Bayesian Networks with Applications to Simulated Macroeconomic Timeseries

Emmet Hall-Hoffarth

June 2020

1 Introduction

Bayesian networks (Judea Pearl & Mackenzie, 2018) are a non-parametric statistical technique for modelling causal probabilistic relationships. Under some conditions this and associated tools can allow for the inference of an interpretable structural model of a Data Generating Process (DGP) directly from some observed data. While there is much research still to be done regarding the theoretical properties of this methodology, the potential of automatically identifying causality already opens the door for numerous useful economic applications. To my knowledge, little work has been done in this area within the econometrics literature. In particular, Imbens (2019) notes a lack of concrete empirical examples demonstrating the usefulness of this methodology in the field of economics. Therefore, the purpose of this paper is to investigate one potential empirical application in the field of macroeconomics.

The application involves modelling simulated data from well-known macroeconomic models. This application has a number of appealing properties. Firstly, because in a simulation the true DGP is known, it will be possible to precisely evaluate to what extent it is possible to identify the true underlying structure of the data using Bayesian networks. Furthermore, in this controlled environment it is possible to ensure that the assumptions required by Bayesian networks are satisfied, and therefore it provides a fair way to evaluate their applicability. Finally, these models are the central building blocks of modern macroeconomics and therefore provide a very relevant example of how Bayesian networks can be applied in economics.

The remainder of this paper will be organised as follows. The second section will explain the fundamental elements of Bayesian networks, with further discussion provided in an appendix. The third section will provide a review of the relevant literature in economics. The fourth section will discuss the empirical methodology employed in this paper. The fifth section will discuss the results that were obtained. The sixth section concludes.

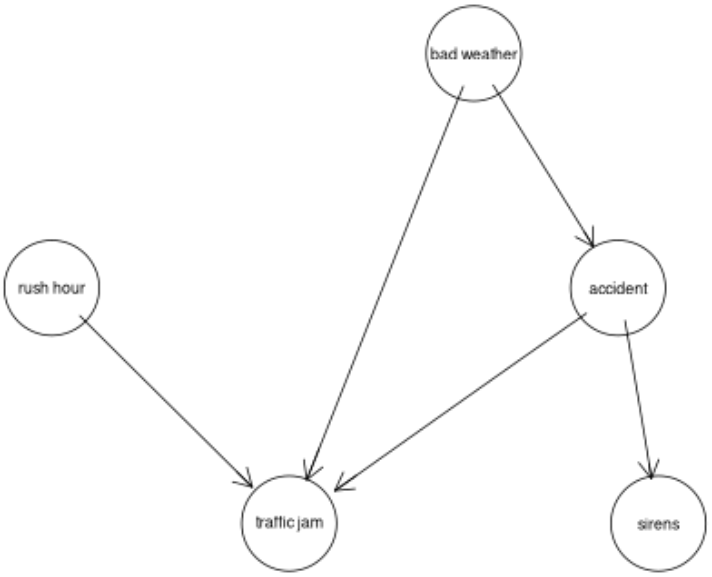


Figure 1: An example of a simple DAG (Liszka, 2013)

2 Bayesian networks

The fundamental assumption of a Bayesian network is that the underlying DGP of some observed data can be represented as a Directed Acyclical Graph (DAG). Figure 1 shows an example of a DAG. Each of the variables in the data forms a node in the graph, and these nodes are connected by arcs. The direction of each of the arcs represents the direction of causality in the sense of conditional probability. For example, if we observe the DAG $B \rightarrow A$, then A's distribution is conditional on B, whereas B's distribution is unconditional. In economic language the analogous interpretation is that B is exogenous while A is endogenous to or determined by B. As the name DAG implies, arcs are assumed to not create any cycles in the graph. This assumption is by no means innocuous, however, it is what gives Bayesian networks the power to identify causal effects. For a given node, the set of nodes which have an arc pointing into that node are known as that node's parents, and the set of nodes that have an arc pointing into them from that node are known as that node's children. A root node is a node that has no arcs leading into itself, and a leaf node is a node that has not arcs leading out of itself.

Each arc represents a conditional probability relationship. Nodes in the graph are assumed to be conditionally independent of all nodes which are not its parents. For example, in figure 1:

$$p(sirens|data) = p(sirens|accident) \tag{1}$$

These conditional probabilities are abstract in the sense that they could be treated as either discrete or continuous, and any distributional assumption of choice could be applied to them. While much of the literature surrounding Bayesian networks focuses on the discrete case, in many

economic applications continuous variables are the primary concern. This is possible as long as we are willing to make some distributional assumption about the nature of the conditional probability. The most common assumption here (and fortunately the most natural economic one), is that the conditional distributions follow a multivariate normal distribution of the conditioning variables. This is implies that conditional distributions are linear functions of conditioning variables with Gaussian errors, which is exactly the assumptions of simple, small-sample OLS regressions common in econometrics. Such models are sometimes known as "Gaussian Bayesian networks." (GBN) For example:

$$sirens|data = sirens|accident = \alpha + \beta accident + \epsilon, \epsilon \sim N(0, \sigma^2) \tag{2}$$

Therefore, this technique is "non-parametric" in the sense that we do not make any assumptions about which underlying relationships exist between variables (indeed, this is what we hope the model will tell us). However, we do make a distributional assumption about the conditional distributions.

