Semi-Parametric State-Space Model Selection for Macroeconomic Timeseries

Emmet Hall-Hoffarth

October 14, 2020

Abstract

This paper presents a set of criteria and an algorithm for agnostic, data-driven selection among macroeconomic DSGE models inspired by structure learning methods for DAGs. The state-space representation of any DSGE model is also a DAG, and therefore concepts used for model selection amongst DAGs can also be used in the context of DSGE models. This implies a set of criteria based on conditional independence tests which can be used to evaluate whether a state-space model, and thereby a DSGE model is consistent with some observed data. I then introduce an algorithm which tests these criteria against the set of possible state-space models producing a subset of allowable models. When combined with likelihood maximisation this algorithm identifies a unique optimal model. The efficacy of this alogorithm is demonstrated for simualted data, and results for real data are also provided and discussed.

1 Introduction

2 Liturature Review

2.1 DAGs

2.1.1 Preliminaries

Formally, a DAG G is a pair (V, E) where V is a set of *nodes*, one for each of k observable variables, and E is a $V \times V$ set of *edges* or *arcs* (Kalisch & Bühlmann, 2007). The presence of an edge (x, y) indicates the presence of an directed edge from node x to node y. As the name suggests, every edge in E is directed such that if $(x, y) \in E$ then $(y, x) \notin E$. E is also assumed to not contain any cycles, that is, there is no set of edges $(i, j) \in E$ containing a directed path starting and ending at the same node. Figure 1 gives a simple example of a DAG.

In general, DAGs can represent either discrete, continuous, or mixed variables, but in the current application only continuous variables will be considered. For simplicity, each arc will hereafter be assumed to define a linear relationship between continuous variables. With this assumption we can more specifically define V as a $(k \times 1)$ vector and E as a $k \times k$ adjacency

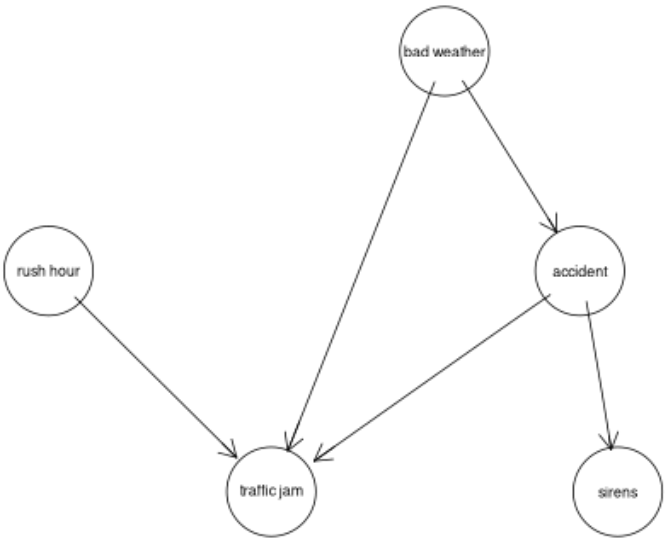


Figure 1: A simple example of a DAG (Liszka, 2013)

matrix containing slope parameters, where $e_{ij} \neq 0$ indicates a directed edge from node i to node j and $e_{ij} = 0$ indicates the lack of an edge. The directedness assumption is analogous, and the acyclic property is equivalent to the statement that E^n has zeros on its diagonal for $\forall n > 0$. The model will now also include a $kx1$ vector ϵ containing mutually independent Gaussian shocks, one for each node.

The set of nodes from which an arc into x originates are known as the *parents* of x ($pa(x)$), and the set of nodes that have an incoming arc from x are known as the *children* of x ($ch(x)$). The set of all nodes from which a directed path into x originates are known as the *ancestors* of x ($ans(x)$) and the set of all nodes that have an incoming path from x are known as the *decendents* of x ($des(x)$).

I will now briefly review some key results pertaining to DAGs that are utilised in this paper. For a more complete treatment see Pearl (2009).

Definition 1. *Faithfulness Let f represent some DGP, and $I(f)$ be the conditional independence relationships implied by f. A DAG G with parameters $\theta \in \Theta$ is said to be* ***faithful*** *to f if and only if the conditional independence relationships implied by G satisfy $I(G(\theta)) = I(G(\theta')) = I(f)\ \forall\ \theta \neq \theta' \in \Theta$. (Pearl, 2009, p.48)*

Outside of the optional assumption of linearity and Gaussian errors that are made here for simplicity, *faithfulness* is the only assumption necessary for the identification of a DAG for a true DGP. It is the assumption that the conditional independence relationships in the DGP are *stable* to perturbations of parameters. Intuitively, if we wish to use conditional independence relationships to identify a model then we must assume the observed conditional independence relationships do not belie the underlying distribution. 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. 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 perturbations 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 equivalently, that the matrix of DGP parameters is of full rank. This is because under this condition, specific combinations of parameters which result in the cancellation of causal effects have has Lebesgue measure 0. 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, capital 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 means that parameters are entirely determined by an economic agent. 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. If parameters are assumed to be come from some distribution with different draws for each population, then homogeneity is the statement that there is little exogenous variation in those parameter values, that is variation outside of the variation caused by selection. 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 cancelling the causal effect of inflationary shocks on output and minimizing exogenous variation.

