DAGs with Applications to Macroeconomic Timeseries

Emmet Hall-Hoffarth

July 2020

1 Introduction

Directed Acyclical Graphs (DAGs) (Pearl, 1995) are a form of graphical model that are particularly useful for modelling causal probabilistic relationships. This framework has many benefits, chief among them that under some conditions one can infer a DAG and thus an causal model of a Data Generating Process (DGP) directly from observed data. This property opens the door for numerous useful economic applications. However, Imbens (2019) notes a lack of concrete empirical examples demonstrating the usefulness of this methodology in the field of economics, citing some potential drawbacks and pitfalls. Therefore, the purpose of this paper is to investigate the applicability of DAGs in economic applications, and to consider an empirical example in the field of macroeconomics.

The application involves modelling simulated data from well-known macroeconomic DSGE models, as well as real macroeconomic timeseries. 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 DAGs. Furthermore, in this controlled environment it is possible to ensure that the assumptions required by DAGs 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 DAGs can be applied in economics.

The remainder of this paper will be organised as follows. Section 2 will provide a review of the relevant liturature on both DAGs and DSGE models, relating the two to each other where appropriate. Section 3 will introduce the methodology, that is, the specific implementation of DAGs used in the empirical study. Section 4 will discuss the data used in the empirical study. Chapter 5 will present the results of the empirical study. Section 6 will conclude and discuss avenues for future research.

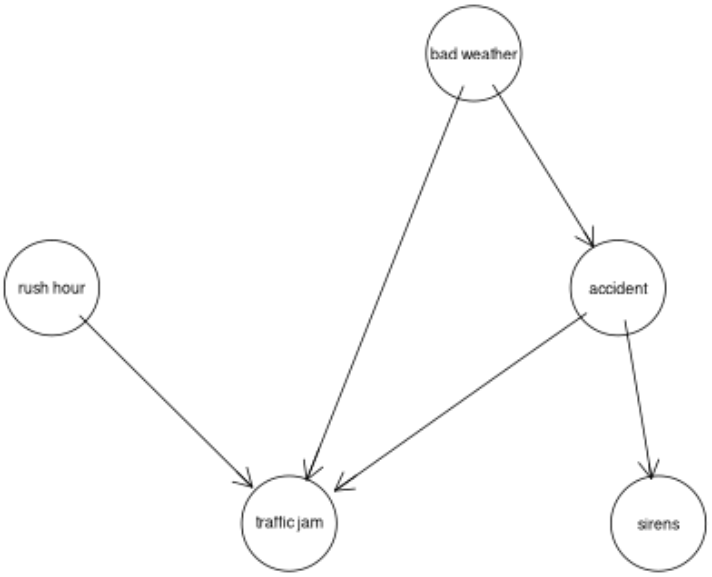


Figure 1: An example of a simple DAG (Liszka, 2013)

2 Liturature Review

2.1 DAGs

The fundamental assumption of a DAG 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. 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 DAGs 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 Bayesina 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.

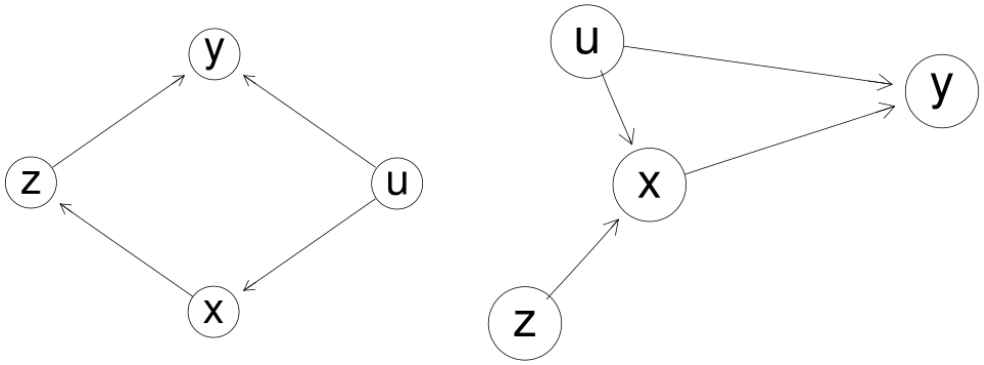
When fully specified, a GBN consists of a system of linear equations that defines the joint distribution of the data. Because of the properties of the normal distribution, this means that we can express a GBN as a single joint normal distribution over the data, where the DAG specifies the exact restrictions that are imposed on the covariance matrix. In order to enhance clarity of exposition, in this paper all DAGs 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.

2.1.1 Identification of Causal Effects

This section considers the situations in which an observationally constructed DAG can correctly identify causal effects. 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), a backdoor path is a path that is not directed, 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 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,



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

even if there is some unobserved variable u that intermediates (lies along) 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 back-door criterion implies that even if the true DGP contains a utility function which is unobservable to and even ignored by the observational DAG 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.

Of course, the backdoor critierion is limited in the same way as any unconfoundedness identification strategy: it is difficult to argue and in many cases implausible that there is no backdoor path (unobserved confounder) between the causal variables of interest. However, 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 (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 DAGs. 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 where assumptions such as uncondoundedness are more plausible rather than the clever arguments and insights usually required by econometric models over a small number of observed variables.

