UNIVERSITÀ CATTOLICA DEL SACRO CUORE

MILAN Campus

Faculty of ECONOMICS

Master in ECONOMICS

UNIVERSITÀ
CATTOLICA
del Sacro Cuore

Learning network dependencies in VAR models: a Bayesian methodology with applications to financial flows

Supervisor

Castelletti Federico

Graduation thesis

Bonfanti Lorenzo

ID Number: 5007917

Academic Year 2022/23

Learning network dependencies in VAR models: a Bayesian methodology with applications to financial flows

Bonfanti Lorenzo[1]

[1]Universitá Cattolica del Sacro Cuore, Milan

Contents

Introduction

Statistical methods based on graphical models are widely employed in several fields such as genomics, physics and biology. The adoption of such models in economics, as claimed by Imbens [10], is still marginal especially for empirical studies. Moreover, graphical models based on Directed Acyclic Graphs (DAGs) are widely employed for causal inference purposes, and provide an alternative, although more recent, tool w.r.t. the *potential outcome framework*. More in general however, directed graphs can be used to infer *dependence* relations between variables from a non-causal perspective. In this work we consider the implementation of DAG models for understanding dependencies between economic variables. Specifically, we develop a graphical Vector Auto Regressive (VAR) model for multivariate time series data that we apply to the analysis of a set of economic variables.

To this purpose we first introduce the theoretical background needed to understand our model construction. Specifically, Chaper 1 will provide the reader with basic concepts regarding graphical structures, introducing the allied notation, as well as VAR models. In particular, we will introduce the important notion of *conditional independence* (CI), representing the distinctive type of dependence relations that can be read-off from a graph. In the last section we also provide a general framework for Bayesian model selection, which will be then extended in Chapter 2 for structure learning of graphical VARs.

Chapter 2 contains the core of this thesis and introduces our graphical VAR model. Starting from the general setting of Bayesian multivariate linear regression, the chapter will detail how the latter can be extended to a multivariate VAR model where dependence relations between variables satisfy the independence constraints imposed by a DAG. Finally, we will introduce the Markov Chain Monte Carlo (MCMC) method that we implement for structural learning of graphical VARs.

In Chapter 3 we will present an application of the proposed methodology. We apply our model to a set of key economic variables for the US economy. After an introductory exploratory analysis, we will implement our model to investigate dependence relations between variables included in the study. We close the same chapter with a discussion of the results, trying to link economic theory with our findings. In Chapter 4, we finally provide a discussion about possible extensions of the proposed graphical VAR model. Here we provide suggestions for further possible applications of DAGs in economics, while discussing the advantages and current limitations of the proposed method. Algorithms implementing our

methodology have been written in `R` and are publicly available at `https://github.com/Lbonfanti1/VAR-DAG-algorithms-and-application`.

Chapter 1

Background

Before deep diving into the core of this thesis, Graphical DAGs, it is necessary to introduce the reader to some basic concepts and notation. The notation we used in this chapter will often recur in demonstrations and functions used to explain both the theory and the algorithm we will implement in the following chapters. Paragraph 1.1 will introduce the reader to the idea of graphical models, explaining what is conditional independence and separation. It will continue by provinding basic definition on theory of graphoids as what is a node, an edge and the difference between directed and undirected graphs. Paragraph 1.2 refreshes the reader with notions of VAR models; we will declare a notation we will use throghout our work together with traditional theory elements. Paragraph 1.3 will deal with a broad introduction to bayesian model selection. It will explain how bayesian models account for uncertainty about the observed data through a prior predictive distribution. It will deal on how to choose a predictive distribution. The patagraph will introduce the concept of marginal likelihood as a measure used to compute the posterior probability of a DAG. In particular, it will define how marginal likelihood is used in order to compare models.

1.1 Graphical models

The idea of graphical models and graphoids derives from a deeper understanding of the so-called properties of conditional independence (CI). After years of study on CI and its relationship to causality, Pearl and Paz [18] recognized the importance of such concepts for probabilistic reasoning; moreover, they were the first who adopted graphs to describe CI relations between triples of variables. Later on, Dawid [7], Pearl and Paz [18] provided the first mathematical notions regarding graphical structures. The first introduced the concept of separation, while the latter conjectured that semi-graphoids coincide with probabilistic CI structures.

Before introducing the main topic of this chapter, we provide a brief introduction to conditional independence, jointly with some basic notation needed for further explanation. Let $\boldsymbol{x} = \{X_1, \ldots, X_q\}$ be

a set of indexes $\{1, \dots, \} = V$. Let $\boldsymbol{x}_A$ be a subset of $\boldsymbol{x}$ with variables indexed by $A \in V$, and consider $p(\cdot)$ to be a continuous probability measure over $\boldsymbol{x}$. Now, given A, B, C are disjoint subsets of V, we say that $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ are (marginally) independent w.r.t. $p(.)$ iff

$$p(\boldsymbol{x}_A, \boldsymbol{x}_B) = P(\boldsymbol{x}_A)p(\boldsymbol{x}_B). \tag{1.1}$$

Also, we say that $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ are *conditionally* independent given $\boldsymbol{x}_C$ w.r.t. $p(\cdot)$ iff

$$p(\boldsymbol{x}_A, \boldsymbol{x}_B|\boldsymbol{x}_C) = p(\boldsymbol{x}_A|\boldsymbol{x}_C)p(\boldsymbol{x}_B|\boldsymbol{x}_C). \tag{1.2}$$

Equation 1.2 states that once the value of $\boldsymbol{x}_C$ is known, $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ do not "influence" each other. We then write $\boldsymbol{x}_A \perp\!\!\!\perp \boldsymbol{x}_B|\boldsymbol{x}_C$.

The condition presented above has also an interpretation in terms of conditional irrelevance: once the value of C is known, variables in A and B do not influence each other. As a consequence, we can conclude that the occurrence of a value a does not influence the probability of occurrence of b.

Graphical models provide a well-established tool to describe and investigate dependency relationships among variables in various scientific domains. There are two main types of graphical models, namely based undirected and directed (acyclic) graphs. Undirected graphs (UGs) were introduced first. Initially developed in statistical physics, these models were used as tools used for describing the relationship of discrete variables. Next, the use of UGs was mainly considered in the area of multivariate statistical analysis. Darroch et al. [6] expanded the field of use of UGs using a special class of undirected graphical models and by interpreting them as CI. Directed acyclic graphs were formalized later; their original application was in decision making theory. Subsequent contributions, as Pearl [16], had significantly influenced the development of fields that stand at the bottom of modern innovations: artificial intelligence.

We now introduce some notation and general concepts that will be used throughout this work.

A graph $\mathcal{G}$ is a pair (V, E) where $V = \{ 1, \dots, q\}$ is a set of vertices/nodes and $E \subseteq V \times V$ a set of edges. Let $u, v \in V, u \neq v$.

If $(u, v) \in E$ and $(v, u) \notin E$, we say that $\mathcal{G}$ contains the directed edge $u \to v$. If instead $(u, v) \in E$ and $(v, u) \in E$, we say that $\mathcal{G}$ contains the undirected edge $u - v$. Moreover, we say that two vertices are adjacent if they are connected by an edge (directed or undirected). If $u - v$, is in $\mathcal{G}$, we say that u is a neighbor of v in $\mathcal{G}$. We denote the set of neighbors of v as $\text{ne}_{\mathcal{G}}(v)$. The common neighbor set of u and v is then $\text{ne}_{\mathcal{G}}(u, v) = \text{ne}_{\mathcal{G}}(u) \cap \text{ne}_{\mathcal{G}}(v)$. For any pair of distinct nodes $u, v \in V$ we say that u is a *parent* of v if $u \to v$.

Moving back to the main distinction between directed and undirected graphs, we say that $\mathcal{G}$ is a *directed* graph if it contains only directed edges. Vice-versa, the same line of reasoning can be applied to undirected graphs.

A sequence of distinct vertices $\{ v_0, v_1, \dots, v_k\}$ in $\mathcal{G}$ is a *path* from v_0 to v_k if $\mathcal{G}$ contains $v_{j-1} - v_j$ or $v_{j-1} \to v_j$ for all $j = \{ 1, \dots, k\}$. A path is directed (undirected) if all edges are directed (undirected). Moreover, we say a path is *partially directed* if it contains at least one directed edge. If there exists a

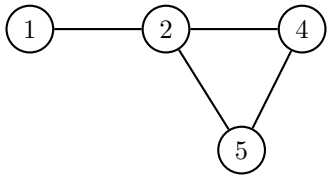


Figure 1.1: Example of Undirected Graphical structure.

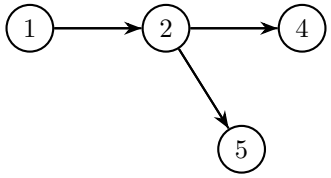


Figure 1.2: Example of Directed Graphical structure.

path from v_0 to v_k , we say that v_k is a descendant of v_0. A sequence of nodes $\{ v_0, v_1, \ldots, v_k\}$ with $v_0 = v_k$ such that $v_{j-1} - v_j$ or $v_{j-1} \to v_j$ for all $j = \{1, \ldots, k\}$ is called a *cycle*. A cycle can be directed or undirected whether it contains only directed or undirected edges respectively.

Let $A \subseteq V$. We denote with $\mathcal{G} = (A, E_A)$ the sub-graph of $\mathcal{G} = (V, E)$ induced by A, whose edge set is $E_A = E \cap (A \times A)$. An undirected sub-graph is complete if its vertices are all adjacent.

A particular class of UGs is that of *decomposable* graphs. A graph is called decomposable if it does not contain induced cycles of length greater than or equal to four without a *chord*, defined as two nonconsecutive adjacent vertices. In this context, we call a *clique* a complete subset that is maximal with respect to inclusion. Let $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$ be a perfect sequence of cliques of the decomposable graph $\mathcal{G}$; consider now, for $k = \{ 2, \ldots, K\}$ three types of sets called *History* (H), *Separators* (S), *Residuals* (R):

$$
\begin{aligned}
H_k &= C_1 \cup \cdots \cup C_K, \\
S_k &= C_k \cap H_{k-1} = C_K \cup H_{k-1} \\
R_k &= C_k \backslash H_{k-1}
\end{aligned}
$$

Notice that $R_1 = H_1 = C_1$, while $S_1 = \emptyset$. Also, $C_1 \cup R_2 \cup \cdots \cup R_K = V$ and also $R_k \cap R_{k'} = \emptyset$. It is then possible to number vertices of a decomposable graph starting from those in C_1, then in C_2 and so on. In completing this process, we obtain a perfect numbering of the vertices [12]. A perfect number of vertices enables us to build its corresponding directed version by orienting the edges from lower to higher numbered vertices.

In this work we focus on Directed Acycic Graphs (DAGs), namely directed graphs with no cycles [13]. Consider a DAG $\mathcal{D}$ and let u, v be two distinct vertices of $\mathcal{D}$; if there exists a directed path from u to v but not paths from v to u, we say that u is an ancestor of v. Otherwise, v is a descendant of u. The set of all ancestors and descendants is denoted with $\text{an}(v)$ and $\text{de}(u)$. Furthermore, given $A \subseteq V$, we define

with An(A) the smallest ancestral set containing A. A subset $C \subseteq V$ is said to be a (u,v)-separator if all trails from u to v intersect C.

A further definition concerns chain graphs; these graphs merge directed and undirected graphs by imposing an acyclic directed structure on disjoint groups of vertices called *chain components*. A chain graph is a simple mixed-graph with directed and undirected edges, such that there are no semi-directed cycles. As a consequence of the absence of semi-directed cycles, a chain graph can be decomposed into ordered chain components $\mathcal{T}$, which are the connected components of the chain graph with no directed edges. There are particular sub graphs that can be derived from a DAG or a chain graph. One is of the form $u \rightarrow z \leftarrow v$ where there are no edges between u and v; this is called a v-structures. Moreover, the skeleton of a graph $\mathcal{G}$ is the undirected graph on the same set of vertices obtained by removing the orientation of all its edges.

Now, consider a DAG $\mathcal{D}$ and q random variables $Y_1, \ldots, Y_q$ each associated to a node in $\mathcal{D}$ The joint distribution of $Y_1, \ldots, Y_q$, $f(\boldsymbol{y}) = f(y_1, \ldots, y_q)$, then factorizes according to $\mathcal{D}$ as:

$$f_{\mathcal{D}}(\boldsymbol{y}) = \prod_{j \in \boldsymbol{V}} f(y_j | \boldsymbol{y}_{pa_D(j)}) \tag{1.3}$$

If Equation (1.3) holds, we say that $f_D(\boldsymbol{y})$ obeys the Markov property of $\mathcal{D}$ [12]. Accordingly, $f_D(\boldsymbol{y})$ encodes a set of marginal and conditional independencies between variables, which can be deduced from the DAG itself by the *d-separation* criterion. Lauritzen [12] proposes an alternative method based on the notion of *moral graph*. A moral graph $\mathcal{D}^m$ is an undirected graph with same skeleton of $\mathcal{D}$ obtained by adding an edge between any pair of vertices with a common child - if not already adjacent - and then dropping the ortientation of all edges. In other words, it joins the two parents in any $v - structure$ $a \rightarrow c \leftarrow b$. The moral graph can be obtained from a DAG by adding an edge between any pair of non-adjacent vertices having a common child and finally dropping the orientation of all edges [13]. The following lemma by Lauritzen et al. [13] then establishes a link between graphical separation and conditional independence

Lemma 1.1 *Let $\mathcal{D} = (V, E)$ be a DAG, $A, B, S \subseteq V$ three disjoint subsets. Then $A \perp\!\!\!\perp B|S$ whenever A and B are separated by S in $(\mathcal{D}_{An(A \cup B \cup S)})^m$, the moral graph of the smallest ancestral set containing $A \cup B \cup S$.*

Let now $\mathcal{D}_1$ and $\mathcal{D}_2$ be two DAGs on the same set of vertices. We say that $\mathcal{D}_1$ and $\mathcal{D}_2$ are *Markov equivalent* if and only if they encode the same conditional independencies [20]. A further important concept is that of *faithfulness*. We say that a probability distribution $f(\cdot)$ is faithful to a DAG $\mathcal{D}$ if the conditional independencies in $f(\cdot)$ are all and only encoded by $\mathcal{D}$ [19]. According to Pearl [17], we say that $\mathcal{D}$ is a perfect map of $f(\cdot)$. If faithfulness holds, then all of the conditional independecies between $Y_1, \ldots, Y_q$ can be directly read off from the DAG.

