Bayesian networks with Applications to Simulated Macroeconomic Timeseries

Emmet Hall-Hoffarth

May 2020

1 Introduction

Bayesian networks (Judea Pearl & Mackenzie, 2018) (Judea Pearl, 2009) 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 this field has in some sense still not reached maturity, the potential of automatically identifying causality opens the door for numerous useful economic applications. To my knowledge, little work has been done in this area within the econometrics literature. Indeed, Imbens, (2019) notes a lack of concrete empirical examples demonstrating the usefulness of this method in the field of economics. Therefore, In this paper I will outline my plan to explore one such potential application in my MPhil thesis. The application will be simulated macroeconomic models.

The first section of this paper will explain what Bayesian networks are, how they are estimated, and what they can be used for. The second section will examine an application that I have constructed using simulated macroeconomic data. Finally, the paper will conclude with some remarks on some areas that I plan to investigate during the course of my research.

2 Bayesian networks

2.1 Primatives

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 is distributed conditional on B, whereas B's distribution is unconditional. In economic language this may be interpreted as meaning that B is exogenous while

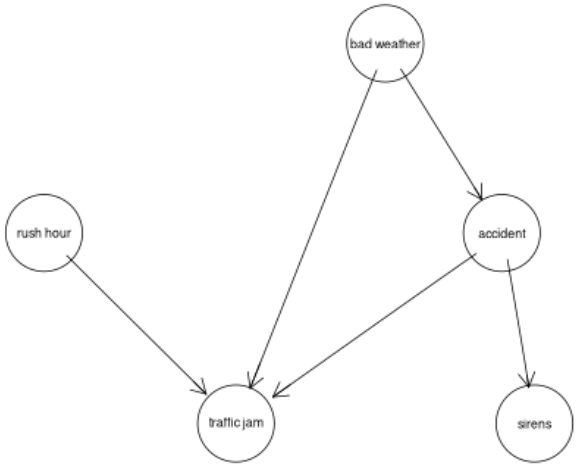


Figure 1: An example of a simple DAG (Liszka, 2013)

A is endogenous to or determined by B. As the name DAG implies, arcs are assumed to not create any cycles in the graph. 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 we are generally dealing with continuous variables. This is possible as long as we are willing to make some assumptions about the nature of the conditional probability (as truly non-parametric estimation of continuous variables implies an intractably large search space). 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 the familiar form that conditional distributions are linear functions of conditioning variables with Gaussian errors, which is exactly the assumptions of simple, small-sample OLS regression. Such models are sometimes known as "Gaussian Bayesian networks." (GBN) For example:

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

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 variance-covariance matrix. For example, if there is no arc between to variables, then this implies the restriction that their covariance is zero.

2.2 Causality

Before diving into the mathematical specifics of Bayesian networks, it is 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, and in the absence of some treatment. 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.

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. This is only a subtle nuance, and indeed, it is my belief that these two concepts are not incompatible. As I will detail in section 2.4 how Bayesian networks can be used to elicit what could rightly be described as a(n average) treatment effect.

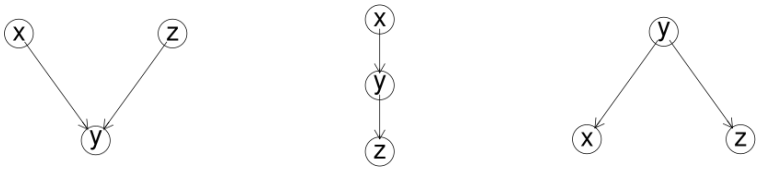
Figure 2: A DAG before structure learning

2.3 Estimation

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 2 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 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



(a) Collider	(b) Chain	(c) Fork
$x = \epsilon_x$	$x = \epsilon_x$	$x = \alpha_x + \beta_{xy} y + \epsilon_x$
$y = \alpha_y + \beta_{yx} x + \beta_{yz} z + \epsilon_y$	$y = \alpha_y + \beta_{yx} x + \epsilon_y$	$y = \epsilon_y$
$z = \epsilon_z$	$z = \alpha_z + \beta_{zy} y + \epsilon_z$	$z = \alpha_z + \beta_{zy} y + \epsilon_z$

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

(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 3. 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). In this case the assumption that the graph must be acyclical implies that every node cannot be directly connected to every other node. 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 3. These are known as the the three canonical "V-structures." (Judea Pearl, 2009) 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 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).