2.1.2 Identification of DAGs

We now consider the assumptions which must be made of a DGP for it to be correctly modelled by as a DAG. Pearl (2009) identifies the relevant condition as *stability* or *faithfulness*. Under this condition there exists a minimal structure of the true DGP which is a DAG. Faithfulness is the assumption that the conditional independence relationships in the DGP are robust to changes in the values of parameters. An equivilent statement of the condition is that a DAG is *faithful* to a given DGP if the statistical independences in the DGP are exactly those implied by the DAG (Spirtes et al., 2000). This assumption is only violated if some causal effects exactly cancel out, resulting in no observed correlation between casually connected variables. Pearl (2009) provides the following example demonstrating how faithfulness fail. Consider the following model: $z = \beta_{zx}x+\epsilon_x$, $y = \beta_{yx}x + \beta_{yz}z + \epsilon_y$. If we impose the parameter restriction $\beta_{yx} = -\beta_{yz}\beta_{zx}$ then x and y are independent. However, this independence relationship is not robust to purturbations of the model parameters and is therefore not stable in the relevant sense.

A sufficient condition for faithfulness is that the DGP parameters are jointly continuous over the parameter space (Steel, 2006), or equivilently, that the matrix of DGP parameters is of full rank. This is because under this condition, specific combinations of parameters which result in the cancelation of causal effects have a probability of 0 of occuring, even if they are possible. If we believe that the true DGP of the macroeconomy is DSGE model, which itself is faithfully represented by a DAG, then this condition is unlikely to be met. DSGE models impose many cross-equation restrictions on parameters that effectively reduce the rank of the parameter matrix. Unfortunately this condition will not allow us to guarantee that DSGE models satisfy the faithfulness assumption. However, if when considering real macroeconomic data we put aside the assumption that the true DGP is a DSGE model, it does not seem *a priori* unreasonable that the parameter relating, for example, captial and output, and the parameter relating consumption and technology vary independently in different populations. Regardless, this condition is merely sufficient, not necessary, and so it does not rule out that DSGE models can be faithfully represented by DAGs.

In another approach to failures of faithfulness, Steel (2006) notes that such failures or near-failures (that is near-zero statistical dependence despite clear causal pathways) are likely occur when parameters are both subject to *selection* and *homogeneity*. In this context, selection has its

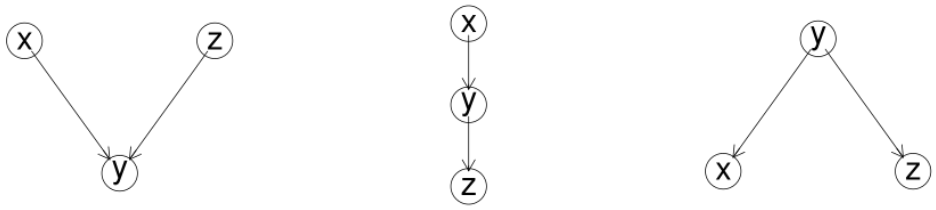
Figure 3: A DAG before structure learning

standard economic meaning. The suggestion is that if the path of a policy variable z is specifically designed as a function of x to counteract the causal effect of x on some outcome y, then it is reasonable to believe that little or no correlation will be observed between x and y despite a clear causal pathway between them. Homogeneity means a lack of exogenous variation in parameter values. If *both* selection and homogeneity occur, failure or near-failure of faithfulness is likely to occur. Within the context of macroeconomics, this seems likely to be the case when considering interest rates and the actions of central banks. Assuming the interest rate is set according to a Taylor (1993) rule, the parameters of that rule are chosen with the specific intent and canceling the causal effect of inflationary shocks on output and minimizing exogenous variation. This example seems to satisfy both selection and homogeneity. On the other hand, many parameters of the macroeconomy, in particular real ones, are plausibly not prone to this type of selection (for example, the autoregressive coefficient on an exogenous technology process). This distinction will be important when considering the empirical results in this paper.

2.1.3 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 3 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



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

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

is for the researcher to specify explicitly which conditional probabilities should be present in the graph, and simply fit the parameters of that graph. How this can be done in the context of DSGE models is discussed in Section 2.2. 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 (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 4. 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 4. These are known as the the three canonical "v-structures." (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 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. There are many that have been implemented using this general approach, the most popular of which is the PC algorithm first developed by Spirtes et al. (2000). This algorithm has been shown to consistently estimate (as $n \to \infty$) the structure of the ground truth DAG of observed data under the assumptions of linear-Gaussian conditional probability functions, stability (as discussed in 2.1.2), and structural complexity that does not grow too quickly relative to n (Kalisch & Bühlmann, 2007).

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 can be used to achieve this. A consistency result for the GES score-based algorithm is given in Chickering (2002). The assumptions are slightly stronger than than that of the PC algorithm — the number of variables must be fixed rather than growing slowly relative to n.

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 DAGs 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. In other words, despite the presence of large sample consistency results for both types of algorithms, it seems likely that small sample bias is more likely to be a problem for score-based methods. 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 or CPDAG, 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.

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 (Scutari et al., 2014) (Friedman et al., 2013). In this method the algorithm maximises a score function, but the number of parents that each node can have is restricted. The main benefit of this is a large gain computation efficiency because the search space is dramatically reduced, and theoretically it has the benefits of both constraint based and score based learning. However, while resulting the graph is always directed, it does not always correctly reflect the observed v-structures because it trades off constraint satisfaction and score maximisation. Nandy et al. (2018) gives an asymptotic consistency result for the ARGES hybrid learning algorithm.