The faithfulness concept plays a crucial role in model selection. More specifically, assume the existence of a (true) model that have generated the observed data. This model corresponds to a probability

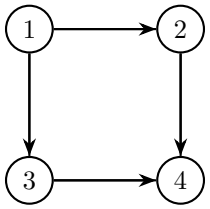


Figure 1.3: A directed acyclic graph

distribution $f_*(\boldsymbol{y}|\cdot)$ among a given model space. Considering that each model is characterized by a specific set of conditional independencies, if faithfulness holds there exists a graph that exactly encodes the conditional independencies of $f_*(\boldsymbol{y}|\cdot)$. Hence, model selection can be performed by searching over a "suitable"graph space. Model selection principles will be explored further later on in Chapter 2. Two DAGs, $\mathcal{D}_1$ and $\mathcal{D}_2$, are Markov equivalent if and only if they have the same skeleton and the same v-structures [2]. Andersson et al. [2] highlights the difficulties related to model selection algorithms that ignore Markov equivalence. The main issues are related to computational difficulties and to the *score equivalence* issue. By using MECs instead of DAGs, we significantly reduce the space dimension that can lead to computational inefficiencies. The *score equivalence* issue instead revolves around the posterior probabilities of Markov equivalent DAGs: these DAGs should represent the same statistical models, hence they should guarantee having equal posterior probabilities. This can be achieved by imposing suitbale constraints on the prior distribution assigned to DAG parameters. All the difficulties mentioned above can be overcomed by treating each MEC as a single model therefore summarizing each MEC with a single graph. Andersson et al. [2] , to this purpose, shows that each markov equivalent class can be uniquely repreented by a partially directed acyclic graph called *essential graph*(EG). A graph $\mathcal{G}$ is the essential graph of a DAG $\mathcal{D}$ if and only if

1. $\mathcal{G}$ is a chain graph;
2. for each chain component $\tau \in$ T, $\mathcal{G}_\tau$ is chordal;
3. $\mathcal{G}$ has no induced subgraphs of the form u $\rightarrow v - z$ (flags);
4. every arrow $u \rightarrow v$ is strongly protected (as illustrated in Andersson et al. [2]).

1.2 VAR models

Vector autoregressive (VAR) models are multivariate statistical models widely used in time series analysis. One interesting feature of such models is the possibility to impose multi directional relationships between variables; in contrast to simpler models, such as AR or ARIMA, VAR enables variables to be treated as endogenous, meaning that all variables can influence each other. Suppose the relationship between a set

of q time-series variables $\boldsymbol{Y}_t = (\boldsymbol{Y}_{1t}, \ldots, \boldsymbol{Y}_{qt})^\top$ and suppose the data generating mechanism of q observed variables can be represented as

$$\boldsymbol{Y}_t = \boldsymbol{\mu}_t + \boldsymbol{X}_t \tag{1.4}$$

In this process, $\boldsymbol{Y}_t$ is composed by the sum of a deterministic coefficient $\boldsymbol{\mu}_t$ and a stochastic term $\boldsymbol{X}_t$ with mean equal to zero. This implies that the expected value of $\boldsymbol{Y}_t$ is $\mathbb{E}(\boldsymbol{Y}_t) = \boldsymbol{\mu}_t$. While $\boldsymbol{\mu}_t$ can accommodate several types of deterministic trends, we can assume for simplicity corresponding to a constant trend. The stochastic component of the DGP instead, $\boldsymbol{X}_t$, is assumed to follow a linear VAR process of order p, referred to as $VAR(p)$, of the form

$$\boldsymbol{X}_t = \boldsymbol{A}_1 \boldsymbol{X}_{t-1} + \cdots + A_p \boldsymbol{X}_{t-p} + \boldsymbol{U}_t, \tag{1.5}$$

where $\boldsymbol{A}_i$, $i = \{\ 1, \ldots, p\}$ are $q \times q$ parameter matrices; the error process $\boldsymbol{U}_t = (\boldsymbol{U}_{1t}, \ldots, \boldsymbol{U}_{qt})^\top$ is instead a K-dimensional zero mean *white noise* process with covariance matrix

$$E(\boldsymbol{U}_t \boldsymbol{U}_t^\top) = \boldsymbol{\Sigma}_{\boldsymbol{U}} \tag{1.6}$$

s.t $\boldsymbol{U}_t \sim (0, \boldsymbol{\Sigma}_{\boldsymbol{U}})$.

The white noise assumption is pivotal in imposing simplicity to the model: because of this assumption it is possible to rule out of the case serial correlations in the errors while allowing for conditional variance dynamics such as generalized auto-regressive conditionally heteroskedastic errors (GARCH). The expression above defines a system of equations where each variable $\boldsymbol{y}_t$ is regressed on its own "lagged versions" as well as lags of other variables up to the lag of order p. According to Kilian and Lutkepohl [11], it is possible to define the matrix polynomial using the lag operator $\boldsymbol{A}(L) = \boldsymbol{I}_Q - \boldsymbol{A}_1 L - \cdots - \boldsymbol{A}_p L^p$, where L^q is the lag we take into considerationand write the process as

$$A(L)\boldsymbol{X}_t = \boldsymbol{U}_t. \tag{1.7}$$

Such model enables the observed variables $\boldsymbol{Y}_t$ to inherit the VAR structure of $\boldsymbol{X}_t$. In particular, by assuming the deterministic term $\boldsymbol{\mu}$ to be constant, we have

$$\boldsymbol{Y}_t = \nu + \boldsymbol{A}_1 \boldsymbol{Y}_{t-1} + \cdots + \boldsymbol{A}_p \boldsymbol{Y}_{t-p} + \boldsymbol{U}_t \tag{1.8}$$

where $\nu = \boldsymbol{A}(L)\mu_0 = \boldsymbol{A}(1)\mu_0 = (\boldsymbol{I}_Q - \sum_{j=1}^{p} \boldsymbol{A}_j)\mu_0$.

The VAR process of $\boldsymbol{X}_t$ and, hence, $\boldsymbol{Y}_t$, is stable if all roots of the determinant polynomial of the VAR operator are outside the complex unit circle, meaning that

$$det(\boldsymbol{A}(z)) = det(\boldsymbol{I}_q - \boldsymbol{A}_1 q - \cdots - \boldsymbol{A}_p q^p) \neq 0 \forall q \in C, |q| \leqq 1 \tag{1.9}$$

C represents the set of complex numbers. Under the assumptions of constant mean and white noise innovations with time-invariant covariance matrix [11], a stable VAR process has time invariant means, variances and covariance structure, hence implying stationarity. Note that a q-dimensional $VAR(p)$ process can be written as a pq-dimensional $VAR(1)$ process by stacking p consecutive $\boldsymbol{Y}_t$ variables in a $p \times q$ dimensional vector, $\boldsymbol{Y}_t = (\boldsymbol{y}_t^\top, \ldots, \boldsymbol{y}_{t-p+1}^\top)^\top$. Also, we can write $\boldsymbol{Y}_t = \nu + \boldsymbol{A}\boldsymbol{Y}_{t-1} + \boldsymbol{U}_t$ where

$$\boldsymbol{A} = \begin{bmatrix} A_1 \; A_2 & \ldots \; A_{p-1} & A_p \\ I_q \; 0 & \ldots \; 0 & 0 \\ 0 \; I_q & \ldots \; 0 & 0 \\ \ldots & \vdots & \vdots \\ 0 \; 0 & \ldots \; I_q & 0 \end{bmatrix} \tag{1.10}$$

The matrix $\boldsymbol{A}$ is also called the companion matrix of the $VAR(p)$ process. Using the stability condition, $\boldsymbol{Y}_t$ is considered stable if

$$\det(\boldsymbol{I}_{qp} - \boldsymbol{A}z) \neq 0 \quad \forall z \in C, \quad |z| \leq 1. \tag{1.11}$$

This condition is equivalent to all eigenvalues of $\boldsymbol{A}$ having modulus less than 1, a property which provides a convenient tool for assessing the stability of the VAR model and for computing the autoregressive roots. By construction, the eigenvalues of $\boldsymbol{A}$ are the reciprocals of the roots of the VAR lag polynomial [11].

VAR models can be estimated with standard methods as least squares (LS), generalized least squares (GLS) and maximum likelihood (ML). For illustrative purposes, we briefly describe the LS method. Consider a $VAR(p)$ model that we can write in the form $\boldsymbol{Y}_t = [\boldsymbol{\nu}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_p]\boldsymbol{Z}_{t-1} + \upsilon_t$ where $\boldsymbol{Z}_{t-1} = (1, \boldsymbol{Y}_{t-1}^\top, \ldots, \boldsymbol{Y}_{t-p}^\top)^\top$ and $u_t \overset{iid}{\sim} \mathcal{N}(0, \boldsymbol{\Sigma}_U)$. The LS estimator is therefore

$$\hat{\boldsymbol{A}} = [\hat{\nu}, \hat{\boldsymbol{A}}_1, \ldots, \hat{\boldsymbol{A}}_p] = \left(\sum_{t=1}^{T} \boldsymbol{Y}_t \boldsymbol{Z}_{t-1}^\top\right)\left(\sum_{t=1}^{T} \boldsymbol{Y}_t \boldsymbol{Z}_{t-1}^\top\right)^{-1} = \boldsymbol{Y}\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{Z}^\top)^{-1} \tag{1.12}$$

where $\boldsymbol{Y} \equiv [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t]$ is $q \times T$ and $\boldsymbol{Z} \equiv [\boldsymbol{Z}_0, \ldots, \boldsymbol{Z}_{t-1}]$ is $(qp+1) \times T$. More precisely, stakcing the colums of $\boldsymbol{A} = [\boldsymbol{\nu}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_p]$ in the $(pq^2 + q) \times 1$ vector $\alpha = vec(\boldsymbol{A})$,

$$\sqrt{T}(\hat{\alpha} - \alpha) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma}_{\hat{\alpha}}) \tag{1.13}$$

where $\mathbf{\Sigma}_{\hat{a}} = plim(\frac{1}{T}\boldsymbol{Z}\boldsymbol{Z}^{\top})^{-1} \otimes \mathbf{\Sigma}_u$ if the process is stable. Under general assumptions, the LS estimator has established properties such as asymptotic Normal distribution [1]. The normality assumption may be however relaxed to allow for conditional/unconditional heteroscedasticity.

1.3 Bayesian Model Selection

When dealing with statistical models, we are typically interested in quantifying the uncertainty around the data generating process. In general, we can consider a collection of possible candidate statistical models, each representing an hypothesis about the true (unknown) data generating process.

In a statistical model, the data generating process is typically indexed by a (vector) parameter $\boldsymbol{\theta}$. Since in what follows we are going to consider a bayesian modelling framework, any statistical model will be completed by assuming a suitable *prior distribution* on $\boldsymbol{\theta}$, representing a prior belief concerning the unknown parameter. Typically, the prior distribution is usually restricted to a specific statistical model itself, represented by a distribution for $\boldsymbol{\theta}$ over the parametric space Θ.

More specifically, a data generating process can be defined as a single completely specified distribution $f(\boldsymbol{x})$. A statistical model can be seen instead as a set $\{f(\boldsymbol{x}|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ of data generating processes, indexed by a parameter $\boldsymbol{\theta}$, i.e. a likelihood function. As mentioned, a bayesian modelling formulation assigns a prior distribution $f(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta$. Bayesian models specify uncertainty about the observed data prior to the observation through a prior predictive distribution:

$$f(\boldsymbol{x}) = \int_{\Theta} f(\boldsymbol{x}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{1.14}$$

Uncertainty around models arises for different possible reasons. Statisticians may question the distribution of the data: a classical example would be testing whether the data follows a normal distribution or a t distribution. Other sources of uncertainty could arise in accurately eliciting a prior distribution for the parameter $\boldsymbol{\theta}$.

Answers to these issues are at the basis of the bayesian approach to model uncertainty, whose main features will be shortly presented below (for a complete presentation see O'Hagan and Foster [14]).

Suppose K possibile distinct models are entertained. One first solution consists in building an *encompassing model*, also called full probability model. This model corresponds to the union of all the candidate models and "contains" all the DGP represented by alternative models. The prior distribution for the parameter of such a model would reflect prior beliefs regarding the true DGP: in practice, this prior is constructed by assigning to each of the alternative models a prior probability. Models will be then compared via their posterior probabilities, which quantify the evidence *a posteriori* in favour of each model.

The second approach consists in formulating a set of possible models without assigning prior probabilities. One can then study how inference on a parameter of interest changes as the model varies within the specified range. If all models lead to the same inferential results, then one can argue that inference

is robust to the uncertainty over model specification.

