Essays in the Application of Machine Learning Methods to DSGE Models

Emmet Hall-Hoffarth

Somerville College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2025

Supervised by Prof. Michael McMahon and Dr. Jeremy Large

Abstract

This thesis comprises three stand-alone papers that relate to the application of machine-learning methods to macroeconomics.

The first chapter, *Non-Linear Approximations of DSGE Models with Neural-Networks and Hard Constraints*, contributes to a growing literature surrounding the use of deep neural-networks to obtain a global and non-linear solution to DSGE models, particularly those featuring rich heterogeneity such as HANK models. This chapter identifies some drawbacks of the commonly recommended approach of implementing model constraints using a penalty function. In particular, simulating states forward may result in divergence, and it is difficult for the obtained solution to capture discontinuities at boundaries, such as agents' budget constraint. These discontinuities are key components of many models, such as Heterogeneous Agent New Keynesian (HANK) models, which are often suggested as a key application for these methods. This chapter then introduces an alternative *hard-constraint* solution method, whereby the outputs of the neural-network are rescaled so as to always satisfy the economic constraints of the model by construction. An exercise solving a simple HANK model demonstrates the quantitative and qualitative benefits of this approach.

The second chapter, *HANK and the Minimum Wage*, applies the method in the first chapter to consider the implications of a minimum wage in a HANK model. The results of this chapter reinforce the classical theories about the issue. Higher minimum wages are found to have an adverse effect on inflation and output, in both the static and dynamic sense. The total effect is decomposed into a *redistributive* effect resulting from low productivity workers increasing demand and labour supply in response to higher wages, and a *distortionary* effect resulting from all other workers having their wages reduced when firms, who now internalise that the average output per unit wage paid is lower, reduce their demand for labour. The model also implies that minimum wages have an ambiguous effect on inequality, as "middle-class" households are the most adversely affected. Poorer households are more likely to benefit directly, whereas richer households are mostly unaffected as they only earn a small portion of their income from labour, and can buy savings from the "middle-class" households who are forced to sell to self-insure against the adverse shock to their labour incomes.

The third chapter, *Causal Discovery of Macroeconomic State-Space Models*, introduces a novel

technique based on *causal-discovery algorithms* that can identify whether the state-space generated by a log-linearised DSGE model is consistent with some observational data. This can be used to test certain modelling decisions in data, such as whether inflation should be modelled as persistent (as a state-variable) or purely forward-looking (as a control). The consistency of a model is determined by treating the models' state-space as a Directed Acyclical Graph (DAG), and testing the thereby implied conditional independence relationships. It is shown that in the limit these conditional independence relationships are sufficient to identify a unique *faithful* state-space, assuming the true data-generating process is some log-linear DSGE model. A statistical test for these conditional independence relationships is then introduced, and a simulation demonstrates that for realistic data sizes, this test identifies the true data-generating process in about 95% of trials. The statistical test is then applied to real macroeconomic data from the US resulting in a model that broadly agrees with the existing literature. In particular, my solution identifies that inflation should be treated as persistent, consistent with the findings of Christiano et al. (2005).

Declaration

Supervision The writing of this thesis was supervised by Prof. Michael McMahon and Dr. Jeremy Large.

Authorship All three papers presented in this thesis are single-authored.

Funding This thesis was written with funding from the University of Oxford Department of Economics.

Total Words This thesis contains approximately 54487 words. This was calculated by counting the words on a representative page (page 10, 529 words) and multiplying by the page count, excluding the preamble, bibliographies, and appendices (103 pages).

Acknowledgements

Firstly, I would like to thank my supervisors, Michael and Jeremy, who provided invaluable guidance and mentorship throughout the writing of my thesis. I would also like to thank various members of faculty in Oxford for their feedback in workshops and elsewhere, specifically, Federica, Francesco, Andrea, Max, and, Kevin. Furthermore, I would like to thank all the other students in the department, in particular, Amelie, Lukas, Momo, Tobi, Matt, Alena, Lovisa, JG, Juliette, and Rosi, whose friendship and community kept me motivated and inspired. Similarly, I would like to thank some of those at the Frankfurt School who accompanied me for part of this journey, namely, Benjamin, Felix, Daniil, Jonas, Malte, Giorgio, Vittorio, Paulina, Sophie, and Robin. I would like to thank my friends from the MPhil (and elsewhere!) for reminding me that there is more to life than writing the next paper. Thank you to my parents, without whose support and empathy none of this would have been possible. And finally, thank you Setare, for always being there when I needed you the most. I couldn't have done it without you.

Contents

Introduction

Machine-learning is not a new field, but it is certainly a rapidly advancing one. Since the writing of this thesis began, on November 30, 2022, OpenAI released ChatGPT, an AI-chatbot that almost immediately obtained widespread popularity due to its ability to answer and (seemingly) even deeply reason about a wide range of topics in a human-like manner (Marr, 2023). Since then many similar models have been released that have iterated on this first popular version. However, these models trace their success back to the attention mechanism proposed by Vaswani et al. (2017), and even more fundamentally, are based upon deep neural-networks, first proposed in the modern sense by Rosenblatt (1958). Within the field of economics, despite the promise of these tools, there has been relatively slow (not to say a lack of) adoption, perhaps due to the discipline's somewhat difficult to align focus on causal and explainable models. Nonetheless, there has also been an increase in interest and applications in recent years in diverse subfields such as central bank communications (Hansen et al., 2018), development (Donaldson & Storeygard, 2016), forecasting (Larsen et al., 2021), and econometrics (Athey & Wager, 2019). This thesis follows this trend and consists of three stand-alone papers, which form the subsequent three chapters. The thesis aims to contribute to the understanding of how various machine-learning techniques, in particular, causal-discovery algorithms and deep neural-networks, can be leveraged in the field of Dynamic Stochastic General Equilibrium (DSGE) modelling.

Chapter 1 introduces a focus that is also continued in Chapter 2. Specifically, Chapter 1 makes a methodological contribution to a new and growing literature on the potential of deep neural-networks for solving and estimating DSGE models, originally proposed by Duarte (2018) and expanded on by Fernández-Villaverde et al. (2020), Maliar et al. (2021), Azinovic et al. (2022), and Kase et al. (2022). The fundamental proposition of all of these papers is a very compelling one. Solving (and to and even greater extent estimating) DSGE models with both a high-dimensional state-space, and where aggregate non-linearities are prevalent poses an intractable computational problem for standard methods, which are broadly only able to deal with one of these two factors at the same time (Kase et al., 2022). This combination of features arises, for example, in models that feature a rich cross-sectional distribution of agents, and where the Zero Lower Bound (ZLB) on nominal interest rates sometimes binds in equilibrium. Furthermore, in many other applications deep neural-networks have shown themselves to be very effective precisely at estimating non-linear functions over high dimensional input spaces (Adcock et al., 2020). Therefore, deep neural-networks seem like a likely candidate to provide a breakthrough in solving DSGE models with fewer simplifying assumptions and increased realism.

However, as this methodology was only relatively recently proposed, it has still (as of time of writing) not gained widespread adoption, and there are a number of details to be worked out in order to facilitate and encourage this. In the first chapter, I contribute an improvement to one such detail, which is how to deal with the constraints implied by the model. Macroeconomic models contain a number of constraints, such as aggregate resource constraints (e.g. $Y_t = C_t + G_t$), and

idiosyncratic constraints like a borrowing constraint $b_t^i \geq \underline{B}$. The neural-network being used to approximate policy functions needs to be made aware of these constraints somehow, because along with the model's first-order conditions, they are essential in pinning down the model's equilibrium. Most of the existing literature suggests doing this by applying a penalty to the loss function that the neural-network seeks to minimise, which increases as the proposed solution drifts further from satisfying the models' constraints.

However, in my research I found that such an approach has a number of drawbacks. In particular, since constraints are not strongly enforced, when new states are generated by simulating the currently estimated policy function forward, and especially at the beginning of training when the current policy is not yet very accurate, it is possible for states to diverge away from the true equilibrium states of the model, resulting in very slow convergence back, or in the states needing to be reset to the deterministic steady-state. Furthermore, in many models including the Heterogeneous Agent New-Keynesian (HANK) model considered in that chapter, a key feature is that at the borrowing constraint behaviour changes sharply, however, a penalty based approach encourages a smooth transition. In an attempt to improve on these limitations, I propose a new solution method that involves rescaling the outputs of the policy neural-network so that all constraints of the model are satisfied by any output, by construction. While there are various machine-learning techniques for restricting outputs to certain domains, none of these are applicable to the specific combination of constraints required by a DSGE model, thus a novel rescaling function is proposed. In an experiment involving a simple, but non-linear HANK model, I was able to obtain qualitatively and quantitatively superior results using this rescaling approach in comparison to the penalty based approach.

The second chapter applies the methodology introduced in the first chapter in order to consider the implications of a minimum wage policy in a HANK model. The application is on one hand a showcase of what is possible with this methodology, as both the rich cross-sectional distribution and non-linearity are key components of the model, but on the other hand it is also a worthy contribution in its own right, as despite the prominence of the topic in the economic literature, relatively few papers have already approached the minimum wage from a DSGE modelling perspective. Leveraging the HANK framework, the model presented in that chapter allows for the measurement of the aggregate effect of a minimum wage change whose direct effect is purely cross-sectional, in other words, a change that only directly affects the proportion of workers for whom the minimum wage is binding. Furthermore, I break down the total effect into two channels: a *redistributive* effect resulting from the benefactors of the minimum wage increasing their consumption demand and labour supply, and a *distortionary* effect caused by firms reducing their demand for labour, as a higher minimum wage decreases the average efficiency of the workers they hire, ultimately resulting in lower wages for workers earning above the minimum wage.

The implications of the model are broadly speaking negative about the minimum wage policy. I find that the effect of a minimum wage is nearly always to decrease output, both in a static and dynamic sense, in many cases it is inflationary, and it also does not unambiguously reduce inequality. While the less-productive (and therefore usually poorest) workers do benefit directly, the costs of the policy fall mostly on the "middle-class" of workers, rather than the richest. These "middle-class" workers generate a large portion of their income from labour, and therefore their income is particularly strongly reduced as firms reduce their demand for labour. These workers tend to sell off their assets to self-insure this adverse shock, and these are bought up by the already wealthy agents, who derive only a small portion of their income from labour, and whose income is therefore relatively unaffected.

Chapter 3 switches to a separate theme from the first two papers and considers the application of causal-discovery algorithms, popularised by Spirtes et al. (2000) and Pearl (2009) to DSGE models. Every DSGE model, once linearised and solved, can be represented by a state-space that divides every observed variable into exactly one of three categories: exogenous state, endogenous state, and policy (or controls). In this chapter I contribute a statistical test and algorithm that can identify a unique state-space that shares the causal structure of some given data in a particular sense. This method can therefore be used for model selection: many DSGE models may generate the same state-space, but the test rules out any model that does not have the correct state-space as the true data generating process. Therefore, the method can be used to weigh in on some fundamental macroeconomic modelling questions, such as whether consumption (J. C. Fuhrer, 2000) or inflation (J. Fuhrer & Moore, 1995) are persistent.