2.1.4 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 DAGs 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 DAG 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, DAGs 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 5. In this model, w is a set of confounders, x is the binary

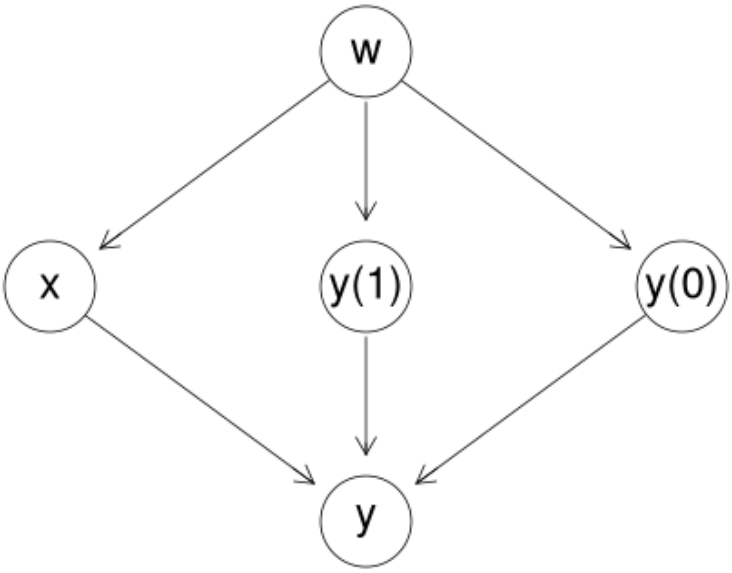


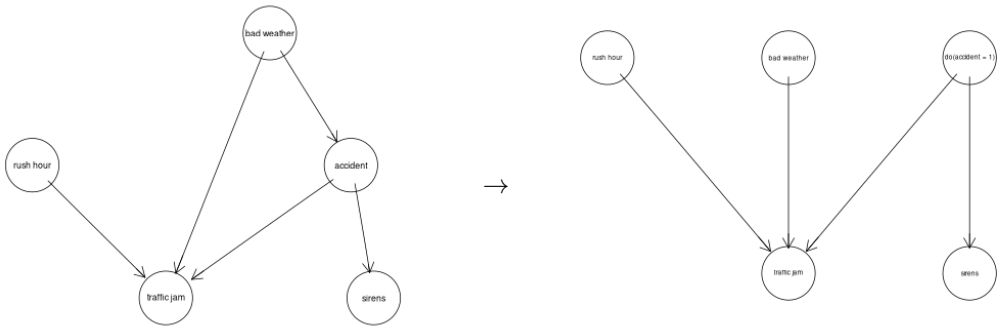
Figure 5: Potential Outcomes as a DAG

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, DAGs 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 6. For example, consider Figure 6. 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



(a) No intervention (b) Intervention

Figure 6: An example of intervention

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.

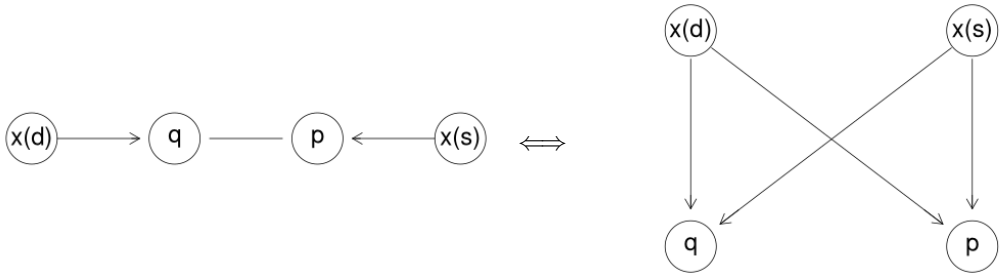
$$p(tj|bw = \bar{bw}, rh = \bar{rh}, do(a = 1)) - p(tj|bw = \bar{bw}, rh = \bar{rh}, do(a = 0)) \tag{4}$$

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.

2.1.5 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 DAG framework. I will propose 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 7, which is inspired by Imbens (2019). Figure 7a shows a simple model of supply and demand where



(a) Explcit simultaneity (b) Directed simultaneity

Figure 7: An example of directing simultaneity

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 7a can just as well be represented by the fully directed graph in figure 7b. To see this consider the following equations which are implied by figure 7a:

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

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

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 7b. At this point I will note the relevance of the Lucas (1976) critique. The model in 7b is a reduced form estimation of the model in 7a, 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 structure 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.

2.2 DSGE Models

Having introduced DAGs as a modelling tool, this section will consider DSGE models and how they can be represented as DAGs. A the solution to a DSGE model is a state space model (Hamilton, 1994), and therefore, all variables can be categorized as either state variables or control variables (Fernandez-Villaverde et al., 2016). Defined as generally as possible, state variables are the variables on whose past the model's current values depend, and control variables are the rest; their past is independent of the current values of the model. In the case of discrete time (the case considered in this paper) DSGE models have the Markov property, meaning that the current value of the states depends only on the history of the model in the previous period. State variables can be further categorized as either endogenous states (such as the capital stock) and exogenous states (such as the state of technology or productivity). As the name suggests, endogenous states are determined simultaneously (endogenously) with contemporanious controls in the model, however, their past is by definition exogenous and relevant to the determination of the current values of the model. Exogenous states, on the other hand, are exogenous in the sense that they are determined independently of any contemporanious variables in the model, and are thus determined entirely by the past of the model or any exogenous innovations (shocks) that might be present. The distinction between endogenous and exogenous states is subtle, but it is important here because the different implications about exogeneity will lead to a slightly different treatment in a DAG.