One last, simpler method consists in studying whether on given model is adequate without considering alternative models. In this approach the aim is just to "prove" against possible criticism that would classify the model as inadequate. Examples of such methods are the resampling methods : cross validation approaches as well as bootstrapping. These models involve repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model.

Considering the model which will be object of discussion for the chapters to come, solely considering the Bayesian framework won't suffice to explain the selection model considered. The graphical structure of the model that will be considered introduces an additional level of difficulty in model selection. In such a framework it comes handy the previously stated concept of Markov property. More graphs can encode the same conditional dependencies, and therefore fall in the case of *Markov equivalence*. Under a broad set of distributional assumptions, Markov equivalent DAGs cannot be distinguished from observational data. Therefore, the whole space of DAG can be partitioned into *Markov equivalent classes*.

As discussed in 1.1, DAGs suffer from inefficiencies: the two main defects are on a computational perspective and on the *score equivalence* issue. Additionally, some of the traditional method that constitute Bayesian Methodology, have been proven fallacious. One example is Hoeting's discovery of inefficiencies in the model averaging approach: pursuing such an approach with DAGs can produce biased results because MECs could have different sizes that cannot easily be computed. The past decades brought under focus the method of Markov chain Monte Carlo (MCMC)- based methods. Such approaches help in reducing the computational complexity of equations to be solved. An alternative to considering DAGs would be working with essential graphs. In particular, while performing model selection, working in the space of MECs instead of DAGs gives a computational advantage (dimension of DAG space is larger). Moreover, due to the larger dimension space, analyzing for equivalent DAG could lead to computational inefficiencies.

In Bayesian methodology, we need to consider the issue of *score equivalence*; the issue raises from posterior probabilities of Markov equivalent DAGS; it should be guaranteed that these models have equal posterior probabilities. The above could be solved by imposing constraints on the priors assigned to DAG parameters.

The rule that will be implemented for model selection is via marginal likelihood; in the Bayesian framework this measure represents a score assigned to models. Subsequently, there will be presented further details on Bayesian model comparison.

Let $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_q\}$ be a collection of real-valued random variables from which we observe n i.i.d. q-variate observations $y_1, \ldots, y_n$ collected in the (n,q) data matrix $\boldsymbol{Y}$. A statistical model, which we can denote as $\mathcal{M}$, consists of a probability density function $f_m(y_1, \ldots, y_q|\boldsymbol{\theta}_{\mathcal{M}})$ where $\boldsymbol{\theta}_{\mathcal{M}}$ is a vector parameter that takes value in the space $\Theta_{\mathcal{M}}$. In this framework, model uncertainty is assumed as follows: the true generating model is one of K different models, $\mathcal{M}_1, \ldots, \mathcal{M}_k$. Given this setting, we can compute the

posterior probability of each model $\mathcal{M}_K$, for $k = 1, \ldots, K$, as follows:

$$p(\mathcal{M}_k|\boldsymbol{Y}) = \frac{m(\boldsymbol{Y}|\mathcal{M}_k)p(\mathcal{M}_k)}{\sum_k m(\boldsymbol{Y}|\mathcal{M}_k)p(\mathcal{M}_k)} \tag{1.15}$$

This is proportional to the product of the *marginal likelihood* of $\mathcal{M}_k$, $m(\boldsymbol{Y}|\mathcal{M}_k)$ and the prior $p(\mathcal{M}_K)$. More specifically, the marginal likelihood is given by

$$m(\boldsymbol{Y}|\mathcal{M}_k) = \int_{\boldsymbol{\theta}_k \in \Theta_k} f_k(\boldsymbol{Y}|\theta_k)p(\boldsymbol{\theta}_k)d\boldsymbol{\theta}_k \tag{1.16}$$

where in particular $f_k(\boldsymbol{Y}|\boldsymbol{\theta}_k) = \prod_{i=1}^n f_k(y_i|\boldsymbol{\theta}_k)$ is the likelihood function and $p(\mathcal{M}_k)$ is instead a prior probability assigned to model $\mathcal{M}_k$.

When comparing two models, $\mathcal{M}_1$ and $\mathcal{M}_2$, we can consider the ratio of their posterior probabilities

$$\frac{p(\mathcal{M}_1|\boldsymbol{Y})}{p(\mathcal{M}_2|\boldsymbol{Y})} = \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} \frac{m_1(\mathcal{M}_1|\boldsymbol{Y})}{m_2(\mathcal{M}_2|\boldsymbol{Y})}. \tag{1.17}$$

The second term of the expression is called the *Bayes factor* of $\mathcal{M}_1$ against $\mathcal{M}_2$ ($BF_{1,2}$) [14]. The BF can be expressed as the product of the prior odds and the ratio of the marginal likelihood of the two models we are comparing.

$$\frac{p(\mathcal{M}_1|\boldsymbol{Y})}{1 - p(\mathcal{M}_1|\boldsymbol{Y})} = \frac{p(\mathcal{M}_1|\boldsymbol{Y})}{p(\mathcal{M}_2|\boldsymbol{Y})} = \frac{p(\mathcal{M}_1)m_1(\mathcal{M}_1|\boldsymbol{Y})}{p(\mathcal{M}_2)m_2(\mathcal{M}_2|\boldsymbol{Y})} = \frac{\pi_1}{1 - \pi_1} \frac{m_1(\mathcal{M}_1|\boldsymbol{Y})}{m_2(\mathcal{M}_2|\boldsymbol{Y})} \tag{1.18}$$

The ratio of marginal likelihood, in other terms, determines how the prior odds are updated to posterior odds in the light of the observed data $\boldsymbol{Y}$.

In model comparison, it is common to consider the case of nested models. Two models are said to be *nested* if the data generating process of the former is a subset of the class of data generating processes defined by the other model. Broadly speaking, we can say that one model is a (simpler) special case of the other, expressed by means of fewer parameters. The comparison of nested models is indeed a frequent practice, such as when investigating the accuracy of an augmented model, obtained by embedding a simpler one with the inclusion of more variables.

An augmented model considers an additional parameter, say ϕ, modifying the original model expressed by $f_0(\boldsymbol{Y}|\boldsymbol{\theta})$ into $f(\boldsymbol{Y}|\boldsymbol{\theta}, \phi = \phi_0)$ when ϕ takes value ϕ_0. The strength of our prior belief on ϕ is incorporated in the prior distribution $f(\phi)$. The matter now is considering how to compare the original model with a likelihood of $f_0(\boldsymbol{Y}|\boldsymbol{\theta}) = f(\boldsymbol{Y}|\boldsymbol{\theta}, \phi = \phi_0)$ with the more general and justified model with likelihood $f(\boldsymbol{Y}|\boldsymbol{\theta}, \phi)$. Starting from the prior distribution, the first model will have a distribution of $f_0(\boldsymbol{\theta})$ while the second model will have a distribution $f(\boldsymbol{Y}|\boldsymbol{\theta}, \phi)$.

For a complete specification of prior beliefs we also need a prior probability $\pi_1 = 1 - \pi_0$ for the second one. In this second case, the numerator of the Bayes factor becomes the value of $f_1(\boldsymbol{Y}|\phi)$ at $\phi = \phi_0$, while the denominator is an average of the values of $f_1(z|\phi)$ for $\phi \neq \phi_0$.

While prior odds ration in favor of the first model is π_0/π_1, posterior odds will be given by multiplying by the Bayes factor $f_0(\boldsymbol{Y})/f_1(\boldsymbol{Y})$, where respectively:

$$f_0(\boldsymbol{Y}) = \int f_0(\boldsymbol{Y}|\boldsymbol{\theta}) f_0(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{1.19}$$

The Bayes factor describes the weight of evidence from the data in favor of the first model, and can be calculated without specifying the prior probabilities π_0 and π_1. Supposing $f_0(\boldsymbol{\theta}) = f_1(\boldsymbol{\theta}|\phi = \phi_0)$. As these two are equal, we can write

The Bayes factor for comparing the two models is finally given by:

$$BF = \frac{f_0(\boldsymbol{Y})}{f_1(\boldsymbol{Y})} = \frac{f_1(\phi_0|\boldsymbol{Y})}{f_1(\phi_0)} \tag{1.20}$$

The equation represents the ratio of posterior to prior densities under the augmented model. If under the augmented model the marginal density is increased, then this can be interpreted as a support for the original model. Note that there are difficulties in specifying a Bayes factor in the case of nested models. Here specifing very weak prior information about ϕ does not let us specify a uniform distribution for $f_1(\phi)$. Note that if the range possible values for ϕ is unbounded, then the uniform prior distribution $f_1(\phi) \propto 1$ is improper, the proportionality constant does not exist. To demonstrate, we could conjecture the scalar parameter ϕ to be on the range $(-\infty, \infty)$ and specify the uniform prior distribution as a limit of the proper uniform distributions:

$$f_1(\phi) = (2c)^{-1} for - c \leq \phi \leq c. \tag{1.21}$$

Now, let $f_1(\phi) = 0$ and let c tend to infinity. Then the equation becomes

$$f_1(\boldsymbol{Y}) = (2c)^{-1} \int_{-c}^{c} f(\boldsymbol{Y}|\phi) d\phi \tag{1.22}$$

For most problems, $f(\boldsymbol{Y}|\phi)$ will tend to zero as ϕ goes to infinity such that the limit of the integral is finite. However, letting c go to infinity induces $f_1(\boldsymbol{Y}) \to 0$ and the Bayes factor tend to infinity. If the continuity condition holds, the Bayes factor will be the ratio of posterior to prior marginal densities at $\phi = \phi_0$.

Suppose that $f(\phi)$ is a prior density, then $c^{-1} f(\phi|c)$ describes a scale family of prior distributions for ϕ. A larger value for c, represents to more diffuse prior distributions.

Such a behavior is observed as the conventional representations of weak prior information about an unbounded parameter ϕ usually has the effect of giving zero prior probability to ϕ lying in any finite region. The discussion falls within the general problem of sensitivity on disturbance of the prior distribution influencing the Bayes factor. The outcome of the evidence brought by theory is that using improper prior distributions would yield to problems in the Bayes factor. This limit of the Bayes factor increases the

difficulty in specifying a prior distribution, specifically when it could be a convenient approximation to exhibit weak prior information.

This same issue regarding sensitivity could be brought out even when the prior information about ϕ is not particularly weak, though data is strong. if the data is stronger than the prior information, $f_1(\phi)$ will vary little over the same region. This would influence posterior odds such that any perturbation of the prior distribution that alters its value will result in a proportionately identical change in the Bayes factor.

While comparing models, no matter how strong the data are, the prior remains important. This statement holds because we are outside of the inference scope, in which as the amount of data increases prior information gets less and less important. We wish now to expand on the concept of sensitivity; consider comparing two arbitrary models where the Bayes factor is in favor of model 1 against model 2. Suppose again that $f_{\mathcal{M}}(\boldsymbol{Y})$ is the marginal distribution of the data $\boldsymbol{Y}$ under model $\mathcal{M}$. Let then $\boldsymbol{\theta}$ be a parameter vector of $p_{\mathcal{M}}$ elements. Suppose that $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ are identical and independently distributed (i.i.d) under model $\mathcal{M}$ with common density. Let $\hat{\boldsymbol{\theta}_{\mathcal{M}}}$ be the maximum likelihood estimate of $\boldsymbol{\theta}_{\mathcal{M}}$ in model $\mathcal{M}$, so that $\hat{\boldsymbol{\theta}_{\mathcal{M}}}$ maximizes $f_{\mathcal{M}}(\boldsymbol{Y}|\boldsymbol{\theta}_{\mathcal{M}})$ for given $\boldsymbol{Y}$.

We can expand the logarithm in a Taylor series around $\boldsymbol{\theta}_{\mathcal{M}} = \hat{\boldsymbol{\theta}}_{\mathcal{M}}$,

$$log f_{\mathcal{M}}(\boldsymbol{Y}|\boldsymbol{\theta}_{\mathcal{M}}) = L_{\mathcal{M}} - (n/2)(\boldsymbol{\theta}_{\mathcal{M}} - \hat{\boldsymbol{\theta}}_{\mathcal{M}})^{\top} \boldsymbol{V}_{\mathcal{M}}^{-1}(\boldsymbol{\theta}_{\mathcal{M}} - \hat{\boldsymbol{\theta}}_{\mathcal{M}}) + \boldsymbol{R} \tag{1.23}$$

where $L_{\mathcal{M}} = log f_{\mathcal{M}}(\boldsymbol{Y}|\boldsymbol{\theta}_{\mathcal{M}})$ is the logarithm of the maximum likelihood, $\boldsymbol{V}_{\mathcal{M}}$ is the modal dispersion matrix, while $\boldsymbol{R}$ is the remainder term, involving higher order derivatives. Remainders for $|\boldsymbol{\theta}_{\mathcal{M}} - \hat{\boldsymbol{\theta}}_{\mathcal{M}}|$ of order $n^{-\frac{1}{2}}$ can be ignored for large n; on the other hand, $f_{\mathcal{M}}(\boldsymbol{\theta}_{\mathcal{M}})$ varies slowly and can be approximated by the constant $f_{\mathcal{M}}(\hat{\boldsymbol{\theta}}_{\mathcal{M}})$. While without model uncertainty $f_{\mathcal{M}}(\hat{\boldsymbol{\theta}}_{\mathcal{M}})$ is a constant which cancels out in Bayes theorem, the same is not true of model comparison; in this case, $f_m(\hat{\boldsymbol{\theta}}_{\mathcal{M}})$ does not cancel out as it varies between models.

