth (1989), Whittaker (1990), Wermuth and Lauritzen (1990) and Edward (1990)). It presents a framework with clarity of interpretation and the ease to analyze seemingly complex interactions. Graphical models approach to causal studies was first introduced by Pearl (2000) and Spirtes et al. (2000) in artificial intelligence. However in recent times, there have been quite a number of applications in economics by Swanson and Granger (1997), Hoover (2001), Raele and Tunnicliffe Wilson (2001), Bessler and Lee (2002), Demiralp and Hoover (2003), Moneta (2008) etc., all of which have proven the reliability of this method.

In contrast to these papers which demonstrate potential applications of graphical models for VAR processes, the current paper considers a Bayesian inference approach to address the identification problem. To empirically asses theoretical hypothesis, it is important to apply techniques that are able to learn models from data. Inference of a graphical model is a model determination problem that has been discussed by Corander (2003), Corander and Villani (2006), Giudici and Green (1999). The intuition that our observed data could have been produced by a model but faced with the uncertainty in the dependence structure and the parameters of the underlying model. The Bayesian approach as has been discussed by Madigan and York (1995), Giudici and Green (1999) and Dawid and Lauritzen (2001) provides a convenient framework to handle the above problem and to recover the model probabilistically from data. Corander and Villani (2006) introduced the Bayesian approach to model graphical VAR processes. Due to limitations on the implementation of the Markov Chain Monte Carlo (MCMC) for graphical modeling of time series data at that time, the paper applied the fractional Bayes approach for inference about the causal structure and the lag length of the process. However, recent development in the application of MCMC for graphical models by Madigan and York (1995), Friedman and Koller (2003), Grzegorczyk and Husmeier (2008), Grzegorczyk and Husmeier (2009), Grzegorczyk (2010) etc., makes it feasible for sampling graphical models for multivariate random variables.

The application of a Bayesian approach provides inferences that are conditional on the observed data without reliance on asymptotic approximation and data transformations. In the current paper, we develop the Bayesian approach used by Giudici and Green (1999) and Grzegorczyk (2010). We addressed the identification problem by decomposing the structural model into multivariate autoregressive and contemporaneous networks that can be represented in the form of a directed acyclic graph. We then simulate the networks with an independent sampling scheme based on a single-move Markov Chain Monte Carlo (MCMC) approach proposed by Giudici and Green (1999), and applied by Grzegorczyk and Husmeier (2009), Grzegorczyk (2010) and Grzegorczyk et al (2011) for time series models. Most of the existing applications of graphical models for VAR uses greedy search, K2 algorithm or the PC algorithm in sampling the network. Bayesian inference via MCMC algorithms have been proved to be more efficient and allows to sample more complicated models that cannot be dealt with using standard approaches.

The rest of the paper is organized as follows. Section 2 introduces the Bayesian approach to graphical vector autoregressive models. Section 3 introduces the independent sampling scheme to the graph posterior computation. Section 4 illustrate the application of the inference procedure with a synthetic data. Section 5 illustrates an empirical assessment for business cycle analysis.

2. Bayesian Graphical Vector Autoregressive Models

A vector autoregressive (VAR) process of order L is of the form

$$X_t = B_0 X_t + B_1 X_{t-1} + \ldots + B_L X_{t-L} + \varepsilon_t \qquad t = 1, \ldots, m \tag{1}$$

where X_t is an n dimensional vector of time series realizations at time t, ε_t is an n dimensional vector independent and serially uncorrelated structural disturbances with mean zero and a diagonal matrix Σ_ε, and $B_0, \ldots, B_L$ are $n \times n$ regression matrices. B_0 is a zero diagonal matrix. Let $X_t = \{X_t^1, X_t^2, \ldots, X_t^n\}$, where X_t^i is a realization of the variable X^i at time t such that $X_t^i \in X_t$.

The above model can be represented in a graphical form which presents a convenient framework for modeling multivariate time series observations. As defined by Brillinger (1996), graphical models are simply statistical models embodying a collection of marginal and conditional independences which may be summarized by means of graphs. It can be shown that there is a one-to-one correspondence between the regression matrices and directed acyclic graphs (DAGs), given as follows (Murphy, 2002);

$$X_{t-s}^j \to X_t^i \iff B_s(i,j) \neq 0 \qquad 0 \leq s \leq L$$

Prominent among the statistical models based on directed graphs are Bayesian networks (BNs). A Bayesian network (BN) is simply a statistical model that combines graph theory and probability theory (based on Baye's Rule) to model interactions between random variables. Formally, a BN is a pair $\mathcal{B} = \{\mathcal{G}, \Theta\}$. The first component, $\mathcal{G}$ is a DAG composed of $\{\mathcal{V}, \mathcal{E}\}$ where $\mathcal{V}$ is the set of vertices (nodes) that represents the random variables and $\mathcal{E}$ are directed edges. The second component, Θ is the set of parameters of the model. In the BN framework, the node where an edge originates is called the *parent node* and where the edge ends is called the *child node*

Dynamic Bayesian network (DBN) is an extension of the BN representation to model stochastic processes. It provides a general formalism for modeling multivariate time series. The DBN assumes a Markovian process, thus the assumption that current realizations of random variables only depends on the immediate past realizations and not on all past observations. Generally, a DBN is a pair $(\mathcal{B}_0, \mathcal{B}_\rightarrow)$, where $\mathcal{B}_0$ is a regular BN which defines the initial state distribution of the variables; and $\mathcal{B}_\rightarrow$ is a the transition network (Murphy, 2002). In the $\mathcal{B}_\rightarrow$, the parents of a node can either be in the same time slice or in the previous time slice. Directed edges within a slice represents instantaneous causation, where as edges between slices represents autoregressive causations.

A DBN composed of Gaussian random variables is referred to as a Dynamic Gaussian network (DGN). The standard representation a VAR process of order L as expressed in (1) is equivalent to a DBN of order L where the conditional probability distributions (CPDs) are linear-Gaussian, (Murphy, 2002). The general concept in the application of DAG for VAR processes assumes that all slices including the initial state distribution have the same structure, in which case the DBN can simply be defined using only the transition network (Friedman et al, 1998). In this paper, we refer to the slice of instantaneous causation as the multivariate instantaneous (MIN) network and the slices of the autoregressive causation as the multivariate autoregressive (MAR) network.

To empirically asses theoretical hypothesis as is the main objective of the application of VAR models, it is important to apply techniques that are able to learn models from data. Consider for example a dynamic model that follows a Markov process of order 1. For structure learning of the MIN network, a recursive formula on the number of possible DAGs that contains n nodes is given as follows (Robinson, 1977);

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-1)$$

where $\binom{n}{i}$ are the binomial coefficients. The structure learning of the possible DAGs that represents the MAR network is given as follows;

$$h(n) = 2^{n^2}$$

Thus the structure learning is a challenging problem since the possible DAGs is super-exponential in the number of nodes. This is referred to as the model determination problem and has been discussed by Corander (2003), Corander and Villani (2006), Giudici and Green (1999). The Bayesian inference approach has been proved to be convenient to handle such problems in the sense that it allows both informative priors so that experts knowledge can be used to inform the current model search. This has been discussed by Madigan and York (1995), Giudici and Green (1999) and Dawid and Lauritzen (2001).

2.1. Bayesian inference

Let $\mathcal{X} = (X_1, \ldots, X_t)_{t=1,\ldots,m}$ be a complete time series dataset of length m with n variables. A complete dataset is a database that contains no missing data. The Bayesian inference approach to the model determination problem is to learn the model structure probabilistically conditional on the observed data. This requires prior elicitation, likelihood estimations and posterior computations.

2.1.1. Priors

According to Friedman and Koller (2003), to define the model prior is to define a discrete probability distribution over graph structures $\mathcal{G}$, and for each possible graph, to define a continuous distribution over the set of parameters Θ. This is given by;

$$P(\mathcal{G}, \Theta) = P(\mathcal{G})P(\Theta|\mathcal{G}) \tag{2}$$

Graph priors

Of the two priors, the graph prior is usually considered less important since its posterior does not grow with the number of data points (Friedman and Koller, 2003). Therefore the commonly used graph prior which is also used in this paper is to assume that all DAGs (directed acyclic graphs) are equally likely, thus assuming a uniform prior over all possible DAGs. This approach is mostly used only for simplicity and can be refined in various ways. Thus if some DAGs are not possible, then the priors can be redefined to assign a zero or very low probabilities to those configurations. The rest of the configurations can then be assigned equal prior probabilities.