Given this definition it is straightforward to characterize the solution to a general DSGE model

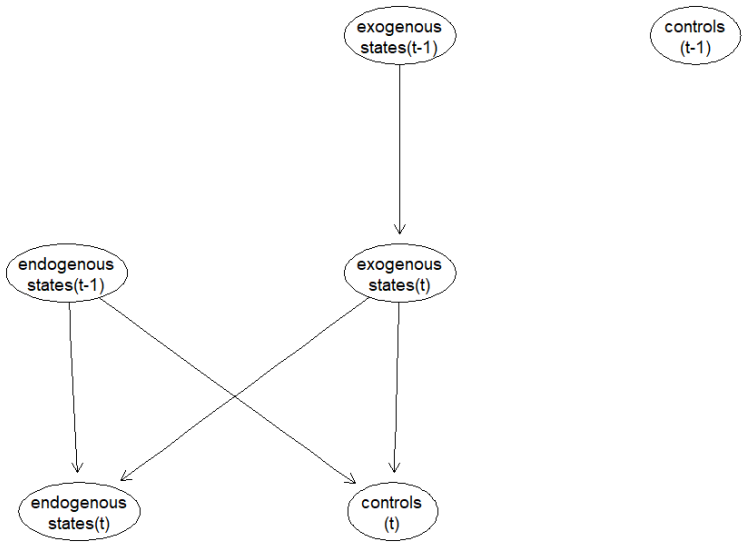


Figure 8: General solution to DSGE model as a DAG

as a DAG. This is demonstrated by Figure 8. This expresses in graphical format all of the assumptions outlined in the previous paragraph about the solution to a DSGE model. There are a few important things to note here. Due to the Markov condition, it is only necessary to consider the first lag of the state variables, although this is done primarily for simplicity as this could easily be generalized. Models of this kind containing temporal data are sometimes known as Dynamic Bayesian networks (Ghahramani, 1997). Section 2.1.5 argued that simultanous relationships can be represented by a DAG by simply expressing the (seemingly) endogenous variables as functions of the relevant exogenous variables. Here we use this to express the simultanously-determined contemporanious values of the controls and endogenous states as descendants of their exogenous determinants, namely, the lag of the endogenous states, and the current value of the exogenous states. Also note that the lag of control variables are completely detached from the rest of the graph.

When data are simulated from a DSGE model the true endogenous states, exogenous states, and controls are known, and therefore it is straightforward to specify the solution to the model as a DAG as in Figure 8. In 5 examples of this kind will be provided that further demonstrate that the solution to any DSGE model can be correctly specified as a DAG by using the DAG to reproduce the simualted IRFs. This DAG is the ground truth that structure learning algorithms seek to identify. Since we have asymptotic consistency results for all of the structure learning algorithms discussed in Section 2.1.3, we should be able to learn this ground truth as long as the relevant assumptions are satisfied. The assumptions of linear condtional probabilities and Gaussian innovations are satisified by the DSGE models employed here by construction. The assumption of faithfulness does not come for free and this may prove problematic for some models

that will be considered here. Finally, the assumption that the sample size n is large relative to the number of variables in the model can be easily satisfied in simulated data, but is a more important consideration in applications to real data. Therefore we will consider simulations with both large and small n.

2.3 Related Modelling Techniques

Having discussed how DSGE models and macroeconomic data more generally can be represented as DAGs this section will discuss how this approach relates to other econometric approaches which are common in the analysis of macroeconomic timeseries. It is possible to draw comparisions with both Structural Vector Autoregression (SVAR) and Autoregessive Distributed Lag (ADL) models, so these will be discussed in turn.

One of the most common and simplest econometric models for this type of data is the vector autoregression (VAR), which was popularized by Sims (1980). This method involves regressing a vector of outcomes y_t on a matrix containing k lags of y in the form $y_t = [y_{t-1}, ..., y_{t-k}]\beta + \epsilon_t$. The primary concern with and limitation of this approach is that the estimated covaraince matrix ϵ_t is unrestricted, so the shocks contained within it are not mutually independent. Therefore, this model can not be used to estimate the effect of a truely exogenous shock on the dynamics of observed variables. In order to address this issue the model is transformed and an assumed causal ordering is imposed in the form of a Cholesky decomposition (Sims, 1980), which has the effect of making the errors of the estimated, transformed model mutally uncorrelated or *structural*. Therefore, such models are known as SVARs. As noted by Demiralp and Hoover (2003), in this context there is an equivallence between SVAR models and DAGs. This is because root nodes are assumed to be mutually uncorrelated, and as a result, any shocks to these will have a structural interpretation.

However, one key difference between a DAG in this context and a SVAR model is that the DAG allows for some variables to depend on contemporanious values of other variables. In particular, the endogenous states and controls depend contemporaniously on the exogenous states. In this sense the DAG is similar to an ADL model. When implementing an ADL model it is necessary for the researcher to choose which contemporanious variables to include as regressors, implicitly assuming that this regressors are at least predetermined relative to the outcomes of interest.

The primary advantage of DAGs is the relatively weak assumptions they require. Both the SVAR and ADL models require the researcher to specify assumptions about the relative exogeneity of observable variables. These assumptions are themselves either derived from a similarly assumption-heavy model such as a DSGE model, or are entirely *ad hoc*. There has been a long tradition within the field of economics including seminal papers by Lucas et al. (1976), Sims (1980), and Jorda, (2005) criticising this type of model building. DAGs, along with structure learning algorithms provide an asymptotically consistent, agnostic, and data-driven alternative to models

of this kind.

3 Methodology

For the application in this paper I have used the "bnlearn" package (Scutari, 2010) for R. For the reasons outlined in section 2.1.3 I used the implementation from this package of the hybrid "rsmax2" algorithm (Spirtes et al., 2000), with the PC algorithm as the structural constraint. This algorithm has a tuning hyperparamerter α, which is the significance level of the Pearson linear correlation test used in the PC algorithm. There is no general rule on the selection of this parameter, however, Kalisch and Bühlmann notes that it should go to zero as n goes to infinity. Therefore, results will be shown for a variety of choices of this parameter. Since the first lags of all variables in the model are added as nodes in the graph since it is not know *a priori* which variables are controls or states. If structure learning succeeds in identifying the ground truth the lags of controls should be left disconnected from the rest of the graph, or at least have parameter values very close to 0.