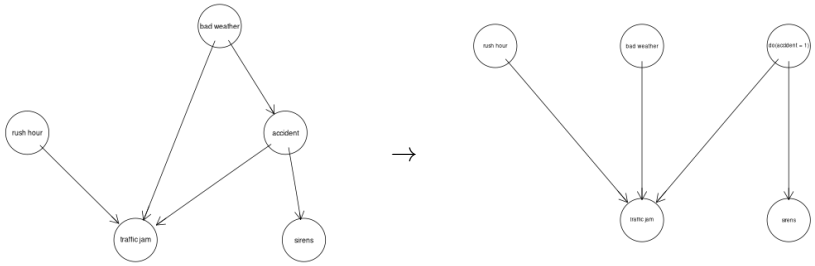
Fortunately, these two methods can be combined into so called "hybrid" structure learning methods which use the strengths of both methods to counter the weaknesses of the other. In this method a score function of choice is maximised over the set of network structures that is allowable given some constraint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

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.

2.4 Inference

Once a DAG has been fit to some data, there are a number of useful inferences that can be made.

Firstly, nodes in the DAG which have no parents are known as "root-nodes." In the context of an economic application we can interpret these as the exogenous variables of the model. Since there is nothing to condition them on set β to 0 in Equation 2, and observe that any such variables are implicitly modelled as i.i.d. Gaussian shocks. In particular, in the application I will seek to identify the exogenous variables in a macroeconomic simulation as the root-nodes of the associated Bayesian network.



(a) No intervention (b) Intervention

Figure 4: An example of intervention

Furthermore, Bayesian networks can be used to estimate counterfactual outcomes, using what Pearl (Judea Pearl, 2009) describes as "do-calculus." Since the DAG specifies conditional independences, it is possible to (exogenously) modify one or more values of input data to some alternative value, and then calculate the predicted values of some other outcomes in the model in order to observe what the model predicts "would have happened" in that scenario. Furthermore, we could predict what might happen in any fabricated future scenario, as long as all of the exogenous variables are given some value. Assuming the DAG is the correct model of the true DGP (in fact, the conditions are weaker than this, see "faithfulness" in (Judea Pearl, 2009)), then the effect on outcomes of such interventions can be given causal interpretations.

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 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 the equation:

$$p(tj|bw = b\bar{w}, rh = r\bar{h}, do(a = 1)) - p(tj|bw = b\bar{w}, rh = r\bar{h}, do(a = 0)) \quad (3)$$

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.

2.5 Limitations

Before continuing on to my application I would like to 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 might be encountered during the course of this research.

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. This problem while important, is probably not a show stopper. Structure learning algorithms do not force a direction onto every arc, it is merely optional. When some arcs in the model are undirected it is known as a, "hybrid network." However, it may be difficult to estimate the parameters of hybrid networks, because in the case of a bidirectional relationship it is unclear what parameters the model should specify. This is the problem of simultaneity that we are familiar with in econometrics. However, we can probably get around this by modelling the joint distribution of the simultaneous variables, conditional on the parent nodes.

Through my experimentation with these methods I have also had a lot of mixed results. I have tried to perform this type of analysis for a few models, and it does not seem to work well for all of them. In the next section I will discuss the results for a baseline RBC model for which the results have been rather promising. On the other hand, I have attempted to model the medium-scale New Keynesian model from (Smets & Wouters, 2007), without much success. In this model there are a large number of variables, many of which seem to actually measure the same thing, so it is hardly surprising that the algorithm becomes confused. Therefore, at this time I will not completely write this one off. On the contrary, I bring this up to point out that while this is potentially a very useful modelling technique it is not a "magic bullet." Some assumptions on input data will be required. It is likely that some substantial part of my thesis will be dedicated to understanding in what situations Bayesian networks work well, as well as what assumptions might be needed to guarantee that a faithful DAG can/will be found.

Upon first being exposed to these ideas many are critical whether it is even conceptually possible to make such causal inferences directly from data. The section on structure learning gives some very brief overview of how we try to go about practically solving this problem. As far as the conceptual problem is concerned I do not yet understand it well enough myself to make a compelling argument. It is clear that going forward with this research will require me to do much deeper reading about the justification of direct structural learning. However, being the pragmatist that I am, it is for this reason that I have chosen to perform some direct applications in order to demonstrate simply whether or not such a strategy is likely to bear fruit, regardless of the strength of its

theoretical underpinnings. The results so far have not been perfect, but I believe it has been successful enough to warrant further investigation into this topic.

3 Application

3.1 Data

In order to demonstrate the capability of the Bayesian network method empirically I have chosen to use simulated data from macroeconomic models. This is not completely arbitrary; I envisage using simulated along with real macroeconomic data in my actual thesis. 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 faithfully represents the underlying DGP. Secondly, in the context of a log-linearized macroeconomic model the parametric assumptions (that conditional distributions are linear with Gaussian errors) are in fact correct. Finally, and more personally, given the choice of an application in microeconomics or macroeconomics I prefer one in macroeconomics because that is what I find more interesting.