The approximation shows that

$$f_{\mathcal{M}}(\boldsymbol{Y}) \approx f_{\mathcal{M}}(\hat{\boldsymbol{\theta}}_{\mathcal{M}}) L_{\mathcal{M}} \left(\frac{2}{\pi}\right)^{p_{\frac{\mathcal{M}}{2}}} n^{-p_{\frac{\mathcal{M}}{2}}} |\boldsymbol{V}_{\mathcal{M}}|^{\frac{1}{2}} \tag{1.24}$$

The Bayes factor will depend asymptotically on the ratio $\frac{f_1(\hat{\boldsymbol{\theta}}_{\mathcal{M}_1})}{f_2(\hat{\boldsymbol{\theta}}_{\mathcal{M}_2})}$ of prior densities under two models and will be sensitive to variation in prior densities no matter the disposal of data. We get the following measure

$$-2 \log \left(\frac{f_1(\boldsymbol{Y})}{f_2(\boldsymbol{Y})}\right) = -2 \log l + (p_1 - p_2) \log n + a \tag{1.25}$$

where $l = L_1/L_2$ so that $-2 \log l$ is the classical likelihood ratio test statistics and a is $O(1)$. Ignoring a or setting it equal to zero gives us the Bayes Information Criterion (BIC). This criterion adjusts the classical likelihood ratio criterion to better favor simpler models with fewer parameters.

If the model $\mathcal{M}$ fully specifies the the distribution of $\boldsymbol{Y}$, in the sense of not having an unknown parameter, $f_{\mathcal{M}}(\boldsymbol{Y})$ is the likelihood of the model. The Bayes factor for comparing this kind of models

could simply be the likelihood ratio. Examining the different approaches further by using the asymptotic distribution derived previously, we can come to further conclusions. As a matter of fact, it is demonstrated that the distribution of $-2log(l)$ has an approximate non-central χ^2 distribution with q degrees of freedom and non-centrality parameter λ. With this structure, as $\phi \neq \phi_0$, the expectation of the ratio tends to infinity as n goes to infinity. In this case the Bayes factor in favor of model 1 tends to zero and the posterior probability of 1 tends to one regardless of prior odds. This behavior demonstrates the Bayesian approach is consistent: as we receive an increasing number of observations, the probability of selecting the correct model tends to one.

In the Bayesian framework, model selection requires for each model parameter to have a prior distribution specified; depending on the approach chosen, the prior could be based on information gathered on the past or expert's knowledge about the matter.

Chapter 2

Bayesian Graphical VARs

This chapter will introduce the statistical model of this thesis, the *Graphical VAR*. Throughout the exposition, we will introduce the general form of a multivariate linear regression model, which will be specialized to the graphical VAR setting. Finally, we will introduce a Markov Chain Monte Carlo (MCMC) method for posterior inference and model selection. Vector autoregressive models are used as a framework to represent serial dependencies of between variables across time; such structures can be represented by graphs with directed or undirected edges depending on the underlying autoregressive process, as already discussed in Chapter 1.

2.1 Multivariate linear regression

Our model is based on a multivariate linear regression framework. In addition, we consider a multivariate normal setting for our variables, that we define as follows. Let $\boldsymbol{y}$ be a q-dimensional random vector. We assume that $\boldsymbol{y} = (y_1, \ldots, y_q)^\top$ is multivariate normal conditionally on a mean vector and covariance matrix, respectively $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ and write accordingly

$$\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{2.1}$$

In particular, $\boldsymbol{\mu}$ is a q-dimensional vector, while $\boldsymbol{\Sigma}$ is a $q \times q$ symmetric positive definite symmetrical matrix. Its p.d.f is given by

$$f(\boldsymbol{y}|\boldsymbol{\Sigma}) = (2\pi)^{-\frac{q}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\} \tag{2.2}$$

Equivalently, we can express $f(\boldsymbol{y}|\boldsymbol{\Sigma})$ in terms of precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ and write

$$\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N_q(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1}), \quad f(\boldsymbol{y}|\boldsymbol{\Omega}) = (2\pi)^{-\frac{q}{2}} |\boldsymbol{\Omega}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Omega}(\boldsymbol{y}-\boldsymbol{\mu})\right\} \tag{2.3}$$

Another distribution that will be adopted in our multivariate model formulation is the *Matrix* Normal (MN). Let $\boldsymbol{Y}$ be a $n \times q$ random matrix. We say that $\boldsymbol{Y}$ has a matrix normal distribution with mean

matrix $\boldsymbol{M}$, row-covariance matrix $\boldsymbol{\Phi}$ and column-covariance matrix $\boldsymbol{\Sigma}$,

$$\boldsymbol{Y}|\boldsymbol{M},\boldsymbol{\Phi},\boldsymbol{\Sigma} \sim N_{n,q}(\boldsymbol{M},\boldsymbol{\Phi},\boldsymbol{\Sigma}), \tag{2.4}$$

if its probability density is given by

$$f(\boldsymbol{Y}|\boldsymbol{M},\boldsymbol{\Phi},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{nq}{2}}|\boldsymbol{\Sigma}|^{\frac{n}{2}}|\boldsymbol{\Phi}|^{-\frac{q}{2}}\exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{M})^{\top}\boldsymbol{\Phi}^{-1}(\boldsymbol{Y}-\boldsymbol{M})\right]\right\} \tag{2.5}$$

In particular, $\boldsymbol{M}$ is a $n \times q$ matrix, $\boldsymbol{\Phi}$ is a $n \times n$ symmetric and positive definite matrix and $\boldsymbol{\Sigma}$ is a $q \times q$ s.p.d matrix.

Following a Bayesian perspective, we need further to specify a prior distribution for the precision matrix $\boldsymbol{\Omega}$ of a multivariate normal model. Let $\boldsymbol{\Omega}$ be a $q \times q$ s.p.d matrix. In this case, we say that $\boldsymbol{\Omega}$ has a Wishart distribution with parameters $a \in \boldsymbol{U}(a > q-1)$ and $\boldsymbol{U}$ a $q \times q$ s.p.d matrix,

$$\boldsymbol{\Omega}|a,\boldsymbol{U} \sim W_q(a,\boldsymbol{U}) \tag{2.6}$$

if its probability density function is given by

$$\begin{aligned} f(\boldsymbol{\Omega}|a,\boldsymbol{U}) &= \frac{1}{2^{\frac{aq}{2}}\Gamma_q\left(\frac{a}{2}\right)}|\boldsymbol{U}|^{\frac{a}{2}}|\boldsymbol{\Omega}|^{\frac{a-q-1}{2}}\exp\left\{-\frac{1}{2}\text{tr}(\boldsymbol{U}\boldsymbol{\Omega})\right\} &(2.7)\\ &\propto |\boldsymbol{\Omega}|^{\frac{a-q-1}{2}}\exp\left\{-\frac{1}{2}\text{tr}(\boldsymbol{U}\boldsymbol{\Omega})\right\}. &(2.8)\end{aligned}$$

Having introduced the previous ingredients, we can now introduce the multivariate Normal linear regression model. Let $\boldsymbol{Y}$ be an $n \times q$ matrix of responses from the q random variables $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_q$ and $\boldsymbol{X}$ an $n \times (p+1)$ matrix of observations from a set of p explanatory variables. Let also $\boldsymbol{B}$ be a $(p+1) \times q$ matrix of regression coefficients describing the effect of the explanatory variables on the responses. A Gaussian multivariate linear regression model can be written in matrix notation as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E} \tag{2.9}$$

where $\boldsymbol{E}$ is an $n \times q$ matrix of error terms, $\boldsymbol{E} \sim N_{n,q}(\boldsymbol{0},\boldsymbol{I}_n,\boldsymbol{\Omega}^{-1})$ with $\boldsymbol{I}_n$ the $n \times n$ identity matrix. Moreover, $\boldsymbol{\Omega}$ is the column precision matrix and $\boldsymbol{0}$ an $n \times q$ matrix. Equivalently, we can write

$$\boldsymbol{Y}|\boldsymbol{B},\boldsymbol{\Omega} \sim N_{n,q}(\boldsymbol{X},\boldsymbol{B},\boldsymbol{I}_n,\boldsymbol{\Omega}^{-1}) \tag{2.10}$$

whose probability density function is given by

$$f(\boldsymbol{Y}|\boldsymbol{B},\boldsymbol{\Omega}) = (2\pi)^{-\frac{nq}{2}}|\boldsymbol{\Omega}|^{\frac{n}{2}}\exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{\Omega}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})^{\top}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})\right]\right\}. \tag{2.11}$$

Moreover, if we let $\hat{\boldsymbol{B}} = (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{Y}$ we can write

$$f(\boldsymbol{Y}|\boldsymbol{B},\boldsymbol{\Omega}) = (2\pi)^{-\frac{nq}{2}}|\boldsymbol{\Omega}|^{\frac{n}{2}}\exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{\Omega}(\boldsymbol{B}-\hat{\boldsymbol{B}})^{\top}(\boldsymbol{B}-\hat{\boldsymbol{B}})\right] + \hat{\boldsymbol{E}}^{\top}\hat{\boldsymbol{E}}\right\}, \tag{2.12}$$

where in particular

$$\hat{\boldsymbol{E}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}}. \tag{2.13}$$

As in a Bayesian statistical framework, we now need to specify priors for the model parameters; in this setting, a conjugate prior for $(\boldsymbol{B}, \boldsymbol{\Omega})$ is available and has the following structure:

$$\boldsymbol{B}|\boldsymbol{\Omega} \sim N_{p+1,q}(\boldsymbol{B}_0, \boldsymbol{C}^{-1}, \boldsymbol{\Omega}^{-1}), \quad \boldsymbol{\Omega} \sim W_q(a, \boldsymbol{U}). \tag{2.14}$$

The corresponding density functions are respectively:

$$\begin{aligned} p(\boldsymbol{B}|\boldsymbol{\Omega}) &= (2\pi)^{-\frac{1}{2}q(p+1)} |\boldsymbol{C}|^{\frac{q}{2}} |\boldsymbol{\Omega}|^{\frac{p+1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}(\boldsymbol{B}-\hat{\boldsymbol{B}})^\top \boldsymbol{C}(\boldsymbol{B}-\hat{\boldsymbol{B}})\right]\right\} \\ p(\boldsymbol{\Omega}) &= \frac{1}{2^{\frac{aq}{2}}\Gamma_q(\frac{a}{2})} |\boldsymbol{U}|^{\frac{a}{2}} |\boldsymbol{\Omega}|^{\frac{a-q-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{U}\boldsymbol{\Omega})\right\}. \end{aligned} \tag{2.15}$$

The joint prior $p(\boldsymbol{B}, \boldsymbol{\Omega}) \propto p(\boldsymbol{B}|\boldsymbol{\Omega})p(\boldsymbol{\Omega})$ is therefore

$$p(\boldsymbol{B}, \boldsymbol{\Omega}) \propto \frac{|\boldsymbol{\Omega}|^{\frac{1}{2}[(p+1)+(a-q-1)]}}{K(\boldsymbol{C}, \boldsymbol{U}, a)} \exp\left\{-\frac{1}{2}\mathrm{tr}[\boldsymbol{\Omega}((\boldsymbol{B}-\hat{\boldsymbol{B}})^\top \boldsymbol{C}(\boldsymbol{B}-\hat{\boldsymbol{B}}) + \boldsymbol{U})]\right\}, \tag{2.16}$$

where $K(\boldsymbol{C}, \boldsymbol{U}, a)$ is the prior *normalizing constant*, corresponding to

$$K(\boldsymbol{C}, \boldsymbol{U}, a) = \frac{(2\pi)^{\frac{q(p+1)}{2}} \Gamma_q\left(\frac{a}{2}\right) 2^{\frac{aq}{2}}}{|\boldsymbol{C}|^{\frac{q}{2}} |\boldsymbol{U}|^{\frac{a}{2}}}. \tag{2.17}$$

Once specified the likelihood function and prior distributions, we can derive the posterior distribution of $(\boldsymbol{B}, \boldsymbol{\Omega})$, $p(\boldsymbol{B}, \boldsymbol{\Omega}|\boldsymbol{Y})$. Using some linear algebra, it can be shown that

$$\begin{aligned} \boldsymbol{B}|\boldsymbol{\Omega}, \boldsymbol{Y} &\sim N_{p+1,q}(\hat{\boldsymbol{B}}, (\boldsymbol{C}+\boldsymbol{X}^\top\boldsymbol{X})^{-1}, \boldsymbol{\Omega}^{-1}), \\ \boldsymbol{\Omega}|\boldsymbol{Y} &\sim W_q(a+n, \boldsymbol{U}+\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}+\boldsymbol{D}), \end{aligned} \tag{2.18}$$

where in particular $\hat{\boldsymbol{B}}$ is the posterior expectation of $\boldsymbol{B}$, and the term $\boldsymbol{D}$ embodies a measure of discrepancy between the prior and the data. In particular, we have

$$\hat{\boldsymbol{B}} = (\boldsymbol{C}+\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{Y}+\boldsymbol{C}\boldsymbol{B}) \qquad \boldsymbol{D} = (\boldsymbol{B}-\hat{\boldsymbol{B}})^\top(\boldsymbol{C}^{-1}+(\boldsymbol{X}^\top\boldsymbol{X})^{-1})(\boldsymbol{B}-\hat{\boldsymbol{B}}) \tag{2.19}$$

From the previous equations, the p.d.f. of the (conditional) posterior distribution of $\boldsymbol{B}$ can be written explicitly as

$$p(\boldsymbol{B}|\boldsymbol{\Omega}, \boldsymbol{Y}) = (2\pi)^{\frac{q(p+1)}{2}} |\boldsymbol{C}+\boldsymbol{X}^\top\boldsymbol{X}|^{\frac{q}{2}} |\boldsymbol{\Omega}|^{\frac{p+1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}(\boldsymbol{B}-\hat{\boldsymbol{B}})^\top(\boldsymbol{C}+\boldsymbol{X}^\top\boldsymbol{X})(\boldsymbol{B}-\hat{\boldsymbol{B}})\right]\right\}, \tag{2.20}$$

while the p.d.f. of the (marginal) posterior of $\boldsymbol{\Omega}$ is

$$p(\boldsymbol{\Omega}|\boldsymbol{Y}) = \frac{|\boldsymbol{U}+\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}+\boldsymbol{D}|^{\frac{a+n}{2}}}{2^{\frac{q(p+1)}{2}}\Gamma_q(\frac{a+n}{2})} |\boldsymbol{\Omega}|^{\frac{a+n-q-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}[\boldsymbol{\Omega}(\boldsymbol{U}+\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}+\boldsymbol{D})]\right\} \tag{2.21}$$

The joint posterior is therefore:

$$\begin{aligned} P(\boldsymbol{B}, \boldsymbol{\Omega}|\boldsymbol{Y}) &= K^{-1}(\boldsymbol{C}+\boldsymbol{X}^\top\boldsymbol{X}, \boldsymbol{U}+\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}+\boldsymbol{D}, a+n) \\ &\quad \cdot |\boldsymbol{\Omega}|^{\frac{p+1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}(\boldsymbol{B}-\hat{\boldsymbol{B}}^\top)(\boldsymbol{C}+\boldsymbol{X}^\top\boldsymbol{X})(\boldsymbol{B}-\hat{\boldsymbol{B}})\right]\right\} \\ &\quad \cdot |\boldsymbol{\Omega}|^{\frac{a+n-q-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}[\boldsymbol{\Omega}(\boldsymbol{U}+\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}+\boldsymbol{D})]\right\} \end{aligned} \tag{2.22}$$

where $K(\cdot)$ is now the posterior normalizing constant.