While the state-space is usually not an object directly of interest to researchers, it does have a number of properties that can be exploited. The state-space can be represented by a Directed Acyclical Graph (DAG), as in Pearl (1995). This DAG summarises the dependencies that exist between modelled variables, and implies that certain conditional independence relationships should exist between them. For example, since time t endogenous states and controls (or policies) are both (linear) functions of time t exogenous states and time $t-1$ endogenous states, the former two should be conditionally independent of each other given the latter two. Conditional independence is testable given observational data, and so-called *causal discovery* algorithms such as Spirtes and Glymour (1991) exploit this fact to find DAGs that are compatible with the causal structure of data. However, these algorithms search over the space of *all possible* DAGs that could be constructed from observed variables. This number scales super-exponentially in the number of variables observed, and therefore leads to an intractable search space. Therefore, I develop a new statistical test and algorithm that reduces the multiple-testing problem by only testing each proposed model once, and a novel algorithm that only searches within the space of possible DSGE state-spaces, which is a vanishingly small subset of all possible DAGs, greatly improving the computational and statistical tractability of the problem.

Using simulated data I show that with a realistic sample size my method is able to identify the true data generating process from thousands of potential candidates about 95% of the time. In an exercise applying the method to real data, the algorithm proposes a reasonable model which is broadly consistent with models that have been proposed in the literature. In particular, I find that the data supports a model in which inflation and consumption are persistent, in other words, enter into the model as state variables.

All three main chapters of this thesis contribute to and explore applications of machine-learning methodologies in macroeconomics. Given the rapid advance of the field in general, it is likely the next years will see a great deal of development in this area. I hope that these contributions will aid future researchers in the furthering of this field.

Bibliography for Introduction

Adcock, B., Brugiapaglia, S., Dexter, N., & Moraga, S. (2020). Deep neural networks are effective at learning high-dimensional hilbert-valued functions from limited data. *arXiv preprint arXiv:2012.06081*.

Athey, S., & Wager, S. (2019). Efficient policy learning. *Econometrica*, *87*(1), 371–406. https://doi.org/10.3982/ECTA14518

Azinovic, M., Gaegauf, L., & Scheidegger, S. (2022). Deep equilibrium nets. *International Economic Review*.

Donaldson, D., & Storeygard, A. (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, *30*(1), 99–118. https://doi.org/10.1257/jep.30.1.99

Duarte, V. (2018). Machine learning for continuous-time economics. *Available at SSRN 3012602*.

Fernández-Villaverde, J., Nuño, G., Sorg-Langhans, G., & Vogler, M. (2020). Solving high-dimensional dynamic programming problems using deep learning. *Unpublished working paper*.

Fuhrer, J., & Moore, G. (1995). Inflation persistence. *The Quarterly Journal of Economics*, *110*(1), 127–159.

Fuhrer, J. C. (2000). Habit formation in consumption and its implications for monetary-policy models. *American Economic Review*, *90*(3), 367–390.

Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation: Evidence from the federal open market committee. *The Quarterly Journal of Economics*, *133*(4), 1787–1841. https://doi.org/10.1093/qje/qjy017

Kase, H., Melosi, L., & Rottner, M. (2022). Estimating nonlinear heterogeneous agents models with neural networks.

Larsen, V. H., Thorsrud, I., & Zhulanova, Y. (2021). News sentiment and inflation expectations. *Journal of Monetary Economics*, *117*, 687–701. https://doi.org/10.1016/j.jmone.2020.06.010

Maliar, L., Maliar, S., & Winant, P. (2021). Deep learning for solving dynamic economic models. *Journal of Monetary Economics*, *122*, 76–101.

Marr, B. (2023). A short history of chatgpt: How we got to where we are today. *Forbes*. Retrieved May 9, 2025, from https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/

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

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

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

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. (2000). *Causation, prediction, and search*. MIT press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Chapter 1

Non-Linear Approximations of DSGE Models with Neural-Networks and Hard Constraints

Abstract

Recently a number of papers have suggested using neural-networks in order to approximate policy functions in DSGE models, while avoiding the curse of dimensionality, which for example arises when solving many Heterogeneous Agent New-Keynesian (HANK) models, and while preserving non-linearity. One important step of this method is to represent the constraints of the economic model in question in the outputs of the neural-network. I propose, and demonstrate the advantages of, a novel approach to handling these constraints which involves directly constraining the neural-network outputs, such that the economic constraints are satisfied by construction. This is achieved by a combination of rescaling operations. These are differentiable and therefore compatible with the standard gradient-descent approach used to train neural-networks. This has a number of attractive properties, and is shown to out-perform the penalty-based approach suggested by the existing literature, which while theoretically sound, has some disadvantages in practice, which are discussed.

1.1 Introduction

Recently, there has been a growth in interest in the literature related to the use of neural-networks for the approximation, and even more recently, also the estimation of Dynamic Stochastic General Equilibrium (DSGE) models. This interest comes as a result of the promise of this new approach, which is to alleviate the classical *curse of dimensionality* of Bellman (1957) that is frequently encountered in the domain of heterogeneous agent models. These benefits are obtained by exploiting the well-documented affinity of neural-networks to high-dimensional and sparse input spaces (Adcock et al., 2020). However, since this technique has only relatively recently been introduced by the likes of L. Maliar et al. (2021), Kase et al. (2022), and Azinovic et al. (2022), there are still details to be worked out before this method can achieve widespread appeal.

This paper makes two main contributions. Firstly, this paper identifies some shortcomings of one aspect of the standard technique that has so-far been proposed in the literature, which involves using a penalty function to enforce economic constraints in the output of the neural-network. In particular, this penalty can be difficult to tune, and when constraints are not strongly enforced states can diverge causing estimation to breakdown, and it can be difficult for the model to capture discontinuities in behaviour at boundaries. Secondly, this paper proposes an alternative *hard-constraint* approach, which involves rescaling neural-network outputs such that the economic constraints are always exactly fulfilled. I argue that this approach resolves many of the shortcomings of the soft-constraint approach: it removes the incentive for the learning algorithm to trade off the constraint with other criteria, it ensures that the future generated states are always at least feasible, and it allows constraints to bind exactly, if that is an equilibrium outcome. This is achieved through a combination of activation functions that constrain the individual outputs of the neural-network such that the idiosyncratic constraints are satisfied, and contextual rescaling (more detail in Section 1.3) of the output vector, so that aggregate constraints are also satisfied. Aside from conceptually addressing many of the shortcomings of the standard *soft-constraint* approach, as demonstrated in Section 1.6, this approach generates significantly more precise results.

In the process of this approximation, one important step is the representation of the constraints of the economic model, and the resulting Karush-Kuhn-Tucker (KKT) conditions, in the approximating function. For example, an important feature of HANK models is an idiosyncratic borrowing constraint experienced by agents, such that their net wealth cannot fall below some lower bound. This results in agents who are close to or are at the budget constraint having a substantially higher Marginal Propensity to Consume (MPC) than those who are further from the constraint. This dispersion in MPC allows for a pathway connecting inequality and monetary policy, and is important to match many empirical facts (Kaplan et al., 2018). Most of the economics literature related to the application of neural-networks to the approximation of DSGE models, with the exception of Azinovic and Žemlička (2023), implement at least some of these economic constraints by allowing the neural-network to select policies that violate the economic constraints, and then applying a

penalty to this violation. The penalty can be proportional to either a *Fischer-Burmeister* (FB) function (defined later in (1.2) in Section 1.2.1) in the case of inequality constraints, or simply the (squared) mean deviation in the case of equality constraints. I describe these as *soft-constraint* methods, in contrast to the *hard-constraint* method that I propose in this paper.

While theoretically sound, in practice this soft-constraint approach has a number of drawbacks. Since the (squared) FB function is quadratic in its arguments, the penalty becomes extremely small and has a relatively flat gradient in a neighbourhood around the constraint. This means that getting precise approximations is very difficult, as economically meaningful constraint violations may still generate very small losses.[1] Also, since the parameters are learned via gradient descent, the flat gradient means that parameter updates are imprecise. Therefore, the neural-network choosing an output value exactly on the constraint is a measure-zero event. This may be problematic if the proportion of agents on the constraint is itself a outcome of interest, or if agents' behaviour is expected to change discontinuously at the constraint.

Furthermore, since the learning algorithm is attempting to minimise a loss containing other optimality conditions derived from the first-order conditions (FOCs) of the economic model, the optimiser might prefer to violate the constraints if this improves the other optimality conditions at the expense of increasing the cost imposed by the penalty functions. For example, the optimiser might choose to consume until the marginal penalty of violating the budget constraint is equal to the marginal utility of consumption. Indeed, the approximated solution should violate the constraints in at least some cases in order for the learning algorithm to be able to get an estimate of the gradient of the loss with respect to the FB penalty function. If, in order to counteract this, a large weight is placed on the FB function, the model may learn to produce outputs that are always far inside the constraints, so as to avoid the penalty entirely. Given these limitations, improving the quality of results relies on either carefully tuning the *hyper-parameters* of the model, for example, the weights placed on penalties related to the various constraints of the model, or training the model for a very long time. Training these neural-networks, while it provides a more general solution, does still take a long time compared to other solution methods, therefore, the necessity of this fine-tuning is an undesirable feature of these techniques.

Perhaps the most salient issue is that when currently approximated policy ($f(X_t, \hat{\theta})$) is used to generate future states (X_{t+h}) by simulating the model forward as in S. Maliar et al. (2011), small violations of the constraints will be propagated, which can lead to states diverging far away from the true ergodic distribution of states. In some cases this can cause the approximation procedure to breakdown entirely, or require states to frequently be reset to initial values. This is problematic because in the high-dimensional setting where neural-network approximation is likely to have an advantage over other approaches, such endogenous generation of states is necessary to achieve sufficient dimensionality reduction so as to make approximation feasible.

[1] To be clear, losses here refer to the penalty function which the parameter optimiser aims to minimise, not welfare losses

To collect more evidence for my hard-constraint approach, in Section 1.6.2 this paper will also consider two intermediate constraint regimes, in which some of the constraints are satisfied by construction, while the others are implemented via a penalty. The results, as expected, rest in between the previously discussed cases. Where the constraints are hard, relevant features of the model are faithfully reproduced. Where the constraints are implemented via a penalty, errors accumulate and are quantitatively and qualitatively meaningful. These results show that choosing between constraint implementations, even if they would seem to be equivalent theoretically, can cause substantially different approximations to be obtained. Furthermore, they show that in order to achieve the best level of precision even making a subset of the constraints hard, while simpler to implement, is not sufficient. Instead, I recommend to treat all constraints with lexicographic preference — that is to restrict the output of the approximating function to the range of feasible values, before only then attempting to find optimal policies within this subspace.