Parameter priors

According to Geiger and Heckerman (1999), the parameter priors must satisfy two important assumptions; global parameter independence and parameter modularity.

Assumption 1. (Global Parameter Independence) *Let the parents set of X_t^i be $\pi^i(t)$. For every DAG model $\mathcal{G}$ for $\mathcal{X}$,*

$$P(\Theta|\mathcal{G}) = \prod_{i=1}^{n} P(\theta_i|\mathcal{G}) \tag{3}$$

where $\theta_i = \theta_{X_t^i|\pi^i(t)}$ is the parameter sub-vector associated with node X_t^i given its parents set $\pi^i(t)$. This means that the parameters are mutually independent apriori given the graph configuration. This assumption allows us to calculate for the likelihood of a single case.

Assumption 2. (Parameter Modularity) *For every two DAG models $G_1, G_2 \in \mathcal{G}$ for $\mathcal{X}$ such that X_t^i has the same parents in G_1 and G_2,*

$$P(\theta_i|G_1) = P(\theta_i|G_2) \tag{4}$$

This means that if X_t^i has the same set of parents in two or more different structures, then the associated parameters must be the same.

Inference of the underlying model involves examining a large number of possible network structures. To avoid having to assign a prior distribution over parameters for each possible structure, the standard approach allows the parameter prior for all structures to be specified using a single network.

2.1.2. Likelihood

For likelihood estimation, Geiger and Heckerman (1999) identified two important assumptions; Likelihood modularity and Random sample condition.

Assumption 3. (Likelihood Modularity) *For every two DAG models $G_1, G_2 \in \mathcal{G}$ for $\mathcal{X}$ such that X_t^i has the same parents in G_1 and G_2, the local distribution for X_t^i in both models are the same;*

$$P(X_t^i|\pi^i(t), \theta_i, G_1) = P(X_t^i|\pi^i(t), \theta_i, G_2) \tag{5}$$

Assumption 4. (Random Sample Condition) *For any $Y \subseteq X$, define $\mathcal{X}^Y$ can be a random sample from $\mathcal{X}$ restricted to Y. Let G be a DAG for any ordering where the variables in Y come first. Then by assumption, global parameter independence and likelihood modularity, (Geiger and Heckerman, 1999);*

$$P(Y|\mathcal{X}, G) = P(Y|\mathcal{X}^Y, G) \tag{6}$$

This means that if our observation $\mathcal{X}$ is from some distribution, then restricting our dataset to random samples $Y \subseteq X$ also follows a similar distribution.

Assumption 5. (Homogeneous Markov Property) *If the model is time homogeneous, the semantics of a DBN of Markov order 1 can be defined by unrolling a two-slice temporal Bayes net (2TBN) until we have m time slices. The resulting joint distribution is given by*

$$P(\mathcal{X}) = \prod_{t=1}^{m}\prod_{i=1}^{n} P(X_t^i|\pi^i(t)) \tag{7}$$

This means that if the model follows a first-order homogeneous Markov, the transition network can be expressed as a 2TBN. By this assumption, the parameters of the CPDs are time invariant and the parents set of X_t^i becomes time invariant.

By the chain rule of probability and the homogeneous Markov chain expression, the joint likelihood is given by;

$$P(\mathcal{X}|\mathcal{G}, \Theta) = \prod_{i=1}^{n}\prod_{t=1}^{m} P(X_t^i|\pi^i, \mathcal{X}_{t-1}, \mathcal{G}, \Theta) \tag{8}$$

where $X_1^{-i} = X_{1\setminus\{i\}}$ is X_1 without X^i, $(X_1 \setminus X^i)$, $X_t^{-i} = X_{t\setminus\{i\}}$ corresponds to $X_t \setminus X^i$ and $\mathcal{X}_{t-1} = (X_1, \ldots, X_s)_{s=1,\ldots,t-1}$.

Given a complete graph G (a graph with no missing edges), by global parameter independence (Assumption 1) and parameter modularity (Assumption 2), the general approach to this problem is to marginalize the likelihood over the space of all parameters. The marginal likelihood can be expressed as follows;

$$\begin{aligned} P(\mathcal{X}|\mathcal{G}) &= \prod_{i=1}^{n} \prod_{t=1}^{m} \int_{\theta_i \in \Theta} P(X_t^i|\pi^i, \mathcal{X}_{t-1}, G, \theta_i)\ P(\theta_i|G)\ d\theta_i \\ &= \prod_{i=1}^{n} \prod_{t=1}^{m} P(X_t^i|\pi^i, \mathcal{X}_{t-1}, G) \end{aligned} \tag{9}$$

By the random sample condition (Assumption 4) and the homogeneous Markov property (Assumption 5), the marginal likelihood expression can be represented invariant of time. Given $\mathcal{X}_{m-1} = (X_1, \ldots, X_s)_{s=1,\ldots,m-1}$, the marginal likelihood can be simplified as follows (Geiger and Heckerman, 1999);

$$\begin{aligned} P(\mathcal{X}|\mathcal{G}) &= \prod_{i=1}^{n} \frac{P(X^i, \pi^i|\mathcal{X}_{m-1}, G)}{(\pi^i|\mathcal{X}_{m-1}, G)} \\ &= \prod_{i=1}^{n} \frac{P(\mathcal{X}_{m-1}^{(X^i,\pi^i)}|G^{(X^i,\pi^i)})}{P(\mathcal{X}_{m-1}^{(\pi^i)}|G^{(\pi^i)})} \end{aligned} \tag{10}$$

where $\mathcal{X}_{m-1}^{(X^i,\pi^i)}$ and $\mathcal{X}_{m-1}^{(\pi^i)}$ are sub-matrices of the data matrix $\mathcal{X}_{m-1}$ consisting only of the rows that correspond to the variable in the subsets $D_1 = \{X^i, \pi^i\}$ and $D_2 = \{\pi^i\}$. $G^{(\cdot)}$ is an arbitrary structure that represents a complete DAG over the variables to which the corresponding dataset $\mathcal{X}^{(\cdot)}$ is restricted. Let $D = \{D_1, D_2\}$, then the dataset $\mathcal{X}_{m-1}^{D} \subseteq \mathcal{X}_{m-1}$ which consists of $n^* \times (m-1)$ realizations, where n^* is the dimensional subset $D \subseteq \{X_1, \ldots, X_n\}$.

Bayesian Gaussian Equivalent (BGE) Score

The expression in (10) is referred to as the Bayesian likelihood metric whose closed-form computation depends on the distributions of the random variables. Based on the assumption that the random variables are samples from a multivariate normal distribution $\mathcal{N}_n(\mu, W)$ (where μ is a vector of unknown means, $W = \Sigma^{-1}$ is the precision matrix and Σ is the covariance matrix). The standard prior is a normal-Wishart distribution. The conditional prior $P(\mu|W)$ is a normally distributed with mean μ_0 and precision κW where $\kappa > 0$; and $P(W)$ is a Wishart distributed with $\alpha > n$ degrees of freedom and a scale matrix T_0. The posterior $P(\mu, W|\mathcal{X})$ is also a normal-Wishart distribution. (Geiger and Heckerman, 1994). Thus, the conditional posterior $P(\mu|W, \mathcal{X})$ is multivariate normal with the first and second moments given as follows;

$$E(\mu|W,\mathcal{X}) = \frac{\kappa\mu_0 + m\overline{X}}{\kappa + m} \qquad Cov(\mu|W,\mathcal{X}) = ((\kappa+m)W)^{-1} \tag{11}$$

where $\overline{X} = (\overline{X}^1, ..., \overline{X}^n)'$ is the sample means of $\mathcal{X}$. The posterior $P(W|\mathcal{X})$ is a Wishart with $\alpha + m$ degrees of freedom a scale matrix given as follows;

$$T_m = T_0 + S_m + \frac{\kappa m}{\kappa + m}(\mu_0 - \overline{X}_m)(\mu_0 - \overline{X}_m)' \tag{12}$$

where $S_m = \sum_{i=1}^{m}(X_t - \overline{X})(X_t - \overline{X})'$ is the sample covariance matrix. A closed form representation of Bayesian metric for random variables from a multivariate normal distribution is given by (Geiger and Heckerman, 1994);

$$P(\mathcal{X}^D|G(D)) = (\pi)^{-\frac{n^*m}{2}}\Big(\frac{\kappa}{\kappa+m}\Big)^{\frac{n^*}{2}}\frac{c(n^*, \alpha+m)}{c(n^*, \alpha)} \cdot |T_0|^{\frac{\alpha}{2}}|T_m|^{-\frac{\alpha+m}{2}} \tag{13}$$

where $|T_0|$ and $|T_m|$ denotes the determinants of the matrices T_0 and T_m respectively. T_0 is defined as the prior scale matrix and T_m is the posterior scale matrix both of which consists of n^* rows and columns that corresponds to the variables in the subset D. κ and α are the equivalent sample size for μ and W respectively and $c(\cdot)$ is a normalization constant given by:

$$c(n, \alpha) = \prod_{i=1}^{n}\Gamma\Big(\frac{\alpha+1-i}{2}\Big) \tag{14}$$

This metric is termed the Bayesian Gaussian equivalent (BGe) metric.