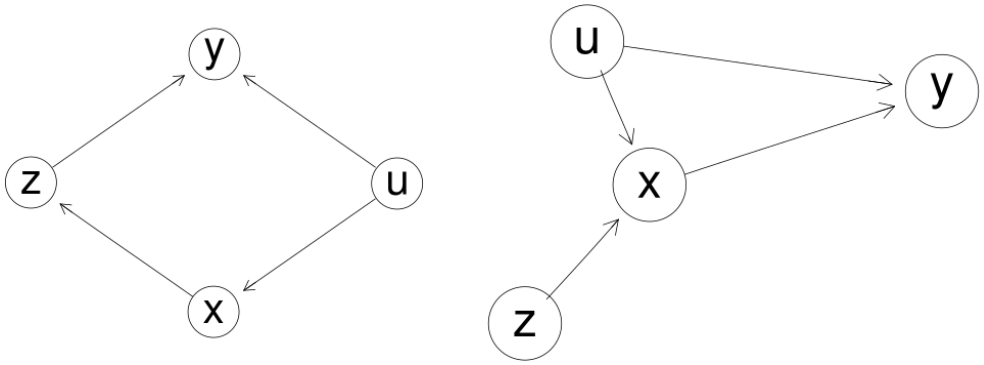
When fully specified, a GBN consists of a system of linear equations that defines the joint distribution of the data. Because of the properties of the normal distribution, this means that we can express a GBN as a single joint normal distribution over the data, where the DAG specifies the exact restrictions that are imposed on the covariance matrix. In order to enhance clarity of exposition,in this paper all Bayesian networks are assumed to be Gaussian unless otherwise specified. The primary benefit of this simplification is the fact that uncorrelatedness implies independence, although it is by no means necessary, and many of the same results hold for arbitrary distributional assumptions.

2.1 Estimation and Identifiabiliy

While the previous section outlined algorithms that can learn the structure of a graph it will be important to characterise the conditions under which this process can be expected to converge to a correct model of the underlying DGP. In other words, in order to believe that the DAG learned from some observed data is the correct model, what do we have to assume about the underlying DGP? Pearl (2009) defines a sufficient assumption as the "back-door criterion." A set of observed variables z is said to satisfy the back-door criterion relative to x and y if:

1. no node in z is a decendent of x
2. z blocks every path between x and y that contains an arrow into x

Here a path is any combination of arcs connecting one node to another (regardless of direction), and a path between x and y is blocked if x and y are independent in the DAG given z. Intuitively, this is the concept in economics commonly described as unconfoundedness. Although



(a) Front-door criterion (b) Instrumental variables

more general, it implies in particular that even if there are variables that are relevant to the true DGP that are unobserved, the DAG can still consistently estimate the causal effect of x on y as long as x and y have no common, unobserved causes (confounders). In particular, this identifying assumption allows for the causal effect of x on y to be observed, even if there is some unobserved variable u that intermediates the causal path. This is profound in many economic applications where models assume that some unobservable function intermediates the relationships between observed variables. For example, a (unobservable) utility function intermediates the path between the observable determinants of demand such as price and preference ordering, and the quantity purchased. In this context, the front-door criterion implies that even if the true DGP contains a utility function which is unobservable to the model it is still possible correctly identify the causal effect of the demand determinants on the quantity purchased as long as all relevant determinants of the utility function are observed. The back-door criterion implies that complex functions that intermediate the relationship between observables are *emergent* in the model without being explicitly assumed.

DAGs can also consistently identify a causal effect if it satisfies the "front-door criterion." A set of variables z is said to satisfy the front-door criterion relative to variables x and y if it satisfies the following three assumptions (Judea Pearl, 2009):

1. z blocks all directed paths from x to y
2. there are no unblocked back-door paths from x to z
3. all back-door paths from z to y are blocked by x

The front-door criterion is demonstrated by figure 2a. Intuitively, what this allow s for is a sort of reverse instrumental variables identification where the instrument z intermediates the causal path from x to y instead of being a parent/determinant of x. The assumptions here are similar in scope to those made in instrumental variables. The first assumption is akin to the exclusion restriction, the second exogeneity, and the third relevance.

Both the back-door and front-door criterion make strong assumptions about unobserved variables, which raises questions about their applicability. However, this problem is hardly specific

to Bayesian networks. Empirical economists are well acquainted with the difficulty of making arguments of unconfoundedness that are also required by many traditional econometric techniques, such as the exogeneity assumption required for instrumental variables. Because DAGs are an easily scalable machine learning technique they can add a sort of "brute force" tool to the empirical economics toolbox. This technique allows for causal inference by appeal to high-dimensional data sets rather than the clever arguments and insights usually required by econometric models over a small number of observed variables.

