Graphical Models For Time Series Analysis

Yann McLatchie

The University of Manchester

January 2021

Abstract

This project is on the application of graphical models to the identification of vector autoregressive (VAR) models for given multivariate time series data.

When building a VAR model, we look to discover the causal structure of the variables. These causal relationships can be derived from the variables' conditional independence structure. In the context of graphical models, we first use the Inverse Variance Lemma to build a conditional independence graph (CIG) from the sample covariance matrix of the multivariate time series data. Nodes in this graph represent the potential variables in our VAR model, and they are connected with an edge if they are significantly conditionally dependent. From this conditional independence graph we achieve a directed acyclic graph (DAG) through demoralisation, reducing the number of edges in our graph and producing a set of possible DAGs. Such DAGs represent the causal relationships between variables, and correspond to structural VAR models. Fitting each of the models represented by these DAGs to data, we find the best one using penalised likelihood methods.

In this project, we apply this methodology to simulated data, and find that we are able to accurately and confidently recover the true model with the help of graphical models. Being confident in the power of this methodology, we move on to apply it to multivariate financial time series data, using the log returns of Standard and Poor's 500 Index, Bond Index, and GSCI Index at different time periods. The R code for the production of a CIG from time series data is provided.

Contents

Chapter 1

Introduction

When faced with time series data, it is interesting to identify structural relationships in our data and ultimately leverage this information into building models for prediction. This paper will investigate the application of ideas from graph theory to time series analysis on stationary processes with the aim of discovering such dependencies between variables. By understanding these relationships, we will be able to suggest an initial set of possible autoregressive models to fit the data which we will then be able to prune down to a best model.

In order to develop the methodology required to so, we begin this paper by laying the probabilistic foundations of independence and conditional independence between variables. This is extended to introducing the Reduction Lemma and Block Independence Lemma to describe conditional independence between variables in a partitioned random vector, which is critical to translating these concepts into graphs. Having covered this, we move onto defining some key concepts in graph theory, before linking them back to independence in the form of conditional independence graphs (CIGs). In this chapter, the graphical representation of conditional independence is introduced. These probabilistic graphical models are used as the foundation for the rest of the paper. They aim to visual the joint distribution of the variables, and to simplify it. In understanding the structural information of the graph, we understand the mutual relationships between the variables (nodes) within. It is also at this stage that we introduce the Inverse Variance Lemma, which allows us to calculate a partial correlation from the precision matrix of a data sample. We use these partial correlations to build CIGs from data samples. If the partial correlation between two variables is statistically significant, then we add these variables to a CIG as nodes, and link them with an edge. Having developed the framework

through which to understand the broader conditional dependencies and represent them in a CIG, we move on to the notion of causal relationships and aim to visualise these with directed acyclic graphs (DAGs). We define the concepts of moral edges in a graph, those linking two parent nodes that share a child node, and the Wermuth Condition. We use this property to build a methodology to obtain a DAG from a CIG, through the process of removing moral edges, "demoralisation". Once we have finished developing the idea of graphical models, we turn to stochastic processes. Defining what it means for a stochastic process to be stationary and Gaussian, we motivate vector autoregressive models (VAR) as a means of describing stationary multivariate time series data. We discuss the difference between canonical VARs, where contemporaneous variables (those occuring at the same time period) have no effect on each other, and structural VARs where there exist such structural dependencies between contemporaneous variables. We provide a methodology to achieve a canonical model from a structural model, and vice versa. Having done so, we move to link the graphical models developed earlier with the identification of appropriate VAR models to fit the data with a simple "algorithm". First we calculate the covariance matrix of the lagged and contemporaneous variables. Using the Inverse Variance Lemma we achieve a partial correlation matrix, and from this build a CIG. We use any *a priori* knowledge of the data to assert the possible direction each edge in our CIG might take, and enumerate all possible DAGs associated with the CIG due to moralisation. We find a direct relationship between a DAG and a VAR model. Thus each DAG represents a possible VAR model which we might fit to our data. Fitting them all iteratively and keeping track of their respective performance using penalised likelihood methods, we are able to determine which model best fits our data.

Once we have covered the theory guiding this methodology, we move to apply it in two examples. In the first, we will simulate data according to a known structural VAR model and aim to recover this true model using our methodology. In the second, we will use five years of data from three Standard and Poor's indices, treating each index as a variable in our stochastic process, and aim to determine which model best fits the data, and what this means intuitively.

The aim of this paper is to demonstrate that this methodology is not only able to directly produce a set of possible models for some given time series data, but further that we are able to fit each model in the set and optimise for BIC and AIC to find the best one.

Chapter 2

Probability Theory

This work is based heavily on the notions of independence and conditional independence. We will use conditional independence graphs, and later directed acyclic independence graphs to represent conditional relationships between variables. Understanding the fundamental concepts of independence is critical in order to achieve these.

2.1 Independence and Conditional Independence

We will take results for this section mostly from Reale (1998). At the most basic level, independence is defined as follows.

Definition 2.1.1. We say that two events A and B are independent, denoted $A \perp B$, if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

We can intuitively understand this as learning something about one of the two events does not provide any information regarding the second. For example, when rolling a die, learning the outcome of the first roll does not provide any information pertaining to the outcome of the second.

The notion of conditional probability is needed to be able to later define conditional relationships between variables. Its definition is as follows.

Definition 2.1.2. The probability of an event A given the event B is given by

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

when $\mathbb{P}(B) > 0$.

We immediately find an equivalent definition of independence using conditional probability. The two events A and B are independent if

$$\mathbb{P}(A \mid B) = \mathbb{P}(A).$$

Using this notion of conditional probability, we are able to define conditional independence of two events given a third.

Definition 2.1.3. We say that two events A and B are conditionally independent given a third event C, denoted $(A \perp B)|C$, if

$$\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C).$$

Intuitively, we can understand this with a common example, often given in literature. Say we wish to predict the size of an individual's vocabulary from their height. These two are likely dependant, since the taller an individual is, the more likely they are to be older and thus to have a broader vocabulary. Suppose now that we fix the individuals' ages, then there is now no reason to believe that height and vocabulary are dependant.

We may generalise these concepts from events in a probability space to random variables.

Definition 2.1.4. A random variable, X, is a function from the probability space Ω to the set of real numbers $\mathbb{R}$, denoted

$$X : \Omega \to \mathbb{R}$$
$$X(\omega) \mapsto x$$

for $\omega \in \Omega, x \in \mathbb{R}$.

Suppose we flip a coin a hundred times, and define a random variable X as the number of heads after those hundred flips. The event that the number of heads after a hundred flips is

less than some integer x, is denoted A, and it's associated probability is

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(X \leq x) \\ &= \mathbb{P}(X = x) + \mathbb{P}(X = x - 1) + \cdots + \mathbb{P}(X = 1) \\ &= \sum_{t \leq x} p_X(t) \\ &= F_X(x)\end{aligned}$$

where we call $F_X(x)$ the cumulative distribution function.

Now say we wanted to compute the probability that after these hundred flips we have exactly some integer x heads. In the example above, we have a finite, discrete set of probability events, and thus this probability in positive and denoted $\mathbb{P}(X = x) = p_X(x)$, the probability mass function.

In the continuous case, we do not have a discrete support, and so we cannot consider a random variable taking a specific value x, since

$$f_X(x) = \lim_{d \to 0} \mathbb{P}(x \leq X \leq x + d) = 0$$

for all x in the probability space. In this case, we consider the probability that x exists in some interval on the real axis, $\mathbb{P}(X \leq x)$, for which we can use an integral to sum all $\omega \in \Omega$ such that $X(\omega) \leq x$, as follows

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{t \leq x} f_X(t)$$

and then $f_X(x) = \frac{\mathrm{d}F}{\mathrm{d}x}(t)$.

Definition 2.1.5. Let X and Y be two continuous random variables, then the conditional density of X given $Y = y$ is defined as

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

Let A be the event that $X \leq x$, and B the event that $Y \leq y$. Recall that A and B are

independent if

$$\begin{aligned}\mathbb{P}(A \cap B) &= F_{X,Y}(x, y) \\ &= \mathbb{P}(X \leq x, Y \leq y) \\ &= \mathbb{P}(X \leq x)\mathbb{P}(Y \leq y) \\ &= F_X(x)F_Y(y) \\ &= \mathbb{P}(A)\mathbb{P}(B).\end{aligned}$$

Then we can further say that the random variables X and Y are independent if there exist two functions g and h such that

$$f_{X,Y}(x, y) = g(x)h(y) \tag{2.1}$$

where,

$$f_{X,Y}(x, y) = f_X(x)f_Y(y).$$

Definition 2.1.6. Let X, Y, and Z be continuous random variables, then Y and Z are said to be conditionally independent given $X = x$ if

$$f_{Y,Z|X}(y, z \mid x) = f_{Y|X}(y \mid x)f_{Z|X}(z \mid x), \tag{2.2}$$

for any x, y, z. This has the same intuition as with Definition 2.1.3.

Proposition 1. The condition (2.2) can be equivalently stated as

$$f_{X,Y,Z}(x, y, z) = \frac{f_{X,Y}(x, y)f_{X,Z}(x, z)}{f_X(x)} \tag{2.3}$$

or,

$$f_{Y|X,Z}(y \mid x, z) = f_{Y|X}(y \mid x). \tag{2.4}$$

In other words, we can cut the conditioning set from X and Z down to X. This can be demonstrated intuitively with an example. Say you have a regression model with a response and two regressors. If the response is independent to one of the regressors given the other, then it is possible to drop this conditionally independent regressor from the model.

Proof. Let us start by proving (2.3) using (2.2).

$$\begin{aligned} f_{X,Y,Z}(x,y,z) &= f_{Y,Z|X}(y,z)f_X(x) \\ &= f_{Y|X}(y \mid x)f_{Z|X}(z \mid x)f_X(x) \\ &= \frac{f_{Y,X}(y,x)}{f_X(x)}\frac{f_{Z,X}(z,x)}{f_X(x)}f_X(x) \\ &= \frac{f_{X,Y}(x,y)f_{X,Z}(x,z)}{f_X(x)} \end{aligned}$$

Each step in this proof is reversible and thus we simultaneously prove (2.2) using (2.3).

Moving onto proving (2.4) using (2.3),

$$\begin{aligned} f_{Y|X,Z}(y \mid x,z) &= \frac{f_{Y,X,Z}(y,x,z)}{f_{X,Z}(x,z)} \\ &= \frac{f_{X,Y}(x,y)f_{X,Z}(x,z)}{f_{X,Z}(x,z)f_X(x)} \\ &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\ &= f_{Y|X}(y \mid x) \end{aligned}$$

Again, each step is reversible and we thus also prove (2.3) using (2.4). □

Proposition 2. Y and Z are conditionally independent given $X = x$ if and only if there exist functions g and h such that

$$f_{X,Y,Z}(x,y,z) = g(x,y)h(x,z).$$

Proof. Assume conditional independence, then from (2.3) we have

$$f_{X,Y,Z}(x,y,z) = \frac{f_{X,Y}(x,y)f_{X,Z}(x,z)}{f_X(x)} = g(x,y)h(x,z)$$

where

$$\begin{aligned} g(x,y) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\ h(x,z) &= f_{X,Z}(x,z) \end{aligned}$$

and $f_X(x) > 0$. Now suppose we have g and h such that $f_{X,Y,Z}(x,y,z) = g(x,y)h(x,z)$. Let

us begin by writing

$$f_{X,Y,Z}(x, y, z) = f_{Y,Z|X}(y, z \mid x) f_X(x).$$

In order to write this in terms of g and h, we must split $f_{Y,Z|X}(y, z \mid x)$ into $f_{Y|X}(y \mid x)$ and $f_{Z|X}(z \mid x)$. By Definition 2.1.6, we thus have that $Y \perp Z \mid X$. □

We will introduce two more propositions before concluding this section with the *Block Independence Lemma*.

Proposition 3 (Reduction Lemma)**.** If (X, Y, Z_1, Z_2) is a partitioned random vector, then

$$Y \perp (Z_1, Z_2) \mid X \Rightarrow Y \perp Z_1 \mid X.$$

Proof. We have that $Y \perp (Z_1, Z_2) \mid X \Rightarrow f_{Y,Z_1Z_2|X}(y, z_1, z_2 \mid x) = f_{Y|X}(y \mid x) f_{Z_1Z_2|X}(z_1, z_2 \mid x)$ by the definition of conditional independence. We can integrate out Z_2 from both sides of this equation, resulting in $f_{Y,Z_1|X}(y, z_1 \mid x) = f_{Y|X}(y \mid x) f_{Z_1|X}(z_1 \mid x)$, which returns $Y \perp Z_1 \mid X$. □

Proposition 4. If (X, Y, Z) is a partitioned random vector, then

$$X \perp (Y, Z) \Rightarrow X \perp Y.$$

In other words, joint independence implies marginal independence.

Proof. The proof of this proposition is similar to the proof of the Reduction Lemma, only with a conditioning set that can be factored out. It is thus a specific case of the Reduction Lemma. □

Proposition 5 (Block Independence Lemma)**.** If (X, Y, Z_1, Z_2) forms a partitioned random vector with positive density $f_X(x)$, then the following two statements are equivalent:

(a) $Y \perp (Z_1, Z_2) | X$

(b) $Y \perp Z_1 | (X, Z_2)$ and $Y \perp Z_2 | (X, Z_1)$

The proof of the Block Independence Lemma is given in Whittaker (1990).

Chapter 3

Graphical Models

In order to link statistical methods with graphical models, we must establish definitions relating to graph theory. We will then move on to define different types of graphical models, and how we can go from one type to the other and vice versa. Finally, we discuss how we can test for significance in conditional independence graphs.

3.1 Basics of Graphs

For this section, we take results from Megyesi (2021).

Definition 3.1.1. A graph is an ordered pair $G = (V, E)$, where V is a non-empty finite set, called the set of *vertices* or *nodes* and E is a set of unordered pairs of elements of V called the *edges*. If $uv \in E$, then we say that u and v are *adjacent* (or neighbours) and they are *incident* with the edge uv.

Definition 3.1.2. A graph $G' = (V', E')$ is called a subgraph of $G = (V, E)$ if $V' \subseteq V$ and $E' \subseteq E$.

Definition 3.1.3. A directed graph, $G^{\succ} = (V, E^{\succ})$, is a graph that has ordered edges represented by arrows.

Definition 3.1.4. A walk in a graph $G = (V, E)$ is an alternating sequence of vertices and edges $v_0 e_1 v_1 \dots v_{n-1} e_n v_n$, where $v_0, v_1, ..., v_n \in V$ and $e_1, e_2, ..., e_n \in E$, and further that the edge e_i connects the two vertices v_{i-1} and v_i.

A path is a walk with no repeated vertices.

A cycle is a closed walk ($v_0 = v_n$) of length $l \geq 3$.

3.2 Conditional Independence Graphs

Conditional independence graphs (CIGs) visually represent the conditional independence relationships between variables. A CIG consists of vertices representing random variables, where two vertices are connected with an edge if the random variables are conditionally dependent given the rest of the graph. Thus if two nodes are not linked with an edge, they are conditionally independent.

Over the course of this section and those that follow also relating to graphical models, we will interest ourselves with results from Reale (1998) and Whittaker (1990).

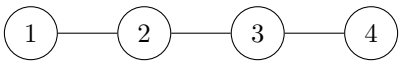


Figure 3.1: An example CIG

Consider the CIG in Figure 3.1 with vertices $1, 2, 3, 4$. We can directly read from it that X_1 and X_3 are independent given X_2 and X_4, or symbolically, $1 \perp 3 \mid \{2, 4\}$. We would like to condense this notation to $1 \perp 3 \mid 2$, as we can see that the node 4 has no effect on the node 1. The Separation Theorem, which we will prove in this section, allows us to do so by concluding that only those nodes that separate two sets of random variables are needed in the conditioning set.

In Figure 3.1, the nodes 1 and 3 are said to be separated by the node 2. We will look to generalise this idea of separation to subsets of vertices before defining some properties of CIGs.

Definition 3.2.1. Two vertices $u, v \in V$ in a graph $G = (V, E)$ are said to be separated by a subset of vertices $S \subset V$ if all paths from u to v include at least one vertex $w \in S$.