A number of general assumptions about DSGE solutions were encoded into the structure learning algorithm in order to reduce the size of the available search space and thus improve performance and accuracy. Firstly, no node that is a lag should have any parent nodes. A blacklist was therefore implemented in the PC algorithm to ensure this is the case in the resulting DAG. Secondly, an arc was manually added from the lag of each variable to its current value. At fist glance this would seem to contradict the fact that the lags of controls should be separate from the rest of the graph, however, during parameter estimation a slope parameter of 0 can be fit to these edges. Therefore this alteration does not remove the ground truth from the set of graphs the algorithm could potentially choose. Even though this alteration is theoretically allowable the primary justification for its inclusion is pragmatic. For example, consider a simpe DSGE model where an exogenous technology process follows an autoregressive process. If structure learning results in no edge between the lag and current value of technology, then any IRFs produced by this DAG would be highly inaccurate as the effect of a technology shock would disappear after only one period, even if the DAG otherwise does a good job of capturing the causal structure of the data. Finally, the general DSGE DAG has an effective depth of exactly three. Including this information in the structure learning algorithm would greatly reduce the size of the search space, and likely greatly improve structure learning performance. Unfortunately no implementation of the PC algorithm or any other structure learning algorithm that I was able to find is able to encode this kind of constraint, so this is left as a suggestion for future research.

In order to evaluate the DAGs that are produced, I will consider two primary metrics. Firstly, I will consider the Structural Hamming Distance (SHD), which is defined as the number of edges that need to be added, removed, or reversed in order to transform the estimated DAG into the ground truth DAG (Tsamardinos et al., 2006). This is a simple and effective measure of how close

the algorithm has come to correctly identifying the underlying causal structure. Secondly I will use the estimated (and ground truth) DAGs to compute IRFs to a shock to one of the model's exogenous states (z). This is achieved according to the following process

1. Set the values of all lags and exogenous states other than z at time $t = 0$ to 0.
2. Set the value of z at time $t = 0$ to some external value, usually 1 standard deviation of the relevant shock.
3. Impute the values of the remaining variables (the endogenous states and controls) implied by the DAG.
4. Set lags in next time period to current values.
5. Impute values implied by DAG in next period.
6. Continue iteratively until period T is reached.

The purpose of this is to demonstrate that even when the structure of the estimated DAG does not appear to be entirely correct, it is still possible that it will reproduce the correct impulse responses, and this demonstrates how DAGs can be a useful small sample heuristic, even if the asymptotic properties do not apply.

4 Data

In order to demonstrate the capability of the DAG methodology empirically I will work with both simulated and real macroeconomic data. Using simualted data has a number of key advantages. Firstly, since the model that simulates the data is known it is possible to evaluate whether the structure learning has succeeded in identifying the ground truth DAG. Secondly, in this context it is possible to ensure to the greatest possible extent that the underlying assumptions of the structure learning algorithms, including linearity and normality are satisfied. Finally, since these models are central to modern macroeconomics it provides a controlled testing environment which is also arguably highly relevant to real data. On the other hand, using real data is an oppurtunity to demonstrate that DAGs are also a powerful heuristic tool that can be implemented outside of a rigorously controlled environment. The remaineder of this section will discuss the various sources and general properties of the data used in this paper.

4.1 Simulations

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

Symbol	Name	Type
eps_g	government spending shock	shock
eps_z	technology shock	shock
g	government spending level	exogenous state
z	technology process level	exogenous state
k	capital	endogenous state
w	wage rate	control
r	return to capital	control
y	output	control
c	consumption	control
l	hours worked	control
i	investment	control

Table 1: Description of Variables

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 DAGs.

4.1.1 RBC

The first model which I have chosen to evaluate is a the baseline RBC model provided by Pfeifer (2020). This model includes 11 variables which are summarized by Table 1. Using the default specification a sample of 100000 observations was generated, from which smaller subsamples are also considered. This model contains two exogenous state variables: technology (z) and government spending (g), and one endogeous state: capital (k). this model was chosen as it is one of the simplest DSGE models and provides a good baseline to demonstrate the effectiveness of DAGs in this context.

4.1.2 Smets and Wouters (2007)

The model from Smets and Wouters (2007) on the other hand is a significantly larger and more complex model. This model was chosen because it is highly influential, and it represents the current state of the art in DSGE models. It also provides a more difficult modelling challenge to approach with DAGs.

This model contains seven exogenous shocks: a productivity shock (ea), a risk premium shock (eb), a government expenditure shock (eg), an investment specific technology shock (eqs), a monetary policy shock (em), a price markup shock (epinf), and a wage markup shock (ew). Furthermore, these shocks all contribute (respectively) 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), which are are the exogenous states of the model. Furthermore, the model contains 13 endogenous states.