Despite these concerns, I would argue that the faithfulness assumption is plausible in most macro-economic contexts. For simulations, whether or not the assumption is violated can be read straight off the structural model. For real data, it seems unlikely that any macroeconomic variable (even the policy rate) is determined in an entirely systematic for deterministic way. Identification of policy rate shocks has been a topic of much scrutiny (Ramey et al., 2016), and this line of research has provided a significant amount of evidence for the existence of such shocks. Given that f is *stable* we can use conditional independence tests in the following way to evaluate whether a DAG G is consistent with f.

Figure 2: A DAG before structure learning

Definition 2. *D-Separation A path P in a DAG G is said to be* ***d-separated*** *or* ***blocked*** *by a set of variables* $\mathbf{z}$ *if and only if:*
A. P contains a chain $x \to m \to y$ or fork $x \leftarrow m \to y$ and $m \in \mathbf{z}$
B. P contains a collider $x \to m \leftarrow y$ and $\{m \cup des(m)\} \cap \mathbf{z} = \emptyset$
A set of variables $\mathbf{z}$ *is said to d-separate x and y if* $\mathbf{z}$ *blocks every path between x and y. (Pearl, 2009, p.16)*

Theorem 1. *D-Separation and Conditional Independence If x and y are d-separated by* $\mathbf{z}$ *in DAG G, and G is faithful to the true DGP f of x and y, then x and y are independent conditional on* $\mathbf{z}$*. (Pearl, 2009, p.16)*

This result is essential for defining the constraint based tests in section 3.1. In particular, it implies the following result that we will leverage:

Definition 3. *Parental Markov Condition Given some DAG G, a node x in G is d-separated from and therefore independent of all its non-decendents by its parents. This is known as the* ***Parental Markov Condition****. (Pearl, 2009, p.16, p.19)*

Corollary 1. *If G is faithful to the DGP f then f admits the following factorisation:*

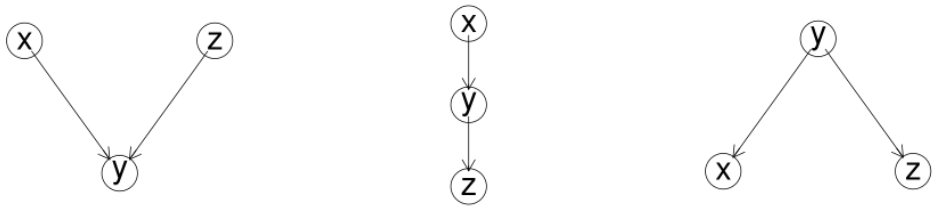
$$f(\mathbf{w};\theta) = \prod_{i=1}^{k} f(w_i|pa(w_i);\theta) \tag{1}$$

2.1.2 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 2, the first task is simply to estimate the parameters of the network, such as the parameter matrices **A**, **B**, **C**, **D**, and **E** in Equation 2 - 4. This is usually done via maximum likelihood.

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 to do 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. How this can be done in the context of DSGE models is discussed in Section 2.2. In this context however, doing so achieves little. This is equivalent to specifying a system of linear regressions to be estimated, probably based on some economic model that was developed by other means, and while this is then automatically encapsulated in a convenient, easily interpreted representation of the underlying assumptions, it seems nothing particularly novel would have been achieved.

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



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

(a) Collider

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

(b) Chain

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

(c) Fork

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

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 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). 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 3. 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 under the assumption of joint-normality using a t-test on (partial) correlations. 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, 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 (CPDAG). The undirected arcs in a CPDAG could face either direction and the graph would still be consistent with both the DAG assumptions 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.

2.2 DSGE Models

Suppose a DSGE model is defined over a set of k variables in a vector $\mathbf{w}$. The solution to a log-linearised DSGE model can be written as a state space model (King et al., 1988) that partitions $\mathbf{w}$ into three mutually exclusive vectors $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$. This state-space model is described by equations 2 - 4:

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}\mathbf{z}_t \tag{2}$$

$$\mathbf{x}_t = \mathbf{C}\mathbf{x}_{t-1} + \mathbf{D}\mathbf{z}_t \tag{3}$$

$$\mathbf{z}_t = \mathbf{E}\mathbf{z}_{t-1} + \epsilon_t \tag{4}$$

Where $\mathbf{x}_t$ is a vector of control variables, $\mathbf{y}_t$ is a vector of endogenous state variables, $\mathbf{z}_t$ is a vector of exogenous state variables, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and, $\mathbf{D}$ are coefficient matrices, and ϵ_t is a vector of shocks. All variables are mean-zero. The shocks in ϵ_t can be interpreted as structural shocks as they satisfy the assumptions $\epsilon_t \sim N(0, \Sigma)$ and Σ diagonal $\implies Cov[\epsilon_{i,t}, \epsilon_{j,t}] = 0 \iff \epsilon_{i,t} \perp\!\!\!\perp \epsilon_{j,t}$ for $i \neq j$. These shocks are assumed to not be observed, both because this is likely true in realistic applications (absent some very clever econometric tricks) and because observing the shocks is not necessary for the inference proposed here.