My hard-constraint approach is similar to the *market-clearing layer* of Azinovic and Žemlička (2023), who developed their ideas independently and concurrently. Fundamentally, both papers attempt to address the limitations of a penalty-based approach by enforcing the economic constraints on neural-network outputs via a rescaling. Relative to that paper, my solution, which includes three steps of rescaling that will be discussed in detail in 3.3 has the benefit that it avoids the need to solve a quadratic optimisation problem during each model evaluation. My application also differs substantially from theirs, as I show how this type of approach can be used for solving HANK models.

The remainder of this paper is organised as follows. Section 1.2 summarises the relevant literature on using neural-networks to solve DSGE models and HANK models (as used in the application in this paper). Section 1.3 outlines the methodological contribution of this paper, which is an algorithm to solve DSGE models with hard-constraints. Section 1.4 introduces the economic model used in the application in this paper, and explains how the suggested method can be implemented to make the constraints hard for this particular economic model. Section 1.5 discusses details of the implementation used to produce the results which are provided and discussed in Section 1.6. Finally, Section 1.7 briefly concludes.

1.2 Literature

1.2.1 Neural-networks and DSGE

The concept of solving control systems using neural-networks has been discussed for a relatively long time, although it has only recently begun to be discussed in the macroeconomics literature. Early papers include Dissanayake and Phan-Thien (1994) and Aarts and Van Der Veer (2001). These papers consider the application of neural-networks to solve partial differential equations, a type of problem that is common in physics and engineering that is in many way analogous to

	High-dimensionality	Idiosyncratic Non-linearity	Aggregate Non-linearity
Perturbation	✓	x	x
Projection	x	✓	✓
Reiter (2009)	✓	✓	x

Table 1.1: The relative strengths and weaknesses of different approximation techniques

solving DSGE models. However, the first paper in this vein in economics was likely Duarte (2018), who solves a Lucas style financial asset model with an arbitrary number of assets, as well as a neoclassical growth model, in continuous time, using a neural-network. Subsequent papers which focus specifically on solving DSGE models include Azinovic et al. (2022), Fernández-Villaverde et al. (2020), and L. Maliar et al. (2021). Particularly relevant to this paper is the Kase et al. (2022), who approximate a 100-agent HANK model, and also show using a simulation that they can recover the true parameters using an estimation technique that combines the particle filter of Fernández-Villaverde and Rubio-Ramírez (2007) with a likelihood surrogate as in Smith et al. (2011). In this case another neural-network is used as the surrogate.

In general, the process of solving DSGE models involves finding a function $f(X_{t-1}, \epsilon_t; \Gamma)$ of the *states* of the model (X_{t-1}), exogenous or structural shocks ϵ_t and, structural parameters Γ which maps to choices of agents Y_t that are both optimal in the sense that they minimise the error in the models' FOCs, satisfy constraints on the agents' choices embodied by the KKT conditions of the model, and clear markets. Since in many interesting cases the function f cannot be found analytically, we instead turn to approximations $\hat{f}(X_{t-1}, \epsilon_t, \Gamma; \theta)$. There are currently two approaches which are commonly used for the approximation of DSGE models. The first, *perturbation*, involves approximating f with a Taylor series expansion around the steady-state of the model. This is particularly applicable when this is a first-order approximation, which is feasible in a high-dimensional context, in other words, when the dimension of X_{t-1} is large. However, even if the approximation is a higher order one, this approach is poor at representing strong non-linearities, especially kinks or jumps, and the quality of the approximation may be poor if the exogenous shocks are large, or the model is far from its steady-state. The other approach, *projection*, involves estimating the policy function over a number of points in the state-space (known as a grid), and then interpolating between these points, for example with a cubic-spline. This approach is better able to deal with non-linearities, however, the number of grid points required increases quickly with the dimension of the state-space, even if techniques such as *sparse grids* are used (Judd et al., 2014). Therefore, this technique suffers from the curse of dimensionality. In practice, many macroeconomic researchers use methods based on Reiter (2009), which combine projection in idiosyncratic variables with perturbation in aggregates. Still, this is not able to represent strongly non-linear dynamics in the aggregate.

Taken together, this implies a dilemma for DSGE solution methods, as discussed in Han and Yang (2021): none of the existing methods are able to offer a compelling solution in a both high-

dimensional and non-linear (in the aggregate) setting. This gap may be important, in particular, in the growing literature regarding HANK models (Kaplan et al., 2018), which contain both a high-dimensional state-space and non-linearities induced by for example an Effective Lower Bound (ELB) on the nominal interest rate.

Recently, advances in machine-learning have pointed towards a potential solution to this problem: neural-networks. The suggestion is simply to use some neural-network functional-form (discussed in more detail later) as the approximating function $\hat{f}(X_{t-1}, \epsilon_t, \Gamma; \theta)$, and then to learn its parameters via Stochastic Gradient Descent (SGD),[2] as is common in machine-learning applications. There is a good theoretical justification for doing so thanks to the highly-influential universal approximation theorem of Hornik et al. (1989), which tells us that any function, even if poorly behaved and highly non-linear, can be approximated to arbitrary precision by some single-layer, fully connected, and arbitrarily wide feed-forward neural-network. Furthermore, it has been widely documented that neural-networks perform very well in *big data* applications, where high-dimensional inputs are a defining characteristic, and where other approaches therefore suffer from the curse of dimensionality, such as in image recognition and natural language processing (Adcock et al., 2020; Bach, 2017).

Each of the referenced papers concerning the application of neural-networks to solving DSGE models follow the same core approach, which I will outline here. The basic idea is to use a fully-connected, feed-forward neural-network as a global parameterisation of the policy functions that are to be approximated. The inputs of this neural-network are the states of the model which I represent here with X_{t-1}, and the *i.i.d.* exogenous or structural shocks ϵ_t. Kase et al. (2022) also suggest adding the structural parameters Γ to the vector of inputs and varying these throughout the training process in order to fit the model for a range of different calibrations at the same time. The outputs of the neural-network are the policy functions Y_t. The neural-network contains *trainable parameters* $\theta = \{W_i, b_i\}_{i=1}^n$ known respectively as weights and biases, which are optimised during the training procedure, as well as some fixed *activation functions* $\left\{\sigma^i\right\}_{i=1}^n$, which are some non-linear vector functions, for example, the ReLu function: $\sigma^i(x) = \max\{0, x\}$. Equation (1.1) shows how all of these components are related.

$$\hat{Y}_t = \sigma^n \left(W_n \times \sigma^{n-1}(\cdots \sigma^0 \left(W_0 \times \{X_{t-1}, \epsilon_t, \Gamma\} + b_0\right) \cdots) + b_n\right) \tag{1.1}$$

The trainable parameters are chosen to minimise a loss function that measures the quality of the fit of the model. In most machine-learning applications this is done by comparing the predictions of the neural-network ($\hat{Y}_t$) to some observed data or *labels* (Y_t) that are treated as a ground-truth.[3] However, this application is somewhat different, as the exact optimal policies are not known, since this is what we are trying to solve for. At this point we could create labels (targets for the neural-

[2] See Appendix A for a brief description of this.

[3] In the machine-learning literature this is known as *supervised learning*.

network to directly try to predict) via value-function iteration, and essentially interpolate between them using the neural-network, however, this would in some sense be equivalent to standard projection methods, and would have all of the same drawbacks, in particular, with respect to the curse of dimensionality. Instead, the quality of the approximation can be evaluated by comparing the predicted policy functions to the optimality conditions of the economic model. This allows for the approximation to remain grid-free. L. Maliar et al. (2021) outline three general ways of constructing the loss function: using the Bellman equation error, the Euler equation error (first-order conditions), and in the case of life-cycle models, minus the lifetime utility. In this paper I focus on the second method using first-order conditions.[4] An example of this will given in later in the application in equation (1.28). This type of loss function has the property that when the loss is exactly zero, the policies satisfy the first-order conditions exactly, so they are the exact solution for that given set of states. In practice, the loss will never reach zero because we are dealing with a global approximation. The challenge then, is to make the loss as small as possible, for as many states as possible, in particular, the states that are likely to be generated by the model in equilibrium.

In order to deal with the constraints of the model, all of the above referenced papers recommend the same strategy. This is to add a penalty to the loss function calculated using the FB function (Fischer, 1992) for inequality constraints, as defined in equation (1.2), and the squared mean difference for equality constraints (i.e. market clearing conditions). Again, these penalties are exactly zero when the constraints are satisfied by the current policy, and otherwise grow quadratically in the size of the violation of the constraint. The FB function is part of a class of functions known as complementarity functions (Chen et al., 2000). These functions are used as continuously differentiable analogues to KKT conditions (which are required for SGD), and are equal to zero if and only if the KKT conditions are satisfied exactly. Thus, minimising a loss function which includes a complementarity function to zero should also imply that any constraints on that model are also satisfied.

$$\Psi_{fb}(a,b)^2 = \left(a + b - \sqrt{a^2 + b^2}\right)^2 \tag{1.2}$$

The parameters of this neural-network are chosen to minimise a loss function via SGD.[5] In general this means that a *batch* of *inputs* (in this context $\{X_{t-1}, \epsilon_t, \Gamma\}$) are randomly sampled, and then the parameters are moved by some small step in the direction of the gradient of the loss with respect to the parameters. In this application the sampling of shocks (ϵ_t) and structural parameters

[4]The application in this paper is not a life-cycle model, so the third approach is not relevant. The first approach, using the Bellman equation, despite being more general, unfortunately did not seem to be as well behaved in my own experimentation. I speculate that this because it may be difficult for the optimiser to target the derivative of the outputs of the neural-network ($V'(\cdot)$), although more research is certainly required on the relative strengths and weaknesses of these approaches.

[5]Note that the gradient of the loss with respect to the neural-network parameters is usually calculated exactly via automatic-differentiation and *backpropogation* (See appendix A for more details) rather than using finite-differences.

(Γ) is straightforward, as the distribution of shocks is usually assumed explicitly, and the structural parameters can be uniformly sampled from some predefined range, however, sampling the states (X_{t-1}) is somewhat more involved.

Initially, the states may be sampled randomly, or simply fixed at some initial condition, however, thereafter the states are generated by simulating forward the state transition of the model using the current approximation of the policy function. In this way the ergodic set of the equilibrium and the policy functions are learned simultaneously. This is particularly beneficial because the set of states that are actually visited in equilibrium is usually a very small subset of the hypercube implied by the entire state-space (S. Maliar et al., 2011), and this difference is particularly pronounced when the dimensionality of the state-space is high. Another key difference to other methods in the vein of Reiter (2009), that allows for strong dimensionality reduction is the fact that the neural-network represents a global functional form for the approximation over the entire state-space, rather than an interpolation between knots. The lack of necessity for discretisation means that when using the neural-network approach, the total dimension of the inputs grows much more slowly in the number of states of the model.