Symbol	Name
pi	CPI Inflation
rm	Federal Funds Rate (Return to Money)
rb	10 Year US Treasury Bond Yield (Return to Bonds)
g	(Real) Government Expenditure
y	(Real) GDP
i	(Real) Private Investment
w	Median (Real) Wage
n	Population
rk	Return to Capital [1]
dk	Change in Capital Stock
z	Total Factor Productivity
u	Unemployment
l	Total Workforce
c	(Real) Personal Consumption

Table 2: Description of US Data

4.2 FRED Data

To provide an example of real macroeconomic timeseries, quarterly data from the US between years 1980-2005 were collected from FRED (2020) for 15 variables outlined in Table 2. All of the variables were detrended by taking the residuals of an estimated first order autoregression, with the exception of capital which displayed unit-root behavior and was therefore first differenced (hence the symbol dk). Total factor productivity and capital stock were provided on an annual basis and were therefore interpolated quadratically. Full details of data preprocessing are availible on GitHub (Hall-Hoffarth, 2020).

5 Results

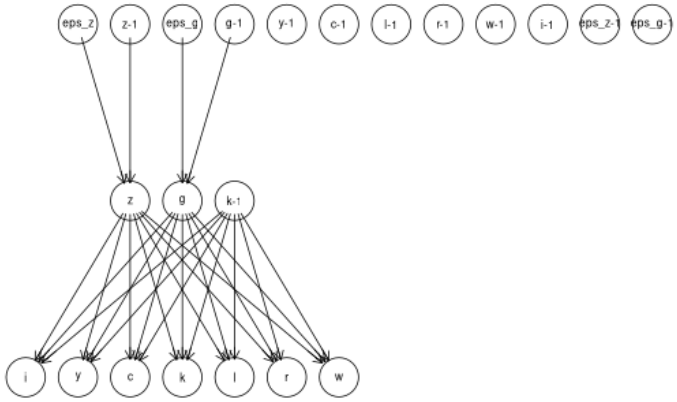
5.1 RBC

5.1.1 Structure Learning

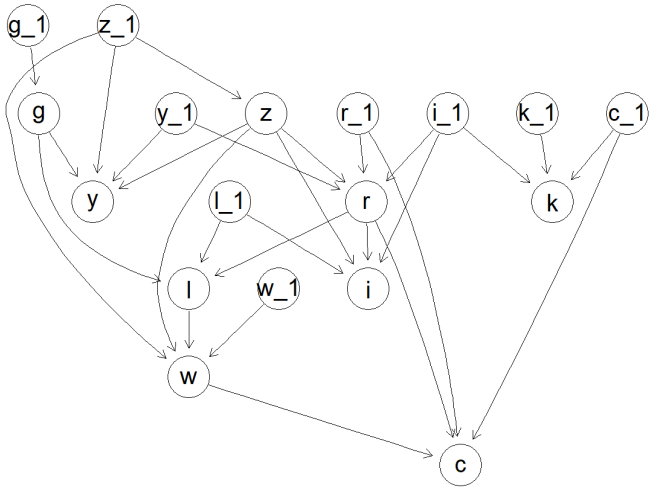
Figure 9 shows the ground truth DAG, as well as a DAG learned using the hybrid structure learning algorithm. This graph has a SHD of 33. There are a few results to note here. Firstly, z and g are successfully identified as exogenous states, whereas k is successfully identified as an endogenous state. On the other hand, no variable is identified as a control, as none of the autoregressive coefficients in the parameterized model fail to reject the null in a t-test at any reasonable significance level.

5.1.2 IRFs

Figure 10 illustrates the IRFs to a technology shock of 1 standard deviation for the original simulation (12a) in comparison with those generated by various DAGs. The ground truth DAG replicates the oringal simulations' IRFs almost perfectly, as it should since all of the necessary assumptions