What we introduced before represents one of the main components of the graphical VAR model that we introduce in the next section. One last essential element we need to introduce is the marginal likelihood of the multivariate linear regression model. As explained in Chapter 1, the latter indeed represents an essential feature for Bayesian model selection techniques. For a multivariate Normal linear regression model with conjugate priors, the marginal likelihood

$$m(\boldsymbol{Y}|\boldsymbol{X}) = \int p(\boldsymbol{Y}|\boldsymbol{B}, \boldsymbol{\Omega})\, p(\boldsymbol{B}, \boldsymbol{\Omega})\, d(\boldsymbol{B}, \boldsymbol{\Omega}) \tag{2.23}$$

can be derived in closed-form expression from the ratio of the prior and the posterior normalizing constants. Specifically:

$$m(\boldsymbol{Y}|\boldsymbol{X}) = \frac{(2\pi)^{-\frac{nq}{2}} K\left(\boldsymbol{C}, \boldsymbol{U}, a\right)^{-1}}{K^{-1}(\boldsymbol{C} + \boldsymbol{X}^\top \boldsymbol{X}, \boldsymbol{U} + \hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}} + \boldsymbol{D}, a + n)}. \tag{2.24}$$

2.2 Graphical VAR

Let $\boldsymbol{y}_t$ be a q-dimensional vector of observations collected at time t, where $t = 1, \ldots, T$. The reduced form of a stable VAR of order k is given by:

$$\boldsymbol{y}_t = \sum_{i=1}^{k} \boldsymbol{B}_i \boldsymbol{y}_{t-i} + \boldsymbol{\epsilon}_t \tag{2.25}$$

In the above notation above, $\boldsymbol{B}_i$ are $q \times q$ matrices of coefficients corresponding to the lags of the model, determining the dynamics of the system, while $\boldsymbol{\epsilon}_t$ represents a white noise q-dimensional process. In this representation, observations at time t are assumed to linearly depend on the previous (past) k realizations of the process, where typically k is known. A constant is omitted for a simpler model specification. However, the latter can be easily included by setting the first variable among the q in the model equal to a unit term.

Let now $\boldsymbol{z}_t = (\boldsymbol{y}_{t-1}^\top, \ldots, \boldsymbol{y}_{t-k}^\top)$ be the kq-vector of lagged observations at time t and $\boldsymbol{B}^\top = (\boldsymbol{B}_1, \ldots, \boldsymbol{B}_k)$ the $q \times kq$ matrix obtained by collecting together the coefficient matrices. The model equation can be written as

$$\boldsymbol{y}_t = \boldsymbol{B}^\top \boldsymbol{z}_t + \boldsymbol{\epsilon}_t \tag{2.26}$$

The likelihood of the *VAR(k)* model for the initial values $\boldsymbol{Y}_0 = (\boldsymbol{y}_0^\top, \boldsymbol{y}_{-1}^\top, \ldots, \boldsymbol{y}_{-k+1}^\top)^\top$ is in particular

$$f(\boldsymbol{y}_1, \ldots \boldsymbol{y}_T|\boldsymbol{B}, \boldsymbol{\Sigma}) = \prod_{t=1}^{T} f(\boldsymbol{y}_t|\boldsymbol{z}_t, \boldsymbol{B}, \boldsymbol{\Sigma}), \tag{2.27}$$

where importantly each conditional distribution $f(\boldsymbol{y}_t|\boldsymbol{z}_t, \boldsymbol{B}, \boldsymbol{\Sigma})$ is a multivariate normal distribution of the form $\boldsymbol{y}_t|\boldsymbol{z}_t, \boldsymbol{B}, \boldsymbol{\Sigma} \sim N_q(\boldsymbol{B}^\top \boldsymbol{z}_t, \boldsymbol{\Sigma})$. The same expression, for convenience, can be parametrized in terms of the inverse covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, as detailed in Section 2.1. By rearranging suitably the coefficient-matrices of the model, we can write in a more compact form

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{B} + \boldsymbol{E}, \tag{2.28}$$

where $\boldsymbol{Y}$ represents a $(T \times q)$ matrix collecting all observations, $\boldsymbol{Z}$ is the $(T \times qk)$ matrix containing all the lagged realizations of the q variables. Finally, $\boldsymbol{E}$ is a $(T \times q)$ matrix of errors following a matrix normal distribution. Specifically,

$$\boldsymbol{E}|\boldsymbol{\Sigma} \sim N_{T,q}(\boldsymbol{0}, \boldsymbol{I}_T, \boldsymbol{\Sigma}) \tag{2.29}$$

Errors are assumed to have zero mean, and cross-covariance between rows of $\boldsymbol{E}$ equal to the identity matrix $\boldsymbol{I}_T$, while cross-covariance between columns equal to $\boldsymbol{\Sigma}$. Based on what has been presented, the likelihood of the model can be expressed as follows:

$$f(\boldsymbol{Y}|\boldsymbol{B}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{Tq}{2}} \boldsymbol{\Sigma}^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}((\boldsymbol{B} - \hat{\boldsymbol{B}})\boldsymbol{Z}^\top \boldsymbol{Z}(\boldsymbol{B} - \hat{\boldsymbol{B}}) + \boldsymbol{E}^\top \boldsymbol{E})\right]\right\}, \tag{2.30}$$

where $\hat{\boldsymbol{B}}$ is the OLS estimator of the coefficient matrix $\boldsymbol{B}$.

To complete the Bayesian model specification, we need to assign priors to model parameters $\boldsymbol{B}$ and $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Similarly to the general multivariate linear regression setting of Section 2.1, we assign

$$\begin{aligned} \boldsymbol{B}|\boldsymbol{\Omega} &\sim N_{p,q}(\boldsymbol{B}_0, \boldsymbol{C}^{-1}, \boldsymbol{\Omega}^{-1}), \\ \boldsymbol{\Omega} &\sim W_q(a, \boldsymbol{U}). \end{aligned} \tag{2.31}$$

Importantly for our purposes, the conditional independencies characterizing the precision matrix will be determined on the basis of a DAG structure, yet unknown and therefore target of our model selection procedure (Section 2.3). Moreover, in the previous equation $\boldsymbol{C}^{-1}$ is a prior variance/covariance matrix between regression coefficients. The joint prior $p(\boldsymbol{B}, \boldsymbol{\Omega})$ is therefore

$$p(\boldsymbol{B}, \boldsymbol{\Omega}) = p(\boldsymbol{B}|\boldsymbol{\Omega})p(\boldsymbol{\Omega}) \tag{2.32}$$

whose p.d.f. can be written as

$$p(\boldsymbol{B}, \boldsymbol{\Omega}) = (2\pi)^{\frac{-pq}{2}} |\boldsymbol{C}|^{\frac{q}{2}} |\boldsymbol{\Omega}|^{\frac{p}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{\Omega}(\boldsymbol{B} - \boldsymbol{B}_0)^\top \boldsymbol{C}(\boldsymbol{B} - \boldsymbol{B}_0)\right]\right\} \left[\Gamma_q\left(\frac{a}{2}\right)^{-1} 2^{-\frac{aq}{2}} |\boldsymbol{\Omega}|^{\frac{a-q-1}{2}}\right] \tag{2.33}$$

We require $a - q + 1 > 0$ so that the Wishart prior on $\boldsymbol{\Omega}$ is proper and the posterior will be proper as well.

By combining the prior with the likelihood, we obtain the posterior distribution, similarly to what was shown in Section 2.1. Specifically, the joint posterior of $(\boldsymbol{B}, \boldsymbol{\Omega})$ will be a Matrix Normal Inverse Wishart $MNIW(\tilde{\boldsymbol{B}}, \tilde{\boldsymbol{C}}, \tilde{a}, \tilde{\boldsymbol{U}})$. Hyperparameters are updated and structured as follows:

- $\tilde{\boldsymbol{B}} = \hat{\boldsymbol{B}}$,
- $\tilde{\boldsymbol{C}} = \boldsymbol{C} + \boldsymbol{Z}^\top \boldsymbol{Z}$,
- $\tilde{a} = a + n$,
- $\tilde{\boldsymbol{U}} = \boldsymbol{U} + \boldsymbol{E}^\top \boldsymbol{E} + \boldsymbol{D}$.

At this stage, we can obtain the marginal likelihood for the model from the ratio of prior and posterior normalizing constants as

$$m(\boldsymbol{Y}|\boldsymbol{X}) = (2\pi)^{-\frac{Tq}{2}} \frac{K^{-1}(\boldsymbol{C}, \boldsymbol{U}, a)}{K^{-1}(\tilde{\boldsymbol{C}}, \tilde{\boldsymbol{U}}, \tilde{a})}, \tag{2.34}$$

with $K(\cdot)$ as in Section 2.1.

It is possible now to introduce the graphical model formulation. Let $\boldsymbol{Y} = \boldsymbol{Y}_t(a), t \in \mathbb{Z}, a = 1, \ldots, q$ be a q-variate stationary stochastic process and let $V = \{1, 2, \ldots, q\}$ be the set of indexes. We define a graph $\mathcal{G} = (V_{TS}, E)$ be a graph with node set $V_{TS} \times \mathbb{Z}$ and an edge set E, whose edges have at most k lags and which is invariant under translation [15]. In a DAG-based graphical VAR, nonzero elements in $\boldsymbol{B}$ corresponds to directed edges between variables at different lags, suggesting the recursive structure of the time series; differently, conditional dependence relations embedded in $\boldsymbol{\Omega}$ correspond to directed edges between nodes/variables at the same lag, i.e. contemporaneous dependencies. To build a DAG model, we need to impose a graph-based structure to our general model.

Let now $\mathcal{D}$ be a DAG with a vertex set V as previously defined. Let the set $pa_{\mathcal{D}}(j)$ specify the parents of vertex j in the graphical structure $\mathcal{D}$ and $\boldsymbol{y}_{ipa_{\mathcal{D}}(j)}$ be a subvector of $\boldsymbol{y}_i$ indexed bu the parent set $pa_{\mathcal{D}}$ The Gaussian multivariate regression sampling density will factorize according to $\mathcal{D}$ as

$$f(\boldsymbol{y}_t|\boldsymbol{\theta}) = \prod_{j=1}^{q} f_{\mathcal{D}}(y_{tj}|\boldsymbol{y}_{t,pa_{\mathcal{D}}(j)}, \theta_j) \tag{2.35}$$

Equivalently, by writing each conditional density as the ratio of suitable marginal distributions, we can write

$$f(\boldsymbol{y}_t|\boldsymbol{\theta}) = \prod_{j=1}^{q} \frac{f(\boldsymbol{y}_{tfa_{\mathcal{D}}(j)}|\boldsymbol{\theta})}{f(\boldsymbol{y}_{tpa_{\mathcal{D}}(j)}|\boldsymbol{\theta})}. \tag{2.36}$$

Furthermore, in the DAG case we would need to further expand the notation, conditioning the distribution on the regressor matrix $\boldsymbol{B}$, covariance matrix $\boldsymbol{\Sigma}$ and the imposed DAG structure $\mathcal{D}$. Following the same factorization, the likelihood function of our VAR model can be written as

$$f(\boldsymbol{Y}|\boldsymbol{B}, \boldsymbol{\Sigma}, \mathcal{D}) = \prod_{j=1}^{q} \frac{f(\boldsymbol{Y}_{fa_j}|\boldsymbol{Z}, \boldsymbol{B}, \boldsymbol{\Omega}^{-1})}{f(\boldsymbol{Y}_{pa_j}|\boldsymbol{Z}, \boldsymbol{B}, \boldsymbol{\Omega}^{-1})}, \tag{2.37}$$

now emphasizing the dependence on the VAR parameters $\boldsymbol{B}$ and $\boldsymbol{\Omega}$. What we need to have for model selection is the marginal likelihood of the model. It can be shown [5] that the DAG marginal likelihood admits the same node-by-node factorization of the likelihood; each term corresponds to the ratio between marginal data distributions restricted to family and parent sets.

$$m(\boldsymbol{Y}|\mathcal{D}) = \prod_{j=1}^{q} \frac{m(\boldsymbol{Y}_{fa(j)}|\boldsymbol{Z})}{m(\boldsymbol{Y}_{pa(j)}|\boldsymbol{Z})} \tag{2.38}$$

All terms can be computed as the marginal data distributions obtained under a complete multivariate linear regression model (Section 2.1).