Proposition 6. Suppose that a set of vertices $V = 1, 2, ..., n$ of a CIG can be partitioned into two sets B and C such that, there are no edges connecting any vertex $b \in B$ with any $c \in C$. Then,

$$b \perp c, \forall b \in B, c \in C$$

Proof. The main idea in this proof is to establish the independence between two random variables $b \in B$ and $c \in C$ given the other nodes in the CIG, and remove all elements of the conditioning set. We begin by fixing two vertices $b \in B$ and $c \in C$ and we take a third vertex $v \in C$. We know that,

$$b \perp c \mid V \backslash \{b, c\} \text{ and } b \perp v \mid V \backslash \{b, v\}$$

By the Block Independence Lemma, we can combine these properties to say,

$$b \perp \{c, v\} \mid V \backslash \{b, c, v\}$$

And by the Reduction Lemma, we can reduce this to,

$$b \perp c \mid V \backslash \{b, c, v\}$$

We have thus succeeded in factoring out our third node v from our conditioning set. Repeating this for all remaining vertices in C, we eventually reduce the set to $C = \{c\}$ as it is a finite set. Doing the same to the set B, we are eventually left with a graph of only the two vertices $V = \{b, c\}$, and no edges connecting them. We are able to do this for any two vertices $b \in B$ and $c \in C$, and thus we eventually conclude that indeed,

$$b \perp c, \forall b \in B, c \in C$$

□

Proposition 7. Let $A \subset V$ be any subset of vertices in a graph $G = (V, E)$ that separates any two vertices $b \in B \subset V$, $c \in C \subset V$, that is that all paths connecting b to c contain at least one vertex $a \in A$, where A, B, and C are disjointed subsets of V. Then,

$$b \perp c \mid a$$

Or, more strongly,

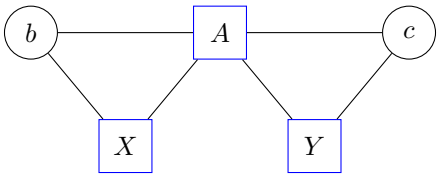
$$B \perp C \mid A$$

Proof. Let $b \in B$ and $c \in C$, then b and c separated by A, and any other vertices in the graph are either connected to $A \cup \{b, c\}$ or not. If they are not, then by construction they do not influence the independence of b and c, and we need not consider them. So we have,

$$X = \{x : x \text{ is separated from } c \text{ by } A\}$$
$$Y = \{y : y \notin X, \text{ and is separated from } b \text{ by } A\}$$

We thus have five nodes in this new representation: b, c, A, X and Y. We link these nodes with an edge if it is possible for any elements within them to be connected in the original CIG.

By construction, there are no edges connecting b and c, b and Y, c and X, or X and Y (as otherwise they would belong to A). This is represented in the following diagram.



We now look to apply the Block Independence Lemma and then the Reduction Lemma, much like in the previous proof, to factor out all elements of X and Y from our conditioning set.

Take the vertices b, c and some $y \in Y$. By construction,

$$b \perp c \mid V \backslash \{b, c\} \text{ and } b \perp y \mid V \backslash \{b, y\}$$

By the Block Independence Lemma,

$$\Rightarrow b \perp \{c, y\} \mid V \backslash \{b, c, y\}$$

And by the Reduction Lemma,

$$\Rightarrow b \perp c \mid V \backslash \{b, c, y\}$$

In the independence graph of $V \backslash \{y\}$, there are no edges connecting b to c.

We thus repeat this process for all $y \in Y$ until we have completely factored out Y. This is a finite process as the set of total vertices V is finite. We then apply a similar logic for the the vertices of X, at which point, we can eventually conclude that indeed,

$$B \perp C \mid A.$$

□

Theorem 1 (The Separation Theorem)**.** Let $\overrightarrow{X}$ be a vector of variables $x_1, x_2, ..., x_n$. If $\overrightarrow{X}_a, \overrightarrow{X}_b$ and $\overrightarrow{X}_c$ are vectors containing disjoint subsets of variables from X, and if, in the independence graph of $\overrightarrow{X}$, each vertex $x_a \in \overrightarrow{X}_a$ is separated from each $x_c \in \overrightarrow{X}_c$ by the vertices of $\overrightarrow{X}_b$, then,

$$\overrightarrow{X_a} \perp \overrightarrow{X_c} \mid \overrightarrow{X_b}$$

Graphically, this is depicted as follows:

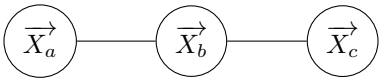


Figure 3.2: A graphical representation of the Separation Theorem

where each edge represents the totality of edges between the variables in each vector.

Proof. The Separation Theorem follows directly from Proposition 7, where we proved that for any two vertices $b \in B \subset V$, $c \in C \subset V$,where all paths connecting b to c contain at least one vertex $a \in A$, then, $b \perp c \mid a$. We can apply this to all $x_a \in \overrightarrow{X_a}$ and $x_c \in \overrightarrow{X_c}$ that are separated by the vertices of $\overrightarrow{X_b}$, and conclude more strongly that indeed, from the previous proposition, $\overrightarrow{X_a} \perp \overrightarrow{X_c} \mid \overrightarrow{X_b}$. □

To return to the first example CIG of this section in Figure 3.1, we can take $\overrightarrow{X_a} = \{1\}, \overrightarrow{X_b} = \{2\}, \overrightarrow{X_c} = \{3\}$, and say, $1 \perp 3 \mid 2$ without need to mention the node 4 as direct result of the Separation Theorem.

3.3 Directed Acyclic Independence Graphs

We will now modify our notation in order to capture information relating to causality, by creating directed acyclic graphs (DAGs). In probability theory, causality is the notion of one event being always followed by another. A causal relationship is one in which one event, which we call the effect, is always preceded by another, the cause. Here we say that the first event *causes* the effect.

To represent this relationship, we will turn our undirected edges into directed edges, representing the causality of one vertex by the one from which the edge originated. Let us take an example to illustrate this.

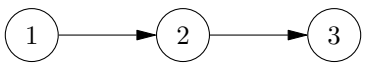


Figure 3.3: An Example DAG

In this example, we can see that the node X_1 causes X_2, and X_2 causes X_3. Probabilistically, this is equivalent to,

$$f_{X_1,X_2,X_3}(x_1, x_2, x_3) = f_{X_3|X_2}(x_3 \mid x_2) f_{X_2|X_1}(x_2 \mid x_1) f_{X_1}(x_1)$$

Clearly, the introduction of cycle to one of these graphs produces a circular causality,

$$f_{X_1,...,X_n}(x_1,...,x_n) = f_{X_n|X_{n-1}}(x_n \mid x_{n-1}) \dots f_{X_1|X_n}(x_1 \mid x_n)$$

that leads to some difficulty. We will thus limit ourselves to only graphs that are acycic.

A major interest of ours is to translate from a CIG to a DAG, and vice versa. Doing so allows us to, in one direction, build hypotheses on variable causality, and in the other, understand the broader independence relationships between variables.

3.3.1 Getting CIGs From DAGs

First, we look to obtain a CIG from a DAG, so that we may understand the dependence relationships between variables. In order to do so, we simply replace all directed edges with undirected edges. For example, take the following DAG, which we will call G_{DAG}, from which

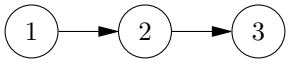


Figure 3.4: G_{DAG}

we can understand the following causality relationship between variables.

$$f_{1,2,3}(\cdot) = f_{3|2}(\cdot) f_{2|1}(\cdot) f_1(\cdot)$$

Replacing directed edges with undirected edges, we achieve the following CIG, called G_{CIG}, in Figure 3.5, from which we understand that $1 \perp 3 \mid 2$.

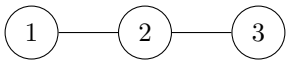


Figure 3.5: G_{CIG}

We wish to show that these two statements are equivalent, and that indeed these two graphs represent the same information. To begin with, we can write the probability densities for our CIG using equation (2.2).

$$1 \perp 3 \mid 2 \Rightarrow f_{1,3|2}(\cdot) = f_{3|2}(\cdot) f_{1|2}(\cdot)$$

And we wish to show that $G_{\text{DAG}} \Leftrightarrow G_{\text{CIG}}$.

$$\begin{aligned} f^{CIG}_{1,2,3}(\cdot) &= f_{1,3|2}(\cdot) f_2(\cdot) && \text{(Rewrite as conditional distributions)} \\ &= f_{1|2}(\cdot) f_{3|2}(\cdot) f_2(\cdot) && \text{(By independence)} \\ &= f^{DAG}_{1,2,3}(\cdot) && \text{(By definition)} \end{aligned}$$

□

In general, we cannot obtain a CIG from a DAG simply by replacing directed edges with undirected edges. An example of this, is DAGs that do **not** satisfy the Wermuth Condition.

Definition 3.3.1 (Wermuth Condition)**.** We say that a DAG satisfies the Wermuth Condition when it has **no** subgraphs where two nodes that are not neighbours both have edges pointing towards the same third node, as in Figure 3.6.

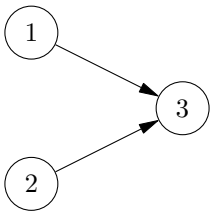


Figure 3.6: A graph that does **not** satisfy the Wermuth condition

Let us show that we indeed cannot find an associated CIG directly for this case, taking the DAG shown above which we will call $G^{\text{W'}}_{\text{DAG}}$. Its associated CIG is the following $G^{\text{W'}}_{\text{CIG}}$ seen in Figure 3.7, and is created by simply replacing directed edges with undirected ones.

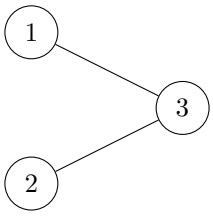


Figure 3.7: $G^{\text{W'}}_{\text{CIG}}$

Let us show that there is indeed a contradiction between $G^{\text{W'}}_{\text{DAG}}$ and $G^{\text{W'}}_{\text{CIG}}$, and that this CIG is not equivalent to its DAG. From $G^{\text{W'}}_{\text{DAG}}$, we understand that,

$$\begin{aligned} f^{\text{DAG}}_{1,2,3}(\cdot) &= f_{3|1,2}(\cdot) f_1(\cdot) f_2(\cdot) \\ &= f_{3|1}(\cdot) f_{3|2}(\cdot) f_1(\cdot) f_2(\cdot) \quad \text{(By independence)} \end{aligned}$$

And from our CIG, we see that $1 \perp 2 \mid 3$, from which we can infer that,

$$\begin{aligned} f^{\text{CIG}}_{1,2,3}(\cdot) &= f_{1,2|3}(\cdot) f_3(\cdot) \\ &= f_{1|3}(\cdot) f_{2|3}(\cdot) f_3(\cdot) \\ &= \frac{f_{3|1}(\cdot) f_1(\cdot)}{f_3(\cdot)} \frac{f_{3|2}(\cdot) f_2(\cdot)}{f_3(\cdot)} f_3(\cdot) && \text{(Bayes' Theorem)} \\ &= \frac{f_{3|1}(\cdot) f_{3|2}(\cdot) f_1(\cdot) f_2(\cdot)}{f_3(\cdot)^2} f_3(\cdot) \\ &= \frac{f^{\text{DAG}}_{1,2,3}(\cdot)}{f_3(\cdot)} \end{aligned}$$

And thus $f^{\text{DAG}}_{1,2,3}(\cdot) \neq f^{\text{CIG}}_{1,2,3}(\cdot)$. If we wish to eventually convert a DAG that does not satisfy the Wermuth condition into an equivalent CIG, we must *moralise* it.

Definition 3.3.2. Moralisation is the process of achieving the *moral* graph associated with the DAG, $G^{\succ} = (V, E^{\succ})$. This is the undirected graph build using the same vertex set, V, converting all directed edges in $E^{\succ}$ to undirected edges linking the same vertices, and then adding all undirected edges needed to satisfy the Wermuth condition for the new graph.

In the case of our DAG in Figure 3.6, the associated moral graph is the following graph seen in Figure 3.8. We call the process of adding an edge between the vertices 1 and 2 to link them, the *marrying* of the vertices.

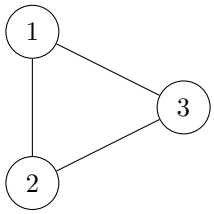


Figure 3.8: The moral graph associated with $G^{\text{W,}}_{\text{DAG}}$

Definition 3.3.3. A graph $G = (V, E)$ is said to be bipartite if its vertices can be partitioned into two non-empty subsets V_1, V_2 such that $V_1 \cup V_2 = V, V_1 \cap V_2 = \varnothing$, and there are no edges connecting two vertices in V_1, and no edges connecting two vertices in V_2. An example is given in Figure 3.9.

Definition 3.3.4. A graph $G = (V, E)$ is said to be k-partite if its vertices can be partitioned into k non-empty subsets $V_1, V_2, ..., V_n$ such that $V_1 \cup V_2 \cup \cdots \cup V_n = V$, $V_i \cap V_j = \varnothing, i \neq j$, and there are no edges connecting two vertices in any V_i.

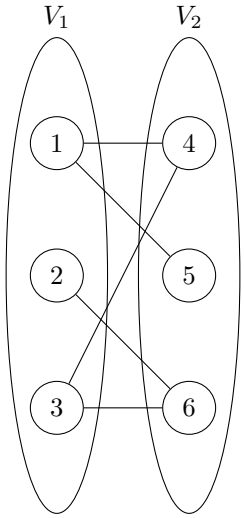


Figure 3.9: A bipartite graph with non-empty subsets V_1 and V_2

The following proposition is, as far as I can tell, a novel interpretation of the Wermuth condition.

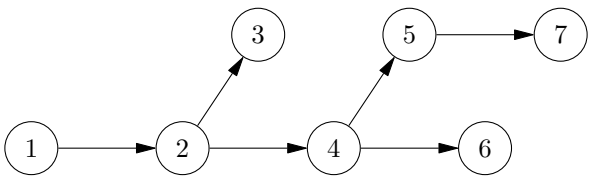
Proposition 8. For any DAG that satisfies the Wermuth condition, its associated CIG is bipartite.

Proof. By construction, we have no cycles in our DAGs and because they satisfy the Wermuth condition, we have no obligation to moralise them. Their equivalent CIGs are thus trees. Taking an arbitrary root node in the tree, v_0, there is only one path from v_0 to all other vertices in the tree. Thus, by colouring all nodes an even distance from v_0 and v_0 itself blue, and colouring all nodes an odd distance from it red, adjacent vertices have different colours.

Assume that this is not the case, then there is exists more than one path from our root vertex to any other vertex, which in turn implies the existence of a cycle in our tree.

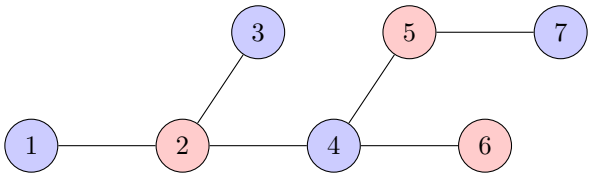
We have thus found a contradiction, and indeed the vertices of the graph can be coloured by two colours such that no two adjacent vertices have the same colour. This is equivalent to saying our graph is bipartite, and we conclude our proof. □

Let us illustrate this result. Take the following DAG.

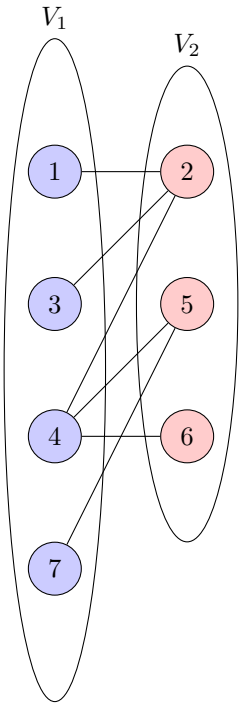


As we can see, there are no cycles by definition, and the Wermuth condition is satisfied

since there are no two nodes both directed towards the same third node. We may thus find its associated CIG directly by replacing all arrows with undirected edges as follows. We choose the node 1 to be our root node and colour it blue. Colouring all nodes an even distance from it blue, and all nodes an odd distance red, we see that no two adjacent nodes are coloured the same.



This graph can in turn be partitioned into two non-empty subsets V_1 and V_2 such that no two vertices in each subset is connected with an edge. This partitioning is as follows.



We find the graph is indeed bipartite, and see it here illustrated in terms of its non-empty subset partitioning.

3.3.2 Getting DAGs From CIGs

We will now look at how we might achieve a DAG from a CIG, a process known as *demoralisation* where we look to convert non-directed edges to directed edges. It is clear that when converting non-directed edges to directed edges, we can achieve many different directed graphs,

for example by reversing the direction of a directed edge. Indeed, there is no unique solution to this problem. In fact, when we demoralise a CIG to achieve a DAG, we are building a set of hypotheses of possible configurations to later test.