Furthermore assume that $\mathbf{E}$ is diagonal ($e_{ij} = 0$ if $i \neq j$) such that the process of each exogenous state depends only on its own past and $|e_{ii}| < 1$ such that the model has a stationary solution. Note that for simplicity the exogenous states are assumed to possess the Markov property, that is, $\mathbf{z}_t$ depends only on $\mathbf{z}_{t-1}$ and not any further lags. As a result, the entire model has the Markov property. However, the framework and algorithm proposed here could easily be generalised to allow for longer lags.

In this setup, all variables can be categorized as either state variables or control variables (Fernandez-Villaverde et al., 2016). Defined as broadly as possible, state variables are the variables whose past is relevant for determining the current value of modelled variables, and control variables are the rest; their past is independent of the current values of the model. 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) (Ravenna, 2007). As the name suggests, endogenous states are determined simultaneously (endogenously) with contemporaneous 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 contemporaneous 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, and it is primarily made because we often wish to give different interpretations to these two types of variables, but the fact that exogenous states enter the prediction of time t endogenous variables at time t rather than at time $t-1$ is an important modelling distinction.

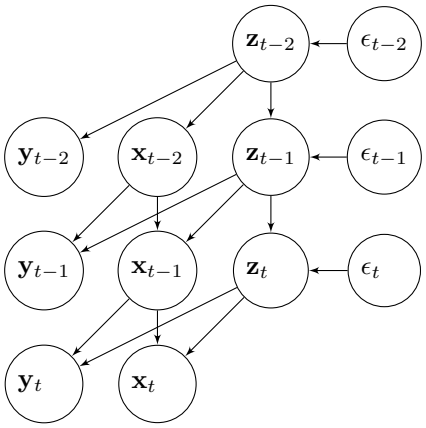


Figure 4: DSGE solution expressed as a DAG

3 Methodology

Given equations 2 - 4 it is straightforward to characterize the general solution to a DSGE model as a DAG. This is demonstrated by Figure 4. This expresses in graphical format all of the assumptions outlined in those equations.

Given this it would seem straightforward to input random samples generated from a DSGE model into the available structure learning algorithms and find the correct solution, given that these algorithms have well established asymptotic convergence properties. Unfortunately, despite extensive experimentation with these tools, I was unable to obtain any results in this way, as these algorithms seem to have a number of important limitations in this context. Constraint-based algorithms rely on conditional independence tests which themselves involve computing the correlation between residuals. In the context of simulated data these residuals may be very small or effectively zero when conditioning on the true parents of a node. In this case the computation of partial correlations may be unstable and lead to spurious results. Furthermore, these results are only asymptotic and it seems that finite-sample bias may be important in economic applications, where sample sizes are small. Particularly problematic is that structure learning algorithms consider all possible DAGs given observed variables as potential candidates, whereas in this context we assume that the solution takes on a particular form, as in 4.

As a result of these limitations, a more effective approach in this context is to design a bespoke algorithm that takes into account the relatively stringent assumptions that can be made about DSGE solutions. For the reasons outlined in section 2.1.2 this will be a hybrid algorithm. Therefore, before introducing the algorithm we will define relevant constraint and score tests in turn.

3.1 Constraint Tests

Figure 4 implies the following four independence relationships among the time t and $t-1$ variables:

$$x_t \perp\!\!\!\perp x'_t \,||\, [\mathbf{x}_{t-1}, \mathbf{z}_t] \text{ for all } (x_t, x'_t) \in [\mathbf{x}_t, \mathbf{y}_t] \tag{5}$$

$$x_{t-1} \perp\!\!\!\perp z_t \,||\, \mathbf{z}_{t-1} \text{ for all } x_{t-1} \in \mathbf{x}_{t-1} \text{ and } z_t \in \mathbf{z}_t \tag{6}$$

$$x_t \perp\!\!\!\perp z_{t-1} \,||\, [\mathbf{x}_{t-1}, \mathbf{z}_t] \text{ for all } x_t \in [\mathbf{x}_t, \mathbf{y}_t] \text{ and } z_{t-1} \in \mathbf{z}_{t-1} \tag{7}$$

$$z_t \perp\!\!\!\perp z'_t || \mathbf{z}_{t-1} \text{ for all } z_t \neq z'_t \in \mathbf{z}_t \tag{8}$$