2.3 Model selection: MCMC and the Metropolis-Hasting algorithm

One key aim of statistical learning is defining a model that fits the data the most. This process is called *model selection*; it is a structural part of every statistician research while picking a model for the data under observation. In Bayesian statistics especially, under model uncertainty, a statistician's interest relies in evaluating the posterior model probability given the observed data. Under all the existing methods for model selection, Markov Chain Monte Carlo (MCMC) methods represent a valid choice for approximating posterior probabilities, especially when the number of models to be compared is large or their enumeration is impractical. Both of these issues, in fact, affect DAG models. Often, the number of DAGs grows super exponentially in the number of variables; often times, even the enumeration of all possible EGs on a given number of nodes becomes infeasible.

A Markov chain on graphs' space considers S_q, a model space represented by a collection of graphs having q nodes. As reported in Castelletti [4], it would be equal considering the ser of all DAGs as the set opf all EGs on q nodes. What then constitutes a Markov chain is the transition between graphs on the finite space considered S_q. Such transition is achieved by using a set of operators that determine a move from a graph $\mathcal{D}$ to $\mathcal{D}'$, where in the space considered (S_q) $\mathcal{D}' \neq \mathcal{D}$. The transition is determined through a local modification of the the intial graph. Specifically, the operator is defined by a type and the modified edges; as for type we refer to the action in use (e.g Insertion or deletion of a directed edge), while for modified edge we refer to the consequence the operation had on the graphical structure of the model to be considered (e.g $x \to y$). For the structure of the model. we call the new graph $\mathcal{D}'$ the *direct successor* of $\mathcal{D}$, as it can be reached using a single transition from the original graph (in simpler terms,by applying one operator from $\mathcal{D}$). Note that the set of operators to be used must satisfy a set of basic properties; a major property to be held is called *distinguishability*. Distinguishability is the property guaranteeing that different operators generate different graphs.

Let now $\mathcal{O}_{\mathcal{D}}$ be a set of operators on $\mathcal{D}$ and let $|\mathcal{O}_{\mathcal{D}}|$ be its cardinality. If distinguishability holds, we can derive the probability of transition to be:

$$p_{\mathcal{D},\mathcal{D}'} = \frac{1}{|\mathcal{O}_{\mathcal{D}}|}. \tag{2.39}$$

The Markov chain induced by $p_{\mathcal{D},\mathcal{D}'}$ can then be used as a proposal distribution $q(\mathcal{D}'|\mathcal{D})$ inside a mcmc scheme having as a target the posterior distribtution $q(\mathcal{D}|Y)$, $\mathcal{D} \in S_q$. An appropriate MCMC algorithm muyst satisfy three properties: irreducibility, reversibility, aperiodicity.

The posterior distribution can be approximated with the Monte Carlo method only when conjugate or semiconjugate prior distributions are used. In situations where a conjugate prior distribution is unavailable or undesirable, the full conditional distributions of the parameters do not have a standard form. In these cases, the Metropolis-Hastings algorithm is a generic method of approximating the posterior distribution corresponding to any combination of prior distribution and sampling model. The MCMC

Metropolis-Hasting algorithm we will consider targets $p(\cdot)$ and the proposal distribution $q(\mathcal{D}'|\mathcal{D})$; if the corresponding Markov Chain is irreducible and aperiodic with stationary distribution $\mathrm{p}(\mathcal{D}|\boldsymbol{Y})$, then

$$\lim_{S\to+\infty}\frac{1}{S}\sum_{s=1}^{S}g(\mathcal{D}^s)=\sum_{\mathcal{D}\in S_q}g(\mathcal{D})p(\mathcal{D}|\boldsymbol{Y}). \tag{2.40}$$

Generally, the transition from $\mathcal{D}$ to $\mathcal{D}'$ is given by an acceptance probability rule, measured as such:

$$\alpha(\mathcal{D}'|\mathcal{D})=min\left(1;\frac{m(\boldsymbol{Y}|\mathcal{D}')}{m(\boldsymbol{Y}|\mathcal{D})}\times\frac{p(\mathcal{D}')}{p(\mathcal{D})}\times\frac{q(\mathcal{D}|\mathcal{D}')}{q(\mathcal{D}'|\mathcal{D})}\right) \tag{2.41}$$

As every bayesian model requires, we need to set a prior probability on $\mathcal{D}$ depending on the number of edges in the graph. We decided to include in our model a specifc set of priors, following the process below. Given our DAG $\mathcal{D}$, let $\boldsymbol{A}^{\mathcal{D}}$ be the symmetric binary adjacency matrix of the skeleton of $\mathcal{D}$. Its elements (u,v) are denoted by $\boldsymbol{A}^{\mathcal{D}}_{(u,v)}$. Conditionally on the edge inclusion probability w, we will assign a Bernoulli prior independently to each element $\boldsymbol{A}^{\mathcal{D}}_{(u,v)}$ belonging to the lower triangular part, that is $\boldsymbol{A}^{\mathcal{D}}_{(u,v)}|w\overset{iid}{\sim}Ber(w), u>v$. We get as consequence:

$$p(\boldsymbol{A}^{\mathcal{D}})=w^{|\boldsymbol{A}^{\mathcal{D}}|}(1-w)^{\frac{q(q-1)}{2}-|\boldsymbol{A}^{\mathcal{D}}|}. \tag{2.42}$$

$|\boldsymbol{A}^{\mathcal{D}}|$ represents the number of edges in the skeleton of the graph $\mathcal{D}$; equivalently, this will correspond to the number of entries equal to 1 in the lower triangular part of $|\boldsymbol{A}^{\mathcal{D}}|$. Finally, we set $p(\mathcal{D})\propto p(\boldsymbol{A}^{\mathcal{D}})$ for $\mathcal{D}\in S_q$.

The algorithm we used for model selection reflects what has been explained upon. We considered a set of DAG S, each holding q nodes. We then define a proposal distribution $q(\mathcal{D}'|\mathcal{D})$ for each $\mathcal{D},\mathcal{D}'\in S$. The operators we introduced are of three types, acting on the starting adjacency matrix: insert, delete and reverse a directed edge. For each element in the space we defined, this algorithm will propose a new DAG by using one of the operators, address prior distribution for the former and the new model, while finally computing the marginal likelihoods given the data.

We will eventually end up with a collection of graphs visited by the chain $\left\{\mathcal{D}^0,\dots,\mathcal{D}^S\right\}$. This range is used to approximate the posterior distribution across models. The approximation comes by the number of visits to each model over the total number of iterations S:

$$p(\mathcal{D}|\boldsymbol{Y})=\frac{m(\boldsymbol{Y}|\mathcal{D})p(\mathcal{D})}{\sum_{\mathcal{D}\in S_q}m(\boldsymbol{Y}|\mathcal{D})p(\mathcal{D})} \tag{2.43}$$

$$\approx\frac{1}{S}\sum_{s=1}^{S}1\left\{\mathcal{D}^{(s)}=\mathcal{D}\right\} \tag{2.44}$$

In the formula above, $1\{\cdot\}$ is the indicator function taking value 1 if $\mathcal{D}^{(s)}=\mathcal{D}$, 0 otherwise.

Once stated the approximation method, it is possible to introduce a method for obtaining a single model estimate, including measures of uncertainty around it. The literature proposes a vast number of approaches; one of the most known being the highest probability model. This method consists of choosing the model with the highest posterior probability associated. Some studies in gaussian linear

regression [3], however, suggest that such model selection method could be non optimal if the model proposed were to be used for prediction. Another common method used is the median probability model; this model, instead of the highest probability model, has been shown to be predictively optimal. The median probability model is obtained by including all variables having a posterior probability of inclusion being greater than 0.5. In this second case, the marginal posterior probability of inclusion $u \to v$ by:

$$p_{u\to v}(\boldsymbol{Y}) = \sum_{\mathcal{D}\in S_{u\to v}} p(\mathcal{D}|\boldsymbol{Y}) \tag{2.45}$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} 1_{u\to v}\left\{\mathcal{D}^{(s)}\right\} \tag{2.46}$$

where $S_{u\to v}$ represents the class of graphs in the graph space S_q containing the directed edge $u \to v$. The indicator function here takes value 1 if and only if the graph $\mathcal{D}$ contains the directed link being searched. Then, a single model estimate is obtained by including all edges $u \to v$ such that $p_{u\to v} > 0.5$.

We need however to disclose a significant different between the median probability model and the other selection method models. The Median Probability model is a PDAG, not a DAG (for instance neither an EG). Furthermore, as an alternative for the classical threshold of inclusion of $k = 0.5$, present literature suggests it is possibe to choose k by considering the expected False Discovery Rate (FDR). Setting for simplicity $p_{u\to v}(\boldsymbol{Y}) = p_{u\to v}$, the expected FDR is expressed as

$$FDR(k) = \frac{\sum_{u=1}^{q}\sum_{v\neq u}(1-p_{u\to})1(p_{u\to}|k)}{\sum_{u=1}^{q}\sum_{v\neq u}1(p_{v\to u}|k)} \tag{2.47}$$

The first step consists in building a grid of trhesholds, for which then will be created a collection of quantile probability models. Each model wiill be associated to a value of $k \in [0, 1]$ and is obtained by including all edges s.t $p_{u\to v} > 1 - k$

The denominator of the function above corresponds to the number of edges in the quantile probability model of order k. The numerator considers for the probability of noninclusion $(1 - p_{u\to v})$. Another important result of the conclusion above is that it is possible to show that FDR(k) is a nondecreasing function of k; we can therefore select k such that the FDR is below a desired finite value.

As the aim of this study does not concern prediction, we decided to pursue as selection model the highest probability method. Once chosen the model selection method, we tested our model accuracy by running it over a series of data with a stated a priori DAG structure.

Chapter 3

Application

The following chapter will present an application of the Bayesian Graphical VAR model to a set of economic variables used for the predicion on inflation in the United States. Paragraph 3.1 will introduce the scope of the example and the variables used, as all the data cleaning process. Paragraph 3.2 presents the results coming from the application of the model itself, while paragraph 3.3 will hold a summary of the conclusions drawn as some possible suggestions for further developement of the study.

3.1 Data presentation

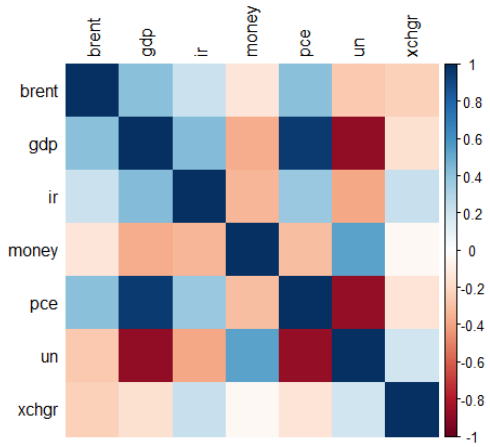
The data considered in this empirical analysis have been gathered from the FRED (Federal Reserve Economic Data). We chose to concentrate on a set of economic variables for the United States, considering data between June 2001 and June 2022. In specific, our aim is to assess how accurately DAG models could trace relationship between a set of variables over time. The usefulness of similar models in the field of economics has been already proven by studies as Giudici and Spelta [9], which focused on undirected graphs. Back to variable choice, we are supported by a vast literature on the topic. Variables used are presented in Table 3.1; we decided to partly adhere from the variables used in the study of Fulton and Hubrich [8].

As Shown in 3.1, the variables considered seem all to be consistent, as no variable seems to show no correlation. Only the exchange rate between Chinese Yuan and US Dollar has a lower degree of correlation with some variables as with broad money or with PCE. From figure 3.1, we can even spot correlations that satisfy expected basic economic theory concept: one brilliant example is the inverse relationship between PCE and unemployment. As PCE is an inflation indicator, this negative relationships reflects the Phillips curve theory. Other correlations suggest less expected relationship; for instance, we could consider the positive relationship that data suggest between Oil price and US GDP. Expanding the underlying theory supporting some of the relationship of the considered economic factors, however, is out of the scope of this chapter. To further support a graphical representation of relationship, one could consider the scatter plot

Table 3.1: Variable description

GDP : *US Gross Domestic Product*
UN : *unemployment rate*
BRENT : *global price of Brent Crude*
IR : *Federal Funds Effective Rate*
XCHGR : *Chinese Yuan Renminbi to U.S. Dollar Spot Exchange Rate*
PCE : *Personal Consumption Expenditure*
MONEY : *Monetary aggregate M3*

Figure 3.1: Correlation plot



of variables, as Figure 3.2 suggest. Besides the graphical representation, we need to remark a substantial difference between what we observe in this exploratory analysis and what we will have as a result as outcome of the model. The correlation we now analyze is static, such measure does not take into account temporal dependence or the effect of an autoregressive model.