2.1.3. Posterior

The model posterior can also be expressed as follows;

$$P(\mathcal{G}, \Theta|\mathcal{X}) = P(\mathcal{G}|\mathcal{X})P(\Theta|\mathcal{G}, \mathcal{X}) \tag{15}$$

where the first component $P(\mathcal{G}|\mathcal{X})$ defines the marginal graph posterior and the second component $P(\Theta|\mathcal{G}, \mathcal{X})$ defines the conditional parameter posterior.

Graph Posterior

Of the two posterior computations, the graph posterior have been described to be a challenging problem (Murphy, 2002). This is referred to as the structure learning problem which has been described by Chickering et.al (2004) to be NP-hard (non-deterministic polynomial-time hard), in the sense that the cardinality of the space of possible structures grows super-exponentially with the number of nodes in the network. By Bayes rule, the graph posterior is given by;

$$P(\mathcal{G}|\mathcal{X}) \propto P(\mathcal{X}|\mathcal{G})P(\mathcal{G}) \tag{16}$$

By assuming a uniform prior over the possible DAGs, the graph priors becomes a constant which does not play a significant role in the posterior computation. Thus the posterior follows the distribution of the marginal likelihood. The standard approach to the above problem is to find the DAG that maximizes the marginal likelihood score;

$$G^* = \arg\max_{\mathcal{G}} P(\mathcal{G}|\mathcal{X}) \propto \arg\max_{\mathcal{G}} P(\mathcal{X}|\mathcal{G}) \tag{17}$$

The Markov Chain Monte Carlo (MCMC) approach is widely used as a standard inference tool for sampling the network structure. Unlike other heuristic algorithms, like the greedy search etc, that attempts to find the highest scoring network, the MCMC approach samples a set of DAGs such that there is no highest scoring network that stands out as significantly unique. The optimal structure is then computed through model averaging. As part of our contribution, we propose an efficient MCMC sampling scheme. This is described in the next section.

Parameter Posterior

Following the assumption on the parameter priors, the parameter posteriors can be shown to satisfy two important properties; Posterior Parameter Independence and Posterior Parameter Modularity (Geiger and Heckerman, 1999).

Lemma 1. (Posterior Parameter Independence) *Given the random sample assumption (Assumption 4), global parameter independence (Assumption 1), and the assumption of no missing data, for every DAG model $\mathcal{G}$ for $\mathcal{X}$,*

$$P(\Theta|\mathcal{X}, \mathcal{G}) = \prod_{i=1}^{n} P(\theta_i|\mathcal{X}, \mathcal{G}) \tag{18}$$

This means that given that the prior and the likelihood factorizes, the posterior parameters also factorizes, thus they are mutually independent given the graph configuration.

Lemma 2. (Posterior Parameter Modularity) *Given the random sample assumption (Assumption 4), global parameter independence (Assumption 1), parameter modularity (Assumption 2) and the assumption of no missing data, if $G_1, G_2 \in \mathcal{G}$ for $\mathcal{X}$ such that X_t^i has the same parents in G_1 and G_2, then*

$$P(\theta_i|\mathcal{X}, G_1) = P(\theta_i|\mathcal{X}, G_2) \tag{19}$$

3. Efficient MCMC Scheme for Structure Learning

In this section we present an efficient inference scheme for learning the structure of a VAR model. To illustrate the inference procedure we define the following;

$$B_s = (A_s \circ \Phi_s) \qquad 0 \leq s \leq L \tag{20}$$

where B_s is $n \times n$ matrix of regression coefficients of the VAR model. A_s is $n \times n$ matrix referred to as the adjacent matrix in Network theory, where a_{ij} represents the directed relationship between X^j_{t-s} and X^i_t. Entries in the matrix A_s are either 1 if $X^j_{t-s} \rightarrow X^i_t$ or 0 if there is no edge between X^j_{t-s} and X^i_t. Φ_s is $n \times n$ matrix of coefficients, where $\phi_{ij} \in \mathbb{R}$ measures the strength of the relationship between X^j_{t-s} and X^i_t and $(\circ)$ is the Hadamard product. Let $b_i = (b_{1i}, ..., b_{ni})'$ be a column vector of B_s, where b_{ji} measures the regression coefficient of the effect of X^j_{t-s} on X^i_t. The relationship between B_s and Φ_s is given by;

$$b_{ij} = \begin{cases} \phi_{ij} & if \quad a_{ij} = 1 \\ 0 & if \quad a_{ij} = 0 \end{cases} \tag{21}$$

By the definition in (20), the standard representation of a VAR of order L in (1) can be expressed as follows;

$$X_t = (A_0 \circ \Phi_0)X_t + \ldots + (A_L \circ \Phi_L)X_{t-L} + \varepsilon_t \qquad t = 1, \ldots, m \tag{22}$$

In the Bayesian inference context, the marginal prior of a_{ij} is a Bernoulli, $a_{ij} \sim Ber(p_{ij})$, where p_{ij} is the probability of $a_{ij} = 1$. The marginal posterior of a_{ij} conditioned on $\mathcal{X}$ is Bernoulli-distributed with the following parameters:

$$a_{ij}|\mathcal{X} = \begin{cases} 1 & if \quad P(a_{ij} = 1|\mathcal{X}) > \tau \\ 0 & otherwise \end{cases} \tag{23}$$

where τ is a threshold set by the user with $\tau \in (0, 1)$. The expression $P(a_{ij} = 1|\mathcal{X})$ is referred to as the confidence score which is interpreted as the posterior probability of the existence of an edge from X^j to X^i.

3.1. Structure Decomposition

The general concept in modeling VAR processes assumes that all within-slice intersection including the initial state distribution have the same structure. By this concept, we find that the marginal likelihood function decomposes according to the structure of the DBN into a multivariate autoregressive (MAR) and a multivariate instantaneous (MIN) network. Let π^i_{t-1} and π^i_t be the parents of X^i_t in the MAR and the MIN network respectively in a DBN of order 1. Thus the parents set of X_t^i in (9) decomposes as $\pi^i = \{\pi_{t-1}^i,\ \pi_t^i\}$. Let $G_{\rightarrow}$ and $G_{\downarrow}$ be the corresponding DAGs. The marginal likelihood in (9) is given by;

$$\begin{aligned} P(\mathcal{X}|\mathcal{G}) &= \prod_{i=1}^{n}\prod_{t=1}^{m} P(X_t^i|\pi_{t-1}^i, \pi_t^i, \mathcal{X}_{t-1}, \mathcal{G}) \\ &= \prod_{i=1}^{n}\prod_{t=2}^{m} P(X_t^i|\pi_{t-1}^i, G_{\rightarrow}) \times \prod_{i=1}^{n}\prod_{t=1}^{m} P(X_t^i|\pi_t^i, G_{\downarrow}) \end{aligned} \tag{24}$$

This decomposition of the structure facilitates the inference procedure such that we can learn the MIN network independently from the MAR network. The posterior graph computation therefore decomposes into searching for the network that maximizes each marginal likelihood score independently.

3.2. The MCMC Sampling Scheme

The MCMC scheme is a standard inference tool for sampling DAGs. This approach was originally proposed by Madigan and York (1995), and later developed by Giudici and Castelo (2003). One of the standard MCMC methods is the Metropolis-Hastings (MH) algorithm, which is based on acceptance-rejection scheme. Thus, given an initial graph, the algorithm samples a new graph using a proposal distribution. The newly sampled graph is then compared with the old graph with a decision rule to either reject or accept the proposed sample. This steps eventually produces a chain of graphs that will convergence to the target distribution with possibly a high number of iterations, though not always guaranteed.