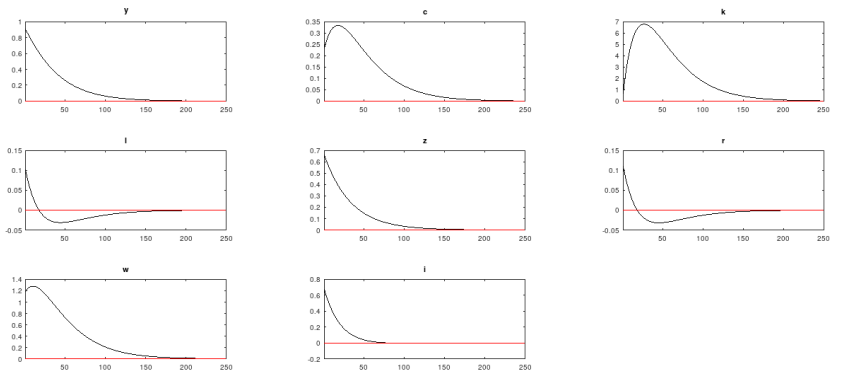


(a) Ground Truth DAG

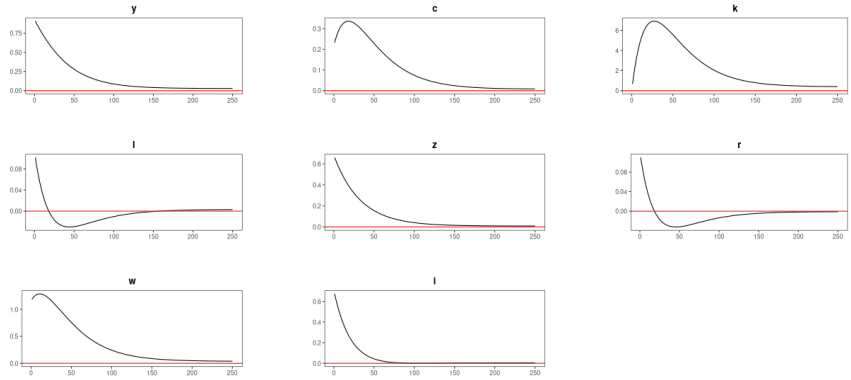


(b) Hybrid DAG

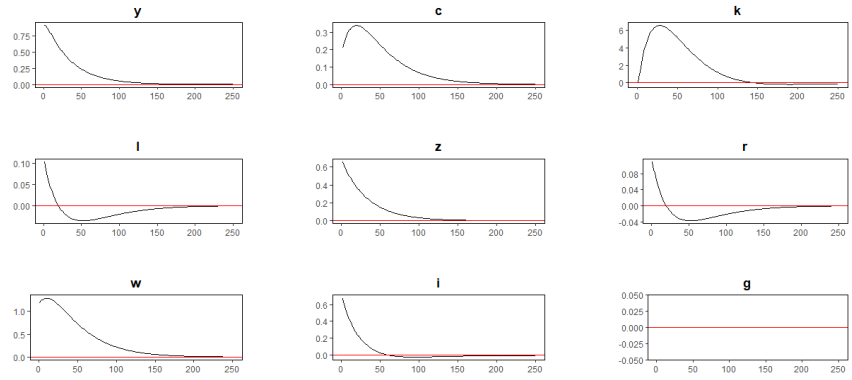
Figure 9: Structure of DAGs manually and algorithmically fit to RBC model data.



(a) Original Simulation



(b) Ground Truth DAG



(c) Hybrid DAG

Figure 10: IRFs generated by the original model and various DAGs generated on the simulated data.

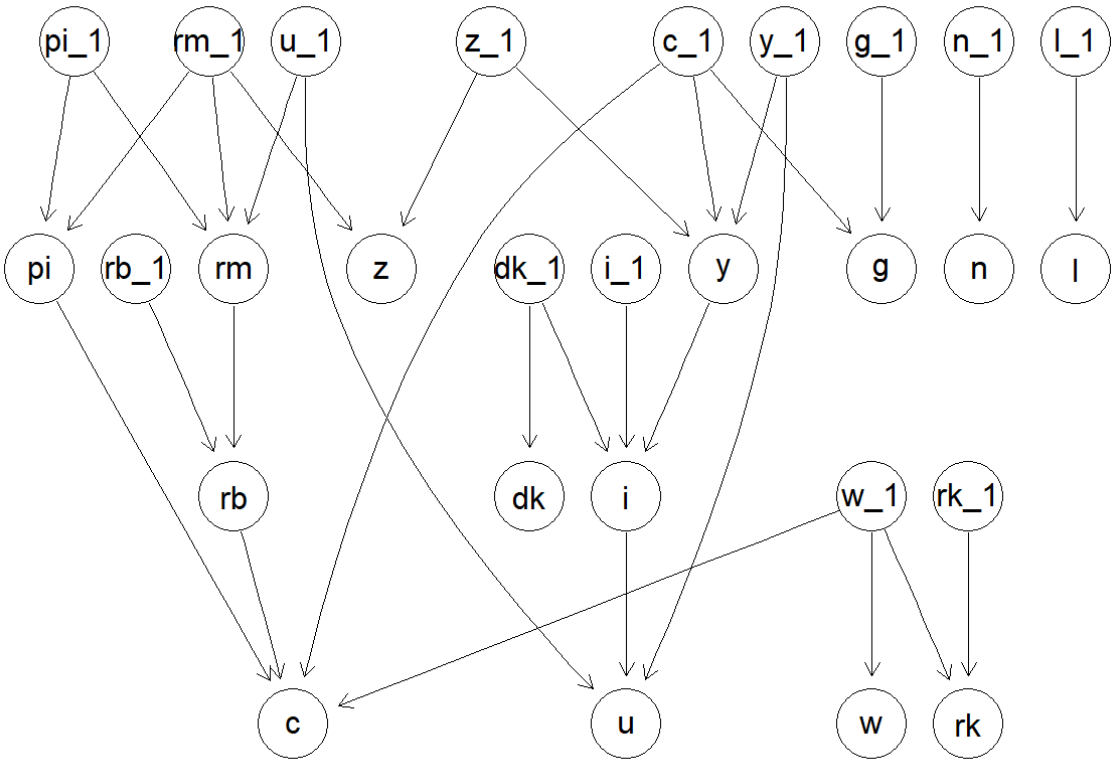


Figure 11: Structure of DAG fit to FRED data.

are satisfied, and by constructi this DAG specifies the same linear system of equations as the solution to the simulated model. More notable however, is that the structure learned using the hybrid structure learning method also produces very accurate IRFs, despite the fact that it does not have all of the correct root nodes.