Nonetheless, this procedure is not without drawbacks. In particular, as noted by Azinovic et al. (2022), at the beginning of the estimation procedure, when the quality of the policy function approximation is poor, following the implied state-transition may lead to very unrealistic states, or even impossible states, if the constraints are violated by that policy. This can slow down convergence significantly, and in some cases cause the approximation procedure to break down entirely. Furthermore, the lack of knots means that we do not get an exact policy for any state, but rather an approximation for all states. This means that we have to rely heavily on the ability of the neural-network to approximate well over the high-dimensional input space in order to get precise estimates. Fortunately, it is possible to measure the quality of the approximation via the loss function, and as demonstrated in for example Azinovic et al. (2022) the approach is highly scaleable, such that additional computational resources can be added to improve the quality of the approximation to an acceptable standard.

1.2.2 Limitations of Penalty-Based Constraints

Using a penalty for constraint violation is not specific to neural-network based solution methods. For example, in Achdou et al. (2022) penalties are applied for violations of constraints in their continuous-time solution method. When a model is solved over some form of grid of points this is an appropriate implementation because it is computationally efficient, straightforward to implement, and can be made arbitrarily accurate by continued iteration. However, what seems to be particularly problematic in practice is the combination of penalties with endogenous sampling of states as in S. Maliar et al. (2011).[6]. Usually, when this forward-simulation is used, the states are

[6]Note that these states are not endogenous in the same sense as the endogenous grid points of Carroll (2006). Those are still grid points that are chosen ex-ante, which are endogenous in the sense that they

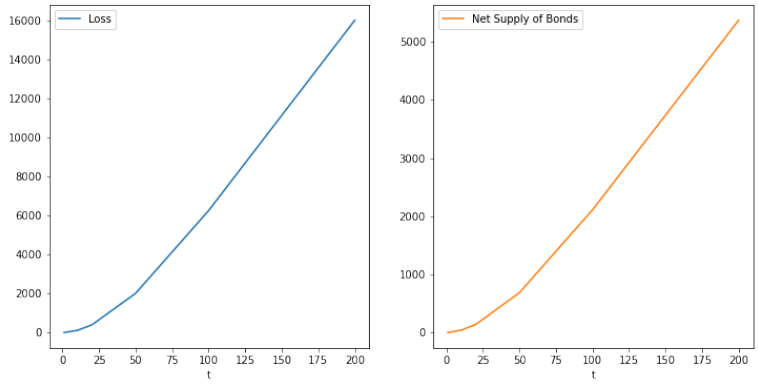


Figure 1.1: Overall loss (left) and net supply of bonds (right) obtained by simulating forward the model (Section 1.4) with randomised initial parameters when constraints are not enforced (soft-constraint method). Over 200 iterations the net supply of bonds grows exponentially away from its equilibrium value (0), and as a result so too does the overall loss.

simulated forward many periods at a time.[7] This is done in order to make sure that states are approximately uncorrelated (L. Maliar et al., 2021), and also in the case of the extended neural-network of Kase et al. (2022) that the states represent a true draw from the ergodic distribution under freshly redrawn structural parameters. As a result, small violations of the constraints such as market clearing conditions have the potential to compound between draws of states.

For example, consider a model in which the initial policy estimates imply an artificially high interest rate, and a net supply of bonds is greater than zero (the equilibrium value). Then, even though the penalty term and subsequent parameter update implies that the net bond positions should reduce, after simulating forward a number of periods from the initial state the return on these net-excess bonds would be so large that the model would have already moved far away from equilibrium. Converging back down to smaller bond positions from here could be slow since we are so far from realistic values, the gradient may be very large and not particularly informative. In principle, if the system is globally convergent to a single equilibrium then it should still be possible to find it despite this compounding, however, in practice, it is common to run into computational issues, if for example bond positions increase to above the largest values that can be represented or consumption goes to zero (which implies division by zero in the FOCs in the case of log-utility). Figure 1.1 clearly demonstrates this divergence by simulating forwards a model with randomised initial parameters for 200 periods, in the case where constraints are not strongly enforced.

As a result of this, in my own experimentation with implementing methods based on the soft-constraint approach, I found that a number of modifications seem to be necessary in order to

are treated as end-of-period states. In this case, on the other hand, the states are truly endogenous in the sense that they change throughout the approximation procedure and depend directly on the current policy function.

[7]For example Kase et al. (2022) simulate forward 20 periods between each parameter update, while L. Maliar et al. (2021) simulate 10 periods forward.

keep states within the set of feasible values and to render the approximation procedure stable. First, the number of forward-simulation between updates must be kept low, at least early in the approximation procedure. This can then be relaxed by gradually increasing the number of forward simulations between parameter updates as long as the constraint-related loss components stay satisfactorily small. Secondly, a larger learning rate can be used, in order to counteract the divergent nature of initial solutions. This learning rate can then subsequently be lowered as training iterations go on. Furthermore, one can train for multiple parameter updates with current estimated states by repeatedly drawing random sub-batches, before performing the next forward simulation.[8] This helps to balance out how often forward simulation takes place relative to parameter updates.

Finally, despite all of these adjustments, for some initial conditions the approximation procedure can still break-down, resulting in incredibly large losses, a corner solution, or null values. In these cases the gradient is ignored, and the states have to be reset to initial values. When approximation diverges, it is not possible to for example simply set the loss to a large or infinite value because doing so would not result in an informative gradient that points the optimiser back towards reasonable parameters. Since there is no obvious relationship between individual neural-network parameters and outputs there is no straightforward way to manually adjust parameters to move away from these extreme values, as this is exactly what we hope the gradient will tell us.

If all of these accommodations are made, it is possible to achieve convergence in policy estimates under soft-constraints, however, these accommodations and refinements would not be necessary if the approximated policies could somehow be manipulated so as to always satisfy the constraints of the model. Furthermore, as I will demonstrate in Section 1.6) using my recommended approach also makes the resulting approximation of policy functions more accurate, especially close to those constraints.

1.2.3 HANK

One clear use-case for these machine-learning techniques is given by HANK models, such as the model proposed by Kaplan et al. (2018). These models combine price rigidity and imperfect competition from standard Representative Agent New-Keynesian (RANK) models with the incomplete markets and heterogeneity of Krusell and Smith (1998) type models in order to provide a role for inequality in the analysis of monetary policy (Acharya et al., 2020). The primary goal of HANK models is to achieve more realistic MPCs, and more realistic monetary policy transmission, when compared with their RANK counterparts. HANK models explicitly take into account the fact that the way agents interact with financial markets may vary significantly depending on the wealth of agents, and allow for the extent to which agents are constrained to be determined endogenously, rather than fixed exogenously, as in Two-Agent New-Keynesian (TANK) models.

[8]This is the approach in for example Azinovic et al. (2022), however, it is not necessary in the hard-constraint approach, which is stable even if forward simulation is performed after every parameter update.

Because heterogeneity is a core feature of these models, they tend to involve high-dimensional state-spaces, especially when a number of different asset classes are modelled. One approach to dealing with this is to assume a limited form of heterogeneity. For example, TANK models exogenously assume a mass of agents are financially excluded and therefore must consume all of their current wealth (Galí et al., 2007). This captures much of the intuition of HANK models, but since the types of agent are exogenous, it fails to capture the fact that severe shocks may meaningfully impact the degree of inequality in the economy and hence aggregate dynamics through this channel. Other papers have developed methodological advances that make solving models with rich and endogenous heterogeneity more feasible. For example, Bayer and Luetticke (2018) and Auclert et al. (2021), and Ahn et al. (2018) augment the methodology of Reiter (2009) using various techniques to achieve greater amounts of dimensionality reduction and thus render larger models possible to solve. However, these methods all rely on linearisation in the response to aggregate shocks, or do not offer a global solution. Therefore, the primary benefit of the neural-network solution method in comparison to these more standard methods is the ability to capture non-linear aggregate dynamics.

1.3 Methodology

The primary methodological contribution of this paper is the utilisation of *hard-constraints* on neural-network outputs, in order to ensure that policy functions are restricted to the space of functions which satisfy all model constraints. Unfortunately, it would be very difficult or impossible to specify a solution of this nature that is completely general. However, it is possible to specify solutions to a number of important cases that cover most heterogeneous agent models used in practice. To fix ideas, consider the standard state-transition equation, where a neural-network is used to approximate the policy function $Y_t \in \mathbb{R}^k$, over a states $X_t \in \mathbb{R}^m$:

$$\hat{Y}_t = \sigma^n \left(W_n \cdot \sigma^{n-1}(\cdots \sigma^0 \left(W_0 \cdot \{X_{t-1}, \epsilon_t, \Gamma\} + b_0 \right) \cdots) + b_n \right) \tag{1.3}$$

$$X_t = S(X_{t-1}, \epsilon_t, Y_t; \Gamma) \tag{1.4}$$

$$\text{s.t. } g(X_t, Y_t) \geq 0 \tag{1.5}$$

Then, the suggestion is to add a scaling function h to the neural-network outputs in order to ensure that the constraints g are always satisfied, regardless of the output of the neural-network. We can write this as:

$$Z_t = \sigma^n \left(W_n \cdot \sigma^{n-1}(\cdots \sigma^0 \left(W_0 \cdot \{X_{t-1}, \epsilon_t, \Gamma\} + b_0 \right) \cdots) + b_n \right) \tag{1.6}$$

$$Y_t = h\left(Z_t\right) \tag{1.7}$$

This can be thought of as a separate operation after the neural-network outputs, or simply as a complex type of output layer. There are a number of common types of constraints for which a straightforward rescaling can be applied, which will now be discussed briefly. If the constraint is of the form $a < Y_t$ then the output can be constructed as $Y_t = a + softplus(Z_t)$ or $Y_t = a + exp(Z_t)$, where Z_t is the output of the last layer in the neural-network, and $softplus$, which is a common activation function used in machine-learning applications, is defined as:

$$softplus(x) = log(1 + exp(x))$$

Likewise, if the constraint is of the form $Y_t < b$ the output can be constructed as $Y_t = b - softplus(X_t)$ or $Y_t = b - exp(X_t)$. Furthermore, if the constraint is of the form $a < Y_t < b$ the output can be constructed as $Y_t = a + (b - a) \cdot sigmoid(Z_t)$, where $sigmoid$ is defined as[9]:

$$sigmoid(x) = \frac{1}{1 + exp(-x)}$$

While these individual constraints are straightforward, more difficult problems can arise when Y_t is a vector, and the constraints imply some relationship between its elements. A prominent, and quite general example, which is particularly relevant for solving heterogeneous agent DSGE models, is the constraint $a^i < Y_t^i < b^i; \sum_i Y_t^i = C$ for $C \in \left(\sum_i a^i, \sum_i b^i\right)$. In the case where $a^i = 0; b^i = C = 1 \forall i$, then this is a well known problem where the outputs can be interpreted as predicted probabilities of various mutually exclusive categories. In machine-learning, predictions for these kinds of *multi-label classification* problems are handled by the *softmax* function, which is defined as:

$$softmax(x) = \frac{exp(x)}{\sum_i exp(x^i)}$$

Another solution, which will be used in parts of Algorithms 1 and 2 is an affine transformation that shifts by the minimum value and scales by the sum: $Y_t = \frac{Z_t - \min_i Z_t^i}{\sum_i Z_t - \min_i Z_t^i}$. If $C \neq 1$, then C can simply be multiplied by these functions to change their sum appropriately. However, when a^i, b^i, and C are unrestricted, this problem is significantly more difficult, and no obvious solution exists. A naive approach might be to first re-scale outputs so that they fall between the upper and lower bounds, and then re-scale again so that the sum of outputs equals C. These steps are summarised

[9]However, in order to avoid a *vanishing gradient problem*, it may be better to simply allow initial values outside of these constraints, and then clip them back inside, see Section 1.5.3.

in the following two equations:

$$z' = \frac{(b-a)x}{\|x\|} + a \tag{1.8}$$

$$z = C\frac{z'}{\|z'\|} \tag{1.9}$$

However, the problem with this is that rescaling done to satisfy the lower and upper bounds a^i and b^i may change the sum of outputs, and thus cause the summation constraint to be violated, and vice versa. Of course, we could always use an iterative algorithm to find a valid solution, however, this is not appropriate in this application, because the outputs of such an algorithm would not be differentiable with respect to the neural-network parameters, at least, not using backpropogation and automatic differentiation. Furthermore, performing such an iteration within every policy evaluation would slow down computation substantially, perhaps enough to render approximation infeasible.