In these two different frameworks, we even have to consider two different types of independence; in a static framework that does not take into account the effect of time, we can say that it holds only a concept of *marginal independence*. Using the bayesian approach, instead, we are able to take advantage of additional information coming from the analysis of joint distributions that valorize time dependency. In the latter case, we are able to work within the concept of *conditional independence*. This helps us in many ways, one of the most important is reducing the impact of possible confounding effects on the variables.

Broadly speaking, we considered variables monitoring the economic activity as well as a measure of inflation, PCE. Specifically, we considered the quarter over quarter percent change in PCE produced by the Bureau of Economic Analysis (BEA). This variable has been chosen as the longer run inflation objective of the Federal Open Markets Committee is stated in terms of PCE inflation. None of the

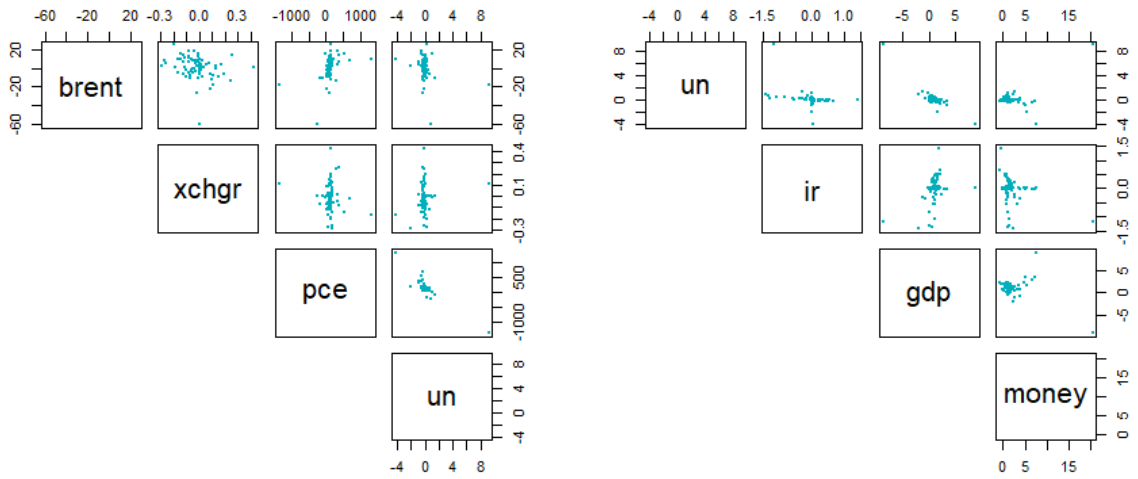


Figure 3.2: Scatter plots

variables suffered from missing values, therefore we did not apply any data manipulation technique. As some of the variables considered are collected with a monthly frequency, we decided to uniform their level of aggregation to quarter over quarter by averaging monthly obserbations.

In order to convert our time series to stationary, we considered now on the series with a first difference applied; figure 3.3 shows the plot of the differenced time series.

. After running a Dickey Fuller test on the selected variables, we were able to determine stationarity of all the variable with a confidence of 95%

3.2 Application

Once finished the data cleaning process, we proceed by building the data frame later to be used in the DAG Model. The first step consists in choosing an optimal lag for the VAR Model we use. For determining the optimal level of lag p for our VAR(p) model we applied a selection method based off of the Aikake information criterion (AIC) on a sequentially increasing order of lag; according to the result, the best lag to apply on the VAR is of order 9.

This result, however, could be stretched considering the autocorrelation function of the variables in the model, as figure 3.4 shows: only the differenced series of interest rate seems to have autocorrelation of level 9 and above; however, we concluded this serial correlation degree is not significant for the analysis. Most of the variables show an autocorrelation of order 2, which we identified being an optimal lag order for our VAR model aswell. Concluding, a lag of 9 in the VAR model woud not satisfy the tradeoff between additional information for the model and increased complexity.

We proceded by creating the matrices later to be used in the DAG model; for notation and procedure, we followed what has already been presented in Chapter 2. Additionally, before training the algorithm,

Figure 3.3: Differenced time series

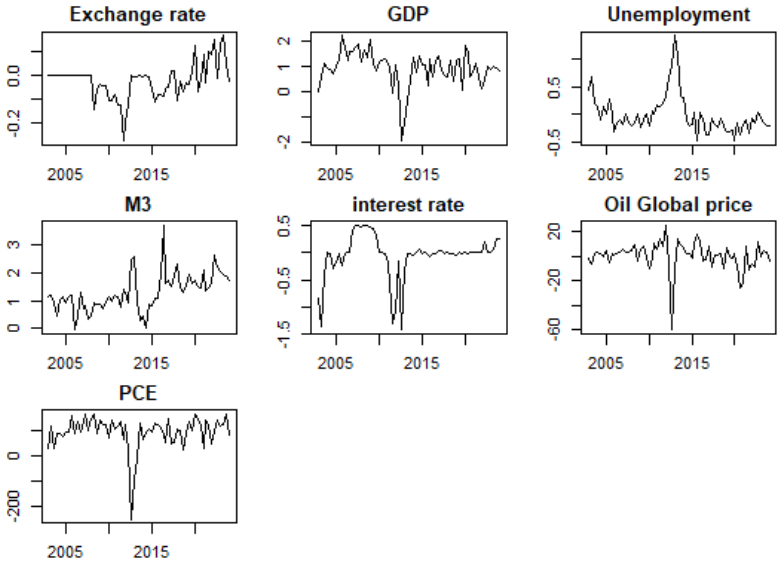
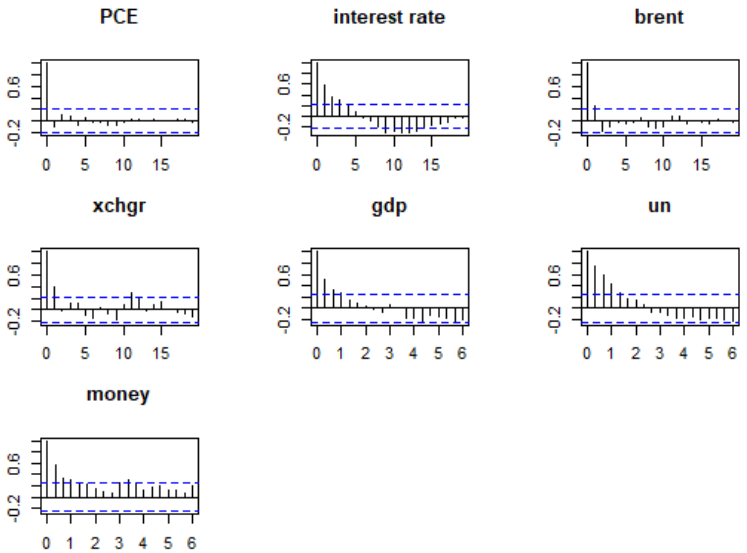


Figure 3.4: Autocorrelation functions of the model variables



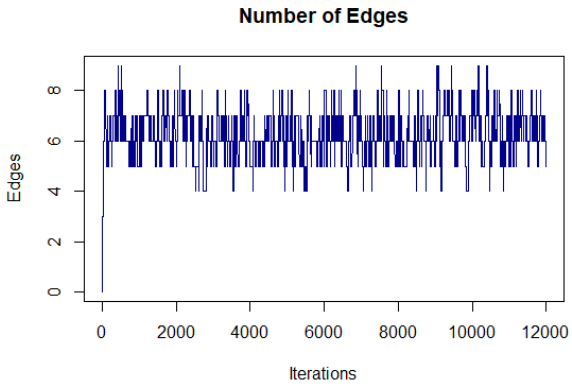


Figure 3.5: Convergence dynamics

we introduced in the coefficient matrix $\boldsymbol{Z}$ an intercept; this has been decided in order to better display the model in the light of a higher variability in some of the variables caused by harsh economic downturns; in the time span we considered for the analysis, in particular, we considered data regarding two major shocks as the 2008 crisis and the Covid-19 pandemic. Regarding the model, we decided to model the edge inclusion probability based on a Bernoulli prior centered at $\pi = 0.5$. This prior value stands as an impartial prior.

As explained in 2.2, our model structure implements an MCMC Metropolis-Hasting algorithm for performing the optimal model choice proposed as an output. For the algorithm to be successful, we need it to reach *stationarity*; this property enables the process of reaching a stationary probability distribution over all state independently of the starting value of the chain. In order to allow for convergence, we need to lenghten the sampling process. Such process allows the model to tune in and allows convergence. We will then discard an inital number of samples s as the chain has not reach its stationary state. This period needed to reach stationarity is defined as *burn-in period*. In this study, we decided to apply a burn in of 5000 iterations, over a total of 25000 iterations (burn-in of 20% of the data). With the aim of analyzing convergence, we can support our decision making over the optimal number of iterations and burn in by performing some diagnostics of convergence. The measure we chose as diagnostic of convergence is the number of edges proposed by the model on each iteration. By sampling a large number of iteration, we can in fact assess the optimal burn in as a sensible iteration measure s. Convergence will show when as the number of iteration increase, the number of edges proposed by the model solidifies in a fixed range, without showing upwards or downwards drifts. By plotting the number of edges over the 25000 iterations, we have the result shown in figure 3.5.

From this graph we can see that after the fist 1000 iteration the algorithm proposal of edges starts to stabilyze around a range between 5 to 8 edges. Some rounds of iteration enlarge the band, bringing the range from 4 to 9 edges.

To further prove whether the chain has reached convergence, we decided to run the algorithm twice,

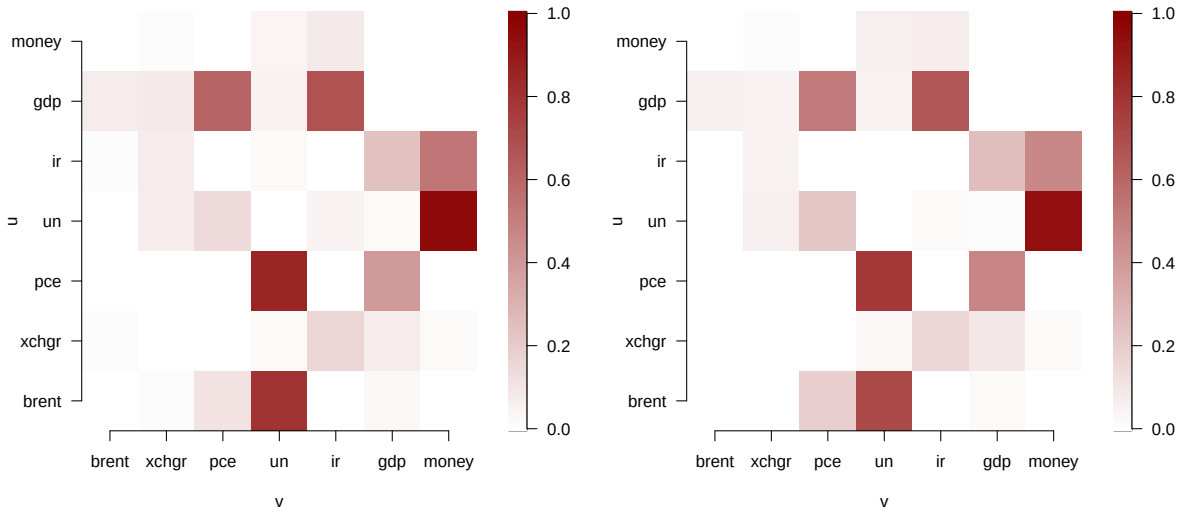


Figure 3.6: Heatmaps with posterior probabilities of edge inclusion obtained from two independent MCMC runs.

keeping fixed the number of iterations, our burn in process and the prior probability w. We called these trials respectively *out_mcmc*1 and *out_mcmc*2. We then proceded by computing the posterior probabilities for both cases. At the end of the process we chose as a measure of comparison a graphical representation: the heatmap of the posterior probabilities resulting from these trials.

Figure 3.6, finally, presents a comparison between the heatmaps of the posterior probabilities of the two trials. As we can see, posterior probabilities do not seem to have a dramatical difference in these two trials. The Heatmap of *out_mcmc*2, however, modeled higher posterior probabilities in some areas. To conclude, as the heatmaps suggest similar DAG structures, we can state that the burn in we ran suffices for the convergence of our MCMC Hasting algorithm. We can present an alternative plot to show a comparison between posterior probabilities in these two cases; we can represent in a scatter plot the result obtained in the two mcmc sampling rounds and see how the distribution of results unravel.

In Figure 3.7, an observation distribution along the 45 degree line would imply a perfect convergence, as the output probabilities coming from a model with the same set of specification should coincide. In our case, we can see that values fall along the dashed red line, although in some cases we see a certain degree of variability around the final output. The degree to which results coincide may be improved by increasing the number of iteration in the algorithm of origin, even though from the Convergence diagnostics graph on the proposal of edges the MCMC Hasting algorithm seemed to have reached convergence.