5.2 Smets and Wouters (2007)

5.2.1 Structure Learning

5.2.2 IRFs

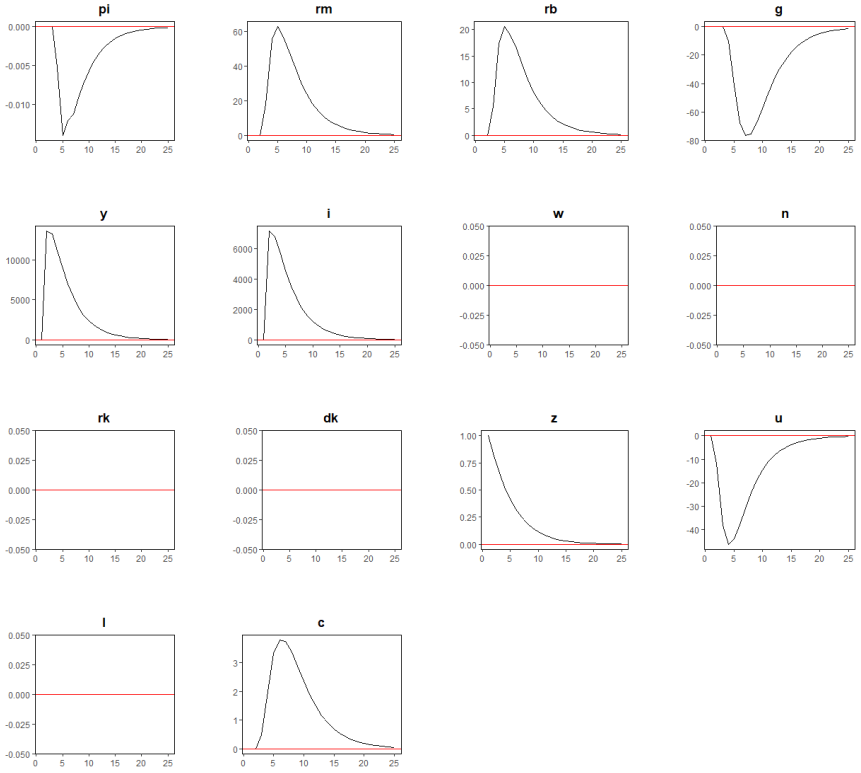
5.3 FRED Data

5.3.1 Structure Learning

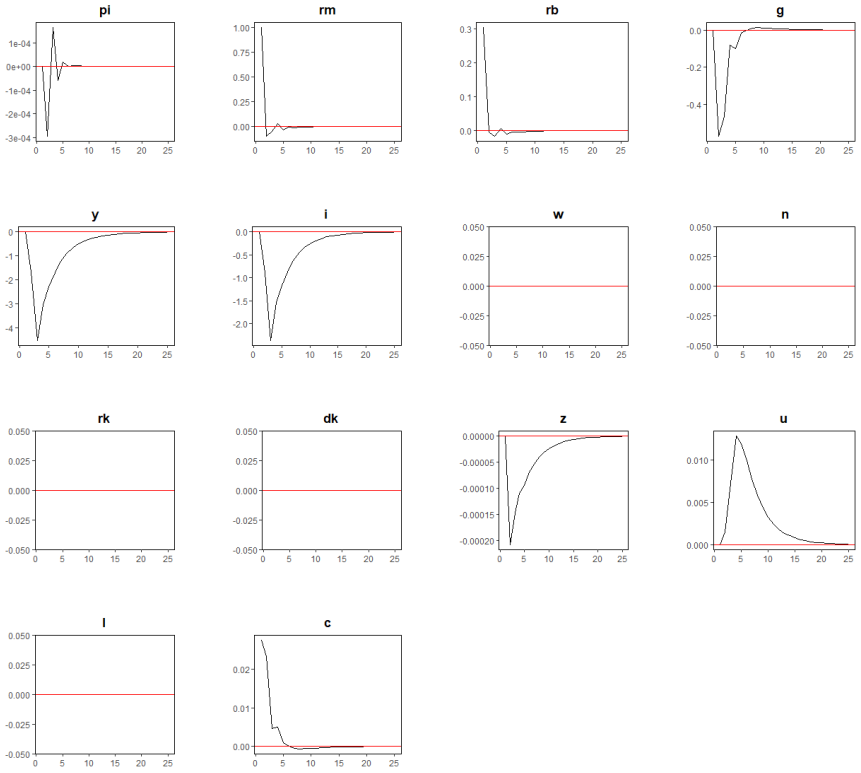
5.3.2 IRFs

6 Discussion and Conclusion

TODO



(a) Technology Shock IRFs



(b) Monetary Policy IRFs

Figure 12: IRFs generated from the DAG learned from FRED data.

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. *Proc. DARPA broadcast news transcription and understanding workshop*, *8*, 127–132.

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

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of machine learning research*, *3*(Nov), 507–554.

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

Fernandez-Villaverde, J., Rubio-Ramirez, J. F., & Schorfheide, F. (2016). Solution and estimation methods for dsge models. *Handbook of macroeconomics* (pp. 527–724). Elsevier.

Friedman, N., Nachman, I., & Pe'er, D. (2013). Learning bayesian network structure from massive datasets: The" sparse candidate" algorithm. *arXiv preprint arXiv:1301.6696*.

Ghahramani, Z. (1997). Learning dynamic bayesian networks. *International School on Neural Networks, Initiated by IIASS and EMFCSC*, 168–197.

Hall-Hoffarth, E. (2020). *Dsge bayesian networks*. Retrieved July 17, 2020, from https://github.com/e-hall-hoffarth/bayesian_networks/

Hamilton, J. D. (1994). State-space models. *Handbook of econometrics*, *4*, 3039–3080.

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.

Jorda, O. (2005). Estimation and inference of impulse responses by local projections. *American economic review*, *95*(1), 161–182.

Kalisch, M., & Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, *8*(Mar), 613–636.

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. *Carnegie-Rochester conference series on public policy*, *1*(1), 19–46.

Nandy, P., Hauser, A., Maathuis, M. H., et al. (2018). High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, *46*(6A), 3151–3183.

of St.Louis, F. R. B. (2020). *Fred economic data*. Retrieved July 12, 2020, from https://fred.stlouisfed.org/

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

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

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

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

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, 1–48.

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.

Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

Steel, D. (2006). Homogeneity, selection, and the faithfulness condition. *Minds and Machines*, *16*(3), 303–317.

Taylor, J. B. (1993). Discretion versus policy rules in practice. *Carnegie-Rochester conference series on public policy*, *39*, 195–214.

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, *65*(1), 31–78.

Verma, T., & Pearl, J. (1991). *Equivalence and synthesis of causal models*. UCLA Computer Science Department. 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.