For example, take the moralised CIG in Figure 3.8. There are three edges, which can each take two directions of flow, meaning we have 2^3 possible configurations of our demoralised DAG. In general, where n is the number of edges in our CIG, there are 2^n possible configurations of its associated demoralised DAG. In our example, we have the following configurations seen in Figure 3.10.

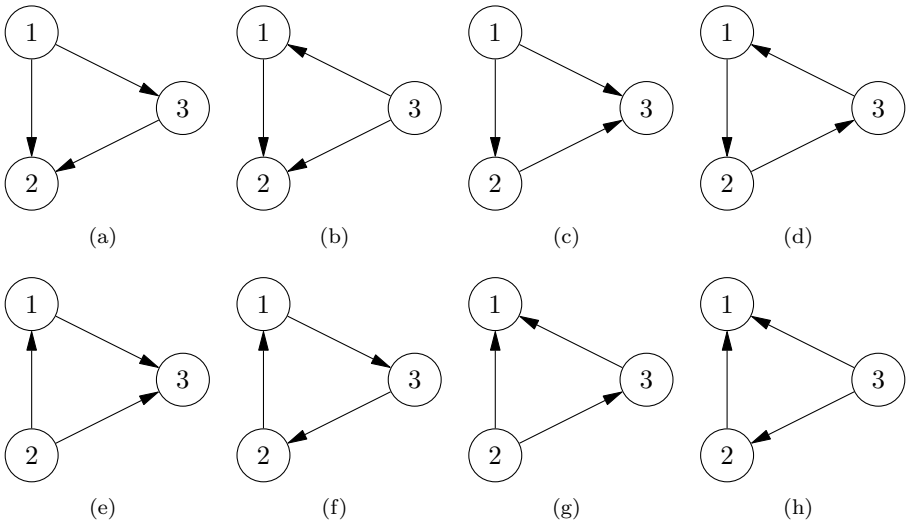
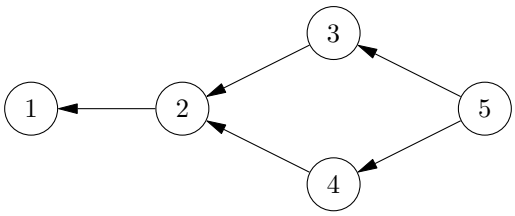


Figure 3.10: Demoralised DAGs associated with the CIG in Figure 3.8.

We can see that the graph in Figure 3.10(d) and 3.10(f) are cycles of length three. We thus remove them from consideration.

This leaves us with six choices from which to choose the most appropriate. In the case of a Gaussian DAG, the joint distribution of the model is the product of its component marginal distributions. For example, taken the following Gaussian DAG.



The joint distribution of this model can be rewritten as

$$f_{x_1,x_2,x_3,x_4,x_5} = f_{x_1|x_2} f_{x_2|\{x_3,x_4\}} f_{x_3|x_5} f_{x_4|x_5} f_{x_5}.$$

Each of these marginal distributions $f_{x_i|\overrightarrow{X}_c}$ can be thought of as a regression of x_i on $\overrightarrow{X}_c$ with uncorrelated error terms, and the error terms of each regression are mutually uncorrelated. The divergence of a DAG model (-2 log-likelihood) can be computed as the sum of the divergence of the marginal distributions (regressions). In our example above, this becomes

$$\text{Div} = \ln S_1 + \ln S_2 + \ln S_3 + \ln S_4 + \ln S_5$$

where S_1 is the residual sum of squares from the regression of x_1 on x_2, and so on. For S_5, this becomes the residual sum of squares of the regression of x_5 on no regressors, or raw sum of squares of the observations of x_5. We may choose the the DAG with the lowest divergence (equivalently, the highest likelihood).

When this isn't possible, or there are multiple models with similar, low divergence, we have three solutions. First, we are able to adjust the threshold for the partial correlation significance test, although this leads to tradeoff between type 1 and type 2 errors. Second, we can choose the most explicative model. This is the model with the most valuable links. Last, we can choose the most parsimonious model. To do this, we remove the least valuable links, as calculated by t value of the regressors in the margin regressions, and look to obtain a model with fewer parameters without incurring a significant loss is likelihood.

We will favour a parsimonious model over an explicative one provided there is not a significant increase in divergence.

Once have chosen the most appropriate model, we must remove those edges which are present due to moralisation. For example, should we choose the graph in Figure 3.10(a), then we would have to establish whether the edge connecting vertices 1 and 3 is present due to moralisation, and if so then remove it.

It is worth noting that there do exist some CIGs from whom it is impossible to derive a DAG, although we will not dwell on these cases too much longer given the scope of this work.

3.4 Gaussian CIGs

The Gaussian (or Normal) distribution is unique among continuous distributions with support from negative infinity to positive infinity due to its Maximum Entropy Characterisation. This is the characteristic that among all continuous distributions with support $(-\infty, +\infty)$, the Gaussian distribution has the highest differential entropy. Thus, if we have data with little prior knowledge, then the normal distribution is a sensible assumption.

We thus move to assert this to our case. Assume we have a data matrix, $\mathbf{X}$, of n samples that is multivariate normally distributed, in dimension d, such that the distribution of each observation can be denoted,

$$\overrightarrow{x} \sim \mathrm{N}_d(\overrightarrow{\mu}, \mathbf{\Sigma})$$

Where $\overrightarrow{\mu}$ is the vector or expectations for each component of $\overrightarrow{x}$,

$$\overrightarrow{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{pmatrix} = \begin{pmatrix} \mathbb{E}(x_1) \\ \mathbb{E}(x_2) \\ \vdots \\ \mathbb{E}(x_n) \end{pmatrix}$$

and $\mathbf{\Sigma}$ is the $n \times d$ covariance matrix, such that,

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{x_1,x_1} & \sigma_{x_1,x_2} & \dots & \sigma_{x_1,x_d} \\ \sigma_{x_2,x_1} & \sigma_{x_2,x_2} & \dots & \sigma_{x_2,x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{x_n,x_1} & \sigma_{x_n,x_2} & \dots & \sigma_{x_n,x_d} \end{pmatrix}$$

Relating this back to CIGs, we are able to identify partially independent componenents of $\overrightarrow{x}$ by calculating the partial correlation matrix, and asserting that two components are independent if and only if their partial correlation is zero. Formally, this is given by,

$$x_i \perp x_j \Leftrightarrow \mathrm{Cor}(x_i, x_j \mid \overrightarrow{x_c}) = 0$$

where $\overrightarrow{x_c} = \overrightarrow{x} \backslash \{x_i, x_j\}$. These partial correlations can be calculated directly from the covariance matrix $\mathbf{\Sigma}$ as a result of the Inverse Variance Lemma.

Theorem 2 (Inverse Variance Lemma)**.** Let $\overrightarrow{x}$ be a normally distributed random vector in d

dimensions, such that,

$$\overrightarrow{x} \sim \mathrm{N}_d(\overrightarrow{\mu}, \mathbf{\Sigma})$$

then we can define the *Precision* (or *Concentration*) matrix of $\overrightarrow{x}$, $\mathbf{\Omega}$, as follows.

$$\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$$

From this, we calculate the pairwise partial correlations of components x_i, x_j via,

$$\rho_{i,j|\mathrm{rest}} = -\frac{\omega_{i,j}}{\sqrt{\omega_{i,i}\omega_{j,j}}}$$

where $\omega_{i,j} \in \mathbf{\Omega}$. Note that this is only defined when $\mathbf{\Sigma}$ is invertible.

Proof. The following proof is adapted from Reale (1998).

We begin by defining the covariance between the two components x_i and x_j,

$$\begin{aligned}\mathrm{Cov}(x_i, x_j \mid \overrightarrow{z}) &= \mathrm{Var}(\overrightarrow{e}) \\ &= \mathbf{\Sigma}_{YY} - \mathbf{\Sigma}_{ZY}\mathbf{\Sigma}_{ZZ}^{-1}\mathbf{\Sigma}_{YZ} = \mathbf{U}\end{aligned}$$

from which we can derive the partial correlation between them via,

$$\mathrm{Cor}(x_i, x_j \mid \overrightarrow{z}) = \frac{u_{i,j}}{\sqrt{u_{i,i}, u_{j,j}}}$$

now let,

$$\mathbf{\Omega} = \mathbf{\Sigma}^{-1} = \begin{pmatrix} \mathbf{\Omega}_{YY} & \mathbf{\Omega}_{YZ} \\ \mathbf{\Omega}_{ZY} & \mathbf{\Omega}_{ZZ} \end{pmatrix}$$

then we want to show that $\mathbf{U}^{-1} = \mathbf{\Omega}_{YY}$. By construction, we have that,

$$\mathbf{\Omega} = \mathbf{\Sigma}^{-1} \Rightarrow \mathbf{\Sigma}\mathbf{\Omega} = \mathbf{I}_d$$

where $\mathbf{I}_d$ is the $d \times d$ identity matrix. We can rewrite this as,

$$\begin{pmatrix} \mathbf{\Sigma}_{YY} & \mathbf{\Sigma}_{YZ} \\ \mathbf{\Sigma}_{ZY} & \mathbf{\Sigma}_{ZZ} \end{pmatrix} \begin{pmatrix} \mathbf{\Omega}_{YY} & \mathbf{\Omega}_{YZ} \\ \mathbf{\Omega}_{ZY} & \mathbf{\Omega}_{ZZ} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d-2} \end{pmatrix}$$

from which we can read,

$$\mathbf{\Sigma}_{YY}\mathbf{\Omega}_{YY} + \mathbf{\Sigma}_{YZ}\mathbf{\Omega}_{ZY} = \mathbf{I}_2 \tag{3.1}$$

$$\mathbf{\Sigma}_{YZ}\mathbf{\Omega}_{YY} + \mathbf{\Sigma}_{ZZ}\mathbf{\Omega}_{ZY} = \mathbf{0} \tag{3.2}$$

From equation (3.2),

$$\mathbf{\Omega}_{ZY} = -\mathbf{\Sigma}_{ZZ}^{-1}\mathbf{\Omega}_{YY}\mathbf{\Omega}_{ZY} \tag{3.3}$$

Substituting equation (3.3) into (3.1), we achieve,

$$\begin{aligned}
&\mathbf{\Sigma}_{YY}\mathbf{\Omega}_{YY} - \mathbf{\Sigma}_{YZ}\mathbf{\Sigma}_{ZZ}^{-1}\mathbf{\Sigma}_{ZY}\mathbf{\Omega}_{YY} = \mathbf{I}_2 \\
&\Rightarrow \underbrace{(\mathbf{\Sigma}_{YY} - \mathbf{\Sigma}_{YZ}\mathbf{\Sigma}_{ZZ}^{-1}\mathbf{\Sigma}_{ZY})}_{\mathbf{U}}\mathbf{\Omega}_{YY} = \mathbf{I}_2 \\
&\qquad\qquad\qquad\qquad \Rightarrow \mathbf{U} = \mathbf{\Omega}_{YY}^{-1}
\end{aligned}$$

So we have successfully shown that indeed $\text{Var}(\overrightarrow{e}) = \mathbf{U} = \mathbf{\Omega}_{YY}^{-1}$ as desired. We now endeavour to show that element-wise,

$$\frac{u_{i,j}}{\sqrt{u_{i,i}, u_{j,j}}} = -\frac{\omega_{i,j}}{\sqrt{\omega_{i,i}, \omega_{j,j}}}$$

where $u_{i,j} \in \mathbf{U}$ and $\omega_{i,j} \in \mathbf{\Omega}$. Let,

$$\mathbf{U} = \begin{pmatrix} \sigma_i^2 & \rho\sigma_j\sigma_i \\ -\rho\sigma_i\sigma_j & \sigma_j^2 \end{pmatrix}$$

so that we have by construction,

$$\rho = \frac{u_{i,j}}{\sqrt{u_{i,i}u_{j,j}}}$$

From this, we may derive,

$$\begin{aligned}
\mathbf{\Omega}_{YY} = \mathbf{U}^{-1} &= \underbrace{\frac{1}{\sigma_i^2\sigma_j^2(1-\rho^2)}}_{\lambda}\begin{pmatrix} \sigma_j^2 & -\rho\sigma_j^2\sigma_i^2 \\ -\rho\sigma_i^2\sigma_j^2 & \sigma_i^2 \end{pmatrix} \\
\Rightarrow \omega_{i,i} = \frac{\sigma_j^2}{\lambda} \quad &, \quad \omega_{j,j} = \frac{\sigma_i^2}{\lambda} \quad , \quad \omega_{i,j} = -\frac{\rho\sigma_i\sigma_j}{\lambda}
\end{aligned}$$

And finally, we conclude that,

$$-\frac{\omega_{i,j}}{\sqrt{\omega_{i,i}\omega_{j,j}}} = -\frac{-\frac{\rho\sigma_i\sigma_j}{\lambda}}{\sqrt{\frac{\sigma_j^2}{\lambda}\frac{\sigma_i^2}{\lambda}}} = \rho = \frac{u_{i,j}}{\sqrt{u_{i,i}, u_{j,j}}}$$

□

3.4.1 Testing Conditional Independence Significance

Say we have n observations on some random variable, $\overrightarrow{X}$, which we assume to be normally distributed in d dimensions,

$$\overrightarrow{X} \sim \mathrm{N}_d(\overrightarrow{\mu}, \mathbf{\Sigma})$$

then from this data matrix, we can estimate the covariance matrix, $\hat{\mathbf{\Sigma}}$, and in turn use this to compute $\hat{\mathbf{\Omega}}$, and thus the sample's partial correlations as per the Inverse Variance Lemma. Under the assumptions, partial correlation and conditional dependence are equivalent. That is to say that if two variables have a partial correlation significantly different to zero, then they are said to be dependant given the rest of the data. These variables will eventually make up the vertex set of a CIG, and the edge set will be defined by these significant partial correlations.

Say we have a random data vector $\overrightarrow{X} = (x_1, ..., x_n)^T$, and we wish to test all partial correlations between each $x_i, x_j \in \overrightarrow{X}$. We are able to find a calculate the t-statistics $t^2_{i,j}$ of these partial correlations $\rho_{x_i,x_j|\overrightarrow{X}_c}$ testing their statistical significance, and find the link

$$\rho^2_{x_i,x_j|\overrightarrow{X}_c} = \frac{t^2_{i,j}}{t^2_{i,j} + \upsilon}$$

where $\upsilon = n - k$ are the residual degrees of freedom in the model. From this relationship, we are able to compute the critical values of ρ and t.

Once we have identified the appropriate CIG from this significance test, we will be able to build and test its associated models.

Chapter 4

Time Series Analysis

In this chapter, we define some basic concepts relating to stochastic processes, before introducing vector autoregressive models as a tool for modelling stationary stochastic processes. We will then discuss the different types of vector autoregressive models and how they relate to each other, before demonstrating how they can be linked back to graphical models. We end the chapter with a worked example of the methodology developed on simulated data.

4.1 Stochastic Processes

We make use of definitions provided in Reale (1998) for this section.

Definition 4.1.1. A stochastic process is a set of random variables indexed by time t, $\{X_t\} = \{X_1, X_2, X_3, ...\}$, where the set of possible values of each X_t is called the state space, denoted S. S may be continuous, discrete, univariate, or multivariate.

We use the following notation to refer to stochastic processes.

$\{X_t\}$	The whole stochastic process
$\{x_t\}$	A sample function (possible outcome of $\{X_t\}$)
X_t	The random variable defined at time t
x_t	A sample value at t

Example 4.1.1 (Markov Chain)**.** A stochastic process is called a Markov Chain if

$$\mathbb{P}(X_t = x_t \mid X_0 = x_0, X_1 = x_1, ..., X_t - 1 = x_t - 1) = \mathbb{P}(X_t = x_t \mid X_t - 1 = x_t - 1)$$

for all $x_0, x_1, ..., x_t \in S$.

Example 4.1.2 (Martingales)**.** A martingale is a stochastic process with the property

$$\mathbb{E}[X_{t+1} \mid X_0 = x_0, X_1 = x_1, ..., X_t = x_t] = x_t$$

for all t. A martingale difference is a martingale with

$$\mathbb{E}[X_{t+1} \mid X_0 = x_0, X_1 = x_1, ..., X_t = x_t] = 0$$

for all t.

A stochastic process is completely determined if the joint distributions of any finite selection of the random variables are known. This is not generally the case. In order to determine these processes more generally, we make two assumptions: that the process is stationary and Gaussian.