Algorithm 1: Rescaling Algorithm

Input: $x \in \mathbb{R}_+^k, a \in \mathbb{R}_+^k, b \in \mathbb{R}_+^k, C \in \mathbb{R}_+, a^i < b^i \forall i, \sum_i a^i < C < \sum_i b^i$
Output: $w \in \mathbb{R}_+^k : a^i < w^i < b^i \forall i, \sum_i w^i = C, w^i \propto x^i$

$z' \leftarrow (b-a)\frac{x}{\sum_i x^i} + a;$
$z \leftarrow C \times \frac{z'}{\sum_i z'^i};$
$A \leftarrow \sum_{z^i < a^i} \left(a^i - z^i\right);$
for $i \in \{0, \ldots, L\}$ **do**
  $\hat{a^i} \leftarrow \max\{0, w^i - a^i\};$
for $i \in \{0, \ldots, L\}$ **do**
  $w^i \leftarrow \min\{a^i, z^i - \frac{\hat{a}}{\sum_i \hat{a}} A\};$
$B \leftarrow \sum_{w^i > b^i} (w^i - b^i);$
for $i \in \{0, \ldots, L\}$ **do**
  $\hat{b^i} \leftarrow \max\{0, b^i - w^i\};$
for $i \in \{0, \ldots, L\}$ **do**
  $w^i \leftarrow \min\{b^i, w^i + \frac{\hat{b}}{\sum_i \hat{b}} B\};$

Instead, Algorithm 1 describes a function that solves this problem using only vectorised operations,[10] which is therefore differentiable and parallelisable. The logic of this algorithm is as follows. The first step is to re-scale each x^i to be within the upper and lower bounds. We then take the result of this and re-scale it again so that the sum of the elements is C. However, as previously noted, this second rescaling may undo the first, in the sense that it can cause some of the elements of the resulting vector to be outside the lower and upper bounds. Therefore, a third and final step is introduced, in which we find the total amount that would have to be added to or subtracted from each element outside the constraint in order to move it exactly to the constraint,

[10]The algorithm shows many steps as for-loops over i but these should be understood to be, and are actually in the implementation, vectorised operations.

and then distribute this sum among the unconstrained elements by an amount that is proportional to their distance to the bounds, such that the total sum of the vector remains unchanged. The weight is distributed proportionally to the distance to the boundaries in order to minimise the probability that any new elements are pushed over their own boundary.[11] We do this operation first for the lower bound a and then for the upper bound b, or vice versa, depending on the context. The difference is that the bounds will be able to bind only for the operation that is done last, for the other constraint, the outputs will satisfy the constraint strongly. Note that although these operations force the outputs within the constraints, it does not force the constraints to bind. In other words, it is still possible that the output of this algorithm can be strictly interior relative to the constraints, if this is the equilibrium outcome.

This type of constraint may seem quite arbitrary, but it is actually very general. In particular, it will occur in any heterogeneous agent model with a discrete number of agents and incomplete markets. In this context the idiosyncratic constraints a^i and b^i relate to the agents' choice of expenditure (income net of savings), which must be strictly positive and also feasible with respect to the borrowing constraint. Notice that this concerns expenditure, not consumption, as it may also include, for example, investments in other assets should they be present in the model. The aggregate constraint C relates to the market clearing condition of the economy. Even if there are multiple goods or assets agents can purchase, at some point it will be necessary to limit the sum of their expenditure in order to ensure market clearing due to Walras' law. See Section 1.5.3 for a particular example of how this is applied.

Algorithm 2: Alternative Rescaling Algorithm

Input: $x \in \mathbb{R}^k_+, a \in \mathbb{R}^k_+, b \in \mathbb{R}^k_+, C \in \mathbb{R}_+, a^i < b^i \forall i, \sum_i a^i < C < \sum_i b^i$
Output: $w \in \mathbb{R}^k_+ : a^i < w^i < b^i \forall i, \sum_i w^i = C, w^i \propto x^i$

$z' \leftarrow (b-a)\frac{x}{\sum_i x^i} + a;$
$z \leftarrow C \cdot \frac{z'}{\sum_i z'^i};$
$T \leftarrow \sum_i \max_i(z^i - b^i, 0) + \sum_i \min_i(z^i - a^i, 0);$
if $T \geq 0$ **then**
    for $i \in \{0, \ldots, L\}$ **do**
        $r^i \leftarrow \max_i(b^i - z^i, 0);$
else
    for $i \in \{0, \ldots, L\}$ **do**
        $r^i \leftarrow \max_i(z^i - a^i, 0);$
$\tilde{r} \leftarrow \frac{r}{\sum_i r^i};$
$w \leftarrow \min(\max(z, a), b) + \tilde{r}T;$

An alternative approach is given in Algorithm 2. This function essentially wraps the final two steps of Algorithm 1 into one. As a result it has the property that both the lower and upper bounds can bind in the output. This may or may not be desirable depending on the context. The

[11] Still, it is possible that with this third step, the outputs end up outside the constraints, however, this is easily detected, and in practice I found this to be uncommon.

purpose of including this is to demonstrate that there is not necessarily a unique solution to this problem, there may be multiple each with their own benefits. As mentioned previously, Azinovic and Žemlička (2023) provide another solution. However, neither of these algorithms are completely robust because they can fail in a particular edge case. In particular, if A or B are particularly large, or $\sum \hat{a}$ or $\sum \hat{b}$ are particularly low, then the weight that is dispersed to the unconstrained elements may be so large that it pushes those elements over their own bounds. In other words, they may be increased or decreased by a proportion of their distance to the bound that is greater than one. These conditions correspond to the case where there is very large wealth inequality between agents, and the proportion of agents who are constrained by the budget constraint is high, such that that the remaining agents save in order to satisfy the market-clearing condition that their consumption is negative. Fortunately, this was not encountered in the application in this paper, because under all of the allowed combinations of parameters the equilibrium does not touch this corner case.

1.4 Model

1.4.1 Households

In order to demonstrate how these constraint regimes can be implemented, I introduce a HANK model, based on that in Kase et al. (2022) and explain how to constrain the outputs of this model. The model contains a discrete number L of heterogeneous agents (which is set to 100 in all of the applications), indexed by i who maximise a utility function U subject to idiosyncratic shocks and budget constraints. The households' optimisation problem can be summarised as:

$$\max_{\{c_t^i, h_t^i\}} U^i = \mathbb{E}_0 \left[\sum_{t=0}^{\infty} \beta^t exp(\Psi_t) \left(\frac{1}{1-\sigma} {c_t^i}^{1-\sigma} - \chi \frac{1}{1+\eta} {h_t^i}^{1+\eta} \right) \right] \tag{1.10}$$

s.t.

$$c_t^i + b_t^i = W_t s_t^i h_t^i + Div_t^i + \frac{R_{t-1}}{\Pi_t} b_{t-1}^i \equiv \omega_t^i \tag{1.11}$$

$$b_t^i \geq \underline{B} \tag{1.12}$$

$$c_t^i > 0 \tag{1.13}$$

$$h_t^i > 0 \tag{1.14}$$

$$\Psi_t = exp\left(\rho_\Psi log(\Psi_{t-1}) + \sigma_\Psi \epsilon_t^\Psi\right)$$

$$s_t^i = exp\left(\rho_s log(s_{t-1}^i) + \sigma_s \epsilon_t^{s,i}\right)$$

All shocks denoted ϵ are distributed $iid\ N(0,1)$. Ψ_t is an aggregate preference shock, while s_t^i is an idiosyncratic labour productivity shock. In the implementation s_t^i are re-scaled to have a mean

of one in all states, so that these idiosyncratic shocks have no direct aggregate effect. However, Greenwood et al. (1988) preferences are not assumed, so it is possible that the distribution of s_t^i can indeed have aggregate implications, through agents' differential responses to income risk depending on their level of wealth. The resulting first-order conditions for the household are:

$$1-\mu_t^i = \beta R_t \mathbb{E}_t\left[\frac{exp(\Psi_{t+1})}{exp(\Psi_t)}\left(\frac{\lambda_t^i}{\lambda_{t+1}^i}\right)^{\sigma}\frac{1}{\Pi_{t+1}}\right] \tag{1.15}$$

$$\lambda_t^i \equiv \left(c_t^i - hC_{t-1}\right)^{-\sigma} = \frac{\chi\left(h_t^i\right)^{\eta}}{s_t^i W_t} \tag{1.16}$$

$$\mu_t^i\left(b_t^i - \underline{\mathrm{B}}\right) = 0, b_t^i \geq \underline{\mathrm{B}}, \mu_t^i \geq 0 \tag{1.17}$$

Where (1.15) is the households' Euler equation, (1.16) is the households' labour supply equation, and (1.17) is the KKT conditions associated with the borrowing constraint.