2.2 Causality and Inference

Now that the mathematical underpinnings of DAGs have been introduced, it will be necessary to discuss the concept of causality that they employ, because it is somewhat different from what we are used to in economics. Most modern empirical work in economics utilises the "Potental Outcome" causal framework (Holland, 1986). In this framework a causal effect or treatment effect is defined as the difference between an outcome for an observational unit in the presence of a treatment $Y_i(1)$, and in the absence of the treatment $Y_i(0)$. This thinking is inspired by the medical and other physical sciences, where for example, the treatment effect of a medication on a patient's blood pressure is defined as the difference between the patient's blood pressure after taking the medication and *what it would have been* if they had not taken the medication. Since in reality we can only ever observe one of these contingencies many statistical techniques have been developed that are able to consistently estimate this amount. Therefore, the potential outcomes framework can be said to make statements about counterfactuals, that is, the difference between outcomes in different states of the world.

The concept of causality that is relevant to Bayesian networks is that of conditional independence. While this may seem unusual, this is actually akin to what is often assumed in macroeconomic theory, where every model has "exogenous shocks" that are the fundamental cause of the model dynamics. If we represent such macroeconomic models as DAGs these exogenous shocks would be the root nodes of the graph, because the root nodes of a Bayesian network are assumed to be distributed independently of all other variables in the graph. In this framework the primary meaning of causality is exogeneity (that is in the literal sense, not being determined by what is observed), rather than treatment effects as in the potential outcomes framework. Since both of these concepts of causality (treatment effects and exogeneity) are commonly used in the field of economics one would like to believe that they are internally consistent, and indeed as I will argue in the remainder of this section, they are not incompatible. Indeed, Bayesian networks are entirely consistent with potential outcomes and can be used to elicit counterfactuals / (average) treatment effects.

Barr (2018) gives an example of how the potential outcomes framework can be represented by a DAG, which is illustrated by figure 3. In this model, w is a set of confounders, x is the binary

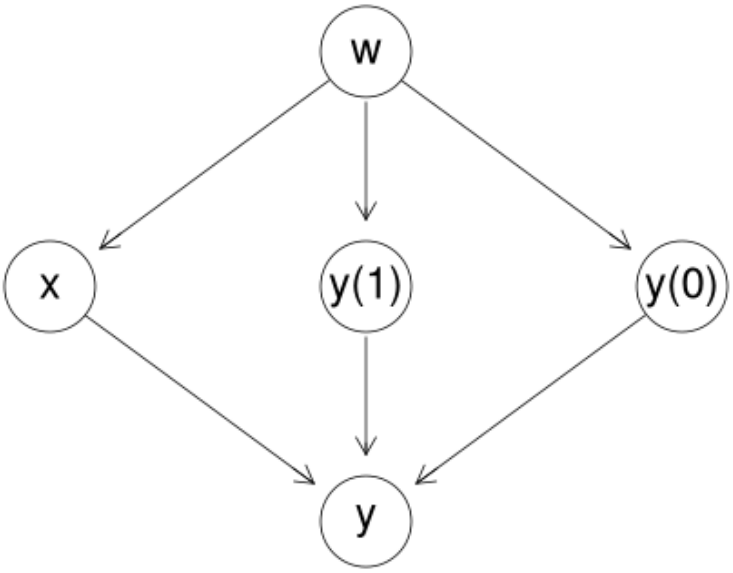


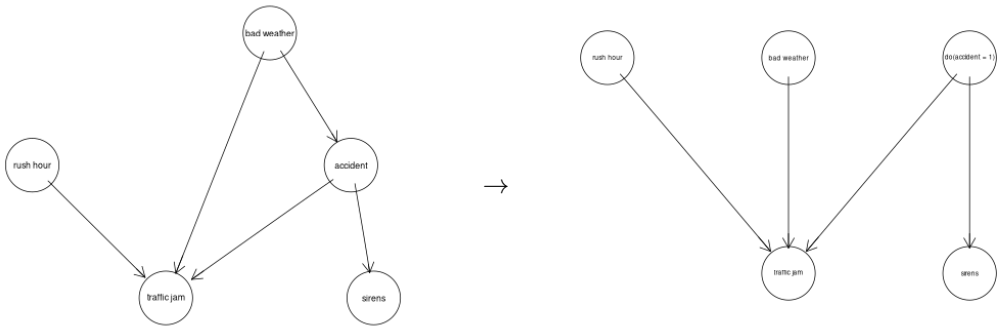
Figure 3: Potential Outcomes as a DAG

treatment of interest, $y(1)$ and $y(0)$ are the potential outcomes with and without the treatment respectively, and y is $y(x)$. The fundamental assumption necessary for the consistent estimation of the average treatment effect here is that that given the confounders w the treatment x is independent of the potential outcomes. This is the assumption of unconfoundedness between the treatment and the treatment effects:

$$x \perp\!\!\!\perp (y(1), y(0))|w \tag{3}$$

In the graph, this assumption is illustrated by the fact that the only paths from x to the potential outcomes are either through y, which is a collider which implies independence, or through w which we have controlled for. This is an example of what Pearl (2018) describes as the "back-door criterion" which he identifies as a necessary conditions for DAGs to have a causal interpretation. This illustrates the deep conceptual similarities between these frameworks.