The first condition (5) is the statement that the model's time t endogenous variables are explained entirely by and are therefore unconfounded conditional on $\mathbf{x}_{t-1}$ and $\mathbf{z}_t$ (the time t states). In DAG parlance, a time t endogenous variable is *d-separated* from and therefore independent of any other time t endogenous variable by the time t states. Condition (6) states that the time t states that every lagged endogenous state is independent of every exogenous state conditional on the lagged exogenous states. This follows from the exogeneity of $\mathbf{z}$ which implies that the only parent of z_t other than the shock is z_{t-1}. Condition (7) holds because the time t states d-separate the time t endogenous variables from the lagged exogenous states. If we were to consider further lags, this conditional independence would apply not only to z_{t-1}, but also to all $t-2$ and earlier variables because of the Markov condition. Finally, Condition (8) holds that all exogenous states are mutually independent conditional on past exogenous shocks. This is a stronger assumption than the other three, and depends crucially on the assumption that the shocks ϵ_t are mutually uncorrelated, and thus structural in nature.

These constraint tests already provide a powerful selection criteria for empirical DSGE models I will call DAG-consistency:

Definition 4. *For a set of variables* $\mathbf{w}$ *a log-linearised DSGE model* M *is valid with respect to a distribution* $f(\mathbf{w})$ *if the exogenous states (*$\mathbf{z}$*), endogenous states (*$\mathbf{x}$*) and controls (*$\mathbf{y}$*) of* M *as defined in equations (2) - (4) satisfy conditions (5) and (6).*

It is proved in an appendix (A) that validity is necessary and sufficient to guarantee that if the ground truth can be characterised as in (2) - (4) then DSGE models that are dag-consistent with the ground truth have the same DAG representation, and therefore, the same state-space model. Constraints (7) and (8) are still applicable because they are implied by the DAG, but there are not required. To be more general we could drop these assumptions as long as the shocks only directly effect the exogenous states, and the other constraints would still hold and be valid tests of the model. However, these constraints (and the associated assumptions) are nonetheless included because they is satisfied by a wide range of DSGE models including all of those considered in the empirical portion of this paper, and all else equal conducting a larger number of tests gives more statistical *power* to reject incorrect models when using finite samples.

3.2 Score Tests

The most basic score function for Gaussian Bayesian networks is the log-likelihood function. The Markov compatibility condition (Definition 3) DAG admits factorisation of the joint

probability distribution into the product of the distribution of each variable conditional on its parents:

$$f(\mathbf{w};\theta) = \prod_{i=1}^{k} f(w_i|pa_i;\theta) \tag{9}$$

Therefore, the log-likelihood can be calculated as:

$$\mathcal{L}(\mathbf{w},\theta) = \sum_{i=1}^{k} ln(f(w_i|pa_i;\theta)) \tag{10}$$

Now consider the assumptions in the current context. $\mathbf{w_i}$ is partitioned into $\mathbf{z}$, $\mathbf{x}$, and, $\mathbf{y}$. We assume that the conditional probabilities are linear functions and follow a mean-zero normal distribution, so the only parameter is the variance-covariance matrix $\mathbf{sigma}^2$. Furthermore, values from period $t-1$ are known at time t and therefore these root nodes are deterministic, so their distribution is trivial. Therefore,

$$\begin{aligned}
\mathcal{L}(\mathbf{w};\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D},\mathbf{E},\sigma^2) &= \sum_{z_{i,t}\in\mathbf{z}_t}\left(\sum_{t=1}^{T} ln(\phi(z_{i,t}|z_{i,t-1}|\mathbf{E},\sigma^2))\right) \\
&+ \sum_{y_{i,t}\in[\mathbf{y}_t,\mathbf{x}_t]}\left(\sum_{t=1}^{T} ln(\phi(y_{i,t}|[\mathbf{x}_{t-1},\mathbf{z}_t]|\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D},\sigma^2))\right) \qquad (11) \\
&= \sum_{y_{i,t}\in\mathbf{y}_t}\left(\sum_{t=1}^{T} ln(\phi(\mathbf{a}_i\mathbf{x}_{t-1}+\mathbf{b}_i\mathbf{z}_t|\mathbf{a}_i,\mathbf{b}_i,\sigma^2_{y_i}))\right)+ \\
&\quad \sum_{x_{i,t}\in\mathbf{x}_t}\left(\sum_{t=1}^{T} ln(\phi(\mathbf{c}_i\mathbf{x}_{t-1}+\mathbf{d}_i\mathbf{z}_t|\mathbf{c}_i,\mathbf{d}_i,\sigma^2_{x_i}))\right) \\
&\quad \sum_{z_{i,t}\in\mathbf{z}_t}\left(\sum_{t=1}^{T} ln(\phi(\mathbf{e}_i z_{i,t-1}|\mathbf{e}_i,\sigma^2_{z_i}))\right) \qquad (12)
\end{aligned}$$