1.4.2 Firms

The firms are standard for a New Keynesian setting. A continuum of identical monopolistically competitive firms produce intermediate goods subject to a quadratic Rotemberg (1987) price-adjustment cost ϕ. This results in a New Keynesian Phillips Curve (NKPC):

$$\phi\left(\frac{\Pi_t}{\Pi}-1\right) = (1-\epsilon) + \epsilon MC_t + \beta\phi\mathbb{E}_t\left[\frac{\Pi_{t+1}}{R_t}\left(\frac{\Pi_{t+1}}{\Pi}-1\right)\frac{\Pi_{t+1}}{\Pi}\frac{Y_{t+1}}{Y_t}\right] \tag{1.18}$$

Where MC_t are real marginal costs. The firms are owned by the households, and distribute profits evenly. Therefore $Div_t^i = Div_t = Y_t - W_t N_t$. However, I assume that the firm discounts future profits as if they were owned by an unconstrained agent, such that their discount factor is $\frac{\Pi_{t+1}}{R_t}$.[12] In equilibrium, it is assumed that intermediate firms make positive profits and therefore I impose $W_t < \frac{Y_t}{N_t}$ as a hard constraint in the output ($W_t \leftarrow \tilde{W}_t\frac{Y_t}{N_t}$, $\tilde{W}_t \leftarrow sigmoid(\cdot)$), such that $Div_t > 0$. Final goods are produced by a representative competitive firm subject to an AR(1) TFP process $A_t = exp\left(\rho_A log\left(A_{t-1}\right) + \sigma_A\epsilon_t^A\right)$. Labour is hired in a competitive market such that:

$$MC_t = W_t/A_t \tag{1.19}$$

1.4.3 Monetary Authority

The monetary authority sets the nominal interest rate R_t according to a dual-mandate Taylor rule, with persistence parameter ρ_R and subject to a monetary policy shock ϵ_t^{mp}:

[12] This is done primarily to improve computational performance, however, I have tested a version instead using stochastic discount factor (average relative marginal utility of each agent) and the results were not meaningfully different.

$$R_t = (R_{t-1})^{\rho_R} \left(R \left(\frac{\Pi_t}{\Pi} \right)^{\theta_\Pi} \left(\frac{Y_t}{Y} \right)^{\theta_Y} \right)^{(1-\rho_R)} exp(\sigma_{mp} \epsilon_t^{mp}) \tag{1.20}$$

$$R = \frac{\Pi}{\beta} \tag{1.21}$$

1.4.4 Resource Constraints, Market Clearing, and Equilibrium

Finally, there are two resource constraints in the economy. The first stipulates that bonds are in zero net-supply, and the second is an output constraint that requires that production of final goods is equal to total effective labour input which is in turn equal to demand for final goods.

$$\frac{1}{L} \sum_{i=1}^{L} b_t^i = 0 \tag{1.22}$$

$$\frac{1}{L} \sum_{i=1}^{L} s_t^i h_t^i = N_t \tag{1.23}$$

$$Y_t = A_t N_t = \frac{1}{L} \sum_{i=1}^{L} c_t^i \tag{1.24}$$

Note that if (1.24) and (1.11) hold for i, t, then (1.22) follows by construction, as a result of Walras' law.

1.4.5 Equilibrium

A Functional Rational Expectations Equilibrium in this model consists of a set of idiosyncratic policy functions $\{c_t^i, h_t^i, \mu_t^i\}_{i=1}^L$, and prices $\{\pi_t, W_t, R_t\}$ such that for all states $\{s_{t-1}^i, A_{t-1}, \Psi_{t-1}, b_{t-1}^i, C_{t-1}, R_{t-1}\}_{i=1}^L$ and shocks $\{\epsilon_t^{s,i}, \epsilon_t^A, \epsilon_t^\Psi, \epsilon_t^{mp}\}_{i=1}^L$ the idiosyncratic policies satisfy the households' optimality conditions (1.15), (1.16) and (1.17) taking prices as given, the prices satisfy the NKPC (1.18) and the monetary policy rule (1.20), and the markets for bonds (1.22), labour (1.23) and goods clear (1.24).

1.5 Machine-Learning Solution Method

1.5.1 Loss Function

In order to solve the model, I approximate the optimal policy functions using a neural-network, as described previously. The neural-network used is fully-connected and feed forward, with a depth of 5 layers, and 128 perceptrons per layer. The inputs of the neural-network consist of the states of the model ($\left\{\Psi_{t-1}, s_{t-1}^i, b_{t-1}^i, C_{t-1}, R_{t-1}\right\}_{i=1}^L$), the exogenous shocks ($\left\{\epsilon_t^{s,i}, \epsilon_t^\Psi, \epsilon_t^a, \epsilon_t^{mp}\right\}_{i=1}^L$), and the structural parameters of the model, which are sampled uniformly over a predefined range

(see Section 1.5.4). The neural-network outputs two aggregate policies Π_t and W_t, and three idiosyncratic policies: $\left\{\tilde{c}_t^i, h_t^i, \tilde{\mu}_t^i\right\}$. The tilde implies that some of these outputs will be further transformed before they can be interpreted as policy functions, as described in Section 1.6.1. In practice when approximating the idiosyncratic policies we feed the inputs into a separate neural-network that in addition to the aggregate inputs also contains the idiosyncratic states $s_{t-1}^i, b_{t-1}^i, \epsilon_t^{s,i}$ of the agent for whom the policy is being predicted. Thus the aggregate policy neural-network has $3L + 31$ inputs and 2 outputs, and the idiosyncratic policy neural-network has $3L + 34$ inputs and 3 outputs.

The loss function that the model aims to minimise consists of the product of the first-order condition errors for a given state and two independent draws of future shocks (indexed here by j) as in L. Maliar et al. (2021). This is done instead of squaring the loss resulting from a single draw of shocks in order to allow for the SGD process to, over a larger number of iterations, integrate over the expectation operators.

$$L_{ee}^{i,j} = \left(1 - \mu_t^i - \beta R_t \left[\frac{exp(\psi_{t+1}^{i,j})}{exp(\psi_t^{i,j})} \left(\frac{\lambda_t^i}{\lambda_{t+1}^{i,j}}\right)^{\sigma} \frac{1}{\Pi_{t+1}^j}\right]\right) \tag{1.25}$$

$$L_{nkpc}^j = \left(\phi\left(\frac{\Pi_t}{\Pi} - 1\right) - (1-\epsilon) - \epsilon MC_t - \beta\phi\mathbb{E}_t\left[\frac{\Pi_{t+1}^j}{R_t}\left(\frac{\Pi_{t+1}^j}{\Pi} - 1\right)\frac{\Pi_{t+1}^j}{\Pi}\frac{Y_{t+1}^j}{Y_t}\right]\right) \tag{1.26}$$

$$L_{ls}^i = \left(\left(c_t^i - hC_{t-1}\right)\right)^{-\sigma} - \left(\frac{\chi\left(h_t^i\right)^{\eta}}{s_t^i W_t}\right)\Bigg) \tag{1.27}$$

$$L = \frac{1}{L}\sum_{i=1}^{L}\left[L_{ee}^{i,1} \cdot L_{ee}^{i,2} + \left(L_{ls}^i\right)^2\right] + L_{nkpc}^1 \cdot L_{nkpc}^2 \tag{1.28}$$

When the outputs are constrained, in the *hard-constraint* approach, these are all the loss components that are necessary. In the alternative case, where the outputs are not constrained, a further 3 loss components are required, to ensure that deviations from the resource and budget constraints are sufficiently penalised. These penalties are given in equations (1.29), (1.30), and (1.31).

$$L_{bc}^i = \Psi_{fb}\left(b_t^i - \underline{\mathrm{B}}, \mu_t^i\right) \tag{1.29}$$

$$L_{oc} = Y - \frac{1}{L}\sum_{i=1}^{L} c_t^i \tag{1.30}$$

$$L_{rc} = \frac{1}{L}\sum_{i=1}^{L} b_t^i \tag{1.31}$$

1.5.2 Learning Algorithm

The *ADAM* learning algorithm (Kingma & Ba, 2014) was used to learn the neural-network parameters. Batches of size mb, each containing states and randomly drawn structural parameters are fed into the neural-network, and proposed policies are calculated. From this, the loss and gradient of the loss are evaluated and the update step is performed. After this, new random structural parameters are drawn, and the model is simulated forward 20 times according to the state-transition rule of the model, and the current approximated policy. The resulting states and structural parameters are then used as the inputs in the next iteration. This procedure is repeated until a desired number of iterations is reached, or until the loss reaches a sufficiently small value.

When constraints are soft, if any of the constraint related losses go above a certain threshold, the states are reset to initial values, and the number of forward simulations between each update is decreased. Conversely, when the constraint related losses go below some low threshold, the number of forward iterations is increased incrementally, up to a maximum of 20. This procedure ensures that the approximation is stable, especially for initial values, where it can easily diverge. These steps are summarised in Algorithm 3 in the appendix.

1.5.3 Constraint Satisfaction

This model has in effect six constraints, four idiosyncratic constraints: (1.11) (1.12), (1.13) and (1.14), and two aggregate or market clearing constraints: (1.22) and (1.24). First, it is ensured that the output for h_t^i is positive by applying a softplus activation function. With these h_t^i, the predicted prices W_t, Π_t, and the states of the model it is possible to calculate N_t, Y_t, MC_t and, Div_t^i, and thus the total wealth of every agent before consumption ω_t^i. Then (1.11), (1.24), (1.22) imply that $\frac{1}{L}\sum_{i=1}^{L}\omega_t^i = \frac{1}{L}\sum_{i=1}^{L}c_t^i + (0) = Y_t$. Therefore, the remaining constraints can all be satisfied by choosing c_t^i for each agent such that:

$$\overbrace{0}^{a_i} < c_t^i \leq \overbrace{\omega_t^i - \underline{B}}^{b_i} \tag{1.32}$$

$$\sum_{i}^{L} c_t^i = \underbrace{\sum_{i}^{L}\omega_t^i}_{C} \tag{1.33}$$

This constraint is of the general type outlined in Section 1.3 and therefore the algorithm outlined there can be used to re-scale $\tilde{c}_t^i$ into valid choices. In particular we use the variation where the constraint on b^i (upper constraint) can bind, so that a mass of agents can hit their budget constraint exactly, and no agent ever has 0 consumption. Given the resulting b_t^i we can evaluate if the budget constraint binds ($b_t^i = \underline{\mathrm{B}}$) for every agent, and then assign their Lagrange multiplier accordingly: $\mu_t^i = \mathbb{1}\{b_t^i = \underline{\mathrm{B}}\}\tilde{\mu}_t^i$.

I also consider two intermediate cases. The first is the case in which the idiosyncratic constraints (in this case, the budget constraint (1.11) and borrowing constraint (1.12)) are satisfied by construction, but no rescaling is done, so market clearing constraints (equations (1.24) and (1.22)) are represented only by the penalties (1.30) and (1.31). Here too, the exact implementation details matter. In order to ensure that it is possible for agents to hit their budget constraint, and to make sure that no agent has 0 consumption, a strictly positive consumption choice is taken for each agent via a softplus activation function, and it is simply clipped at the budget constraint of each agent, as shown in equation (1.34).

$$c_t^i = \min\left\{\tilde{c}_t^i, \omega_t^i - \underline{\mathrm{B}}\right\} \tag{1.34}$$

$$\tilde{c}_t^i = log(1 + exp(Z_t)) \in (0, \infty) \tag{1.35}$$

The second intermediate case is one in which the aggregate constraints (1.22) and (1.24) are satisfied by construction, and the borrowing constraint (1.12) is implemented via the FB penalty. This is achieved by a rescaling in which each agent picks a strictly positive consumption level $\tilde{c}_t^i$, and their sum is re-scaled such that:

$$\begin{aligned} c_t^i =& \frac{1}{L}\sum_{i=1}^{L} \omega_t^i \cdot \frac{\tilde{c}_t^i}{\frac{1}{L}\sum_{i=1}^{L}\tilde{c}_t^i} \\ \implies \frac{1}{L}\sum_{i=1}^{L} c_t^i =& \frac{1}{L}\sum_{i=1}^{L}\omega_t^i = Y_t \end{aligned} \tag{1.36}$$

In the soft-constraint approach, there is no need to calculate wealth before calculating consumption, so the labour supply FOC (1.16) can be substituted in directly, such that the labour supply is exactly optimal for any given consumption choice. This is a key advantage of the soft-constraint approach vis-à-vis the hard-constraint approach. However, despite this, the results seem to indicate that this trade-off is worth making.