The standard proposal distribution is a single-move conditional on the neighborhood structure of a node. In the simplest term, the algorithm randomly selects a node from the current graph (G_{old}) and proposes an action to either add or delete a single edge to produce a new graph (G_{new}). The proposed graph G_{new} is either accepted and added to the chain of graphs or rejected in which case the previous graph G_{old} is maintained. The decision to accept or reject a proposed graph depends on an acceptance probability given by:

$$A(G_{new}|G_{old}) = \min\left\{\frac{P(\mathcal{X}|G_{new})}{P(\mathcal{X}|G_{old})}\frac{P(G_{new})}{P(G_{old})}\frac{Q(G_{old}|G_{new})}{Q(G_{new}|G_{old})}, 1\right\} \tag{25}$$

By assuming a uniform graph prior implies that $P(G_{new}) = P(G_{old})$. The proposal moves are symmetric which implies that $Q(G_{old}|G_{new}) = Q(G_{new}|G_{old})$ and hence, the acceptance ratio is given by:

$$A(G_{new}|G_{old}) = \min\left\{\frac{P(\mathcal{X}|G_{new})}{P(\mathcal{X}|G_{old})}, 1\right\} \tag{26}$$

Having made the proposal changes to the DAG, a draw U from a uniform distribution on $(0, 1))$ is compared with the outcome of the acceptance probability. If $U < A(G_{new}|G_{old})$ the new proposal is accepted and added to the chain, otherwise the current DAG is retained. Thus the mechanism automatically accepts samples showing improvements (i.e when $A(G_{new}|G_{old}) = 1$) and accepts the rest with the acceptance probability $A(G_{new}|G_{old})$.

With a non-zero probability, the proposal distribution of the sampler will propose a change to any edge in the current graph G_{old}, which guarantees irreducibility. Thus it is possible to reach other state $G_{new} \in \mathcal{G}$ with $P(G_{new}|G_{old}) > 0$ in finite time regardless of the present state. Also with a non-zero probability the chain will remain in the current state for any edge implying aperiodicity. The above properties are sufficient conditions for *ergodicity* of the Markov Chain, which is also a sufficient condition for stationarity of the distribution as $iterations \to \infty$. Thus the MH sampler will in the limit return realizations from the posterior distribution $P(\mathcal{G}|\mathcal{X})$.

A technical approach in the structure inference is to impose a restriction on the neighborhood of a node. This is referred to as the *fan-in.* The neighborhood is simply the possible number of other variables a node can be connected to. The idea is that there is often a trade-off between how densely a node is connected with the effectiveness of the search process. The fewer the neighbors (parents), the fewer the search options at each iteration hence the search procedure can afford evaluate each option carefully. On the other hand, if each node has many neighbors, the search procedure takes a longer time to explore all options and to carefully evaluate them. Therefore, a reasonably few number of neighbors are set to allow effective evaluation of the search options, reduce computational time and to improve convergence. In the current work, we assumed no prior knowledge on the maximal number of neighbors of the nodes.

3.3. Sampling MAR Network

In sampling MAR network, Grzegorczyk (2010) recommended to transform the time series into $1 \times n$ data cells with each cell composed of $(n+1) \times (m-1)$ time series matrix. Each cell corresponds to a transformed data by extracting the series of the i^{th} variable and shifting it one time ahead of the other variables. That is the last observation of the original data is deleted to obtain a matrix of $n \times (m-1)$. This means the extracted series will contain realizations from $t = 2, ..., m$ and the others from $t = 1, ..., m-1$. The extracted series is then added as a new row to the $n \times (m-1)$ matrix to obtain $(n+1) \times (m-1)$ time series matrix. Thus the data in the i^{th} cell is given as follows:

$$\mathcal{X}_{m-1}(i) = \begin{pmatrix} X_1^1 & X_2^1 & \cdots & X_{m-1}^1 \\ \vdots & \vdots & \ddots & \vdots \\ X_1^n & X_2^n & \cdots & X_{m-1}^n \\ X_2^i & X_3^i & \cdots & X_m^i \end{pmatrix}$$

where $i = 1, ..., n$ and $\mathcal{X}_{m-1}(i)$ is the $(n \times (m-1))$ data matrix of the i^{th} cell, i.e $\mathcal{X}_m \setminus X_m$ with a forward shift of X^i. Given a complete graph $G_{\rightarrow}$, the marginal likelihood for the MAR network of order 1 is given as;

$$P(\mathcal{X}|G_{\rightarrow}) = \prod_{i=1}^{n} \frac{P\Big(\mathcal{X}_{m-1}^{(X_t^i,\pi_{t-1}^i)}\Big|G_{\rightarrow}^{(X_t^i,\pi_{t-1}^i)}\Big)}{P\Big(\mathcal{X}_{m-1}^{(\pi_{t-1}^i)}\Big|G_{\rightarrow}^{(\pi_{t-1}^i)}\Big)} \tag{27}$$

The posterior of the scale matrix (T_m) in (12) is computed for each cell by replacing n and m with $n^* = n + 1$ and $m^* = m - 1$;

$$T(i)_{m^*} = T_0 + S_{m^*}(i) + \frac{\kappa m^*}{\kappa + m^*}(\mu_0 - \overline{X}(i))(\mu_0 - \overline{X}(i))' \tag{28}$$

where $\overline{X}(i) = (\overline{X}^1, ..., \overline{X}^{n^*})'$ and $S_{m^*}(i)$ are the sample mean and sample covariance matrix respectively for the transformed time series dataset $\mathcal{X}_{m-1}(i)$.

Search Procedure

The search procedure in sampling MAR network involved *Addition* or *Removal* of an edge at each iteration. Thus at each iteration, we randomly select a row and a column and either add one (edge) if there is initially a zero or delete an existing edge. In these scheme only operations that results in in legal networks (i.e acyclic networks) are considered. Thus edges only flow forward but not backward. That is past observations can only affect current realizations and not the reverse.

3.4. Sampling MIN Network

Sampling MIN network is exactly the same manner as learning a BN. Given a complete graph $G_{\downarrow}$, the marginal likelihood for the MIN network is given by;

$$P(\mathcal{X}|G_{\downarrow}) = \prod_{i=1}^{n} \frac{P\Big(\mathcal{X}_{m}^{(X_t^i,\pi_t^i)}\Big|G_{\downarrow}^{(X_t^i,\pi_t^i)}\Big)}{P\Big(\mathcal{X}_{m}^{(\pi_t^i)}\Big|G_{\downarrow}^{(\pi_t^i)}\Big)} \tag{29}$$

The posterior of the scale matrix (T_m) is the same as expressed in (12).

Search Procedure

Unlike the MAR network, the standard search procedure for a BN involves *Addition*, *Removal* or *Reversal* of an edge. Reversal of an edge is a two-step action, which involves a removal of an existing edge and an adding an edge in the opposite direction. Since it is difficult to check for cycles in such networks, Grzegorczyk and Husmeier (2008) recommended a new edge reversal move which involves pre-computation of the score for all parent configurations. Though it has been proved effective, the procedure depends on the fan-in restriction, in most cases considered to be reasonably samll. Since the reversal move involves removal and/or addition, we considered only *Addition* or *Removal* in our sampling of the MIN network. Removal of an edge always produces legal networks since it does not induce cycles.

Proposition 1. *(Addition in MIN Network)* Let X^i and X^j be two nodes in a MIN network. $X^j \to X^i$ is legal if and only if the intersection between descendants of X^i and the ancestors of X^j is empty.

Following the idea by Giudici and Castelo (2003), the above can be described as a necessary and sufficient condition to produce legal MIN networks. In graphical models, the descendants of a node X^i in a network are the nodes that can be reached following a directed edge from X^i. Thus the descendants of X^i consists of the children, grand children and great grand children etc, of node X^i. The ancestors of X^j consists of the parents, grand parents and great grand parents etc, of node X^j. From the above proposition, it can be shown that if the intersection between descendants of X^i and the ancestors of X^j is non-empty, then adding an edge from X^j to X^i will produce cycles (non-legal networks).