As final step, we proceed by building a DAG onto the adjacency matrix obtained by the mcmc algorithm. The outcome DAG is visible in Figure 3.8

Our graph outlined an interesting set of relationships. An immediate visible conclusion we can draw

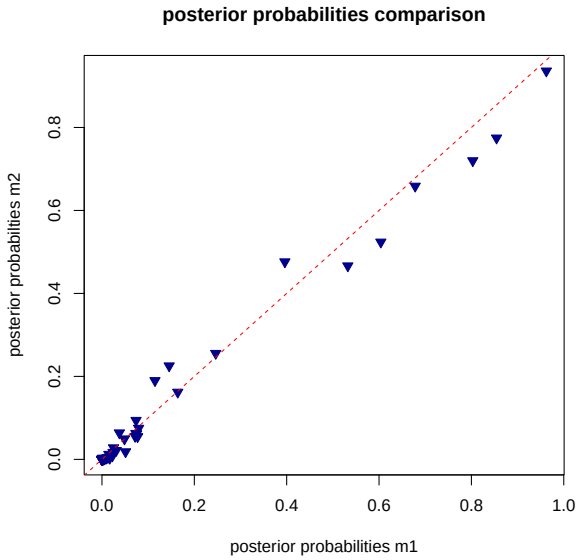


Figure 3.7: Scatter plot confronting posterior probabilities o our two mcmc trials

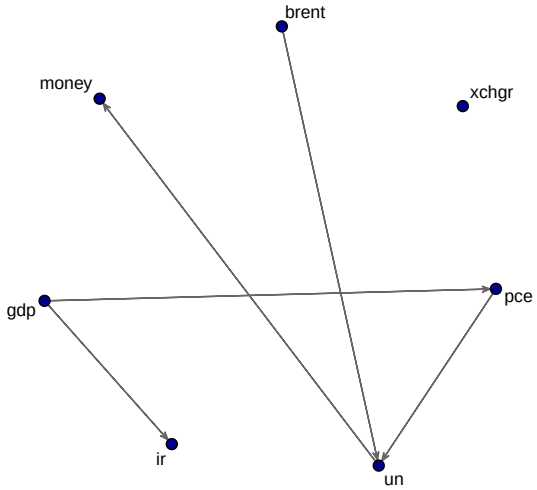


Figure 3.8: graphical representation of our DAG model

from this DAG is the absence of edges pointing towards the Yuan - USD echange rate. As the model suggests, the variable is not apparently related to any of the other in the model. Moving on, the graph contains every type of graphical structure: we do have forks, chains and colliders. We call a forked path a structure of the type $u \leftarrow v \rightarrow x$; the only forked path we have in this model is the path that connects GDP, PCE and Interest rates. In particular, according to the graph, GDP has been identified as the *parent* of both PCE and Interest rates. As PCE and interest rate have a mutual parent, they can be considered as associated; this conclusion could reflect theory, as in fact interest rates are a tool policy makers use for controlling for inflation. GDP being a moving factor could be a sensible conclusion. looking at economic theory, as in the short run economic otuput could be stimulated by increase in aggregate demand, this could potentially influence inflationary pressures (e.g producers increase prices) in the long run. Considering the specific fork path, however, it is more difficult to determine causality for the open nature of the path; in this case, there is the possibility of having confounding effects that condition the observed relationship. Going on to present other structures, we can spot in our graph the presence of different chains, being respectively:

- $brent \rightarrow unemployment \rightarrow money$,
- $gdp \rightarrow interestrate \rightarrow money$,
- $gdp \rightarrow PCE \rightarrow unemployment$.

In these directed path, each node is causal on the next. Accordingly the model captured that, in the US economy. global oil prices have a direct influence over the unemployment rate. Unemployment, on the other hand, reportedly influences the monetary aggregate M3. Giving an economic explanation to the following relationship, the consequence of volatility/shocks in oil prices on US unemployment rate has been studied in different papers; an underlying theory explaining the former relationship could be the real options theory, which reveals that the uncertainty about goods' prices leads firms to postpone or abandon their production and investment and they seem to be robust to the use of different real oil price measures. Proceeding through the chain, the relationship between unemployment and money itself could be interpreted as the policy makers decision making process: once understated the economic outlook through key economic indicators (one of them being unemployment, which in the US is a key variable for monetary policy as the mandate of the FED declares), policy makers act by affecting money velocity either through interest rates or monetary aggregates.

We define inverted forks, or path with colliders, the paths of the form $u \rightarrow v \leftarrow x$. In this case, two arrowheads meet a node; in technical terms, we say that a children node has two parent nodes. This path is not an open path, as it is blocked at the collider: in such case, we don't need to account for v to assess for the causal effect of u on x as the"backdoor path" is blocked by v. Pearl defines x and u as *d-separated* (direction separated) as there is no effect on one by another. Applying the concept on our graph, we can state that Global Oil Price and PCE are *d-separated*, meaning there is no causal effect

related between these variables. The same conclusion could not be drawn in the case unemployment rate and interest rate; these nodes have a common ancestor, GDP, which in our graphical structure impedes to exclude possible confounding effects.

Once explored the existing relationships and path in our DAG, we can briefly discuss the absence of causality between variables. The absence of a traced causality between Oil prices and PCE seems unexpected, as energy prices contribute to inflation (in fact, lately they have been a major inflation push factor in most economies). The absence of a significant relationship, however, may be traced in the nature of the inflation measure we chose. PCE, according to the Bureau of Economic Analysis (BEA), measures the goods and services purchaesed by persons: such category is composed by households and non profit institutions serving households (NPISHs) who are resident in the United States. Furthermore, energy costs are incorporated in this measure in its section related to *Nondurable goods*. It is possible that in the computation of the PCE measure, energy prices have been considered having a small impact in the composition of the measure itself. The result may have been different while considering the other US inflation measure, CPI. This second measure may incorporate more weight related to consequences of Crude price increases. Reportedly, the relative importance of Energy cost in 2022 was of 7.54%.

3.3 Conclusions

This last chapter of the study helped in demonstrating practically how a DAG model works in a set out of its original field: economy and finance. After having performed a thorough variable exploration activity, we decided to choose 7 economic factors representing the main movers for the United states economy. We performed an exploratory analysis on these variables, checking for missing values, errors in the data. After having cleaned the data, we pursued on by performing tests on stationarity of our series; after having applied differencing, we got rid of nonstationarity and performed other tests on studying the autocorrelation level of the variables. Once prepared the data, we collected all variables in a dataframe, leaving out the time string. We used this initial dataframe to question the optimal value p for the VAR model we needed to build and feed to the DAG algorithm. Once decided for the optimal value of lags to consider, in our case 2, we created the variables needed for computing marginal likelihoods, in general for running the MCMC metropolis Hasting algorithm that would propose the optimal final DAG. At this point, we tried different sampling rounds with a range of iteration s. We did this twice saving separately the posterior probabilites draws in order to compare the cases and actually determine if the chain had reached convergence. We then determined an optimal burn in rate and proceded by running the algorithm one last time; our model demonstrated to capture all of the edges between the variables, attributing significant economic importance to variables that, given the exploratory analysis, would not seem as significant as the model finally suggested.

Chapter 4

Conclusions and future lines of research

In this work, we presented a Vector Auto Regressive (VAR) Bayesian model for structure learning of dependence relations between time-dependent variables. In Chapter 1, we provided the reader with basic theory and notation needed for understanding the basic concepts of graphical models and VAR models. Here, we introduced graphical concepts terminology useful to read conditional independence relations from graphs. In particular, we considered directed and undirected, cyclic, and acyclic graphs, although we based our modeling framework on Directed Acyclic Graphs (DAGs). We then introduced the reader to Bayesian model selection methods; we explained how to perform model choice in a Bayesian framework Specifically, we focused on Bayesian model choice through marginal likelihoods and Bayes Factors (BFs).

We also introduced Markov Chain Monte Carlo (MCMC) techniques as a computational tool for practical implementation of model selection methods, which were then applied for structure learning of graphical DAG-based VARs. In Chapter 2 we then presented our Bayesian graphical VAR model. To this purpose, we first introduced the notation and framework of Bayesian multivariate linear regression, which represented the starting point for the development of our graphical VAR model.

Next, we extended the model to time series data whose dependence structure satisfy the conditional independencies imposed by a DAG and write the implied likelihood function. . Following a Bayesian framework, we specified priors for the model parameters Markov w.r.t. a DAG. By combining the likelihood with the prior we obtained the posterior distribution of the model parameters. In addition, we derived the marginal (i.e. integrated w.r.t. the prior) data distribution (also named marginal likelihood), which in force of prior conjugacy was obtained from the ratio of prior and posterior normalizing constants. Next, we showed how a multivariate linear regression model can be generalized to incorporate a VAR structure satisfying the independence constraints imposed by a directed network. In the last part we also detail how to perform model selection using the notion of Fractional Bayes Factor (FBF); the

latter provides an effective and objective solution to the issue of prior elicitation when substantive prior knowledge is not available. Finally, we introduced an MCMC scheme for model selection of graphical VARs based on a Metropolis Hastings algorithm.

The last chapter focused on the application of the proposed model in the field of economics and international finance; we performed our empirical analysis on selected economic variables that are regarded as relevant indicators of the US economy. Data were collected from the Federal Reserve Economic Data (FRED) repository. After a brief exploratory analysis we fixed prior hyperparameters, the number of MCMC iterations and then run our algorithm for model selection of graphical VARs. To assess the robustness of the so-obtained results we finally performed graphical convergence diagnostics. Our inferential results supported the economic theory which establishes dependence relationships between the variables we considered. Our model, however, did not trace some relationships we expected to gather from the available theory; this may depend on the specific set of variables that have been included in the study.

Graphical VAR models can provide an effective tool for economic research. Besides this field, such models are widely employed in medicine, e.g. for diagnostic purposes, causal reasoning, as well as decision making under uncertainty and prediction via automated insights. More specifically, DAG-based models provide a natural generalization of simpler models such as VARS. In particular, they allow to perform model selection and parameter learning. In addition, DAGs can be implemented to several data types and in particular both discrete/categorical and continuous variables. For the former, one could for instance implement unsupervised algorithms to automatically determine a discrete latent variable which can model hidden patterns. Alternatively, it is possible to implement constraint-based algorithms which run independence tests to determine the underlying network structure. Still in the framework of VAR models and time series analysis, it could be possible to improve the proposed model presented for different tasks, one being for example prediction. Bayesian networks in particular can handle efficiently missing data (i.e. unavailable observations to predict) due to their probabilistic foundations. Differently, missing data could not be easily estimated via imputation techniques or other methods because of the complex multivariate structure behind the data. Even more useful would be the possibility to simultaneously predict multiple outputs using a single graphical model.

Besides economic analysis and international finance, DAG models can find a fertile area of application as statistical methods for dealing with systems of digital payments. Directed networks are employed in cryptocurrencies as tools for recording and tracking transactions, as an alternative to the traditional block-chain. In the cryptocurrency world, they were indeed mentioned the distributed ledger technologies (DLT). Its improvement point is that, unlike blockchain, dag networks process transactions individually wihout grouping them together. DAG networks, on top, are able to instantly approve transaction and adding it to the ledger. To approve transaction, generally a DAG network uses previous transactions on the network to confirm the new one; instead of a traditional blockchain, thereofore, each transaction is peer reviewed. The DAG structure allows the network to process new transactions even when others

are not yet completed. Additionally more than one transaction can get processed at the same time. Moreover, since transactions are peer approved, as more users join the network the model can handle many transactions and the network can scale in principle to support an infinite amount of transactions. Such transaction tracking method, due to its scalability, could result in a possible improvement if used in a major payment system as those of traditional currencies.

Bibliography

[1] Andersson, S., Madigan, D., and Perlman, M. (1943). "On the Statistical Treatment of Linear Stochastic Difference Equations." *Econometrica*, 11: 173–200.

[2] — (1997). "A characterization of Markov equivalence classes for acyclic digraphs." *Ann. Statist.*, 25(2): 505–541.

[3] Barbieri, M. and Berger, J. (2004). "Optimal predictive model selection." *Ann. Stat*, 32: 870–897.

[4] Castelletti, F. (2020). "Bayesian Model Selection of Gaussian Directed Acyclic Graph Structures." *International Statistical Review*, 88(3): 752–775.
URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12379`

[5] Consonni, G., La Rocca, L., and Peluso, S. (2017). "Objective Bayes Covariate-Adjusted Sparse Graphical Model Selection." *Scandinavian Journal of Statistics*, 44: 741–764.

[6] Darroch, J., Lauritzen, S., and Speed, T. (1980). "Markov fields and log-linear interaction models for contingency tables." *Ann. Statist.*, 8(3): 522–539.

[7] Dawid, A. (2001). "Separoids: a mathematical framework for conditional independence and irrelevance." *Ann. Math. Artif. Intell.*, 31(1): 335–372.

[8] Fulton, C. and Hubrich, K. (2021). ""Forecasting US inflation in real time", Finance and Economics Discussion Series 2021-014." *Washington: Board of Governors of the Federal Reserve System*.

[9] Giudici, P. and Spelta, A. (2016). "Graphical Network Models for International Financial Flows." *Journal of Business & Economic Statistics*, 34(1): 128–138.

[10] Imbens, G. W. (2019). "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics."
URL `https://arxiv.org/abs/1907.07271`

[11] Kilian, L. and Lutkepohl, H. (2017). *Structural Vector Autoregressive Analysis*.

[12] Lauritzen, S. (1996). *Graphical Models*.

[13] Lauritzen, S., Maathuis, M., Drton, M., and Wainwright, M. (2020). *Handbook of Graphical Models*.

[14] O'Hagan, A. and Foster, J. (2004). *Kendall's advanced theory of statistics: Bayesian inference.*.

[15] Paci, L. and Consonni, G. (2020). "Structural learning of contemporaneous dependencies in graphical VAR models." *Computational statistics & Data Analysis*, 144.

[16] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*.

[17] — (2000). *Causality: Models, Reasoning, and Inference*.

[18] Pearl, J. and Paz, A. (1987). "Graphoids, graph-based logic for reasoning about relevance relations." *Advances in Artificial Intelligence*, 2: 357–363.

[19] Sprites, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction and Search (2nd edition)*.

[20] Verma, T. and Pearl, J. (1991). "Equivalence and synthesis of causal models." *Proceedings of the sixth annual conference on uncertainty in artificial intelligence.*, 255–270.