1.5.4 Calibration

Table 1.2 displays the parameters used for approximating the model. Many of these are taken from Kase et al. (2022), with the exception of $\underline{\mathrm{B}}$ and σ_s, which were increased to ensure that many agents hit their borrowing constraint in equilibrium, and ρ_A and σ_A, which were added because I model the TFP process as deviations from steady-state rather than trend growth. As in that paper, the economic parameters are added as inputs to the neural-network, and during training they are occasionally uniformly re-sampled between the given bounds. When the lower and upper bounds are the same, then that parameter is calibrated to a specific value. Although this paper

Parameter	Baseline Value	Min	Max
β	0.9975	0.9975	0.9975
σ	1	1	1
η	1	1	1
ϵ	11	11	11
χ	0.91	0.91	0.91
ϕ	1000	700	1300
θ_Π	2	1.5	2.5
θ_Y	0.25	0.05	0.5
Π	1.005	1.005	1.005
Y	1	1	1
$\underline{\mathrm{B}}$	-0.05	-0.5	-0.01
ρ_Ψ	0.7	0.5	0.9
ρ_s	0.8	0.7	0.9
ρ_A	0.8	0.7	0.9
ρ_r	0.25	0.1	0.5
σ_Ψ	0.03	0.01	0.05
σ_s	0.05	0.01	0.08
σ_A	0.008	0.003	0.012
σ_{mp}	0.005	0.001	0.008

Table 1.2: Calibration of parameters used to approximate the model in Section 1.4, along with upper and lower bounds between which parameters are uniformly drawn during training.

does not deal directly with the estimation of structural parameters, they are still included and varied as part of policy approximation in order to keep this possibility open for extensions to this work and also because approximating the model over a range of parameters in this way makes it possible to query the model in different ways over ranges of parameters without refitting the model, which saves a considerable amount of time. This does however slightly decrease the precision of the results for any given combination of parameters, given the same amount of training.

1.5.5 Calculation of Impulse Responses

Figures 1.6 and 1.9 display generalised IRFs (Koop et al., 1996) generated by first taking a sample of states from the ergodic distribution and then for each of these states a large number of simulations are performed, in each of which the same shock (i.e. $\epsilon_t^A = 2$) is introduced at $t = 0$, and the state-transition is simulated forward in the standard way, where other shocks are not turned off, but are instead drawn from their respective distributions, as usual. The generalised IRF is then the average response of the outcome variables over this large number of draws of shocks. The generalised IRFs in these figures were generated using 4000 draws for each of 256 states, resulting in over 1 million simulated impulse responses in total. Despite this, there is still a noticeable amount of noise in the impulse responses. This could be attenuated by simulating even more impulse responses, however, the approximation error is likely to remain meaningful for any feasible number of simulations, since there are a large number of shocks to integrate over.

Type		All Hard Constraints	All Soft Constraints
	Total Loss	3.60e-04	1.13e-02
Optimality	Euler Equation Loss	2.94e-04	5.64e-03
	Phillips Curve Loss	1.17e-07	5.14e-03
	Labour Supply Loss	6.64e-05	1.10e-32
Constraints	KKT Loss	3.25e-35	2.74e-04
	Output Constraint Loss	3.96e-32	1.03e-04
	Net Supply of Bonds	1.36e-32	1.56e-04

Table 1.3: Average total loss and its constituent components evaluated over the last 50 iterations of the fitting procedure for the models fit with all constraints hard and all constraints soft (penalty-based). The soft-constraint version was fit with weights of $1e2$ on each of the three constraint-related loss components.

1.6 Results

1.6.1 Hard and soft-constraints

Table 1.4 compares the losses obtained by the hard and soft-constraint solution methods. In this average over a range of different calibrations, the hard-constraint method outperforms the soft-constraint one. Note that losses that are less than $1e-30$ can be attributed to floating-point imprecision; these are the conditions that are fulfilled exactly for each solution method. For the hard-constraint version this is all three constraints, and for the soft-constraint version this is the labour supply loss, which can be fulfilled by construction using the labour-consumption FOC. The hard-constraint implementation achieves an overall loss almost two orders of magnitude smaller than the soft-constraint approach. However, in the soft-constraint case, the larger overall loss does not come from the constraint penalties themselves, which are all at least an order of magnitude smaller than the other loss components. Rather, the presence of relatively small, but still meaningful errors in the constraint conditions makes it more difficult for this method to satisfy the FoCs of the model. In particular, the soft-constraint has its worst relative performance on the Phillips curve loss, which is an equation that involves only aggregate variables. As discussed, this could be because the states that are generated as inputs are further from the true ergodic equilibrium, or because the optimiser is trading off these optimality conditions with the essentially arbitrary penalties that it incurs.

Not only does the hard-constraint version provide quantitatively better results, but this solution is also qualitatively better in a number of critical ways. For example, it is reasonable to expect that the proportion of constrained agents is increasing in the borrowing constraint. The hard-constraint version of the model captures this correctly, while the soft-constraint version struggles. This can be seen in Figure 1.2. In this figure the yellow line depicts the average proportion of agents from a sample of draws from the ergodic set who fall at or below a given level of wealth when the budget constraint is set to its default value of -0.05, in other words, the cumulative distribution of wealth when $\underline{B} = -0.05$. The blue line depicts the average proportion of agents who are at the constraint

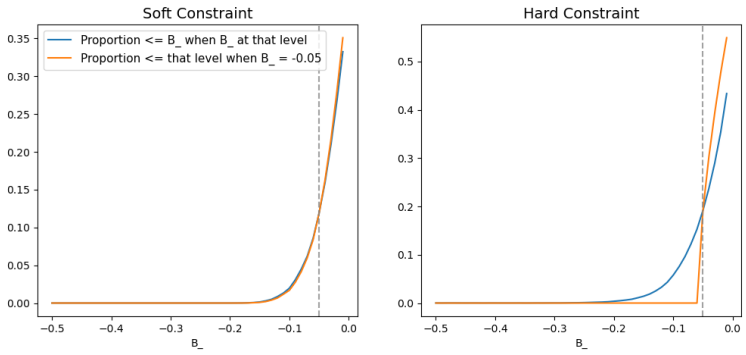


Figure 1.2: The average proportion of agents whose end of period savings are at or below a given value of the budget constraint $\underline{B}$ (blue line) and the average proportion of agents whose end of period savings fall below a given value when the budget constraint is set to the default value of -0.05 (yellow line) for the soft-constraint model (left) and the hard-constraint model (right).

$\underline{B}$ as the constraint varies across values on the x-axis.

For the soft-constraint version, the two lines are almost identical. This tells us that this model fit with soft-constraints has learned relatively little about how agents' behaviour should change around the constraint, as the shape of the distribution of wealth is essentially invariant to the value of $\underline{B}$. Therefore, although the soft-constraint model does predict less agents are constrained when the constraint is weaker, this is simply a by-product of sliding the value of $\underline{B}$ along an unaltered wealth distribution. Conversely, the hard-constraint version does not allow any agents to have wealth below -0.05 when $\underline{B} = -0.05$ (yellow line), while it does predict a smoothly decreasing proportion of constrained agents as the budget constraint is weakened (blue line). This figure also illustrates another important outcome: imposing hard-constraints does not force the constraint to bind. It is clearly possible for exactly zero of the agents to be at the budget constraint, and indeed this is the case when the budget constraint is weak. This difference in the handling of the budget constraint is important because the high MPC of agents near to the budget constraint is exactly what causes the implications of the HANK model to deviate from those of a RANK model (Kaplan et al., 2018). MPCs generated by each approach are displayed slightly later in Figure 1.5.

Figure 1.3 visually demonstrates the distribution of errors related to the three constraints in the model combined from a batch of states drawn from the approximated ergodic distribution for the two different methods. The top row displays the results for the hard-constraint method, and we can see that the market clearing error and net bond supply are imperceptible when put on the same scale as the error from the soft-constraint model. The rightmost column shows the distribution of idiosyncratic bond positions for all agents across the drawn states where the borrowing constraint is indicated by the vertical black line. There are striking differences in the shape of the wealth distribution generated by both methods. For the hard-constraint method, around 20% of the observations lie exactly on the borrowing constraint, and the distribution is shaped as a decay to the right. For the soft-constraint method, many observations lie to the

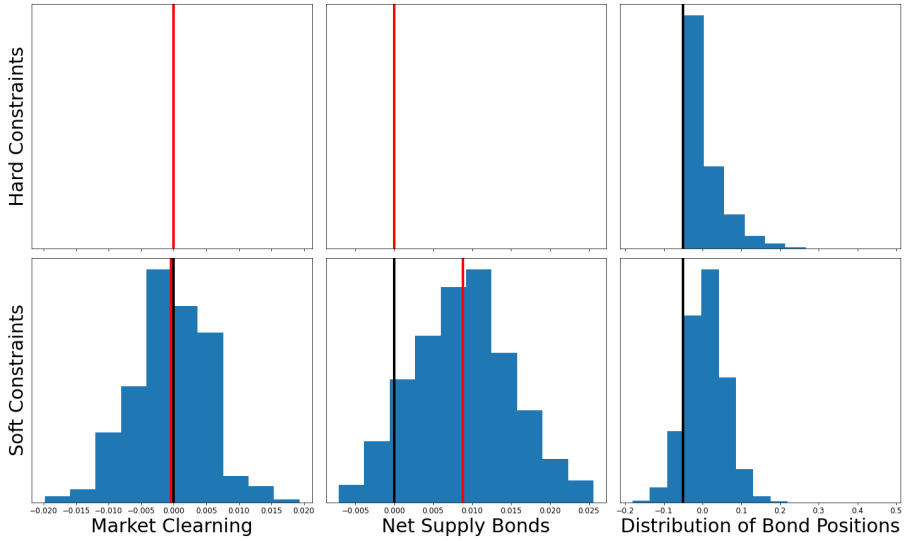


Figure 1.3: Cross-Sectional distribution of output net of consumption (leftmost column), net supply of bonds (middle column), and individual bond positions (rightmost column) for the hard-constraint implementation (top row) and soft-constraint implementation (bottom row). Histogram spans observations from a batch of 256 simulated draws from the ergodic set. Red lines indicate means over the batch.

left of the borrowing constraint, and the distribution has a significantly more symmetrical shape. Furthermore, we see that the soft-constraint version seems have a notable positive bias in the net supply of bonds. The optimiser could have chosen to do this in order to trade off the resulting penalty with a reduced penalty from the budget constraint, as the positive net supply of bonds makes it less likely that agents will violate the budget constraint, all else equal. These notable issues are despite the fact that the errors related to the model constraints are of the order $1e-4$ for the soft-constraint method, which is small relative to the errors in the first-order conditions, even for the hard-constraint model. So while this error is quantitatively small, it is still clearly economically relevant. The difference in the distribution of bond holdings can be seen even more clearly when comparing the distribution of bond positions for all agents in a sample of individual states for the hard and soft-constraint methods, as shown in Figure 1.4.