Definition 4.1.2 (Stationarity)**.** A stochastic process is called stationary if its probabilistic structure remains constant with time. We call it strongly stationary if the collections of random variables

$$X_k, X_{k+1}, ..., X_{k+n} \quad \text{and} \quad X_{k+h}, X_{k+1+h}, ..., X_{k+n+h}$$

have the same joint distribution funciton for all k, n, and for $h > 0$. The process is called weakly stationary if

$$\mathbb{E}[X_t] = \mu_x \quad \text{and} \quad \text{Var}(X_t) = \sigma_x^2 \quad \text{for all } t$$

and the covariance between $\{X_t\}$ and $\{X_{t+k}\}$ depends only on the lag, in other words

$$\text{Cov}(X_t, X_{t+k}) = \mathbb{E}[(X_t - \mu_x)(X_{t+k} - \mu_x)]$$

for all t and for $k \in \mathbb{Z}$. Strong stationarity implies weak stationarity if the second order moment exists, although the opposite is not true in general.

Example 4.1.3 (White Noise)**.** This is a purely random process $\{\varepsilon_t\}$, where all random variables in the process have zero mean, and constant and uncorrelated variances. We thus have

$$\mathbb{E}[\varepsilon_t] = 0 \quad \text{for all } t$$

Example 4.1.4 (Random Walk)**.** A random walk, given by $X_t = X_{t-1} + \omega_t$, where $\omega_1, \omega_2, ..., \omega_t$ are independent and identically distributed with $\mathbb{E}[\omega_i] = \mu_\omega$ and $\text{Var}(\omega_i) = \sigma_\omega^2$, is a non-

stationary process. We can show this by expanding the process as follows.

$$\begin{aligned} X_t &= X_{t-1} + \omega_t \\ &= (X_{t-2} + \omega_{t-1}) + \omega_t \\ &= x_0 + \sum_{k=1}^{t} \omega_k \end{aligned}$$

Where thus

$$\mathbb{E}[X_t] = x_0 + t\mu_\omega \quad \text{and} \quad \text{Var}(X_t) = t\sigma_\omega^2$$

are both in terms of t, and thus variable with time.

Definition 4.1.3 (Gaussianity)**.** A stochastic process $\{X_t\}$ is called Gaussian if for all sets of k random variance in the process, their joint density function is Gaussian, that is

$$f(X_{t1}, X_{t2}, ..., X_{tk}) = (2\pi)^{-\frac{k}{2}} \mid \mathbf{\Sigma} \mid^{-\frac{1}{2}} \exp(-\frac{1}{2}(\overrightarrow{x} - \overrightarrow{\mu})^T \mathbf{\Sigma}^{-1}(\overrightarrow{x} - \overrightarrow{\mu}))$$

where $\overrightarrow{x}$ is a vector of state values, $\overrightarrow{\mu}$ the vector of means, and $\mathbf{\Sigma}$ the covariance matrix of the random variables.

A Gaussian process is determined if $\overrightarrow{\mu}$ and $\mathbf{\Sigma}$ are known to us. If the process is weakly stationary then $\overrightarrow{\mu}$ remains constant across time, and thus we can determine that for Gaussian processes, weak stationarity implies strong stationarity.

Definition 4.1.4 (Autocovariance Function)**.** For a stochastic process $\{X_t\}$, the autocovariance function is the covariance between variables at different times, formally given by

$$R(k) = \text{Cov}(X_t, X_{t-k}).$$

This will be used in the autocorrelation function to follow.

Definition 4.1.5 (Autocorrelation Function)**.** The autocorrelation function (ACF) of a stochastic process $\{X_t\}$ is the correlation between two variables at different times, formally given by

$$\begin{aligned} r(k) &= \text{Corr}(X_t, X_{t-k}) \\ &= \frac{\text{Cov}(X_t, X_{t-k})}{\sqrt{\text{Var}(X_t)\text{Var}(X_{t-k})}} \\ &= \frac{R(k)}{\sqrt{\text{Var}(X_t)\text{Var}(X_{t-k})}} \end{aligned}$$

Considering stationary stochastic processes, we have that $\text{Var}(X_t) = \text{Var}(X_{t-k})$ for all t, k and thus,

$$r(k) = \frac{R(k)}{R(0)}$$

for all k.

Properties of the Autocorrelation Function:

1. $r(0) = 1$
2. $R(k) = R(-k)$
3. $r(k) = r(-k)$
4. $\mid r(k) \mid \leq 1$
5. The ACF is invariant to linear transformations of the stochastic process, in other words,

$$r_X(k) = r_{a+bX}(k)$$

The following definition is taken from Romer (2020).

Definition 4.1.6 (Partial Autocorrelation)**.** Partial Autocorrelation is a measure of conditional correlation between the values of a stochastic process at times t and $t-k$, given the values of the process at all other values of t. Formally, this is expressed as

$$\frac{\text{Cov}(X_t, X_{t-k} \mid \overrightarrow{X_c})}{\sqrt{\text{Var}(X_t \mid \overrightarrow{X_c})\text{Var}(X_{t-k} \mid \overrightarrow{X_c})}}$$

where $\overrightarrow{X_c} = X_{t-1}, X_{t-2}, ..., X_{t-k+1}$.

4.2 Autoregressive Models

A univariate autoregressive model (herein referred to as an AR model) of order p is a stochastic process of the form

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t$$

where ε_t is uncorrelated with $X_{t-1}, X_{t-2}, ...$ and with $\varepsilon_{t-1}, \varepsilon_{t-2}, ...$.

Example 4.2.1 (AR(1))**.**

$$X_t = \phi X_{t-1} + \varepsilon_t$$

To ensure that we have finite first and second order moments, and thus satisfy the stationarity condition, the roots of the polynomial in ϕ must lie outwith the unit circle.

Example 4.2.2. Take the AR(1) model once more and rewrite it, expanding each Z_i term. Then we achieve,

$$\begin{aligned} X_t &= \phi X_{t-1} + \varepsilon_t \\ &= \phi(\phi X_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2 X_{t-2} \\ &= \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \cdots + \phi^p\varepsilon_{t-p} + \phi^{p+1}X_{t-p-1} \end{aligned}$$

where

$$\mu_X = \phi^{p+1}X_{t-p-1} \quad \text{and} \quad \sigma_X^2 = \sum_{i=0}^{p} \phi^{2i}\sigma_\varepsilon^2 + \phi^{p+1}\sigma_{X(t-p-1)}.$$

For stationarity, we need these to be finite, and thus require that $|\phi| < 1$.

Vector autoregressive models (VAR models) relate each variable at time t to their values at previous time periods up to a certain point. We refer to variables that occur further back in time as *anterior variables* over the course of this paper. For example, x_{t-1} is anterior to x_t. This look-back period is called the lag, and is denoted p. There are two types of VAR models: canonical and structural. The structural VAR models makes assertions about the causal relationship between contemporaneous variables, while the canonical do not.

4.2.1 Canonical VAR Models

The canonical VAR models takes the form

$$\overrightarrow{X}_t = \mathbf{\Phi}_1\overrightarrow{X}_{t-1} + \mathbf{\Phi}_2\overrightarrow{X}_{t-2} + \cdots + \mathbf{\Phi}_p\overrightarrow{X}_{t-p} + \overrightarrow{E}_t \tag{4.1}$$

which maybe rearranged for $\overrightarrow{E}_t$ using the polynomial $\mathbf{\Phi}(z), z \in \mathbb{C}$ as

$$\begin{aligned} \mathbf{\Phi}(z)\overrightarrow{X}_t &= \overrightarrow{E}_t \\ \Rightarrow \overrightarrow{X}_t &= \mathbf{\Phi}(z)^{-1}\overrightarrow{E}_t. \end{aligned}$$

In order for the process described in (4.1) to be a stationary process, we need $\mathbf{\Phi}(z)^{-1}$ to converge for $|z| \leq 1$. Or equivalently, $\det(\mathbf{\Phi}(z))$ not equal to zero for $|z| \leq 1$.

Example 4.2.3. The canonical VAR(2) process is given by

$$\overrightarrow{X}_t = \mathbf{\Phi}_1 \overrightarrow{X}_{t-1} + \mathbf{\Phi}_2 \overrightarrow{X}_{t-2} + \overrightarrow{E}_t$$

or more verbosely,

$$\begin{pmatrix} X_{1,t} \\ X_{2,t} \end{pmatrix} = \begin{pmatrix} \phi_{1,1,1} & \phi_{1,2,1} \\ \phi_{2,1,1} & \phi_{2,2,1} \end{pmatrix} \begin{pmatrix} X_{1,t-1} \\ X_{2,t-1} \end{pmatrix} + \begin{pmatrix} \phi_{1,1,2} & \phi_{1,2,2} \\ \phi_{2,1,2} & \phi_{2,2,2} \end{pmatrix} \begin{pmatrix} X_{1,t-2} \\ X_{2,t-2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix}.$$

4.2.2 Structural VAR Models

These models account for dependencies between the elements of contemporaneous variables, that is variables for the same time index, and are of the form

$$\mathbf{\Phi}_0^* \overrightarrow{X}_t = \mathbf{\Phi}_1^* \overrightarrow{X}_{t-1} + \mathbf{\Phi}_2^* \overrightarrow{X}_{t-2} + \cdots + \mathbf{\Phi}_p^* \overrightarrow{X}_{p-1} + \overrightarrow{E}_t^*$$

where $\mathbf{\Phi}_0^*$ has some non-zero off-diagonal elements.

A structural VAR model is borne from some hypothesis we have about the relationship between variables in our model. The elements of $\overrightarrow{E}_t^*$ are uncorrelated both among themselves and with the elements of $\overrightarrow{X}_t$. In order for our structural model to be uniquely defined, some elements of $\mathbf{\Phi}_0^*$ must be zero. The covariance matrix of our structural model is diagonal, whereas in a canonical model it may take any which form.

Note that the canonical VAR model is a special case of the structural, where $\mathbf{\Phi}_0^* = \mathbf{I}$, and there are no constraints on $\overrightarrow{E}_t^*$.

We may wish to convert a structural VAR to a canonical VAR or vice versa, since the canonical VAR model is uniquely parametrised, and thus makes prediction easier. Equally, the structural VAR is expected to be more parsimonious, and as a result, easier to interpret.

Lemma 1. Any structural VAR can be transformed into a canonical VAR by multiplying through by $(\mathbf{\Phi}_0^*)^{-1}$ as follows

$$\begin{aligned} \overrightarrow{X}_t &= (\mathbf{\Phi}_0^*)^{-1} \mathbf{\Phi}_1^* \overrightarrow{X}_{t-1} + (\mathbf{\Phi}_0^*)^{-1} \mathbf{\Phi}_2^* \overrightarrow{X}_{t-2} + \cdots + (\mathbf{\Phi}_0^*)^{-1} \mathbf{\Phi}_p^* \overrightarrow{X}_{p-1} + (\mathbf{\Phi}_0^*)^{-1} \overrightarrow{E}_t^* \\ &= \mathbf{\Phi}_1 \overrightarrow{X}_{t-1} + \mathbf{\Phi}_2 \overrightarrow{X}_{t-2} + \cdots + \mathbf{\Phi}_p \overrightarrow{X}_{t-p} + \overrightarrow{E}_t. \end{aligned}$$

It is worth noting that the same canonical VAR models can be attained from multiple

different structural VAR models. In order to go from a canonical VAR to a structural in that case, we require prior information about our model.

Lemma 2. For any given canonical VAR, there exist multiple associated structural VARs. Let us decompose $\text{Var}(\overrightarrow{E}_t) = \mathbf{\Sigma}_\varepsilon$ as

$$\mathbf{\Sigma}_\varepsilon = \mathbf{TDT}^T$$

where $\mathbf{T}$ is triangular, $\mathbf{D}$ is diagonal, and the two are not unique. Let $\mathbf{\Phi}_0^* = \mathbf{T}^{-1}$, and multiply our canonical VAR through by this, yielding a structural VAR with

$$\mathbf{\Phi}_j^* = \mathbf{\Phi}_0^* \mathbf{\Phi}_j \quad \text{and} \quad \overrightarrow{E}_t^* = \mathbf{\Phi}_0^* \overrightarrow{E}_t$$

where, by construction, $\mathbf{\Sigma}_\varepsilon = \mathbf{D}$ so that the components of $\mathbf{\Sigma}_\varepsilon$ are uncorrelated.

Example 4.2.4. We will provide a worked example where we will begin with a structural VAR model, convert this to its canonical equivalent, and then make the reverse transformation only to achieve a different structural VAR model. Take the following system of equations

$$\begin{cases} x_t = 0.8x_{t-1} + \varepsilon_{xt} \\ y_t = 0.5y_{t-1} + 0.2x_t + \varepsilon_{yt} \end{cases}$$

where $\varepsilon_x, \varepsilon_y \sim N(0,1)$ and are independent of each other. By construction, the covariance of our errors $\mathbf{\Sigma}_\varepsilon = \mathbf{I}$.

We rearrange this system to obtain a structural VAR representation as follows.

$$\begin{cases} x_t = 0.8x_{t-1} + \varepsilon_{xt} \\ y_t - 0.2x_{t-1} = 0.5y_{t-1} + \varepsilon_{yt} \end{cases}$$
$$\Rightarrow \begin{pmatrix} 1 & 0 \\ -0.2 & 1 \end{pmatrix} \begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} 0.8 & 0 \\ 0 & 0.5 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{xt} \\ \varepsilon_{yt} \end{pmatrix}$$

By multiplying through by $(\mathbf{\Phi}_0^*)^{-1} = \begin{pmatrix} 1 & 0 \\ -0.2 & 1 \end{pmatrix}^{-1}$, we achieve the associated canonical

VAR,

$$\begin{aligned}\begin{pmatrix} x_t \\ y_t \end{pmatrix} &= \begin{pmatrix} 1 & 0 \\ -0.2 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0.8 & 0 \\ 0 & 0.5 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ -0.2 & 1 \end{pmatrix}^{-1} \begin{pmatrix} \varepsilon_{xt} \\ \varepsilon_{yt} \end{pmatrix} \\ &= \begin{pmatrix} 0.8 & 0 \\ 0.16 & 0.5 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0.2 & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_{xt} \\ \varepsilon_{yt} \end{pmatrix} \end{aligned} \tag{††}$$

where the covariance of the errors is now

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0.2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0.2 & 1 \end{pmatrix}^T = \begin{pmatrix} 1. & 0.2 \\ 0.2 & 1.04 \end{pmatrix}.$$

In order to transform this canonical VAR back to a structural representation, we first decompose the covariance of its errors.

$$\underbrace{\begin{pmatrix} 1 & 0.2 \\ 0.2 & 1.04 \end{pmatrix}}_{\mathbf{\Sigma}} = \underbrace{\begin{pmatrix} 1 & \frac{0.2}{1.04} \\ 0 & 1 \end{pmatrix}}_{\mathbf{T}} \underbrace{\begin{pmatrix} \frac{1}{1.04} & 0 \\ 0 & 1.04 \end{pmatrix}}_{\mathbf{D}} \underbrace{\begin{pmatrix} 1 & 0 \\ \frac{0.2}{1.04} & 1 \end{pmatrix}}_{\mathbf{T}^T}$$

And we thus find directly that

$$\mathbf{\Phi}_0^* = \mathbf{T}^{-1} = \begin{pmatrix} 1 & -\frac{0.2}{1.04} \\ 0 & 1 \end{pmatrix}$$

and by multiplying our canonical VAR model given in (††) through by this, we achieve a second associated structural VAR model.

$$\begin{aligned}\begin{pmatrix} 1 & -\frac{0.2}{1.04} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_t \\ y_t \end{pmatrix} &= \begin{pmatrix} 1 & -\frac{0.2}{1.04} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0.8 & 0 \\ 0.16 & 0.5 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} 1 & -\frac{0.2}{1.04} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0.2 & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_{xt} \\ \varepsilon_{yt} \end{pmatrix} \\ &= \begin{pmatrix} 0.769 & -0.1 \\ 0.16 & 0.5 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} 0.962 & -0.192 \\ 0.2 & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_{xt} \\ \varepsilon_{yt} \end{pmatrix} \end{aligned}$$

Note that this structural VAR associated with the canonical VAR in (††) is different to our initial structural VAR. This example shows that a canonical VAR model encapsulates the information of all of its associated structural models, and that many structural models can be

used to obtain the same canonical model.

Within structural VAR models, there are two subclasses: causal models, and simultaneous equations models. When it is possible to order our system of equations such that $\mathbf{\Phi}_0^*$ is triangular, we say that our structural VAR is causal, or recursive.