3.2 Methodology

In order to collect this data I have found a repository of Dynare code for simulating many well-known macroeconomic DSGE models (Pfeifer, 2020). I then modified this code slightly such that the output would be a data file containing simulated values of i.i.d. shocks and endogenous variables. I then load this data into R, and using the "bnlearn" package (Scutari, 2020), I fit a Bayesian network, using a hybrid algorithm to learn the structure because there are theoretical advantages that I outlined in the section on structure learning and it seemed to perform best in my short experimentation. Finally, I estimate the parameters of the model using maximum likelihood. Once the model is fit I perform a few experiments to evaluate the model performance.

3.3 Results

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.

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

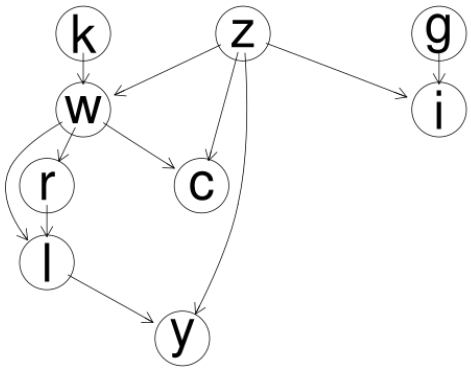


Figure 5: Structure of DAG fit to RBC data

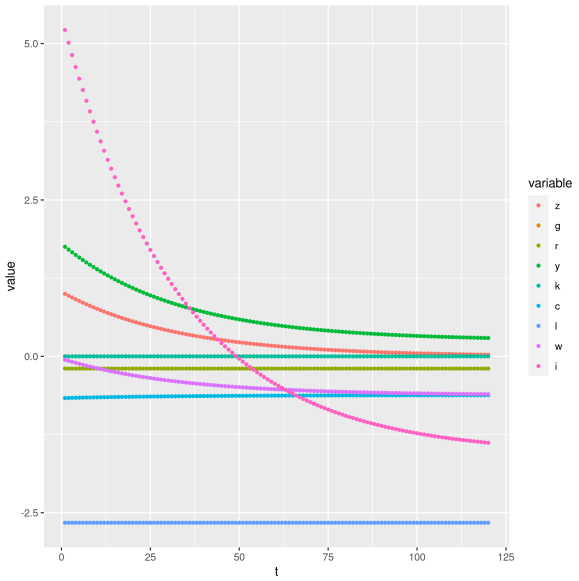


Figure 6: IRFs to technology shock predicted by model

Figure 5 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 ($R^2 \approx 1$), 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 6 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[1]. 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.

[1]Note that "l" is excluded because the model thinks it is exogenous.

4 (Sketch) Research Plan

There are a few ways that I can see taking this research, both theoretically and empirically. Imbens (2019) gives a detailed account of many of the limitations to the application of DAGs in economics, and therefore, I see a great opportunity to make a significant contribution by addressing some or all of these issues in my thesis. Some of the main issues he raises, which I have already outlined my plan to address in this paper are the lack of empirical examples, inability to perform inference, and inability to handle simultaneity.

In addition to this, I would like to look at trying to identify under which conditions structure and parameter learning algorithms are likely to succeed in finding an accurate model for the data. I will investigate the various structure learning algorithms in each category and compare them theoretically and empirically. Ideally I would be able to specify precise conditions under which there theoretical guarantees of convergence can be given. This need not be done from scratch because many of these theoretical properties are already understood in the statistics literature. The key would therefore be to translate these properties into relevant constraints on economic models.

In general, I will need to do a good amount of work in order to relate these new concepts to existing ones in the economics literature. For example, we can think of the Bayesian network as imposing certain restrictions on the covariance matrix of the assumed joint normal distribution of observed variables. Therefore, in order to get a better understanding of the economic content of these assumptions it would be a worthwhile exercise to describe in general how these restrictions compare to a Cholesky decomposition, especially in the case where the Bayesian network is dynamic.

If it is impossible to guarantee structure learning algorithms will perform reliably, and they do not seem to do so empirically, I could continue by fully or partially (by blacklisting certain arcs), manually specifying the network structure. While I think this would be less interesting (because I have already argued that structure learning is probably the most exiting thing about Bayesian networks), I still think this research would be worth pursuing because there are quite a few things that can be achieved once a faithful DAG is fit to the data. Since it is a structural model we can use it to calculate IRFs and compare to the original simulations to test the accuracy of the model. We can also generate random samples from the network and compare the statistical properties of these data to those of the original data.

Finally, regardless of what form the final thesis takes I intend to take it to real data. I have decided to start with simulated data because knowing the true DGP gives the opportunity to test whether structure learning is working. However, if I can succeed in making a strong argument that these algorithms can learn effectively under some conditions, then obviously the most interesting thing will be to take it to real data and see what sort of implications the model gives us.

References

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

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

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. (2020). Bnlearn. Retrieved April 8, 2020, from https://www.bnlearn.com

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