The above condition is quite strong to implement in practice, especially for large networks. A weaker but seemingly necessary condition, is to consider verifications of the acyclicity condition at levels. By levels we mean the following; a first level (three-nodes level) verification is to check for the intersection of immediate relations of each of the two nodes, thus parents X^j and children of X^i; a second level (four-nodes level) verification involves grand parents of X^j and children of X^i or grand children of X^i and parents of X^j, etc. We recommend to begin with the first level, then if the final network reports cycles among a minimum of 4-nodes, then the second level can be implemented etc. This condition is necessary but not sufficient to produce legal networks and users have to run initial simulations to monitor the level of verification required to produce DAGs.

4. Simulations Results

We illustrate our inference procedure on a five ($n = 5$) dimensional system of variables with $m = 100$ time series data points generated by;

$$B_0 = \begin{pmatrix} 0 & -.8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ .5 & 0 & 0 & 0 & 0 \\ 0 & -.5 & .6 & 0 & .8 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad B_1 = \begin{pmatrix} -.9 & 0 & 0 & 0 & 0 \\ 1.2 & 0 & .8 & 0 & 0 \\ -1.6 & 0 & .8 & 0 & 0 \\ 0 & 0 & 1.2 & .9 & 0 \\ 0 & -1.5 & 0 & 0 & .9 \end{pmatrix} \tag{30}$$

The simulation procedure was initialized by setting the hyperparameters μ_0, κ, α, T_0 and the graph priors. Following Grzegorczyk (2010), we set $\mu_0 = (0, ..., 0)'$, $\kappa = 1$, $\alpha = n + 2$ and $T_0 = 0.5 \cdot I_{n\times n}$, where $I_{n\times n}$ is an $n \times n$ identity matrix. A fan-in $= n - 1$ was set for the MIN network and n for the MAR network. Thus we assumed no prior knowledge on the maximal number of neighbors of the nodes in the network. The prior over the DAGs is assumed to be uniform. We initialize the graph with an empty DAG without directed edges. The idea is to assume that the variables are not connected apriori. A burn-in sample of 20,000 with total iteration of 220,000 was implemented for both network search. Our simulation was implemented on a Sony VAIO E Series machine with Intel Core i7-3612QM 2.10 GHz processor. The algorithm used for our simulation is a modification of the of algorithm used by Grzegorczyk et al (2011). An average run time of 10 (12) seconds for every 10,000 iterations was recorded for the MIN (MAR) network simulation.

4.1. Convergence Diagnostics

Several approaches have been discussed to speed up and analyze the convergence of the chain toward the target distribution by improving the mixing properties of the chain. Kass et al (1998) recommends to run multiple chains through parallel independent chains. We implemented 3 parallel independent chains and to ensure that we sample independent and identical distributed (i.i.d) DAGs, we set a thinning interval of 100 to reduce the autocorrelation of the sampled chains. This resulted in 20,000 sampled DAGs for our analysis.

To monitor convergence, we used the single-chain measure by Geweke (1992), and the convergence diagnostics for multiple chains by Gelman and Rubin (1992). For each chain, the posterior distribution of the edge in the network of the sampled chain was divided into 2 windows, containing the first 10% and the last 50%. According to Geweke (1992), if the whole chain is stationary, the means of the values early and late in the sequence should be similar. The output of this measure is a p-value, such that if the p-value is greater than 5%, then we accept the hypothesis that the sample is i.i.d, thus the MCMC converged. For multiple parallel chain convergence, the Gelman-Rubin diagnostics tests whether all the chains converged to the same posterior distribution. The scale parameter ($\hat{R}$) tests between-chain variance and within-chain variance for each edge of the network. The chains are said to converge if $\hat{R} \approx 1,\ (\hat{R} < 1.2)$.

Geweke (p-values) for Single Chain Convergence

	MIN Network					MAR Network				
	X_t^1	X_t^2	X_t^3	X_t^4	X_t^5	X_{t-1}^1	X_{t-1}^2	X_{t-1}^3	X_{t-1}^4	X_{t-1}^5
X_t^1	-	0.475	0.453	0.482	0.484	0.499	0.492	0.495	0.479	0.484
X_t^2	0.475	-	0.485	0.474	0.488	0.499	0.472	0.499	0.489	0.490
X_t^3	0.458	0.472	-	0.478	0.483	0.499	0.487	0.499	0.484	0.488
X_t^4	0.466	0.476	0.468	-	0.457	0.489	0.490	0.499	0.499	0.480
X_t^5	0.470	0.465	0.479	0.475	-	0.487	0.499	0.473	0.489	0.499

Gelman-Rubin ($\hat{R}$) for Multiple Chain Convergence

	MIN Network					MAR Network				
	X_t^1	X_t^2	X_t^3	X_t^4	X_t^5	X_{t-1}^1	X_{t-1}^2	X_{t-1}^3	X_{t-1}^4	X_{t-1}^5
X_t^1	-	0.998	1	1	0.999	1	1	1	0.998	0.999
X_t^2	0.998	-	1	1	1	1	1	1	0.998	0.999
X_t^3	1	0.998	-	1	0.998	1	1	1	1	0.998
X_t^4	1.002	1	1	-	1	0.999	1	1	1	1
X_t^5	0.998	1	1	1	-	0.999	1	1	0.998	1

Table 1: Convergence diagnostics of the I-MCMC scheme. The top (down) panel presents the Geweke $p-$values (Gelman-Rubin $\hat{R}$) for the MIN and MAR network chains.

We implemented the Geweke convergence for each of the chains, and we noticed that the results was similar in all cases. For the sake of space, we report the p-values of only one of the chains. Table 1 shows convergence diagnostics of the independent MCMC (I-MCMC) scheme for the MIN and MAR networks. From the table, we can see that the Geweke p-values are greater than 5% for all the edges in both networks. The Gelman-Rubin ($\hat{R}$) values are also ≈ 1 and below 1.2. With these results, we are confident that the sampled chains are i.i.d and the I-MCMC converges.

Table 2 displays the confidence scores of the presence of edges in the MIN and MAR networks of model (30) and Figure 1 displays the network of the data generating process (DGP). The confidence scores are obtained through model averaging over the sampled DAGs. The edges are directed from the variables in the column labels to the variables in the row labels. The left (right) panel of the table shows the confidence scores in the MIN (MAR) network. From the table, we can see that $P(X_t^1 \rightarrow X_t^2|\mathcal{X}) = 0.591$ and $P(X_t^2 \rightarrow X_t^1|\mathcal{X}) = 0.409$. This means that the posterior probability of the effect of X_t^1 on X_t^2 is stronger than that of reverse effect of X_t^2 on X_t^1. From the MIN network panel, we obtain evidence that $X_t^1 \to (X_t^3, X_t^4)$; $X_t^2 \to X_t^4$; and $X_t^3 \to X_t^4$. In comparison with the network of the DGP in Figure 1, all these observations are consistent with the exception of the effect of X_t^1 on X_t^4. From the MAR network panel, we obtain strong evidence that $X_{t-1}^1 \to (X_t^1, X_t^2, X_t^3)$; $X_{t-1}^2 \to X_t^5$; $X_{t-1}^3 \to (X_t^2, X_t^3, X_t^4)$; $X_{t-1}^4 \to X_t^4$; and $X_{t-1}^5 \to X_t^5$ all of which are consistent with the network of the DGP.

Posterior Probabilities of Edges

	MIN Network					MAR Network				
	X_t^1	X_t^2	X_t^3	X_t^4	X_t^5	X_{t-1}^1	X_{t-1}^2	X_{t-1}^3	X_{t-1}^4	X_{t-1}^5
X_t^1	-	0.409	0.373	0.064	0.223	1	0.236	0.375	0.216	0.192
X_t^2	0.591	-	0.195	0.116	0.248	1	0.197	1	0.203	0.196
X_t^3	0.627	0.185	-	0.095	0.189	1	0.189	1	0.193	0.172
X_t^4	0.809	0.884	0.880	-	0.862	0.217	0.674	1	1	0.278
X_t^5	0.465	0.370	0.268	0.138	-	0.445	1	0.243	0.171	1

Table 2: Marginal posterior probabilities of the presence of edges in the MIN and MAR networks. The edges are directed from the variables in the column labels to the variables in the row labels.