Furthermore, Bayesian networks can be used to estimate counterfactual outcomes, using what Pearl (2014) describes as "do-calculus." This uses the notation $P(y|do(x = \bar{x}))$. The difference between this and $P(y|x = \bar{x})$ is that $do(x = \bar{x})$ reflects an exogenous change in x to $\bar{x}$, whereas $P(y|x = \bar{x})$ suggests that the model is in a state that would predict that the variable x takes on the value $\bar{x}$. Under some conditions, computing $P(y|do(x = \bar{x}))$ can achieved by breaking the links of x with its parents and setting it to $\bar{x}$, and then observing y in the model. This is demonstrated by Figure 4. For example, consider Figure 4. Suppose for simplicity that all of the variables are binary (1 in the presence of the event, 0 otherwise). On the LHS of the diagram we have the model for observed values of all of the variables. On the RHS we intervene on "accident." Notice



(a) No intervention (b) Intervention

Figure 4: An example of intervention

that doing so breaks the link between "bad weather" and "accident." We can now estimate the causal treatment effect of an accident on the probability of a traffic jam given some values of "bad weather" and "rush hour" (in other words, all else equal) according to equation 4. This equation is very familiar, and is effectively the same as the calculation of an average treatment effect in the potential outcomes framework.

$$p(tj|bw = \bar{bw}, rh = \bar{rh}, do(a = 1)) - p(tj|bw = \bar{bw}, rh = \bar{rh}, do(a = 0)) \tag{4}$$

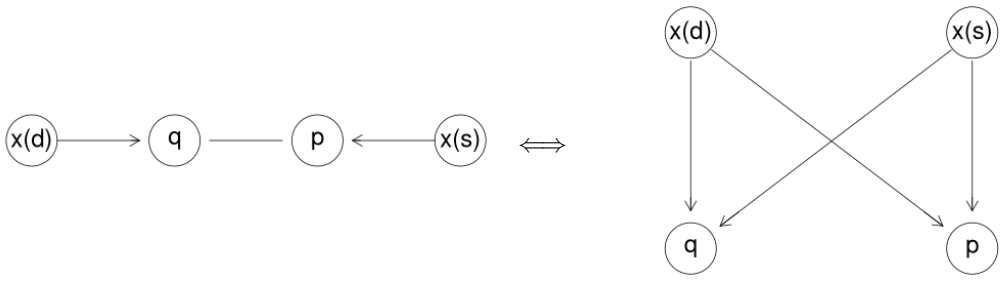
In addition, there are other kinds of possible prediction exercises that we may be interested in with some economic interpretation. Since the model defines every endogenous variable as a (linear) function of the exogenous shocks it can be interpreted as a structural model of the data. Therefore, we might compute impulse response functions (IRFs) for each endogenous variable in the model to one or more shocks.

2.3 Limitations

Before continuing I will point out some of the limitations both of this methodology, and of my own knowledge in order to give some idea of the pitfalls that needed to be taken into account in the course of carrying out this research.

2.3.1 Simultaneity

The concept of a DAG, while a powerful tool, is not a perfect model for all data. The strongest assumption is that it is directed. In many economic applications, while we may believe that some variables are truly exogenous such that they must be causes of movement in endogenous variables and not the other way around, we usually also assume that some or all of the endogenous variables are determined in general equilibrium, that is to say there is not necessarily a directionality to every relationship between endogenous variables. The problem of simultaneity is important, but there are ways which we can work with it in the Bayesian network framework. I will propose



(a) Explcit simultaneity (b) Directed simultaneity

Figure 5: An example of directing simultaneity

two solutions to this problem: the first is that many relationships that we commonly think of as simultaneous have a mathematically equivalent fully directed model, and the second is that it is possible to relax the assumption that the graph is fully directed.

In order to see why explicitly modelling simultaneity may not be necessary, consider figure 5, which is inspired by Imbens (2019). Figure 5a shows a simple model of supply and demand where quantity q and price p are determined simultaneously in the presence of demand shock $x(d)$ and supply shock $x(s)$. The relationship between quantity and price is simultaneous because changes in each one affect the other. However, the relationships implied by figure 5a can just as well be represented by the fully directed graph in figure 5b. To see this consider the following equations which are implied by figure 5a:

$$p = \alpha_p + \beta_{ps}x(s) + \beta_{pq}q + \epsilon_p \tag{5}$$

$$q = \alpha_q + \beta_{qd}x(d) + \beta_{qp}p + \epsilon_q \tag{6}$$

By substituting p into the equation for q and vice versa it can be shown that this system of equations is equivalent to:

$$p = \frac{1}{1 - \beta_{pq}\beta_{qp}}[(\alpha_p + \beta_{pq}\alpha_q) + \beta_{ps}x(s) + \beta_{pq}\beta_{qd}x(d) + (\epsilon_p + \beta_{pq}\epsilon_q)] \tag{7}$$

$$q = \frac{1}{1 - \beta_{qp}\beta_{pq}}[(\alpha_q + \beta_{qp}\alpha_p) + \beta_{qd}x(d) + \beta_{qp}\beta_{ps}x(s) + (\epsilon_q + \beta_{qp}\epsilon_p)] \tag{8}$$