Example 4.2.5. Consider the system of equations

$$\begin{cases} x_t = \alpha_1 x_{t-1} + \alpha_2 y_t + \alpha_3 + u_t \\ y_t = \beta_1 y_{t-1} + \beta_2 x_{t-1} + v_t \\ z_t = \gamma_1 z_{t-1} + \gamma_2 y_t + \gamma_3 x_{t-1} + e_t \end{cases}$$

where $\{u_t, v_t, e_t\}$ are uncorrelated errors. We may reorder this system as

$$\begin{cases} y_t = \beta_1 y_{t-1} + \beta_2 x_{t-1} + v_t \\ z_t = \gamma_1 z_{t-1} + \gamma_2 y_t + \gamma_3 x_{t-1} + e_t \\ x_t = \alpha_1 x_{t-1} + \alpha_2 y_t + \alpha_3 z_t + u_t \end{cases}$$

or in matrix form,

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ -\gamma_2 & 1 & 0 \\ \alpha_2 & \alpha_3 & 1 \end{pmatrix}}_{\mathbf{\Phi_0^*}} \underbrace{\begin{pmatrix} y_t \\ z_t \\ x_t \end{pmatrix}}_{\vec{X}_t} = \underbrace{\begin{pmatrix} \beta_1 & 0 & \beta_2 \\ 0 & \gamma_1 & \gamma_3 \\ 0 & 0 & \alpha_1 \end{pmatrix}}_{\mathbf{\Phi_1^*}} \underbrace{\begin{pmatrix} y_{t-1} \\ z_{t-1} \\ x_{t-1} \end{pmatrix}}_{\vec{X}_{t-1}} + \underbrace{\begin{pmatrix} v_t \\ e_t \\ u_t \end{pmatrix}}_{\vec{E}_t}$$

The fact that $\mathbf{\Phi_0^*}$ is lower triangular suggests that each variable is effected only by antierior variables, that is those occuring before, and so we can deduce a causal path between them and represent this in a DAG.

When it is not possible to achieve a triangular $\mathbf{\Phi_0^*}$, we have a simultaneous equations model, and may not need a DAG to represent the causality between variables.

4.3 Graphical Modelling for Time Series Analysis

We now look to represent our VAR models as DAGs, taking a set of values from our stationary time series $\{X_t, X_{t-1}, \ldots, X_{t-k}\}$ as variables in our graphical models. Under Gaussianity, we

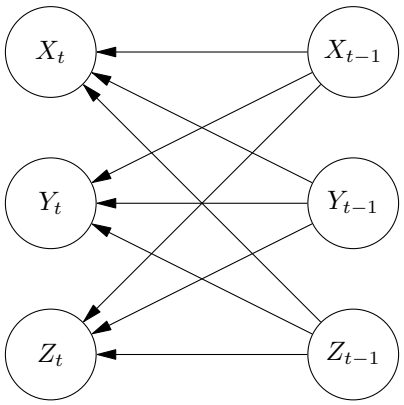
can calculate the covariance matrix $\mathbf{\Sigma} = \text{Cov}(X_t, X_{t-1}, \ldots, X_{t-k})$, derive the partial correlations $\rho_{i,j}$ of each pair of variables given the rest, and from these build a CIG.

From this CIG, we can derive the appropriate lag for the model VAR(p) we look to estimate by seeing which furthest back variable one of the variables at time t is connected to. However, to do so we must also derive the DAG structure for both the canonical and structural VAR models, since when we moralise them we attain all possible CIGs.

Consider the canonical VAR(1) model

$$\begin{pmatrix} X_t \\ Y_t \\ Z_t \end{pmatrix} = \begin{pmatrix} \varphi_{1,1,1} & \varphi_{1,2,1} & \varphi_{1,3,1} \\ \varphi_{2,1,1} & \varphi_{2,2,1} & \varphi_{2,3,1} \\ \varphi_{3,1,1} & \varphi_{3,2,1} & \varphi_{3,3,1} \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \\ Z_{t-1} \end{pmatrix} + \begin{pmatrix} e_t \\ f_t \\ g_t \end{pmatrix},$$

its related DAG is



where each link is the graphical representation of the $\varphi_{i,j,k}$ in the canonical model. Note that contemporaneous variables in the canonical model are not linked, since by the definition of the canonical model, they do not effect each other.

Building upon the earlier Proposition 8, I offer an elementary implication of the structure of the $\mathbf{\Phi_1}$ matrix in canonical VAR(1) models on the DAG and CIG associated with it.

Remark 1. If a canonical VAR(1) model has $\mathbf{\Phi_1}$ such that each row has at most one non-zero elements, then the DAG associated with the canonical VAR models satisfies the Wermuth condition. Further, its associated CIG is bipartite.

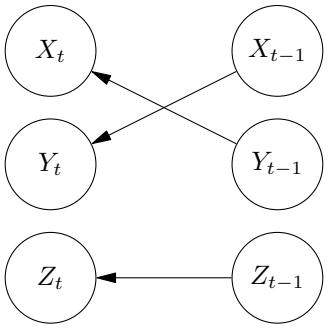
Proof. Assume we have a canonical VAR(1) model with $\mathbf{\Phi_1}$ with no more than one non-zero element in each row, then each variable at time t is caused by at most one anterior variable and by consequence each variable at time t in the DAG has an indegree (the number of edges

flowing into it) of at most one. Since no two anterior variables both cause a variable at time t, then the Wermuth condition is satisfied by definition. The second statement then follows directly from Proposition 8. □

Example 4.3.1. Take the VAR(1) model

$$\begin{pmatrix} X_t \\ Y_t \\ Z_t \end{pmatrix} = \begin{pmatrix} 0 & \varphi_{1,2,1} & 0 \\ \varphi_{2,1,1} & 0 & 0 \\ 0 & 0 & \varphi_{3,3,1} \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \\ Z_{t-1} \end{pmatrix} + \begin{pmatrix} e_t \\ f_t \\ g_t \end{pmatrix}$$

with $\mathbf{\Phi_1} = \begin{pmatrix} 0 & \varphi_{1,2,1} & 0 \\ \varphi_{2,1,1} & 0 & 0 \\ 0 & 0 & \varphi_{3,3,1} \end{pmatrix}$. Using this sparse $\mathbf{\Phi_1}$ matrix, we can construct the following DAG using the same method as described above.



This clearly satisfies the Wermuth condition, since no two anterior variables both cause the same posterior variable, and thus the associated CIG is determined simply by exchanging directed edges with undirected edges, since no demoralisation is necessary. The resulting CIG is clearly bipartite.

In a structural VAR model, however, dependence between contemporaneous variables is allowed for. For example, take the structural VAR model

$$\begin{pmatrix} X_t \\ Y_t \\ Z_t \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ \varphi_{2,1,1} & 0 & 0 \\ 0 & \varphi_{3,2,1} & 0 \end{pmatrix} \begin{pmatrix} X_t \\ Y_t \\ Z_t \end{pmatrix} + \begin{pmatrix} \varphi_{1,1,2} & \varphi_{1,2,2} & 0 \\ 0 & \varphi_{2,2,2} & 0 \\ 0 & 0 & \varphi_{3,3,2} \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \\ Z_{t-1} \end{pmatrix} + \begin{pmatrix} e_t \\ f_t \\ g_t \end{pmatrix}.$$

This may be represented by the following DAG, where some contemporaneous variables are linked.

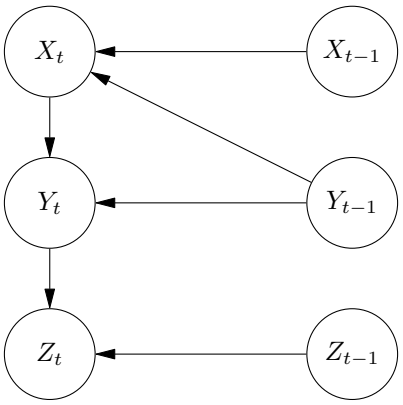


Figure 4.1: The DAG associated with the SVAR above

From this DAG, we understand the causal relationships between the variables in our model. We can also gain insight into the broader conditional independence of the variables by computing its associated CIG as described earlier in this paper. Moralising the DAG, adding links between mutual parents of a node at time t, we obtain its related CIG as follows.

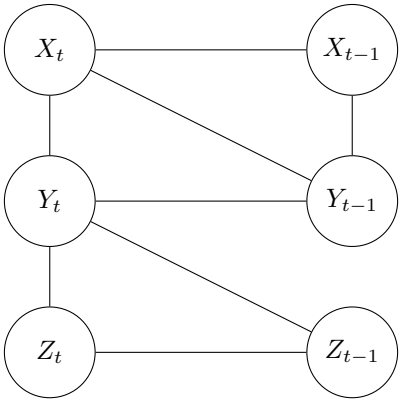


Figure 4.2: The moralised CIG associated with the DAG

Knowing the CIG of an autoregressive model is valuable for us when we wish to remove moral edges and compare possible different DAG models. Note that in the example given above, our true DAG associated with the VAR model includes the edge from Y_{t-1} to X_t which is a moral edge. Removing this edge in the demoralisation process would be wrong. In the following subsection we will look at an example where we start with a CIG estimated from simulated data and, by removing potentially moral edges iteratively, we compare all possible DAGs given our *a priori* knowledge of the data to determine which are congruent. Fitting these models to data and comparing them using penalised likelihood methods, we can determine which VAR model, and thus which combination of moral edges, fits our data best. This will show us which moral edges we should remove, and those we should not.

In order to find a DAG associated with the CIG of a VAR model, we need only to consider the links to the variables at time t. We must first ascertain all possible combinations of directed edges between the variables at time t. In the CIG above, these are as follows.

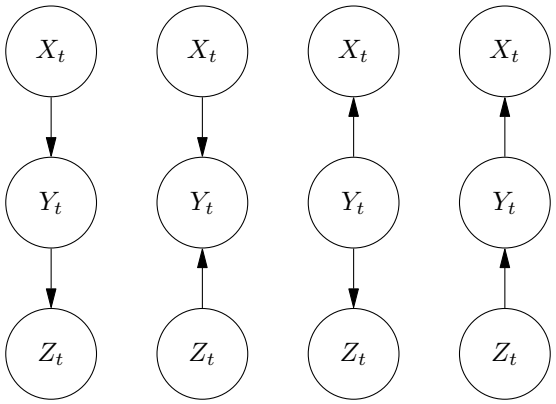


Figure 4.3: Different combinations of contemporaneous directed edges.

Taking each of these in turn, we determine which of the edges in the associated DAG could be a result of moralisation, knowing that edges connecting anterior variables (variables that occur further back in time) effect more recent variables given the passing of time. We then consider all possible combinations of moral edges in these DAGs and fit the model to our data, calculating their associated BIC and AIC whose definitions below are adapted from Strimmer (2021).

Definition 4.3.1. The Bayesian Information Criterion (BIC) for a model with K parameters and n observations is defined as

$$\text{BIC} = -2 \log L + K \log(n)$$

where L is the maximised likelihood function of the model. Similarly, the Akaike Information Criterion (AIC) is defined as

$$\text{AIC} = -2 \log L + 2K.$$

We may determine which models fit the data best from these criteria.

An upper bound for the number of combinations we will need to check can be found. Knowing that a cycle of length three in the CIG associated with a VAR model may have an edge caused by moralisation, let n be the number of such cycles in the CIG. Then for each

combination of edges connecting contemporaneous variables, see Fig. 4.3, we have at most 2^n different combinations of moral edges in the graph, and thus at most 2^n possible DAGs to fit for that initial combination of contemporaneous edges.

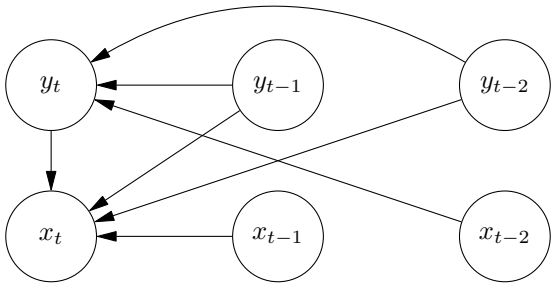
In the following subsection, we look to apply the theory developed over the course of this paper to simulated data.

4.3.1 Worked Example with Simulated Data

We will now simulate data for the structural VAR(2) model

$$\begin{pmatrix} 1 & 0 \\ 0.5 & 1 \end{pmatrix} \begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} 0.5 & 0 \\ -0.2 & 0.6 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} 0.25 & 0.5 \\ 0.1 & 0 \end{pmatrix} \begin{pmatrix} y_{t-2} \\ x_{t-2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{yt} \\ \varepsilon_{xt} \end{pmatrix}$$

where the $\varepsilon_{xt}, \varepsilon_{yt} \overset{\text{i.i.d}}{\sim} \text{N}(0,1)$, using R. The true DAG associated with this model is



Our simulated data will have the values of y_t and x_t at each time period t, and by looking back we build lagged variables y_{t-1}, x_{t-1}, y_{t-2}, x_{t-2}, y_{t-3}, x_{t-3}. Each of these becomes a variable in our graphical model. We will use the simulated data to estimate the partial correlations between each two variables given the rest, and using these partial correlations and a threshold by which we judge significance, build a CIG of our model. Once we have the CIG, we will find the equivalent DAGs and find the most parsimonious model with the highest log-likelihood to represent our data, using Bayesian and Akaike Information Criteria. We will then compare this resulting DAG with the true DAG given above, and see if they are the same.

In R, we first define the coefficient matrices of our model in order to generate data as follows.

```
set.seed(234)
```

```
# Number of time series observations
tt <- 1000

# Structural coefficients
Phi <- diag(1, 2)
Phi[lower.tri(Phi)] <- c(0.5)

# Coefficient matrices
Phi1 <- matrix(
  c(0.5, -0.2, 0, 0.6), nrow=2, ncol=2
)
Phi2 <- matrix(
  c(0.25, 0.1, 0.5, 0), nrow=2, ncol=2
)
```

We will now turn our structural model into the canonical form

$$\begin{aligned}
&\mathbf{\Phi}_0 \overrightarrow{X}_t = \mathbf{\Phi}_1 \overrightarrow{X}_{t-1} + \mathbf{\Phi}_2 \overrightarrow{X}_{t-2} + \overrightarrow{\varepsilon}_t \\
&\Rightarrow \overrightarrow{X}_t = \mathbf{\Phi}_1^* \overrightarrow{X}_{t-1} + \mathbf{\Phi}_2^* \overrightarrow{X}_{t-2} + \overrightarrow{\varepsilon}_t^*
\end{aligned}$$

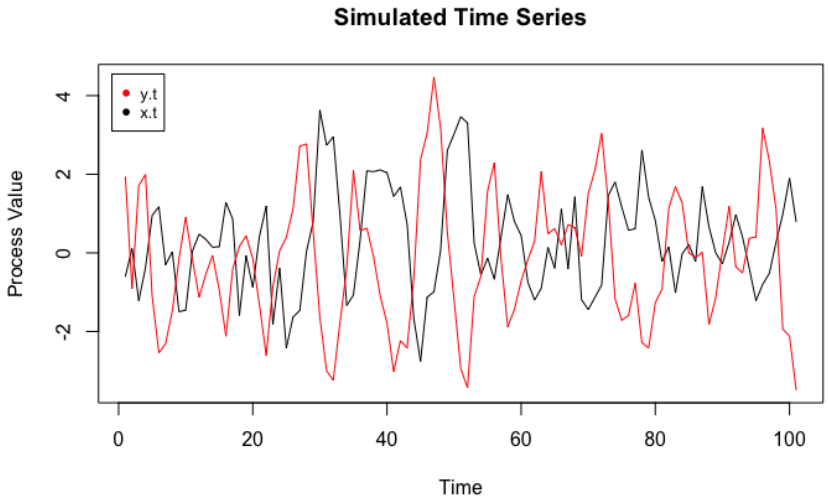
with $\mathbf{\Phi}_i^* = \mathbf{\Phi}_0^{-1}\mathbf{\Phi}_i$ and $\overrightarrow{\varepsilon}_t^* = \mathbf{\Phi}_0^{-1}\overrightarrow{\varepsilon}_t$.

```
# Structural matrices
Phi1.star = solve(Phi)%*%Phi1
Phi2.star = solve(Phi)%*%Phi2

# Generate series
series <- matrix(0, 2, tt + 1)
for (i in 3:(tt + 1)){
  series[, i] <- (
    Phi1.star%*%series[, i-1] +
      Phi2.star%*%series[, i-2] +
      solve(Phi)%*%rnorm(2, 0, 1)
  )
}
series <- tail(series, n=tt)
series <- ts(t(series)) # Convert to time series object
```

```
dimnames(series)[[2]] <- c("y", "x") # Rename variables
```

We now have the values of our process for 1,000 time periods. We can visualise the final 100 of these as follows.