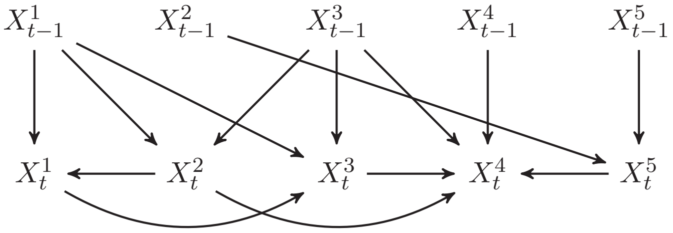

Figure 1: The graphical representation of the interactions among the variables of the data generating process of (30).

4.2. Network Evaluation

The presence of an edge between any two nodes (variables) is defined by the choice of a threshold (τ), where $0 < \tau < 1$. By comparing the confidence scores in Table 2 with the DGP in (30), we count the number of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) edges. TP - Real Positive edges correctly Predicted Positive; FP - non-existing (Real Negative) edges Predicted as Positive; TN - non-existing (Real Negative) edges correctly Predicted Negative; and FN - Real Positive edges Predicted as non-existing.

Recall (Sensitivity) measures the proportion of the Real Positive edges that are correctly Predicted Positive. It is referred to as True Positive Ratio (TPR).

Precision measures the proportion of the Predicted Positive edges that are correctly Real Positive. It is also known as True Positive Accuracy (TPA). Specificity measures the proportion of the Real Negative edges correctly Predicted Negative. This is referred to as the True Negative Ratio (TNR). The Accuracy (ACC) measures the proportion of the Real edges (both Real Positive and Real Negative) that are correctly Predicted. The above terminologies are estimated as follows;

$$TPR = \frac{TP}{TP + FN} \qquad TPA = \frac{TP}{TP + FP}$$
$$TNR = \frac{TN}{FP + TN} \qquad ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Network Evaluation

MIN Network MAR Network

τ	0.5		0.7		0.5		0.7	
	$Real+$	$Real-$	$Real+$	$Real-$	$Real+$	$Real-$	$Real+$	$Real-$
$Pred+$	4	2	3	1	9	1	9	0
$Pred-$	1	13	2	14	0	15	0	16

TPR	80%	60%	100%	100%
TNR	86.67%	93.33%	93.75%	100%
TPA	66.67%	75%	90%	100%
ACC	85%	85%	96%	100%

Table 3: Contingency table and performance of the I-MCMC scheme. The left (right) panel displays the MIN (MAR) network evaluation. ($Real+$) Real Positive, ($Real-$) Real Negative, ($Pred+$) Predicted Positive and ($Pred-$) Predicted Negative.

Table 3 displays the results of the network evaluation with a top panel showing the contingency table and the down panel showing the performance of the I-MCMC scheme. $Real+$ means Real Positive edges, $Real-$ means Real Negative edges, $Pred+$ means Predicted Positive edges and $Pred-$ means Predicted Negative edges. With $\tau = 0.5$, we can see from the contingency table that our inference scheme sampled 4 TP, 2 FP, 1 FN and 13 TN for the MIN network; 9 TP, 1 FP, 0 FN and 15 TN for the MAR network. The performance shows that the sampled MIN network recovered 80% of the real positive edges with 66.7% precision, a specificity of 86.67% with 85% accuracy. That of the MAR network recovered 100% of the real positive with 90% precision, a specificity of 93.75% with 96% accuracy. For $\tau = 0.7$, the MIN network experienced a reduced recall from 80% to 60% but a higher specificity and precision from 86.67% to

93.33% and 66.67% to 75% respectively with the same level of accuracy. That of the MAR network recovered an accurate description of the DGP.

We conducted a comparative evaluation by comparing the performance of the I-MCMC with a joint MCMC (J-MCMC) scheme. The J-MCMC scheme samples both the MIN and the MAR networks in a joint simulation. The simulation procedure follows the same logic as implemented in the I-MCMC. The only difference is that the I-MCMC scheme samples both networks independently whereas the J-MCMC samples them jointly. The same convergence diagnostics used for the I-MCMC was used for the J-MCMC. The results are not statistically different from those reported for the I-MCMC. For purpose of space, we report only the posterior probabilities of the edges in the joint network.

Posterior Probabilities of Edges

	I-MCMC MIN Network					I-MCMC MAR Network				
	X_t^1	X_t^2	X_t^3	X_t^4	X_t^5	X_{t-1}^1	X_{t-1}^2	X_{t-1}^3	X_{t-1}^4	X_{t-1}^5
X_t^1	-	0.409	0.373	0.064	0.223	1	0.236	0.375	0.216	0.192
X_t^2	0.591	-	0.195	0.116	0.248	1	0.197	1	0.203	0.196
X_t^3	0.627	0.185	-	0.095	0.189	1	0.189	1	0.193	0.172
X_t^4	0.809	0.884	0.880	-	0.862	0.217	0.674	1	1	0.278
X_t^5	0.465	0.370	0.268	0.138	-	0.445	1	0.243	0.171	1

J-MCMC Joint Network

	X_t^1	X_t^2	X_t^3	X_t^4	X_t^5	X_{t-1}^1	X_{t-1}^2	X_{t-1}^3	X_{t-1}^4	X_{t-1}^5
X_t^1	-	0.359	0.230	0.124	0.199	1	0.200	0.279	0.153	0.179
X_t^2	0.109	-	0.117	0.098	0.059	1	0.158	1	0.178	0.160
X_t^3	0.109	0.132	-	0.128	0.117	1	0.154	1	0.175	0.143
X_t^4	0.108	0.137	0.120	-	0.439	0.172	0.685	1	1	0.388
X_t^5	0.332	0.545	0.155	0.088	-	0.393	1	0.203	0.168	1

Table 4: Marginal posterior probabilities of the presence of edges in the networks samples with I-MCMC and J-MCMC. The edges are directed from the variables in the column labels to the variables in the row labels. The top (down) panel shows the I-MCMC (J-MCMC) results.

Table 4 displays the confidence scores of the presence of edges in the networks sampled with I-MCMC and J-MCMC schemes. The edges are directed from the variables in the column labels to the variables in the row labels. The top (down) panel shows the confidence scores in the I-MCMC (J-MCMC) networks. We quickly notice that a greater number of the confidence scores in the I-MCMC MIN network are significantly higher than their counterparts in the J-MCMC joint network except for $P(X_t^2 \to X_t^5|\mathcal{X}) = 0.545$ which seems higher in the J-MCMC network than in the I-MCMC MIN network $P(X_t^2 \to X_t^5|\mathcal{X}) = 0.370$. However, verification with the network of the DGP in Figure 1 does not show the existence of such edge. This results increases our confidence that the I-MCMC scheme performs better than the J-MCMC scheme especially in recovering the contemporaneous relationship among the variables of the model. To ascertain this claim we evaluate the two schemes by comparing the sampled networks with the network of the DGP.

Comparative Network Evaluation

	I-MCMC Network				J-MCMC Network			
τ	0.5		0.7		0.5		0.7	
	Real+	*Real−*	*Real+*	*Real−*	*Real+*	*Real−*	*Real+*	*Real−*
Pred+	13	3	12	1	9	2	9	0
Pred−	1	28	2	30	5	29	5	31

TPR	92.86%	85.71%	64.29%	64.29%
TNR	90.32%	96.77%	96.77%	100%
TPA	81.25%	92.31%	90%	100%
ACC	91.11%	93.33%	86.67%	88.89%

Table 5: Contingency table and comparative performance of the I-MCMC and the J-MCMC schemes. The left (right) panel displays the I-MCMC (J-MCMC) results. (*Real+*) Real Positive, (*Real−*) Real Negative, (*Pred+*) Predicted Positive and (*Pred−*) Predicted Negative.

Table 5 displays the output of the comparative network evaluation in the form of a contingency table (top panel) and the comparative performance of the two sample schemes. The left (right) panel shows the results the I-MCMC (J-MCMC) scheme. For $\tau = 0.5$, the I-MCMC recovered a total of 13 TP, 3 FP, 1 FN and 28 TN whereas the J-MCMC recovered 9 TP, 2 FP, 5 FN and 29 TN. We notice that all the TP of the J-MCMC are edges in the MAR relations. The same is true for $\tau = 0.7$. This however confirms our claim that the I-MCMC outperforms the J-MCMC especially in the predictive accuracy, sensitivity and precision of the contemporaneous relations among the random variables the model. Overall, the I-MCMC network attains a higher sensitivity, specificity and more accurate than that of the J-MCMC network.