Which is (a version of) what is represented by figure 5b. At this point I will note the relevance of the Lucas (1976) critique. The model in 5b is a reduced form estimation of the model in 5a, and as such, it will be impossible to identify the policy parameters β_{pq} and β_{qp}. In general, DAGs are a statistical technique that rely only on observed data, and therefore, they will not be immune to the Lucas critique. However, we *can* identify the impact of exogenous shocks to the model. In the context of this example, this means that while the DAG cannot consistently estimate the

supply and demand elasticities, it can consistently estimate the effect of a demand shock $x(d)$ or supply shock $x(s)$ on the equilibrium of the model. In the context of macroeconomic models we are often interested in computing IRFs which is the equilibrium effect of an exogenous shock to the model. This argument illustrates why even though we believe that many of the variables in these macroeconomic models are simultaneously determined, we can still estimate IRFs using DAGs. More generally, although DAGs might not be able to identify all structural parameters that economists might be interested in, they are nonetheless able to identify causal effects and in many applications this is likely sufficient. For this reason, the discussion in the applications in this paper will focus on identifying the root nodes of a graph and their effects, while less emphasis will be placed on the relationships further down the causal tree.

However, it is also possible to explicitly model simultaneity in the context of graphical models. As discussed earlier, constraint based structure learning algorithms do not force a direction onto every arc, so it is entirely possible for sturcure learning to result in a Partially Directed Acyclical Graph (PDAG). Such models are known as hybrid networks or chain graphs, originally proposed by Wermuth and Lauritzen (1990). Recall that the DAG assumption can be characterised as a set of constraints on the covariance matrix of the joint normal distribution. In a hybrid network then there are no constraints on the partition of the covaraince matrix for the variables in the model that are assumed to be simultaneous. Note however, that while this approach may be more comfortable from the economic point of view it will still not immunise the model to the Lucas critique. Unfortunately, I was unable to find any convincing implementations which allow for hybrid networks. Therefore, all of the graphs that I use in my application are fully directed and used with appeal to the previous argument.

3 Liturature Review

4 Methodology

There are two fundamental problems to solve when estimating a DAG. The first is known as "parameter learning," and the other "structure learning." Given a DAG as in Figure 1, the first task is simply to estimate the parameters of the network, such as α and β in Equation 2. This is usually done via maximum likelihood, however, other "score" functions are available such as the Bayesian Information Criterion (BIC) (Chen, Gopalakrishnan, et al., 1998).

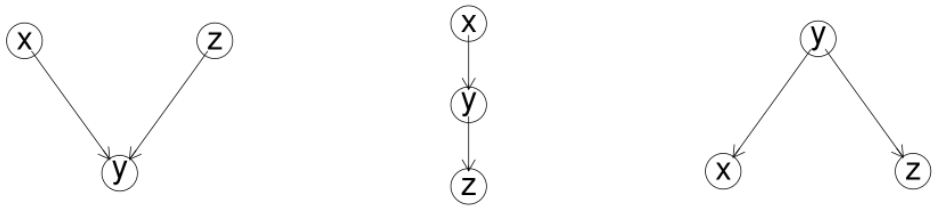
The second task, as demonstrated by Figure 6 is that if we just start with some data it is not obvious which conditional probabilities to estimate in the first place. One way of achieving this is for the researcher to specify explicitly which conditional probabilities should be present in the graph, and simply fit the parameters of that graph. This however, is not what I am particularly interested in. If this is done the researcher has effectively specified a system of linear regressions to be estimated, probably based on some economic model that they already had in mind, and

Figure 6: A DAG before structure learning

while this is then automatically encapsulated in a convenient, easily interpreted representation of the underlying assumptions, it seems nothing of profound economic significance is achieved in this case.

A more more exciting approach is to algorithmically learn the structure of the graph, that is to learn a structural model, directly from observed data. One "brute force" method to solving this problem is to compute the posterior likelihood of every possible network, however, this number is super-exponential in the number of variables such that it becomes very computationally expensive, very quickly (Chickering, 1996). As a response to this, many heuristic approximation techniques have been developed. These can be grouped into two categories: constraint-based and score-based structure learning algorithms (Spirtes & Glymour, 1991) (Verma & Pearl, 1991).

Constraint-based algorithms rely on the fact that changing the direction of an arc changes the conditional independences implied by the graph, the presence of which can be tested for in the data. To see how the DAG assumptions can be sufficient to learn a causal model in this way, consider the example in figure 7. Suppose we have a graph with three nodes, such that no one node is completely independent from the other two (as this would make the graph trivial, and we could in any case rule out this case with an independence test). Furthermore, the graph cannot have all three possible arcs because it would either contain a cycle, or the third arc would imply a relationship which is redundant given the other two. Then the graph must have exactly two arcs. Given this, there are exactly three possible permutations of the network, which are the three shown in figure 7. These are known as the the three canonical "v-structures." (Judea Pearl, 2014) These structures are partially identifiable from observational data because they imply different testable hypotheses about conditional independence. While the chain and fork imply that x and z are unconditionally



$x = \epsilon_x$
$y = \alpha_y + \beta_{yx}x + \beta_{yz}z + \epsilon_y$
$z = \epsilon_z$

(a) Collider

$x = \epsilon_x$
$y = \alpha_y + \beta_{yx}x + \epsilon_y$
$z = \alpha_z + \beta_{zy}y + \epsilon_z$