Where $\mathbf{x_i}$ is the i_{th} row of $\mathbf{X_i}$. Notice that we can calculate the variances separately in each regression because the Markov compatibility condition implies that they are independent. Finally, we can substitute in for the maximum likelihood estimate of σ^2 for each regression and the functional form of ϕ to arrive at an expression for the log-likelihood function:

$$\mathcal{L}(\mathbf{w}) = -\frac{T}{2}\left(k(1+ln(2\pi)) + \sum_{i=1}^{k} ln(\hat{\sigma}_i^2)\right) \tag{13}$$

$$\hat{\sigma}_i^2 = \frac{1}{T}\sum_{t=1}^{T}(w_{i,t}-\hat{w}_{i,t})^2 \tag{14}$$

Where $\hat{w}_{i,t}$ is are of predicted values of some w_i in $\mathbf{w}$ implied by estimates of equations (2) - (4) using the maximum-likelihood estimates of the coefficient matrices.

Since maximising the log-likelihood does not penalise complexity, it often favours models with many more edges than exist in the ground truth. In other words, maximising log-likelihood over a space of candidate DAGs may lead to *overfitting*. As a result there are many penalised score functions based on the log-likelihood that are available. Here we consider two

of the most popular, the Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarz et al., 1978), which are calculated as follows:

$$AIC(\mathbf{w}) = 2k - 2\mathcal{L}(\mathbf{w}) \tag{15}$$

$$BIC(\mathbf{w}) = kln(T) - 2\mathcal{L}(\mathbf{w}) \tag{16}$$

Where $k = 2[|\mathbf{z}| + (|\mathbf{z}| + |\mathbf{x}|)(|\mathbf{y}| + |\mathbf{x}|)]$ is the total number of estimated parameters. Given that the algorithm terminates when a minimal valid set of states has been found, it may seem that this bias towards complexity is irrelevant. However given the minimal number of states, it is still possible to reallocate between exogenous and endogenous states. In this context the bias towards complexity means a we are likely to choose more exogenous states than truly exist, since these involve the estimation of more parameters than endogenous states, and since they enter at time t instead of time $t-1$ they likely contain more relevant information about time t endogenous variables. Therefore, it is still beneficial to score valid models in a penalised way.

3.3 Algorithm

Algorithm 1: Brute force hybrid state-space estimation algorithm

Input: *alpha*: significance level

Output: *all_valid_states*: A set of minimal sets of exogenous and endogenous states whose implied conditional independences are valid relative to the observed data

```
begin
    continue = true
    n_states = 0
    max_states = #observables − 1
    all_valid_states = list()
    while continue and n_states <= max_states:
        all_potential_states = get_potential_states(n_states)
        for potential_states ∈ all_potential_states:
            constraint_tests = get_constraint_tests(potential_states)
            score_tests = get_score_tests(potential_states)
            sig_level = alpha / length(constraint_tests)
            if every constraint_test .p_value > sig_level for constraint_test ∈
              constraint_tests:
                append potential_states to all_valid_states
                continue = false
    sort all_valid_states by score_tests
    return all_valid_states
```

The algorithm is very simple and is designed to reflect a few model selection heuristics.

The algorithm also assumes that the constraints validity is more important than score maximisation. The scores of models that are not valid relative to the constraints are irrelevant because these models are thrown out. The justification for this heuristic is outlined int 2.1.2. Essentially, unlike score functions, constraints directly encode information about a relevant sense of causality.

The algorithm applies the Minimal State Variable (MSV) criteria (McCallum, 1999). Once a model is found that is valid relative to the conditional independence tests, models with more states are not considered because these are inferior solutions. There are a number of relevant justifications for the application of this criteria. Firstly, this can be seen as the application of *Occam's Razor* to state-space models, wherein state variables have more complex dynamics than controls. Consider equations (2) - (4). Exogenous states are involved in all three equations, endogenous states two, and controls only one. Another way to see this is in figure 4. Among time t and $t-1$ variables, adding an exogenous state results in the addition of edges in four places, an endogenous state in three places, and a control in only two. Therefore, models with with fewer states, especially exogenous states are more parsimonious and are therefore preferable, all else equal. The MSV criteria also allows for a potentially very large increase in the speed of the algorithm. Without it we must consider every possible combinations of states. Since the choice of states is multinomial with three categories the complexity of this algorithm is $\mathcal{O}(3^k)$. However, if the ground truth has only $m < k$ states then we can skip $\sum_{r=m}^{k} 2^r \binom{k}{r}$ iterations, which is potentially many orders of magnitude if $m << k$. This algorithm is nonetheless highly inefficient, however, given that macroeconomic data is usually of relatively low dimension, it is still feasible in many important cases. There are undoubtedly many performance improvements which could be made to this algorithm, but this is left as a topic for for future research.