5. Application to Business Cycle Analysis

We illustrate the our inference scheme on a six ($n = 6$) dimensional US macroeconomic variables. The dataset used in this section consists of 188 quarterly observations from 1947:2 to 1994:1, used by Moneta (2008) (an updated version of the data used by King et al. (1991)). The time series are the logarithms of the per capita consumption expenditure (C), real per capita gross private domestic fixed investment (I), per capita real balances (M), real per capita private gross domestic product (Y), interest rate (R) - a three-month US nominal Treasury bill rate, and price inflation ($\mathit{\Pi}$) - log of the implicit price deflator at the time t minus log of the implicit price deflator at the time $t-1$.

Following the hyperparameters as used in the simulation example, we set the following; $\mu_0 = (0, ..., 0)'$, $\kappa = 1$, $\alpha = n + 2$ and $T_0 = 0.5 \cdot I_{n \times n}$, where $I_{n \times n}$ is an $n \times n$ identity matrix. A fan-in $= n - 1$ was set for the MIN network and n for the MAR network. Thus we assumed no prior knowledge on the maximal number of neighbors of the nodes in the network. The prior over the DAGs is assumed to be uniform. We initialize the graph with an empty DAG without directed edges. The idea is to assume that the variables are not connected apriori. A burn-in sample of 20,000 with total iteration of 220,000 was implemented for both network search. Our simulation was implemented on a Sony VAIO E Series machine with Intel Core i7-3612QM 2.10 GHz processor. The algorithm used for our simulation is a modification of the of algorithm used by Grzegorczyk et al (2011). An average run time of 14 (16) seconds for every 10,000 iterations was recorded for the MIN (MAR) network simulation.

5.1. Convergence Diagnostics

To analyze the convergence of the chain toward the target distribution, we implemented 3 parallel independent chains following the recommendation by Kass et al (1998). To ensure that we sample independent and identical distributed (i.i.d) DAGs, we set a thinning interval of 100 to reduce the autocorrelation of the sampled chains. This resulted in 20,000 sampled DAGs for our analysis. To monitor convergence, we used the single-chain measure by Geweke (1992), and the convergence diagnostics for multiple chains by Gelman and Rubin (1992). The results of the single-chain convergence is shown in the top 2 panel of Table 6. The down 2 panel of Table 6 shows the result of Gelman-Rubin convergence diagnostics for multiple chains. From the table, we can see that the Geweke p-values are greater than 5% for all the edges in both networks. The Gelman-Rubin ($\hat{R}$) values are also ≈ 1 and below 1.2. With these results, we are confident that the sampled chains are i.i.d and the I-MCMC converges.

Table 7 presents the confidence scores of the presence of edges in the MIN and MAR networks. The confidence scores are obtained through model averaging over the sampled DAGs. The edges are directed from the variables in the column labels to the variables in the row labels. The top (down) panel of the table shows the confidence scores in the MIN (MAR) network of the variables. From the table, we can see that $P(I_t \to C_t|\mathcal{X}) = 0.5850$ and $P(C_t \to I_t|\mathcal{X}) = 0.4060$. Thus we have evidence that the effect of investment (I_t) on consumption (C_t) is stronger than the reverse effect of C_t on I_t. From the MIN network panel, we also notice evidence of; the effect of investment (I_t) on money (M_t) and GDP (Y_t); the effect of money (M_t) on consumption (C_t); the effect of GDP (Y_t) on consumption (C_t); and the effect of interest rates (R_t) on price inflation ($\mathit{\Pi}_t$).

Geweke (p-values) for Single Chain Convergence

MIN Network

	C_t	I_t	M_t	Y_t	R_t	Π_t
C_t	–	0.4520	0.4476	0.4665	0.4615	0.4610
I_t	0.4513	–	0.4724	0.4519	0.4946	0.4700
M_t	0.4479	0.4696	–	0.4741	0.4840	0.4811
Y_t	0.4665	0.4519	0.4590	–	0.4344	0.4831
R_t	0.4716	0.4633	0.4760	0.4279	–	0.4801
Π_t	0.4932	0.4773	0.4798	0.4781	0.4801	–

MAR Network

	C_{t-1}	I_{t-1}	M_{t-1}	Y_{t-1}	R_{t-1}	Π_{t-1}
C_t	0.4999	0.4945	0.4746	0.4999	0.4694	0.4790
I_t	0.4849	0.4999	0.4755	0.4804	0.4783	0.4759
M_t	0.4925	0.4707	0.4999	0.4846	0.4682	0.4620
Y_t	0.4999	0.4666	0.4693	0.4999	0.4726	0.4738
R_t	0.4936	0.4915	0.4616	0.4857	0.4999	0.4526
Π_t	0.4812	0.4825	0.4883	0.4832	0.4862	0.4999

Gelman-Rubin ($\hat{R}$) for Multiple Chain Convergence

MIN Network

	C_t	I_t	M_t	Y_t	R_t	Π_t
C_t	–	0.9998	0.9999	1	1	1
I_t	0.9998	–	1	0.9999	1	1
M_t	0.9999	0.9999	–	1	1	1
Y_t	1	0.9999	1	–	1	1
R_t	0.9998	1	1	1	–	0.9999
Π_t	1	1	0.9999	0.9999	0.9999	–

MAR Network

	C_{t-1}	I_{t-1}	M_{t-1}	Y_{t-1}	R_{t-1}	Π_{t-1}
C_t	1	1	1	1	1	1
I_t	1	1	1	0.9999	0.9999	0.9998
M_t	0.9999	1	1	0.9998	0.9998	1
Y_t	1	1	1	1	0.9998	0.9999
R_t	0.9998	0.9998	1	0.9998	1	0.9999
Π_t	0.9998	0.9998	0.9998	1	1	1

Table 6: Convergence diagnostics of the Independent MCMC sampling scheme. The top (down) 2 panel presents the Geweke single-chain convergence $p-$values (Gelman-Rubin $\hat{R}$) for the MIN and MAR network chains.

Posterior Probabilities of Edges

MIN Network

	C_t	I_t	M_t	Y_t	R_t	Π_t
C_t	–	0.5850	0.5098	0.5765	0.2257	0.3390
I_t	0.4060	–	0.4448	0.4243	0.3847	0.2958
M_t	0.4880	0.5477	–	0.3285	0.4898	0.1308
Y_t	0.4235	0.5757	0.3713	–	0.4868	0.1223
R_t	0.1565	0.4022	0.4795	0.4638	–	0.3847
Π_t	0.2455	0.2782	0.1398	0.1517	0.6153	–

MAR Network

	C_{t-1}	I_{t-1}	M_{t-1}	Y_{t-1}	R_{t-1}	Π_{t-1}
C_t	1	0.5395	0.8008	1	0.1648	0.3227
I_t	0.5413	1	0.9375	0.8468	0.1307	0.1815
M_t	0.7767	0.2137	1	0.3150	0.2288	0.1547
Y_t	1	0.9492	0.3348	1	0.7422	0.1465
R_t	0.2687	0.3810	0.1802	0.2422	1	0.6332
Π_t	0.1790	0.1745	0.1573	0.1602	0.9830	1

Table 7: Marginal posterior probabilities of the presence of edges in the MIN (top panel) and MAR (down panel) networks. The edges are directed from the variables in the column labels to the variables in the row labels.

From the MAR network panel, we notice strong evidence of the following relations; $C_{t-1} \rightarrow (C_t, I_t, M_t, Y_t)$; $I_{t-1} \rightarrow (C_t, I_t, Y_t)$; $M_{t-1} \rightarrow (C_t, I_t, M_t)$; $Y_{t-1} \rightarrow (C_t, I_t, Y_t)$; $R_{t-1} \rightarrow (Y_t, R_t, \Pi_t)$; and $\Pi_{t-1} \rightarrow (R_t, \Pi_t)$. Thus there is a strong evidence of quarterly lagged effect of money on consumption as well as a reverse effect of consumption on money. We also observe a similar relationship between consumption and GDP, consumption and investment etc.

From Table 7, we notice that some of the confidence scores are quite close to 0.5. To ensure our confidence, we conducted a 95% credibility interval of all edges whose confidence scores exceed $\tau = 0.5$, to statistically validate the edges. The decision rule considers edge as statistically valid, if the 2.5% quantile of the posterior distribution of an edge is greater than 0.5. This test was conducted for both networks. However, we noticed that all the edges considered from the MAR network are statistically valid. The results of the MIN network is presented in Table 8. The validated network that defines the dynamic interaction among the macroeconomic variable under a homogeneous Markov process of order 1 is depicted in Figure 2.