(b) Chain

$x = \alpha_x + \beta_{xy}y + \epsilon_x$
$y = \epsilon_y$
$z = \alpha_z + \beta_{zy}y + \epsilon_z$

(c) Fork

Figure 7: The three possible v-structures of a 3 node DAG. Error terms ϵ are all i.i.d. Gaussian shocks.

dependent and only independent conditional on y, the collider implies exactly the opposite; that x and z are unconditionally independent and dependent conditional on y. Given some observed data we can easily test for the presence of conditional and unconditional independence using a χ^2 test. The results of these tests can be used to rule out certain network structures which would be inconsistent with the observed data. Although for every set of three variables the network is only partially identifiable, full identification can (but will not always) be achieved when more variables are observed, by comparing overlapping triplets of variables and progressively reducing the set of network structures that are consistent both with the DAG assumptions and with the observed conditional independences.

Score-based methods as the name implies assign some score to every network based on its predictive accuracy and then use gradient-decent to identify the optimum network structure. There are a number of scoring functions and hill climbing algorithms that one can use to achieve this.

The major benefit of the constraint based method is that it directly utilises conditional independence as a primitive, which is the concept of causality that Bayesian networks seek to identify. This is in contrast to score base methods, which effectively maximise the predictive accuracy of the model, and there is seemingly no guarantee that the most predictive model is the most likely causal explanation. The major benefit of score based methods on the other hand is that they will always converge to a single fully directed graph as a solution whereas constraint based methods, because V-structures are only partially identifiable, may not be able to identify a unique solution. Instead, when the graph is only partially identifiable, the algorithm will return an undirected graph, because that arc could take on face either direction and the graph would still be consistent with both the DAG assumption and the observed conditional independences. By permuting the two possible directions of each undirected arc we arrive at a set of graphs that are said to be "observationally equivalent." This is problematic because it is difficult or impossible to fit parameters to graphs that are not fully directed (see limitations section).

Fortunately, these two methods can be combined into so called "hybrid" structure learning methods which use the strengths of both methods to counter the weaknesses of the other (Marco Scutari, Howell, Balding, & Mackay, 2014). In this method a score function of choice is maximised over the set of network structures that is allowable given some constraint based algorithm. This has the benefit of using conditional independence to identify causality, while also always converging to a unique, fully directed graph.

I believe that structure learning is the greatest contribution of the Bayesian network framework. Many of the topics in this paper should seem quite familiar to any econometrician because they are fundamentally the same as some basic econometric concepts, although perhaps expressed through different language. However, the novel benefit of using this method is that we are effectively able to estimate a structural model directly from the data, without first having to specify which relationships we believe should be present. This is a very powerful concept because it removes researcher bias by allowing the data to speak for itself.

For the application in this paper I have used the "bnlearn" package (Scutari, 2010) for R. I used the provided hybrid structure learning algorithm, rsmax2 with the pc.stable structure restriction algorithm and the default score function which is the Bayesian information criterion. The parameters of the model are then fit via maximum likelihood.

5 Data

In order to demonstrate the capability of the Bayesian network methodology empirically I have chosen to use simulated data from macroeconomic models. There are a few key reasons why I have chosen to work with this data. Firstly, since the model that simulates the data is known it is possible to evaluate whether the structure learning has succeeded in identifying the underlying relationships in the data. In other words, since the true DGP is known it is possible to infer whether the estimated DAG correctly represents the underlying DGP. Secondly, in the context of a log-linearized macroeconomic model the distributional assumption that conditional distributions are linear with Gaussian errors, and structural assumption that the DGP is a DAG are in fact correct. Finally, since these models are central to modern macroeconomics it provides an excellent opportunity to demonstrate the applicability of Bayesian networks in economic applications.

In order to collect simulated data I consulted a github repository containing Dynare code to replicate a number of well known macroeconomic models (Pfeifer, 2020). In particular, I chose to model the baseline RBC model as a simple case and the Smets and Wouters (2007) model for a more difficult and complex modelling challenge. I modified the simulation code slightly such that rather than impulse response functions the output of the simulation would be a file containing (n) observations of i.i.d. draws of the exogenous shocks, and the associated observed values of the other variables in the model. This file was then used as the input for fitting Bayesian networks.

Symbol	Name	Exogenous?
g	government spending	yes
z	technology process	yes
k	capital	yes
w	wage rate	no
r	return to capital	no
y	output	no
c	consumption	no
l	hours worked	no
i	investment	no