Given the fact that this algorithm considers every possible state-space model, and the result in 3.1, if conditional independence was completely observable this algorithm would always identify the ground truth state-space model. Unfortunately, in empirical applications with finite samples we have to rely on conditional independence tests that are not always correct. Since observed variables are assumed to follow a normal distribution, zero (partial) correlation is the same as (conditional) independence. So the conditional independence test that is used in this application is a t-test on the (partial) correlation of the relevant variables. It is important to note that these conditional independence tests are used in a somewhat unusual manner. Unlike most economic applications, the case we are interested in is the case in which we *do not reject* the null hypotheses that the correlation is 0. In each test, the significance level α is the probability of rejecting the hypothesis that the true (partial) correlation is 0, and therefore, rejecting the model under consideration.

For any given model we may consider a large number of conditional independence tests (the number is polynomial order in the number in each category), and so there is a multiple comparisons problem. In an attempt to correct for this a Bonferroni (1936) correction is applied. The Bonferroni correction adjusts the significance level assuming the independence

of tests. In this context tests are clearly not independent, therefore, given the correction, α should be interpreted as an upper bound on the probability of rejecting a correct model.

When investigating a statistical methodology there are two broad categories of properties that we may wish to establish. First are the asymptotic properties relating to the identifiability of the parameters of interest. In this context the goal is to establish that the algorithm selects the correct ground truth state-space model as the number of observations n goes to infinity. It is straightforward to (at least informally) demonstrate the asymptotic consistency of this algorithm. Consider data generated from a DGP characterised by (2) - (4) that satisfies (5) - (8). Sample (partial) correlation is a consistent estimator of (partial) correlation, so the partial correlations estimated for (5) - (8), and thus the probability of rejecting the true model goes to zero as $n \to \infty$. On the other hand, consider a DSGE model that does not have the same state-space representation as the DGP. The result in A implies that at least one of these the partial correlations in (5) - (6) is not zero, and therefore, the associated sample partial correlation does not converge to zero as $n \to \infty$. The t-statistic for this partial correlation goes to infinity and the chance of rejecting this incorrect model goes to 1 as $n \to \infty$.

The other properties to consider are the finite sample or statistical inference properties. These may be much more difficult to establish because the probability that the correct model is not chosen depends both on the probability that some other incorrect model is *valid*, and that that incorrect model has a higher score than the correct model. Both of these probabilities, in turn, depend on the alternative incorrect model under consideration. Therefore, instead of considering any theory in this regard, I will use simulations to demonstrate performance on finite samples in 5.

3.4 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 introduced 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 covariance 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.

3.5 Model Evaluation

Since the state-space models that are estimated here are fundamentally reduced-form models they will not circumvent the Lucas (1976) critique. When the state-space model is estimated directly from a micro-founded DSGE model these coefficient matrices are functions of the structural parameters of that model. However, when these matrices are estimated directly from data they can not, and should not be interpreted as structural parameters. The methodology in this paper is one of *model selection*, that is, evaluating whether a given DSGE model is consistent with observational data.

3.6 IRFs

One very common way of evaluating DSGE models is to compare the Impulse Response Functions (IRFs) they imply and to compare those with the IRFs of reduced form models such as VAR models (Ramey et al., 2016, p.83). This is also possible when directly estimating state-space models, and the results of this will be considered in the empirical section of this paper. IRFs are calculated, starting with a vector of initial values (shocks), by iteratively using the estimated matrices $\hat{\mathbf{A}} = \hat{\mathbf{E}}$ to calculate current time step values using past values. Note that this can be done for either exogenous or endogenous states, but not for controls, as changes to these are by construction not propagated through to future time steps.

Symbol	Name	Type
g	government spending	exogenous state
z	technology	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 for the baseline RBC model.

4 Data

In order to demonstrate the capability of the DAG methodology empirically I will work with both simulated and real macroeconomic data. Using simulated 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 opportunity to demonstrate that DAGs are also a powerful heuristic tool that can be implemented outside of a rigorously controlled environment. The remainder 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 more difficult and complex modelling challenge. I modified the simulation code slightly such that simulations would output 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 Baseline RBC

The first model which I have chosen to evaluate is the baseline RBC model provided by Pfeifer (2020). This model includes 11 variables which are summarised by table 1. This model contains two exogenous state variables: technology (z) and government spending (g), and one endogenous state: capital (k). There are two shocks in the model: eps_z that affects only technology directly and eps_g that affects only government spending directly, but as explained in section 2.2 these are dropped from the data.

Symbol	Name	Type
nu	policy rate	exogenous state
a	technology	exogenous state
z	preferences	exogenous state
p	price level	endogenous state
y	output	control
i	nominal interest	control
pi	inflation	control
y_gap	output gap	control
r_nat	natural interest rate	control
r_real	real interest rate	control
n	hours worked	control
m_real	real money balances	control
$m\_nominal$	nominal money balances	control
w	nominal wages	control
c	consumption	control
w_real	real wages	control
mu	mark-up	control

Table 2: Description of Variables for the baseline New Keynesian model.