We now look to calculate the partial correlations between variables using the Inverse Variance Lemma described earlier. We will include variables going back to time period $t-3$ to show that our model will recognise that this process is indeed of order 2, and not 3.

```
# Build design matrix
X.t = cbind(series[,"y"], series[,"x"])
X = X.t
for (i in 1:3) {
  temp = stats::lag(X.t, i)
  X = cbind(temp, X)
}
colnames(X) <- c("y.t", "x.t", "y.t1", "x.t1", "y.t2", "x.t2", "y.t3", "x.t3")
X <- X[complete.cases(X),]
# estimate the covariance matrix
S = cov(X)
# derive the precision matrix
Omega = solve(S)
# calculate partial correlations of observations with the inverse variance lemma
pcor <- function (Omega, i, j) {
  return(-(Omega[i,j])/(sqrt(Omega[i,i] * Omega[j,j])))
```

```
}
pcor.list <- c()
for (i in 1:nrow(Omega)) {
  for (j in 1:ncol(Omega)) {
    part.cor <- pcor(Omega, i, j)
    pcor.list = c(pcor.list, part.cor)
  }
}
pcor.mat <- matrix(pcor.list, nrow=nrow(Omega))
diag(pcor.mat) = -diag(pcor.mat)
```

which returns the following partial correlations between variables.

	y_t	x_t	y_{t-1}	x_{t-1}	y_{t-2}	x_{t-2}	y_{t-3}	x_{t-3}
y_t	1							
x_t	-0.4507	1						
y_{t-1}	0.2710	-0.1332	1					
x_{t-1}	0.2291	0.5297	-0.2773	1				
y_{t-2}	0.2522	0.1098	0.1968	-0.1839	1			
x_{t-2}	0.3561	0.0075	0.0469	0.4134	-0.3489	1		
y_{t-3}	-0.0143	-0.0239	0.2442	0.1042	0.3553	-0.1468	1	
x_{t-3}	0.0120	-0.0229	0.3992	0.0169	0.2362	0.4137	-0.4647	1

We need to calculate the threshold by which we measure statistical significance of partial correlation. Using the link between partial correlation and t-statistics discussed before, we calculate the critical quantile for the Student's t distribution at the $\alpha = 0.05$ significance level, which we denote t, and from this calculate the critical correlation coefficient via

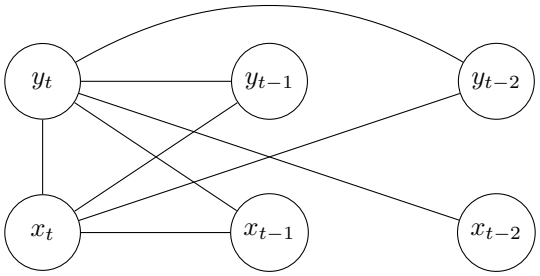
$$r_{\text{critical}} = \frac{t^2}{t^2 + \upsilon}$$

where υ is the residual degrees of freedom. In R, this is implemented as follows.

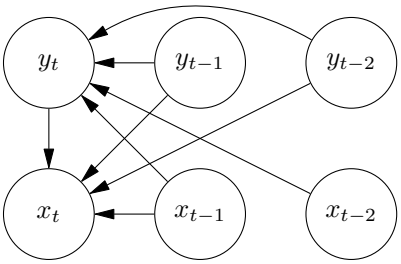
```
critical.r <- function(n, alpha = 0.01) {
  df <- n - 2
  critical.t <- qt(alpha/2, df, lower.tail = FALSE)
  critical.r <- sqrt( (critical.t^2) / ( (critical.t^2) + df ) )
  return( critical.r )
}
crit.r = critical.r(nrow(X))
```

from which we find that $r_{\text{critical}} = 0.0621$. Taking the most statistically significant partial

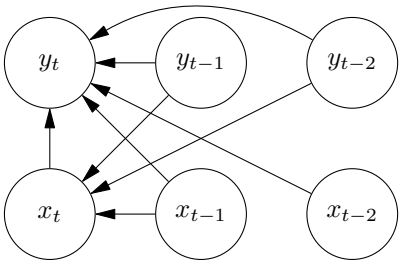
correlations between variables as edges, we build the following CIG.



Since we are dealing with time series data, we have the *a priori* knowledge that the present cannot affect the past. Using this knowledge we restrict ourselves to the following two DAG possibilities.



(a) Model A



(b) Model B

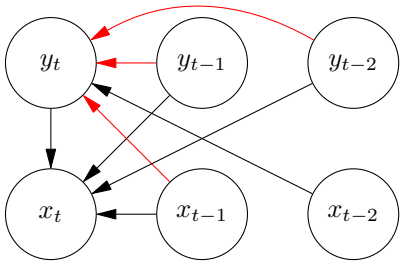
In model A, the edge connecting x_t and y_t flows from y_t to x_t, and in model B it flows the opposite way. These two models will serve as our more explicative models. That it, they are the most complex models we will work with, that explain more of the variation in our data with more dependencies than the more parsimonious we will achieve through demoralisation. Using the definition of moralisation, we can consider three remarks, given in Reale (1998), to help us determine which of these edges may or may not be moral in order to achieve some more parsimonious models.

1. If a node has only one outward edge, it is not a moral edge because in order for it to be moral two parents are required, and each parent needs at least two outward edges.

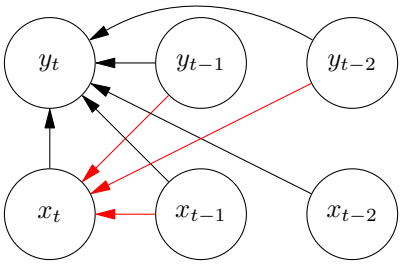
2. If a node has multiple outward edges, at least one is not a moral edge.

3. If a node has no outward edges, then none of its incoming edges is moral.

From these criteria, we identify three edges in each of the two models that may be moral. These edges are coloured red in the diagram below.



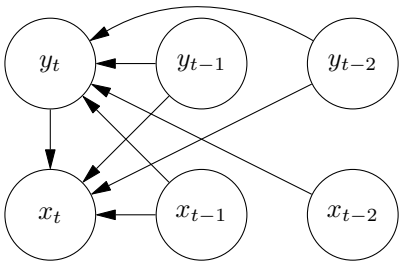
(c) Model A with potentially moral edges coloured red



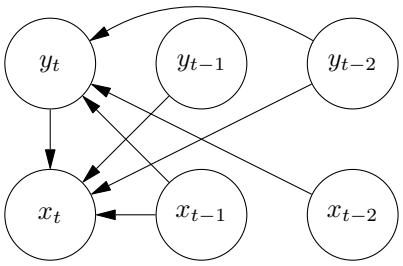
(d) Model B with potentially moral edges coloured red

We thus have 8 possible combinations of these potentially moral edges per combination which we investigate in turn to make 8 different DAGs per model, 16 DAGs in total. Once we have made these DAGs, we will look to discount those that are incongruent with the assumptions of our VAR model, and select the best option from the remaining models.

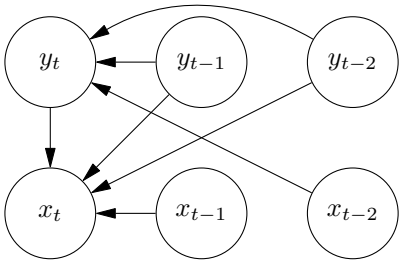
We will begin by enumerating all such possibilities for Model A, numbering them from 1 to 8, in Figure 4.4, and then repeating the process for Model B in Figure 4.5.



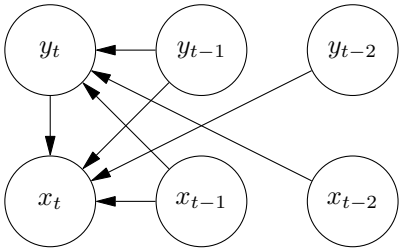
Model A (1).



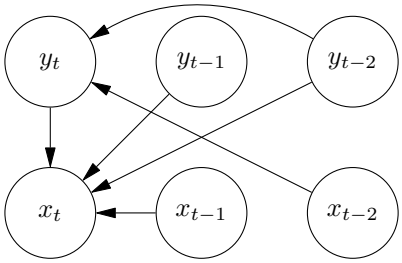
Model A (2).



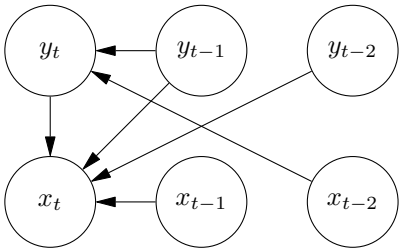
Model A (3).



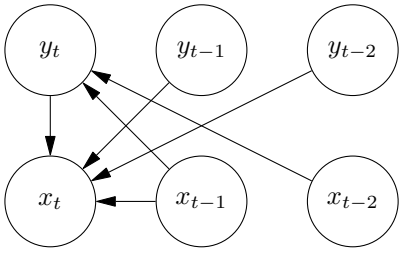
Model A (4).



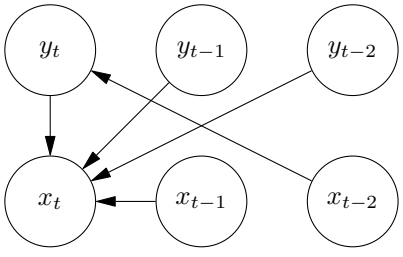
Model A (5).



Model A (6).

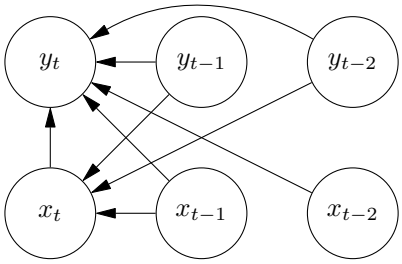


Model A (7).

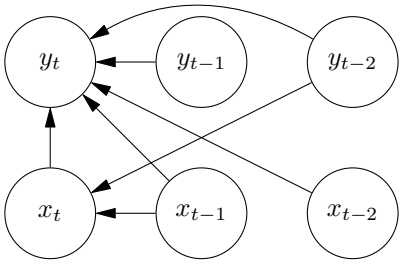


Model A (8).

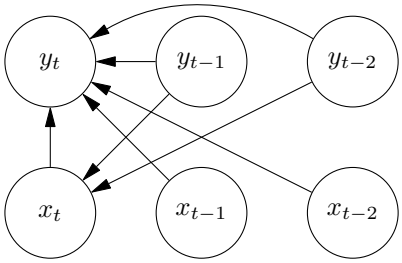
Figure 4.4: Possible DAGs from Model A



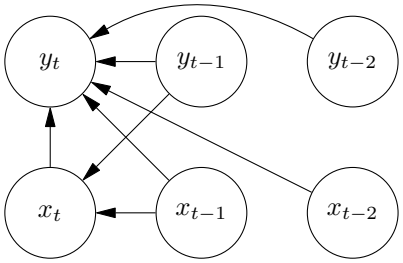
Model B (1).



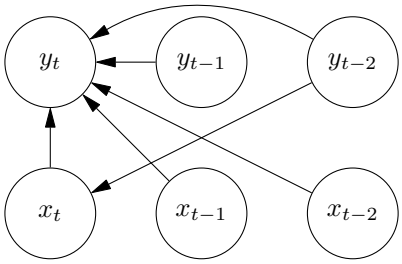
Model B (2).



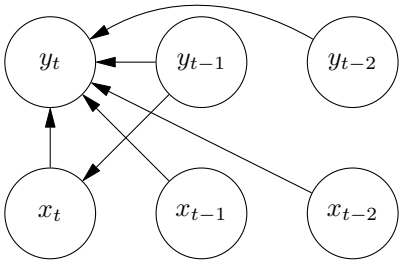
Model B (3).



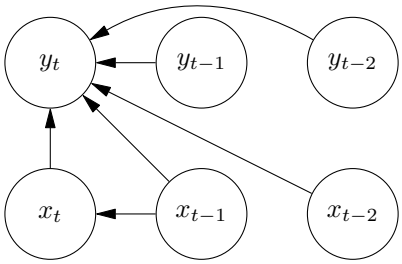
Model B (4).



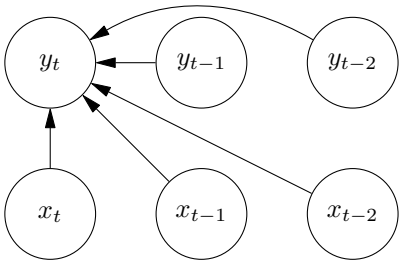
Model B (5).



Model B (6).



Model B (7).



Model B (8).

Figure 4.5: Possible DAGs from Model B

We note that Model B (8) is incongruent with the assumptions of our VAR, as x_t is no longer caused by any anterior variables, while still causing the contemporaneous y_t. This is clearly impossible, and we thus discount it.

For the remaining 15 DAG possibilities, we estimate a VAR of order 2 on the data including only those variables included in the DAGs using the `vars` R package (Pfaff, 2008b) in order to measure the performance of the models in turn. We will use as metrics, the Bayesian Information Criterion and Akaike Information Criterion.

Once we have calculated the BIC and AIC of all models, we will chose the model with the lowest of both if possible, otherwise we will consider BIC more than AIC when choosing the most appropriate DAG. By construction, as the number of observations grows, so too does the set of models AIC deems to be suitable, whereas the opposite is true for BIC where the number of suitable models decreases with the number of observations. Below are the BIC and AIC of each of the models.

	Model A		Model B	
Model Number	BIC	AIC	BIC	AIC
1	5754.791	5720.443	5754.791	5720.443
2	6034.646	6005.206	6129.215	6099.775
3	5745.673	5716.233	6020.609	5991.168
4	5797.822	5768.381	5759.04	5729.599
5	6081.738	6057.204	6425.187	6400.653
6	5784.665	5760.131	6054.971	6030.437
7	6288.89	6264.356	6140.522	6115.988
8	6534.093	6514.466	Discounted	

Table 4.1: The BIC and AIC for all 16 possible DAGs emerging from the empirical CIG.

From this, we find that Model A (3) has the lowest BIC and AIC. The next best model is the full model, that is the model with none of the potentially moral edges removed. We are working with a large number of observations in this dataset, and since we will always favour the more parsimonious model, we take Model A (3) as the best model for the data. Reminding ourselves what this model looks like, we find that it is indeed the true model of our DAG presented at the beginning of the example.

We may thus conclude that on these simulated data, our graphical modelling technique was able to identify not only the correct lag in the model, but also the correct causality relationships

between variables. Note that just because an edge is "moral", does not mean that it is not included in the correct model, as seen in Figure 4.1. In fact, the true model for this data generating process, includes the moral edge from y_{t-1} to y_t. This graphical model approach was able to identify that indeed this edge was significant, and that it was included in the final model.

We will now endeavour to show that the success of this method extends to real-life data, in the form of Standard and Poor's Index returns.

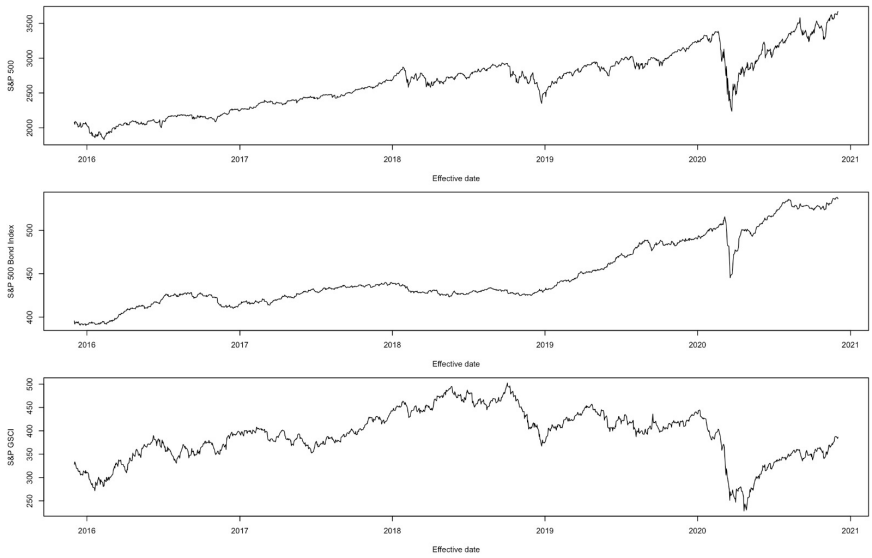
Chapter 5

Application to Econometrics

Much of the literature surrounding autoregressive models is applied to econometrics. In order to render this paper relevant, we will demonstrate how our methodology can be applied to financial data.