Table 1: Description of Variables

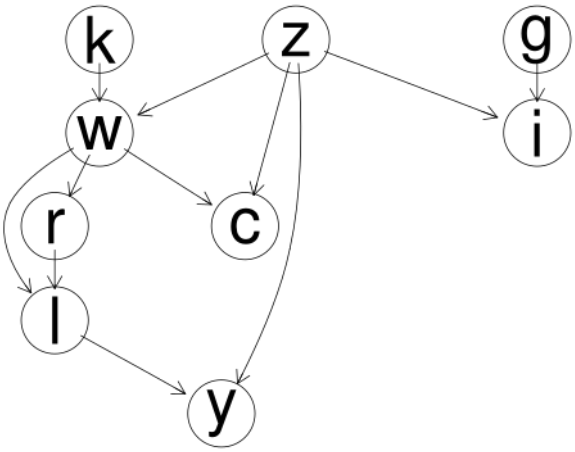


Figure 8: Structure of DAG fit to RBC data

6 Results

6.1 RBC

The data that I have used in this comes from a relatively detailed RBC model with a good number of variables to study. Table 1 gives a summary of the variables in the data. In this model "z" and "g" are exogenous and each follow and independent i.i.d. Gaussian process. Capital is also exogenous in period t because it is chosen (endogenously) in period t-1. In general, we could include lags of some or all variables in order to take dynamics into account, but I will leave that for future investigation. For the sake of space, I will forego any further discussion of the relationships between endogenous variables, stating simply that these are of the standard RBC nature.

Figure 8 shows the structure of the DAG that was fit to the RBC data, with a sample size of 200. Here we note that structure learning successfully identified the three exogenous variables in the model as root nodes. Since the graph is fully directed we are able to fit parameters to the model in order to perform predictions. For example, we can perform pure prediction on output. In the spirit of best practice for machine learning we split the data into training and test data sets, and use only the training data to fit the parameters of the model. When doing this we find that the accuracy of predicting any endogenous variable in the test data is extremely high ($R^2 \approx 1$),

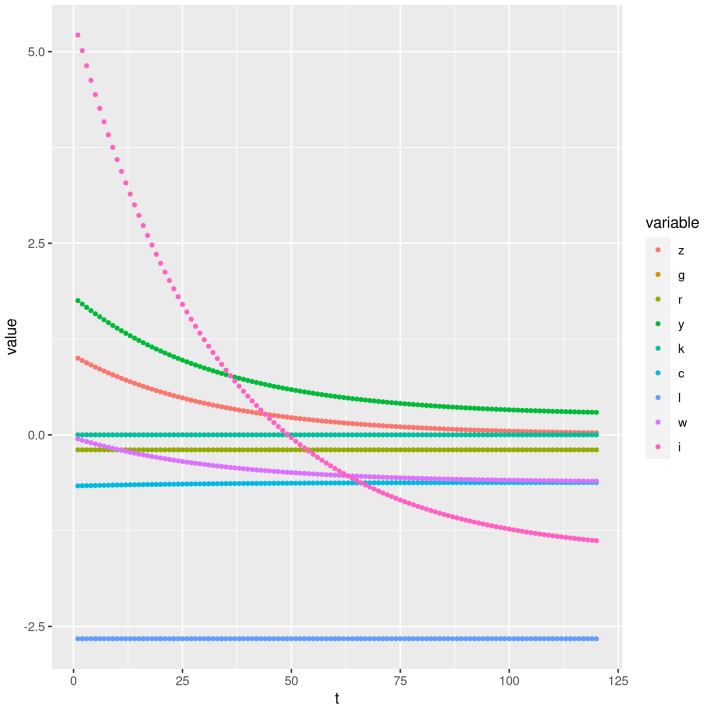


Figure 9: IRFs to technology shock predicted by model

although it should be noted that the variables are perfectly co-linear, so this prediction exercise is not exactly challenging.

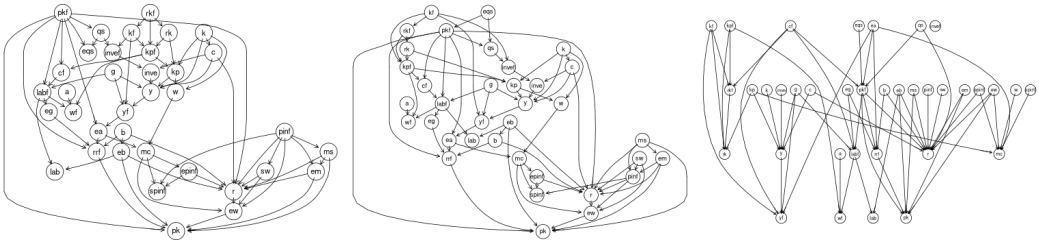
A more interesting exercise is to compute IRFs for the endogenous variables. Figure 9 shows the result of computing IRFs for endogenous variables to a positive $AR(1)$ technology shock of one standard deviation. This is calculated by setting the value of "z" to decay by α_z every period from an initial value of 1 standard deviation, setting the other shocks to 0, and then calculating the predicted values for all of the endogenous variables. Therefore, each period we are calculating the estimated treatment effect on every endogenous variable of an exogenous technology shock, all else equal. While not perfect, qualitatively, the results are much what we would expect to see from an RBC model. For example, we can observe the standard result that investment reacts much more strongly than other variables to a technology shock.

6.2 Smets and Wouters (2007)

6.2.1 Model

This model from this seminal paper is quite complex and contains a large number of variables and thus it provides difficult challenge with which to demonstrate the capabilities of the Bayesian network methodology.