This model was chosen as it is one of the simplest DSGE models and provides a good baseline to demonstrate the effectiveness of this methodology. In particular, the default calibration of this model which was used has autoregressive coefficients on the exogenous technology and government spending processes that are very close to one, and as a result there is a high degree of persistence in all variables in the model. this model will test the algorithms performance when the assumption of stationarity is challenged.

4.1.2 Baseline New Keynsian

New Keynesian models are extremely popular in modern macroeconomics and are thus a worthy test for this methodology. In particular, I use a model from Gali (2015) as provided by Pfeifer (2020). The variables in this model are summarised in table 2 [1]. This model has a total of six states: three exogenous (policy rate, technology and, preferences) for which there is one i.i.d. shock each, and one endogenous (price level).

4.2 US Data

To provide an example of real macroeconomic time-series, quarterly data from the US between years 1985-2005 were collected from FRED (2020) for 15 variables outlined in Table 3. All of the variables were detrended by taking the residuals of an estimated first order autoregression. Total factor productivity and capital stock were provided on an annual basis and were therefore interpolated quadratically. Full details of data preprocessing are available in the project repository (Hall-Hoffarth, 2020).

[1] Some control variables which were just linear functions of another variable were dropped, for example, annualised rates.

[2] Estimated as average return to the NASDAQ over the relevant period

Symbol	Name
pi	CPI Inflation
rm	Federal Funds Rate (Return to Money)
g	(Real) Government Expenditure
y	(Real) GDP
i	(Real) Private Investment
w	Median (Real) Wage
rk	Return to Capital [2]
z	Total Factor Productivity
u	Unemployment
l	Total Workforce
c	(Real) Personal Consumption

Table 3: Description of Variables for US Data

Index	**Exogenous States**	**Endogenous States**	**Wins**	**Valid**
1	g z	k	888	997
2	c l	k	109	109
3	g r	k	2	941
4	g w	k	1	986
5	g y	k	0	974
6	g i	k	0	996
7	c g	k	0	200
8	g l	k	0	4

Table 4: Small-sample results for RBC model

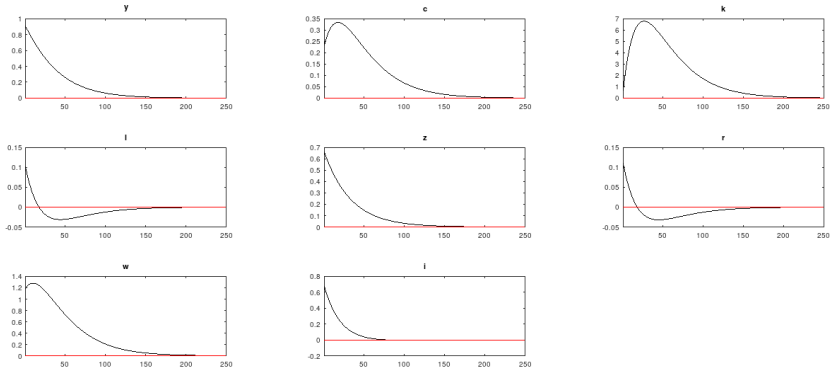
5 Results

In this section many of the properties of the proposed algorithm will be thoroughly investigated. Using simulated data allows for the possibility of many experiments to test these properties in a controlled environment. In particular, for the models under consideration two tests will be considered, one for each of the algorithmic properties outlined in section 3.3. To demonstrate asymptotic consistency, results from the algorithm for a very large number of samples (100,000) will be provided. To demonstrate the finite sample properties, results from a large number of runs of the algorithm (1000) with a relatively small sample size (100, the sample size of the real data) will be provided and discussed.

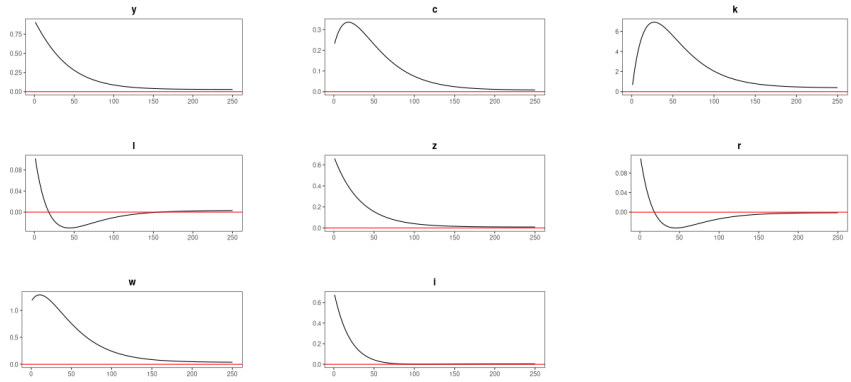
5.1 Baseline RBC

Using a the entire sample for the RBC model the algorithm successfully identifies the correct states, and no other (incorrect) models are valid. Figure 5 shows the impulse responses to a technology shock generated by the original simulation and the estimated model. There are almost identical, as they should be. This is a simple validation of the state-space representation and that fact that we have recovered the correct parameters using observational maximum likelihood.