In comparison with the results by Moneta (2008), we notice some similarity in the sense that our network shows an instantaneous causal effect of investment on consumption and GDP. However a significant difference is that our result shows

Statistical Validation

	$Q.(2.5\%)$	$Mean$	$Q.(97.5\%)$
$I_t \to C_t$	0.5617*	0.5850	0.7053
$M_t \to C_t$	0.4749	0.5098	0.6599
$Y_t \to C_t$	0.5521*	0.5765	0.6805
$I_t \to M_t$	0.4976	0.5477	0.5911
$I_t \to Y_t$	0.5501*	0.5757	0.7270
$R_t \to \Pi_t$	0.6062*	0.6153	0.6650

Table 8: Statistical Validation of the edges in the MIN network whose confidence scores are greater than $\tau = 0.5$. (*) indicates significant in 95% credibility interval.

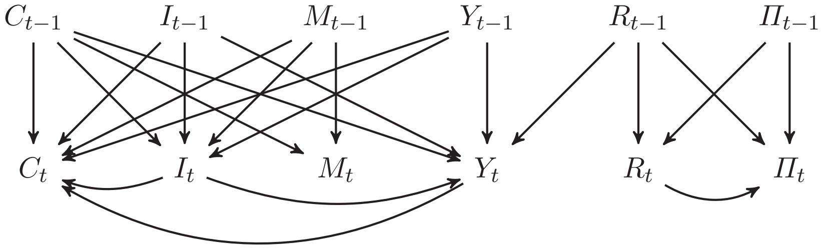

Figure 2: The graphical representation of the interactions among macroeconomic variables for quarterly observations from 1947:2 to 1994:1.

evidence of a contemporaneous effect of GDP on consumption whereas that of Moneta (2008) captures the reverse. It is evident from our network that money has no quarterly lagged effect as well as contemporaneous effect on GDP. However money has a quarterly lagged effect on consumption and investment. From our results, money can only possibly affect GDP through its quarterly lagged effect on investment which has a contemporaneous effect on GDP.

6. Conclusion

Bayesian vector autoregressive models have widely been applied in macroeconomics and macroeconometrics to estimate economic relationships and to empirically assess theoretical hypothesis. As our contribution to the identification problem in structural VAR models, we presented a Bayesian graphical approach to model and estimate the interaction among random variables. Our approach simply decomposes the Bayesian VAR into a multivariate autoregressive and multivariate instantaneous interactions. We then proposed an efficient Markov Chain Monte Carlo (MCMC) scheme that samples the two networks independently. We evaluated the efficiency of our inference procedure with a synthetic data and an empirical assessment of the real business cycles hypothesis.

Our result shows that the independent MCMC scheme outperforms the joint MCMC scheme especially in the predictive accuracy, sensitivity and precision of the contemporaneous relations among random variables in a Bayesian VAR model. The structure evaluation shows that our inference procedure is able to recover an accurate description of the network of the underlying dynamics among the random variables. This presents a convenient framework for modeling and estimating contemporaneous relationships among macroeconomics variables.

References

Bernanke, B. S. (1986). *Alternative explanations of the money-income correlation.* Carnegie-Rochester Conference Series on Public Policy 25: 49–100.

Bessler, D. A and Lee, S. (2002). *Money and prices: US data 1869–1914, A study with directed graphs.* Empirical Economics, 27: 427–446.

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

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

Cooley, T. F and Leroy, S. F. (1985). *A theoretical macroeconometrics: A critique.* Journal of Monetary Economics, Elsevier, 16(3), 283–308.

Corander, J. (2003). *Bayesian graphical model determination using decision theory.* Journal of Multivariate Analysis, 85, 253–266.

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

Dawid, A. P and Lauritzen, S. L. (2000). *Compatible prior distributions.* In Bayesian methods with application to science, policy and official statistics; Selected papers from ISBA 2000, Eurostat.

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

Edwards, D. (1990). *Hierarchical interaction models (with discussion).* Journal of the Royal Statistical Society, Series B 52, 3–20 and 51–72.

Fawcett, T. (2006). *An introduction to ROC analysis.* Pattern Recognition Letters, 27(8): 861–874.

Friedman, N and Koller, D. (2003). *Being Bayesian about network structure.* Journal of Machine Learning, 50(1-2): 95–125.

Friedman, N., Murphy, K. and Russell, S. (1998). *Learning the structure of dynamic probabilistic networks.* In Proc. Fourteenth Conf on Uncertainty in Artificial Intelligence, 139–147 (Morgan Kaufmann, San Francisco, CA).

Geiger, D. and Heckerman, D. (1994). *Learning Gaussian networks.* Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, 235–243.

Geiger, D. and Heckerman, D. (1999). *Parameter Priors for Directed Acyclic Graphical Models and the Characterization of Several Probability Distributions.* Annals of Statistics, 30, 1412–1440.

Gelman, A., and Rubin, D. B. (1992). *Inference from Iterative Simulation Using Multiple Sequences*, (withdiscussion), Statistical Science, 7: 457–511.

Geweke, J. (1992). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments.* In Bayesian statistics 4, (Eds.). J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith, Oxford, U.K: Oxford University Press, 169–193.

Giudici, P. and Castelo, R. (2003). *Improving Markov chain Monte Carlo model search for data mining.* Machine Learning, 50(1-2):127–158.

Giudici, P. and Green, P. J. (1999). *Decomposable graphical Gaussian model determination.* Biometrika 86, 785–801.

Grzegorczyk, M and Husmeier, D. (2008). *Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move.* Journal of Machine Learning, 71, 265–305.

Grzegorczyk, M and Husmeier, D. (2009). *Non-stationary continuous dynamic Bayesian networks.* In Bengio, Schuurmans, Lafferty, Williams and Culotta (editors) Proceedings of the Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS), Curran Associates, 682–690.

Grzegorczyk, M (2010). *An introduction to Gaussian Bayesian networks.* In Yan Qing (Ed.): Systems Biology in Drug Discovery and Development (Springer Series: Methods in Molecular Biology, Vol. 662). Humana Press. ISBN 978-1-60761-799-0.

Grzegorczyk, M., Husmeier, D., and Rahnenführer, J. (2011). *Modelling non-stationary dynamic gene regulatory processes with the BGM model.* Computational Statistics, 26(2), 199–218.

Hoover, K. D. (2001). *Causality in macroeconomics.* Cambridge University Press, Cambridge.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998). *Markov Chain Monte Carlo in Practice: A Roundtable Discussion.* The American Statistician, 52: 93–100.

King, R. G., Plosser, C. I., Stock, J. H and Watson, M. W. (1991). *Stochastic Trends and Economic Fluctuations.* American Economic Review, 81, 819–840.

Lauritzen, S. L and Wermuth, N. (1989). *Graphical models for associations between variables, some of which are quantitative and some qualitative.* Annals of Statistics 17, 31–57.

Madigan, D and York, J. (1995). *Bayesian graphical models for discrete data.* International Statistical Review, 63 (2): 215–232.

Moneta, A. (2008). *Graphical causal models and VARs: an empirical assessment of the real business cycles hypothesis*, Empirical Economics, 35(2), 275–300.

Murphy, K. P. (2002). *Representation, Inference and Learning.* PhD thesis, Computer science division, University of California, Berkeley, CA, USA.

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

Pearl, J. (2000). *Causality: Models, Reasoning and Inference.* Cambridge University Press, London, UK.

Reale, M and Tunnicliffe Wilson, G. (2001). *Identification of vector AR models with recursive structural errors using conditional independence graphs.* Statistical Methods and Applications 10, 49–65.

Robinson, R.W. (1977). *Counting unlabeled acyclic digraphs.* Lecture Notes in Mathematics, Combinatorial Mathematics V, 622: 28–43.

Sims, C. A. (1980). *Macroeconomics and Reality.* Econometrica, Econometric Society, 48(1), 1–48.

Spirtes, P., Glymour, C and Scheines, R. (2000). *Causation, Prediction, and Search.* MIT Press, Cambridge, MA.

Swanson, N. R and 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.

Wermuth, N and Lauritzen, S. L. (1990). *On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion).* Journal of the Royal Statistical Society, Series B, 52, 21–72.

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

Zellner, A. (1962). *An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias.* Journal of the American Statistical Association, 57 (298), 348–368.