This model contains seven exogenous shocks: a productivity shock (ea), a risk premium shock (eb), a government expendiature shock (eg), an investment specific technology shock (eqs), a mone-



(a) Unconstrained (b) Observationally Equivalent (c) Forced Exogenous Roots

Figure 10: Three different approaches to structure learning with the Smets and Wouters model.

tary policy shock (em), a price markup shock (epinf), and a wage markup shock (ew). Furthermore, these shocks all contribute to the AR/MA process of the stock of technology (a), risk premium (b), government spending (g), investment specific technology (qs), and, money supply (ms), price shock (spinf), and wage shock (sw), all of which are exogenous from a cross sectional perspective. The model contains a number of state variables [1] that are exogenous from the cross sectional perspective: flexible economy capital services (kf), capital stock (kpf), investment (invef), consumption (cf), as well as nominal rigidity economy capital services (k), capital (kp), investment (inve), consumption (c), inflation (pinf), and, wages (w). Taken together the model contains 24 variables that are exogenous from the cross sectional perspective.

6.2.2 Structure Learning

We now consider the extent to which Bayesian network structure learning is able to identify these exogenous variables. Before beginning some light data cleaning was performed. Difference variables were removed because they contain information about the past, as for now, we wish to consider cross-sectional behavior. Finally, "observed" values were dropped as these are redundant. After these simplifications 36 variables remained in the data, for which 10000 observations were made available. Figure 10a shows the result of performing the hybrid structure learning algorithm on this data. This graph has 7 root nodes: rkf, pkf, k, pinf, a, b, and, g, 5 of which are in fact exogenous. The two endogenous variables that are identified as root nodes are rkf and pkf. This result is not optimal, however, there are some tricks we can use to arrive at a more sensible model.

One simple way to improve the model is to consider the class of observationally equivalent models which can be achieved by changing the direction of some arcs in such a way that the resulting graph is still consistent with the observed v-structures. For example we can reverse the arc $b \rightarrow eb$ because clearly the causality runs the other way around. Making all of the possible corrections we arrive at figure 10b. The root nodes of this graph are kf, k, a, g, ms, sw, eb, eqs, all of which are in fact exogenous. This is a much better result in the sense that it achieves no "false-positives" for exogeneity, however, it does not identify all exogenous variables. Intuitively, this is possible because there are variables in the model which are nearly colinear with

[1] As the state variables of the model are never explicitly stated in the original paper, I take them to be the set of all variables whose value is a function of its own lagged value in the linear model.

the exogenous shocks (for example the correlation between eb and b is 0.99, because preference shocks demonstrate very little time series behavior in the calibration of the model). If this is the case then the variables effectively measure the same quantity, and thus lack inherent causality, so the choice of arc direction between them is essentially arbitrary. Therefore, when using Bayesian networks one should be weary of colinearity in the observed data. Through some experimentation I have found that selectively excluding variables that appear to be colienar with another can greatly improve the structure learning results, however, I have excluded these results until I can implement this feature selection in a systematic way.

Figure 10c shows a stucture that was learned after implementing a blacklist that prevents any of the exogenous variables from having parents, effectively guaranteeing that the set of root nodes will be exactly the set of exogenous variables in the model. Since this is rather trivial with simulated data the purpose of including this model is to provide a benchmark to evaluate the performance of the algorithmically developed structures.

6.2.3 Evaluation

7 Conclusion

References

Barr, I. (2018). Causal inference with python. Retrieved June 2, 2020, from http://www.degeneratestate.org/posts/2018/Jul/10/causal-inference-with-python-part-2-causal-graphical-models/

Chen, S., Gopalakrishnan, P. et al. (1998). Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proc. darpa broadcast news transcription and understanding workshop* (Vol. 8, pp. 127–132). Virginia, USA.

Chickering, D. M. (1996). Learning bayesian networks is np-complete. In *Learning from data* (pp. 121–130). Springer.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945–960.

Imbens, G. W. (2019). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. arXiv: 1907.07271 [stat.ME]

Liszka, J. (2013). Bayesian networks and causality. Retrieved April 7, 2020, from http://blog.jliszka.org/2013/12/18/bayesian-networks-and-causality.html

Lucas, R. E. et al. (1976). Econometric policy evaluation: A critique. In *Carnegie-rochester conference series on public policy* (Vol. 1, pp. 19–46).

Pearl, J. [Judea]. (2009). *Causality*. Cambridge university press.

Pearl, J. [Judea]. (2014). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Elsevier.

Pearl, J. [Judea], & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Pfeifer, J. (2020). Dsge_mod. Retrieved April 8, 2020, from https://github.com/JohannesPfeifer/DSGE_mod

Scutari, M. (2010). Bnlearn: Bayesian network structure learning. *R package*.

Scutari, M. [Marco], Howell, P., Balding, D. J., & Mackay, I. (2014). Multiple quantitative trait analysis using bayesian networks. *Genetics*, *198*(1), 129–137.

Smets, F., & Wouters, R. (2007). Shocks and frictions in us business cycles: A bayesian dsge approach. *American economic review*, *97*(3), 586–606.

Spirtes, P., & Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, *9*(1), 62–72.

Verma, T., & Pearl, J. [J.]. (1991). *Equivalence and synthesis of causal models*. UCLA Computer Science Department. Retrieved from https://books.google.co.uk/books?id=ikuuHAAACAAJ

Wermuth, N., & Lauritzen, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *Journal of the Royal Statistical Society: Series B (methodological)*, *52*(1), 21–50.