Table 4 shows the results for the 1000 small sample tests. These results are promising for a number of reasons. Firstly, we see that with a sample size of 100, the algorithm still manages to select the correct states nearly 90% of the time, and the correct states were only rejected



(a) Original Simulation



(b) Ground Truth DAG / Estimated DAG

Figure 5: IRFs to a one standard deviation technology shock generated by the original simulation and estimated model.

in only 3 iterations or 0.3%. Therefore, it seems that the true probability of rejecting the correct model is much lower than the 5% upper bound. On the other hand, the conditional independence tests seem to have a good deal of power to reject incorrect models. Out of the 835 candidates tested in each iteration, only 8 were ever valid in any iteration, despite the fact that the true model was almost never rejected. Furthermore, sorting by the score function (bic) also seems to have had the intended effect. Models 3 through 7 were all valid for a high number of iterations but managed to be picked by the algorithm only 3 times combined because the scoring favoured the true model.

There was only one model other than the true model that was selected a substantive number of times which is the model with exogenous states c and l, and endogenous state k. This model had the highest score every time it was valid. It seems likely that this model score so highly because these c and l have very high autoregressive coefficients (0.994 and 0.972 respectively), and as a result they get a high likelihood score when treated as exogenous states compared with g and z, while themselves being highly correlated with g and z such that the prediction while treating g and z as controls also gets a relatively high likelihood score. The conclusion here is that the algorithm may run into difficulties in small samples if there is a high degree of multicollinearity between variables. However, this problem is benign because if variables are highly collinear then they effectively measure the same thing and so little is lost by dropping one of them. This is why this was done for the New Keynesian model.

5.2 Baseline New Keynesian

5.3 US Data

6 Discussion and Conclusion

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, *8*, 3–62.

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*.

Gali, J. (2015). *Monetary policy, inflation, and the business cycle: An introduction to the new keynesian framework and its applications*. Princeton University Press.

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

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.

King, R. G., Plosser, C. I., & Rebelo, S. T. (1988). Production, growth and business cycles: Ii. new directions. *Journal of Monetary Economics*, *21*(2-3), 309–341.

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.

McCallum, B. T. (1999). Role of the minimal state variable criterion in rational expectations models. *International finance and financial crises* (pp. 151–176). Springer.

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. (2009). *Causality*. Cambridge university press.

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

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

Ramey, V. A., West, K. D., Taylor, J. B., & Woodford, M. (2016). Handbook of macroeconomics. *by JB Taylor and H. Uhlig. North-Holland. Chap. Macroeconomic Shocks and Their Propagation*, 71–161.

Ravenna, F. (2007). Vector autoregressions and reduced form representations of dsge models. *Journal of monetary economics*, *54*(7), 2048–2064.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

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.

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

A Proofs

Theorem 2. *Let M be a log-linearised DSGE model that generates a distribution $f(\mathbf{w})$ over a set of variables $\mathbf{w}$, which can be partitioned into $\mathbf{z}$, $\mathbf{x}$, and $\mathbf{y}$ (exogenous states, endogenous states, and controls). Further suppose that M is faithfully represented by some DAG g of the form of figure 4 according to the partitioning of $\mathbf{w}$. Then g is the unique faithful DAG that satisfies both (5) and (6).*

Proof. Suppose not. Then M is faithfully represented by a DAG h which is different to g. Since M is still a log-linear DSGE solution it must still have a faithful DAG representation of the general form in figure (4). Therefore, the difference must be that h classifies one or more of the variables in $\mathbf{w}$ differently than g. Define the following notation: g_x is the set of variables that are categorised as endogenous states in DAG g and likewise for h and other categories.
Continue by considering cases:
Case 1: $a \in g_y$ and $a \in h_x$
(6) fails because in g $\mathbf{z}_t$ has a direct path to a in g. Contradiction.
Case 2: $a \in g_y$ and $a \in h_z$
(6) fails because there is a direct path from $\mathbf{x}_t$ to a in g. Contradiction.
Case 3: $a \in g_x$ and $a \in h_y$
(5) fails because a is not in the conditioning set for this test in h and therefore there is an unblocked backdoor path from a to the other time t endogenous variables in g. Contradiction.
Case 4: $a \in g_x$ and $a \in h_z$
(6) fails because there is a direct path from $\mathbf{x}_t$ to a in g. Contradiction. Case 5: $a \in g_z$ and $a \in h_y$
(5) fails because there is a direct path from a to any time t endogenous variable in g. Contradiction.
Case 6: $a \in g_z$ and $a \in h_x$
(5) fails because there is a direct path from a to any time t endogenous variable in g. Contradiction. □