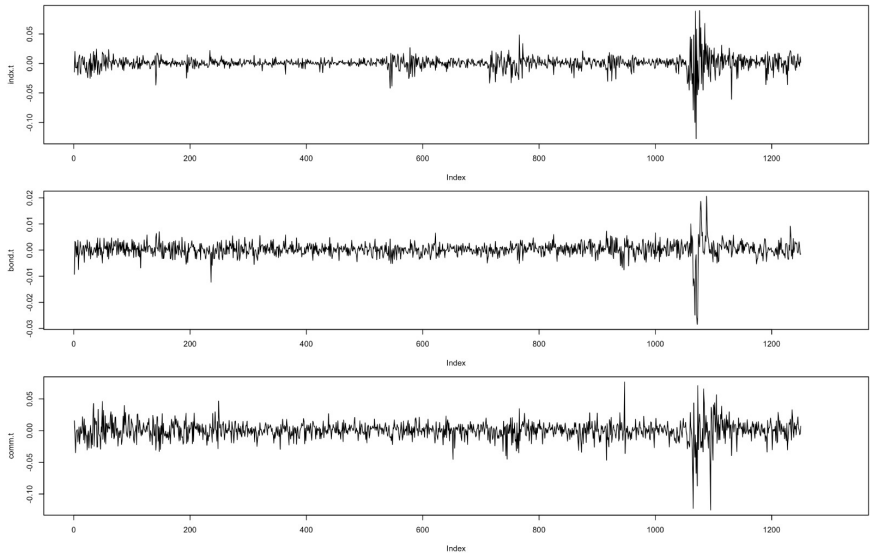
5.1 Worked Example with Financial Data

In order to apply this methodology to a financial case study, we first consider the index prices of the Standard and Poor's 500 Index, Bond Index, and GCSI index over the past five years, plotted below.



This is clearly not a stationary process, so in their current state these data cannot be used. We thus take the log of the returns of the three indices, plotted below, and find that this

process is indeed stationary. These are the natural logarithms of the changes in the index price from one day to the next.



We will now consider these log returns as variables in our stochastic process representing a portfolio, and look to determine the underlying relationship between them. We will first estimate the sample covariance matrix of the observed log returns of these indices over the past five years, and from this compute the partial correlations between lagged and contemporaneous variables. Then calculating a critical correlation coefficient we will build a CIG. From this CIG, we impute all possible DAG structures of the variables. Fitting each VAR model associated with these DAGs to data and calculating their respective BIC and AIC, we will finally conclude the best fitting VAR model for our process. The R code necessary for these calculations is more or less identical to that used in the previous worked example, and so is omitted in the interest of space, although the main functions can be found in McLatchie (2020). We remove all time periods with null data since, given the amount of data available, our model will likely converge despite these missing data.

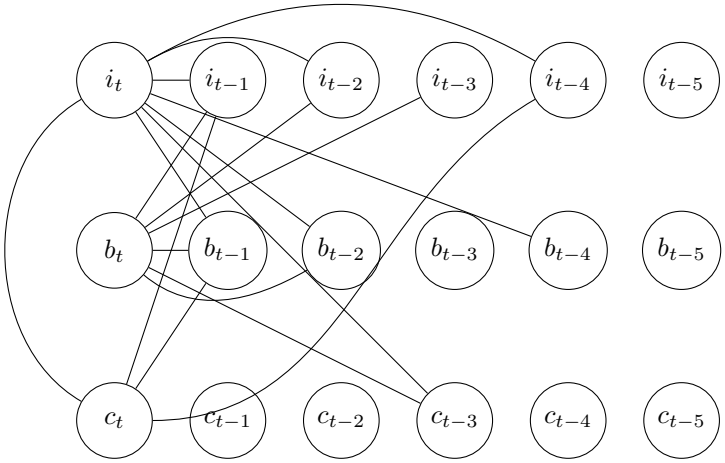
We run our analysis using five lagged variables, and find that when we do, our model only find four of them to be statistically significant. Our model will thus make use of the first four lagged variables. The sample covariance matrix of our data leads to the following partial correlation matrix.

	i_t	b_t	c_t	i_{t-1}	b_{t-1}	c_{t-1}	i_{t-2}	b_{t-2}	c_{t-2}	i_{t-3}	b_{t-3}	c_{t-3}	i_{t-4}	b_{t-4}	c_{t-4}	i_{t-5}	b_{t-5}	c_{t-5}
i_t	1																	
b_t	-0.028	1																
c_t	0.435	0.009	1															
i_{t-1}	-0.222	0.216	0.094	1														
b_{t-1}	0.075	0.195	-0.077	-0.051	1													
c_{t-1}	0.063	0.063	-0.049	0.410	-0.008	1												
i_{t-2}	0.087	0.089	0.004	-0.214	0.189	0.082	1											
b_{t-2}	0.127	0.089	-0.027	0.080	0.163	-0.088	-0.070	1										
c_{t-2}	0.032	0.035	0.010	0.065	0.058	-0.053	0.381	-0.016	1									
i_{t-3}	-0.03	0.075	-0.001	0.056	0.078	0.003	-0.218	0.184	0.079	1								
b_{t-3}	0.042	0.03	-0.06	0.124	0.081	-0.029	0.051	0.156	-0.069	-0.069	1							
c_{t-3}	0.091	0.074	-0.012	0.036	0.013	-0.001	0.057	0.040	-0.063	0.375	-0.026	1						
i_{t-4}	-0.145	0.006	0.134	-0.068	0.087	0.015	0.044	0.092	0.019	-0.221	0.209	0.083	1					
b_{t-4}	-0.086	0.011	0.040	0.013	0.033	-0.050	0.106	0.092	-0.009	0.045	0.180	-0.068	-0.061	1				
c_{t-4}	-0.027	0.012	0.028	0.081	0.072	-0.012	0.035	0.016	0.003	0.054	0.043	-0.060	0.367	-0.030	1			
i_{t-5}	0.049	0.009	-0.025	-0.146	-0.007	0.142	-0.041	0.086	-0.021	0.048	0.070	0.024	-0.228	0.213	0.083	1		
b_{t-5}	-0.047	-0.029	0.039	-0.089	0.018	0.052	-0.036	0.050	-0.002	0.100	0.139	-0.002	0.070	0.232	-0.063	-0.095	1	
c_{t-5}	-0.035	0.032	0.039	-0.034	0.012	0.023	0.075	0.070	-0.006	0.030	0.020	-0.001	0.044	0.038	-0.062	0.355	-0.039	1

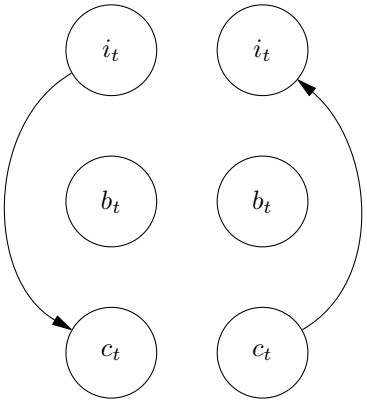
Calculating the critical partial correlation coefficient similarly to in the worked example, we find that

$$r_{\text{critical}} = \frac{t^2}{t^2 + v} = 0.073.$$

Using this, we build the following CIG.



Using once more the *a priori* knowledge that the present cannot effect the past, we find that there are only two possible contemporaneous structures.



Unfortunately, we have no further *a priori* knowledge that can restrict these choices further. We will call the model on the left, with edge flowing from i_t to c_t, Model A, and the model on the right, Model B. We equally identify that in Model A, the edges from i_{t-1} to c_t, b_{t-1} to c_t, and i_{t-4} to c_t are all potentially moral edges. In Model B, edges i_{t-1} to i_t, b_{t-1} to i_t, and i_{t-4} to i_t are potentially moral.

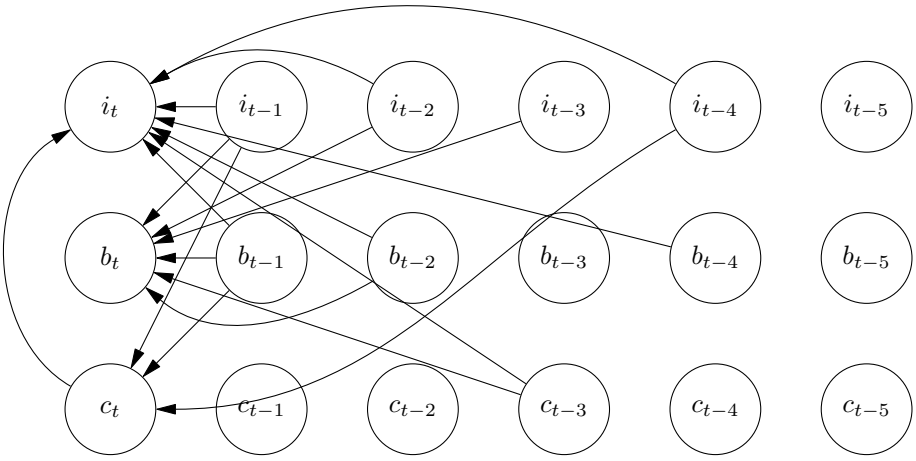
Enumerating all possible combinations of moral edges for Model A, we have the following

DAGs shown in Figure 5.1. We may immediately discount Model A (8), since c_t is not caused by any anterior variables, although still causes i_t. For Model B, we have the following possibilities shown in Figure 5.2.

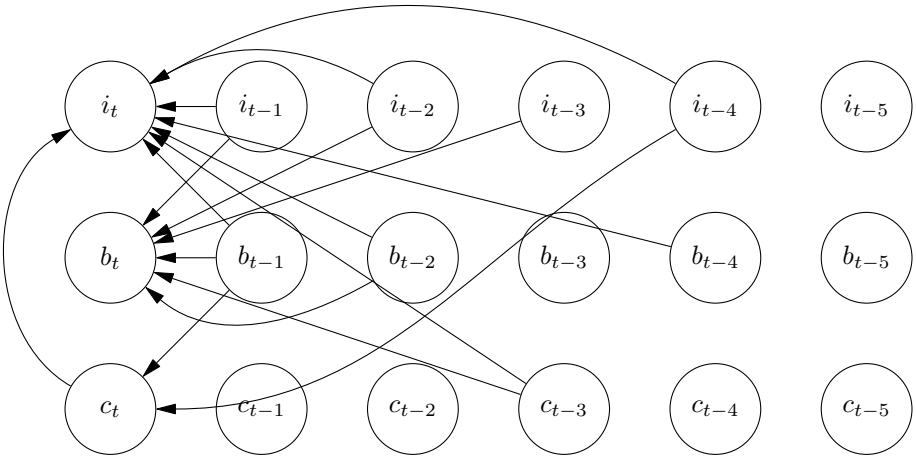
Fitting each of these models to data, we achieve the following BIC and AIC scores for each.

	Model A		Model B	
Model Number	BIC	AIC	BIC	AIC
1	-26012.22	-26094.27	-26012.22	-26094.27
2	-26017.82	-26094.73	-25940.2	-26017.12
3	-26017.96	-26094.88	-26014.88	-26091.8
4	-26002.36	-26079.28	-25998.03	-26074.94
5	-26023.63	-26095.42	-25943.36	-26015.15
6	-26008	-26079.79	-25924.7	-25996.49
7	-26009.3	-26081.08	-26002.87	-26074.66
8	Discounted		-25929.96	-25996.62

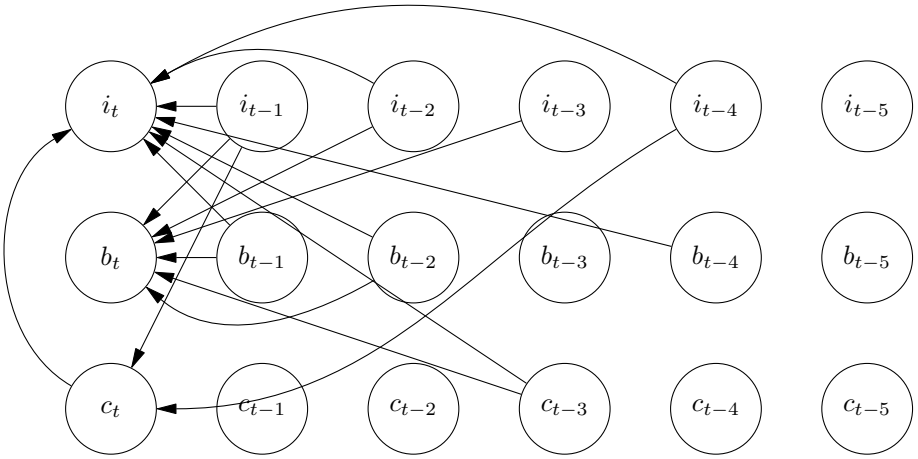
We conclude by choosing model A (5) as the best fitting model. The intuition drawn from this model, where price increases in the commodity futures market cause a price increase in the S&P 500 is that the change of commodity prices thus impacts the corporations' operations. This relationship is positively correlated, as shown in the partial correlation table, which may suggest that there are more commodities producers than consumers in the index, and thus an increase in commodities prices positively impacts their valuation.



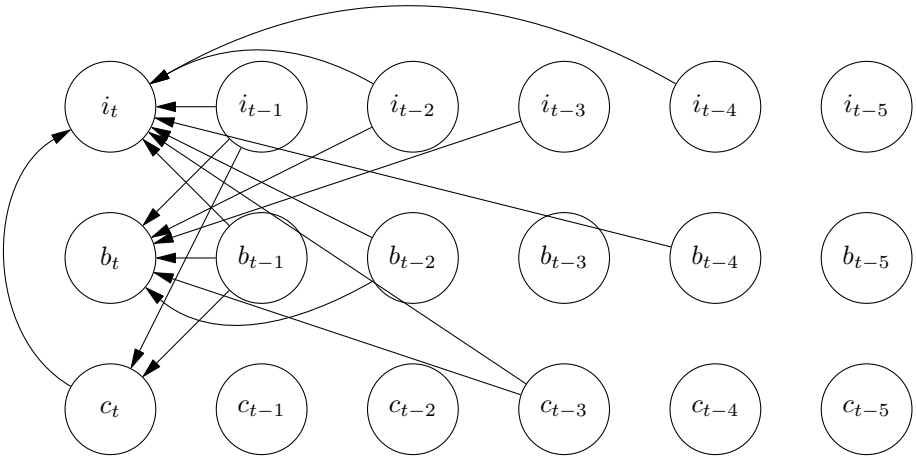
Model A (1).



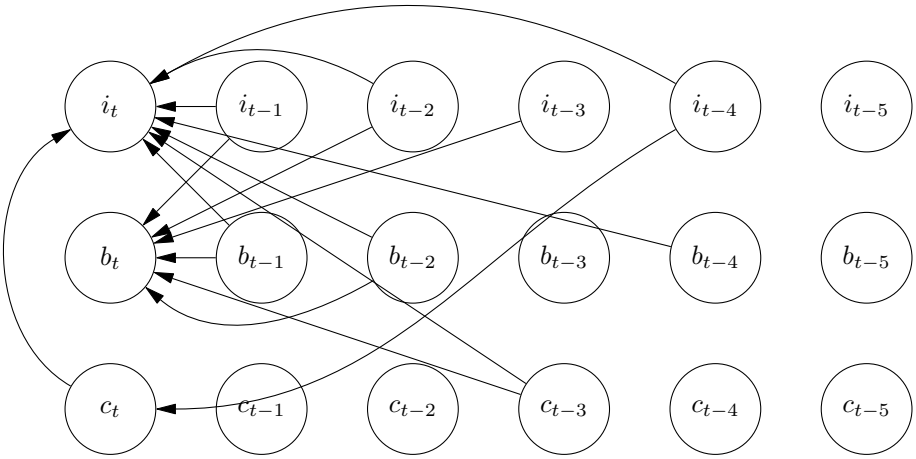
Model A (2).



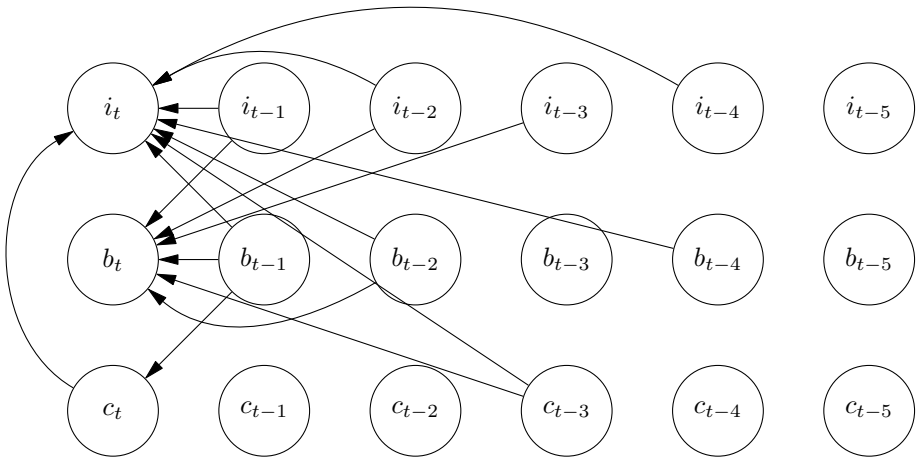
Model A (3).



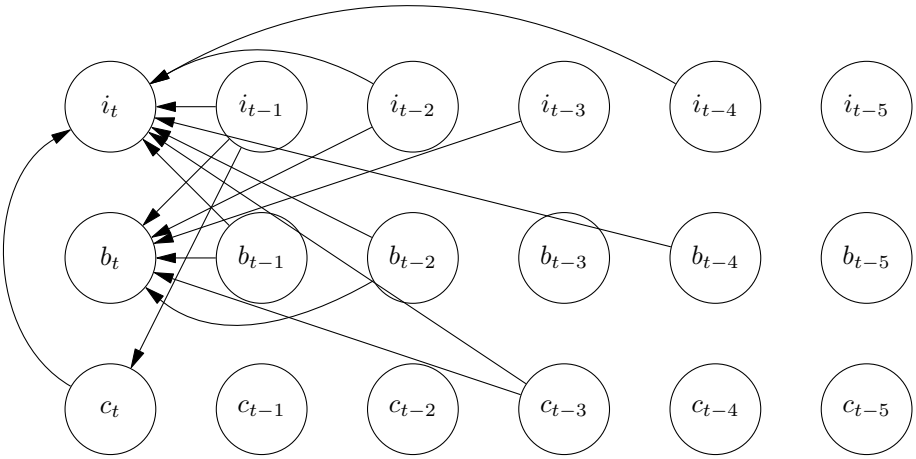
Model A (4).



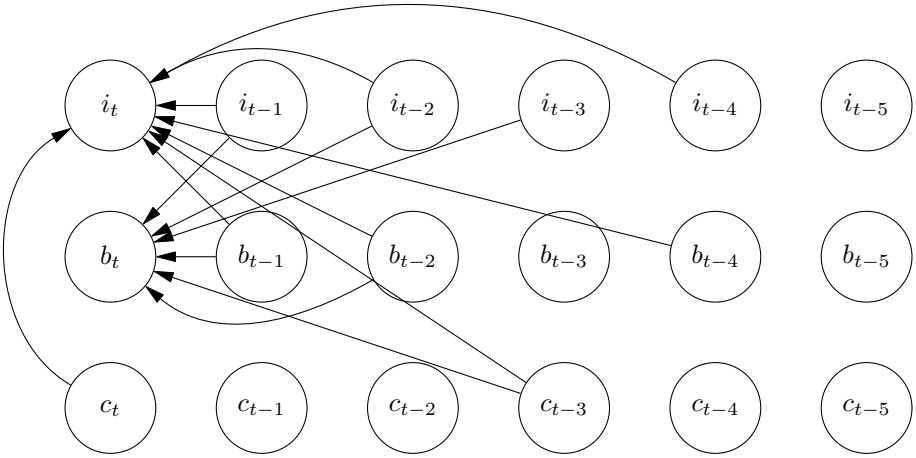
Model A (5).



Model A (6).

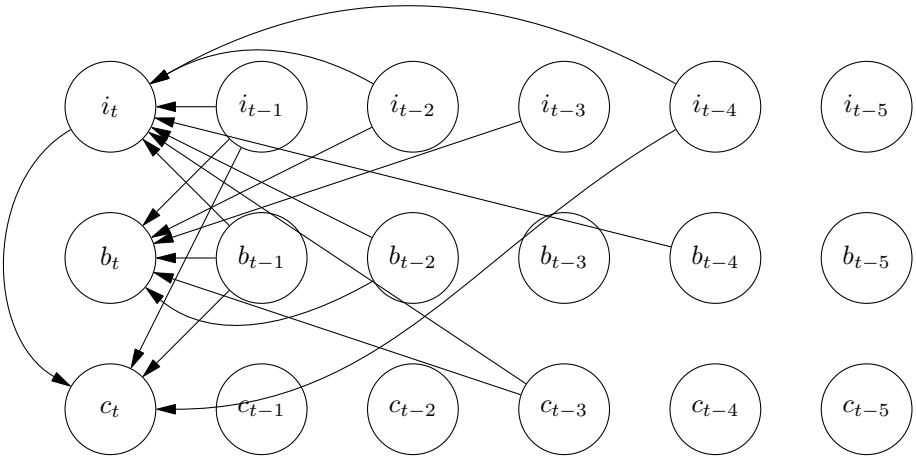


Model A (7).

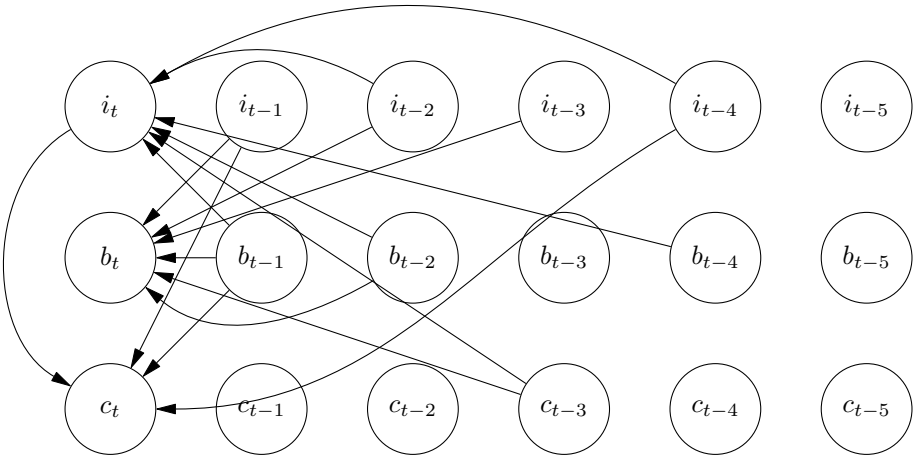


Model A (8).

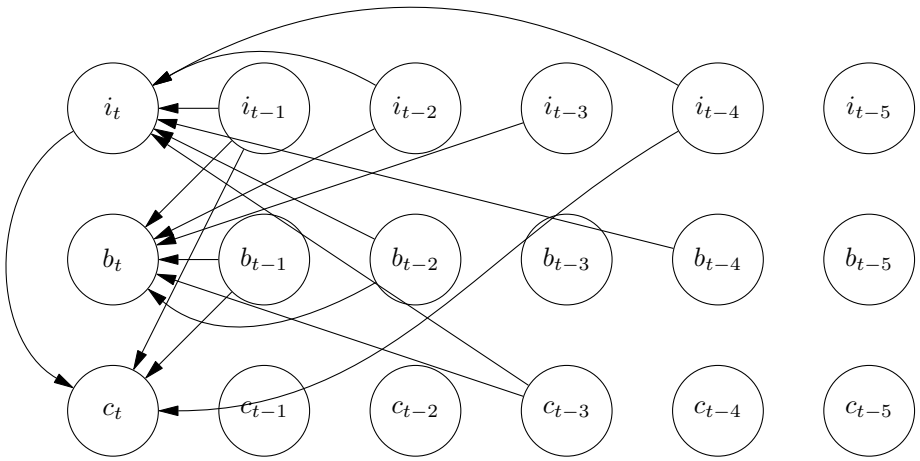
Figure 5.1: Possible DAGs from Model A



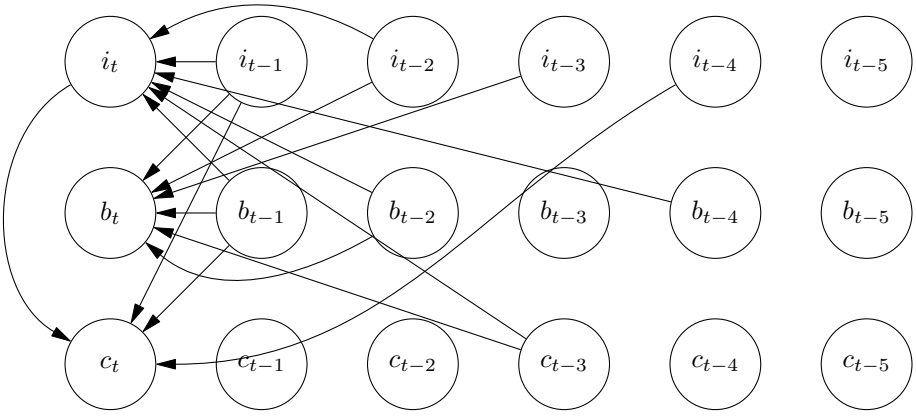
Model B (1).



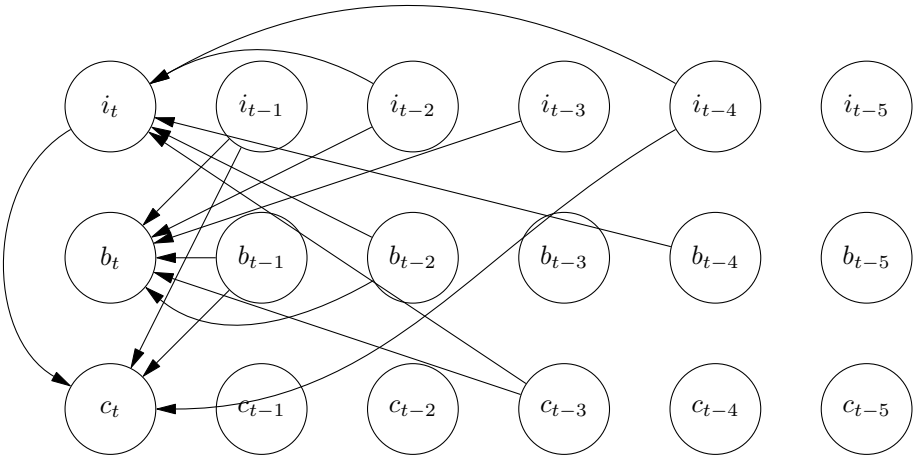
Model B (2).



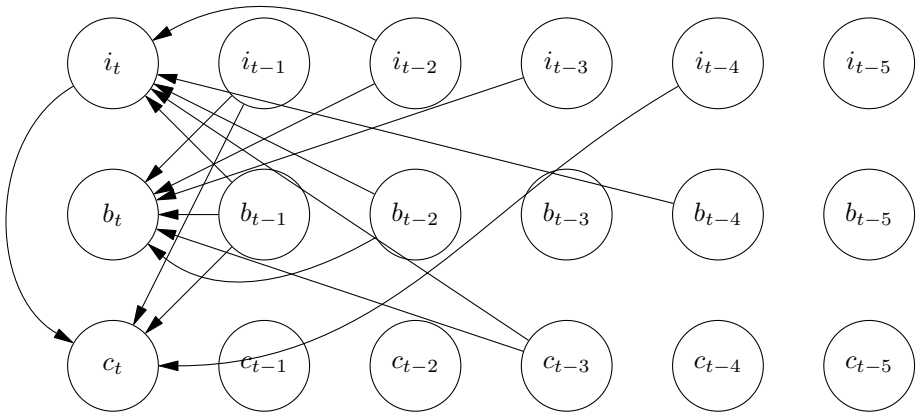
Model B (3).



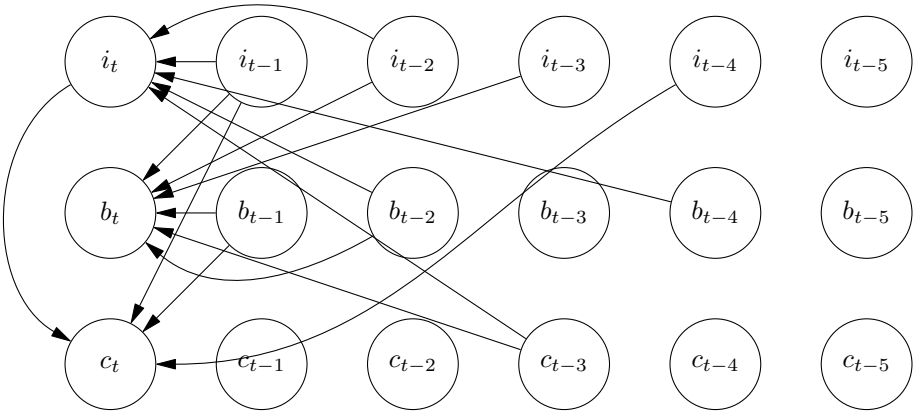
Model B (4).



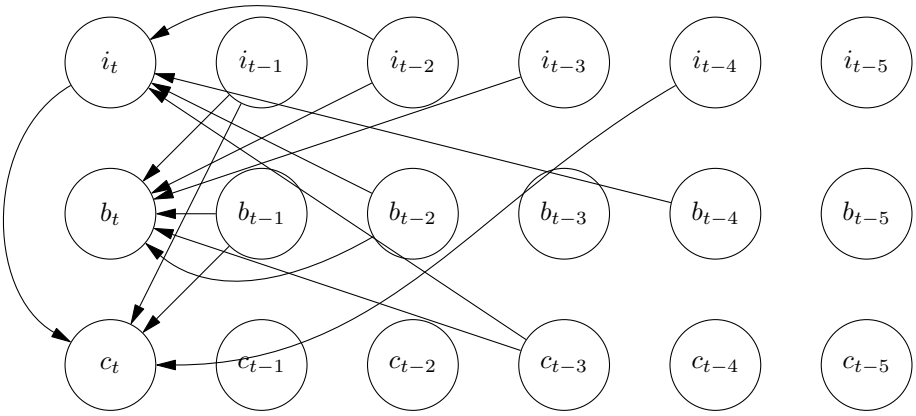
Model B (5).



Model B (6).



Model B (7).



Model B (8).

Figure 5.2: Possible DAGs from Model B

Chapter 6

Conclusion

We have thus demonstrated how graphical models can be used to determine the order of an autoregressive model on multivariate time series, as well as the causal relationships between variables. Using this information, we were able to determine the best VAR model for the data from its associated DAG. In order to achieve this result, we discussed how we build a CIG from the sample covariance matrix of a lagged variables using the Inverse Variance Lemma, and how we demoralise this resulting CIG iteratively in order to determine which moral links are necessary in our final DAG model.

We have shown through a worked example on simulated data that this methodology, while being expensive when we have many potentially moral edges in our graph, is able to determine with confidence the best performing model for data from a starting set of possible models. Using data from Standard and Poor's indices, we showed how this methodology can be applied to real-world cases.

The primary result of this report has been to show how we can use graphical models to reveal contemporaneous and noncontemporary dependencies between variables, and from these determine an initial set of possible models in a time efficient manner. The iterative fitting of DAGs with all possible combinations of moral edges, an innovation on the methodology provided in Reale (1998), allowed us to find the best model using penalised liklihood methods.

6.1 Further Work

Given more time, the methodology followed in this paper could be written into code, and eventually fully automated. My attempt at this can be found in McLatchie (2020).

Bibliography

Yann McLatchie. cig. https://github.com/yannmclatchie/cig, 2020.

Gabor Megyesi. Lecture notes in graph theory and combinatorics, 2021.

B. Pfaff. *Analysis of Integrated and Cointegrated Time Series with R.* Springer, New York, second edition, 2008a. URL http://www.pfaffikus.de. ISBN 0-387-27960-1.

Bernhard Pfaff. Var, svar and svec models: Implementation within R package vars. *Journal of Statistical Software*, 27(4), 2008b. URL http://www.jstatsoft.org/v27/i04/.

Marco Reale. *A Graphical Modelling Approach to Time Series.* PhD thesis, Lancaster University, 1998.

Megan Romer. Lecture notes in applied time series analysis. https://online.stat.psu.edu/stat510/, 2020.

Juliane Schafer, Rainer Opgen-Rhein, Verena Zuber, Miika Ahdesmaki, A. Pedro Duarte Silva, and Korbinian Strimmer. *corpcor: Efficient Estimation of Covariance and (Partial) Correlation*, 2017. URL https://CRAN.R-project.org/package=corpcor. R package version 1.6.9.

Korbinian Strimmer. Lecture notes in multivariate statistics and machine learning, 2021.

Joe Whittaker. *Graphical Models in Applied Multivariate Statistics.* John Wiley & Sons, 1990.