Figure 1.5 shows a stark difference in the immediate MPCs generated by the hard and soft-constraint approaches.[13] The hard-constraint method (first row) faithfully reproduces consumption behaviour which is qualitatively consistent with the mechanisms of the HANK model. The consumption function is kinked at the bottom as a result of the borrowing constraint[14] and there

[13]For the hard-constraint approach these MPCs can be calculated directly from the rescaling function, which is a differentiable function that takes wealth as an input and outputs consumption. On the other hand, for the idiosyncratic and soft-constraint versions, the consumption policy is only a function of states. Therefore, immediate MPC is calculated as $\frac{dc_t^i}{d\omega_t^i} = \frac{dc_t^i}{db_{t-1}^i} / \frac{d\omega_t^i}{db_{t-1}^i}$. In order to maintain consistency, MPCs for both approaches are calculated in this latter way.

[14]This is a standard feature of heterogeneous agent models with incomplete markets, such as Krusell and Smith (1998), for example.

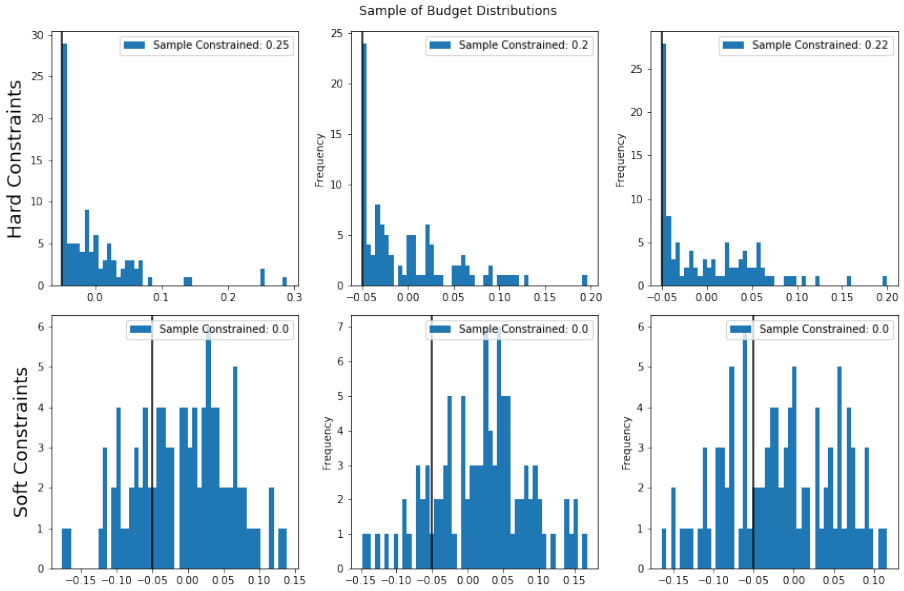


Figure 1.4: Bond holding distributions for 3 randomly sampled states (across columns) for the hard-constraint method (top row) and soft-constraint method (bottom row).

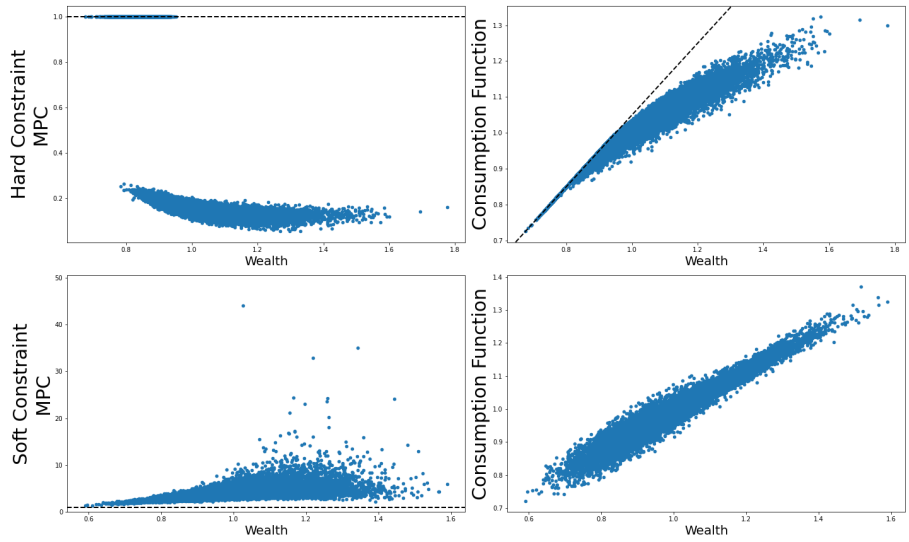


Figure 1.5: Immediate MPCs (right column) and consumption functions (left column) for the hard-constraint method (top row) and soft-constraint method (bottom row) across all agents in a sample of states drawn from the ergodic distribution.

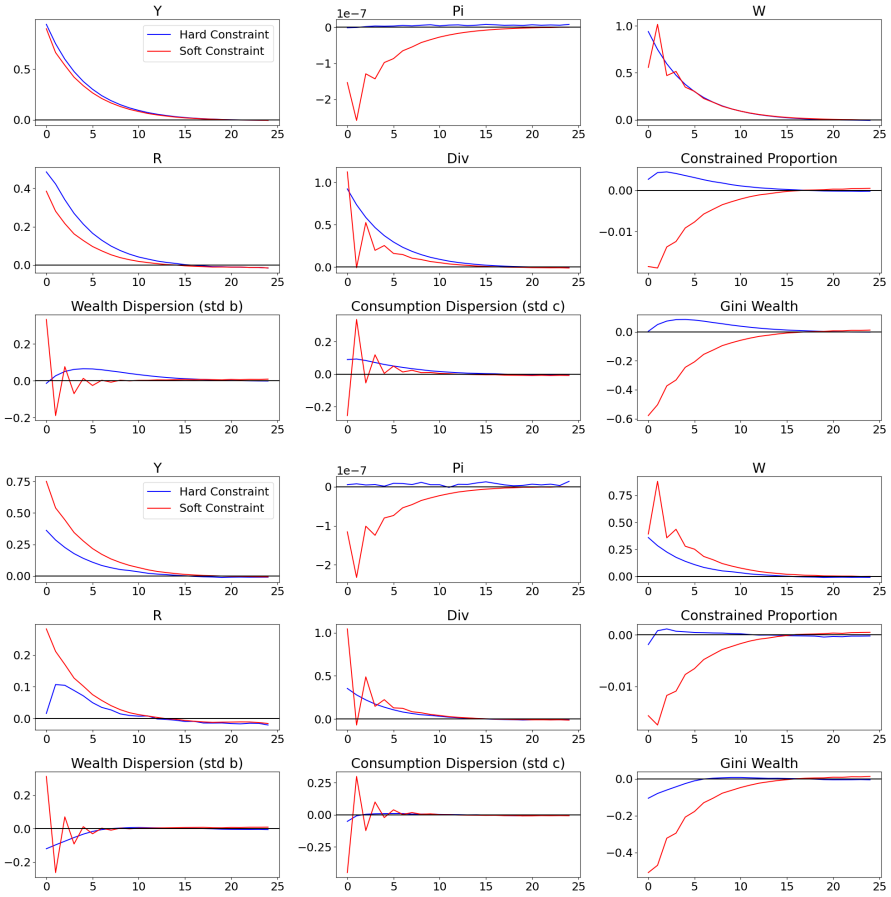


Figure 1.6: Impulse Responses to a 2 standard deviation expansionary TFP shock (ϵ_t^A) generated from the hard-constraint and soft-constraint models. Results in the top three rows are an overall average over all states, and the bottom three rows show only ZLB states (where $R_{t-1} = 0$). Note that for the soft-constraint version the constrained-proportion is the proportion that fall outside the budget constraint.

is considerable heterogeneity in immediate MPCs: constrained agents, that is to say agents who ended the period with bond positions at the budget constraint, and who therefore mostly have lower total wealth, have MPCs of 1, while unconstrained agents have MPCs decreasing in wealth at a much lower mean of around 0.2.

This discontinuous change in behaviour is exactly the kind of behaviour that the soft-constraint approach struggles to model correctly. Indeed, the soft-constraint (bottom row) version does not reproduce any of the qualitative features that we expect a solution to display. The consumption function appears to be nearly linear, and many of the MPCs are much greater than 1, and appear to be increasing in wealth, which is inconsistent with optimising behaviour in this setup. This comes as a direct result of the failure of the model to strongly enforce the budget constraint.

Figure 1.6 shows the generalised impulse responses to a two standard deviation expansionary TFP shock under each of the two main constraint regimes. The top three rows show an average over all states, whereas the bottom three rows show responses for states that start with the nominal

interest rate at the ZLB only. The response of wealth dispersion for the hard-constraint model, for example, demonstrates one of the major benefits of the neural-network solution method. The response has a different direction and shape in the ZLB states compared to the average states. This shows how the solution is able to capture responses that are non-linear in the aggregate.

When comparing these, the hard-constraint version (top) should be regarded as closer to the ground-truth, given its much lower loss, and thus, approximation error. While the soft-constraint version does manage to match the hard-constraint version in some aggregate variables, notably output and wages, it struggles in particular to match the impulse responses of the hard-constraint solution in cross-sectional variables, and produces in some cases erratic behaviour, such as for consumption and wealth dispersion. This could in part be because it predicts that the zero lower bound is encountered far too frequently. This can be seen clearly in the distribution of prices and aggregate variables shown in Appendix B. Furthermore, for cross-sectional observables where the soft-constraint version produces smooth responses, such as the Gini-coefficient of wealth, it seems to greatly overstate the effect of the shock, relative to the hard-constraint version.

Of course, it is still possible that the results for the soft-constraint method could be made better by using a different neural-network architecture, using different step-sizes and weights on loss components, training longer, or picking the initial conditions differently. To more closely consider one of these possibilities, Appendix D provides results obtained by training the soft-constraint version of the model with varying weights on the constraint-related penalties. The results vary relatively little over this dimension when models are allowed to train for a large number iterations, and all versions still do not outperform the hard-constraint version. Notwithstanding this, it is quite possible that there is some combination of weights that performs even better. The main problem is that these neural-network fits do require a substantial amount of time to run, so it is undesirable to have hyper-parameters such as the weights on various loss components that need to be fine-tuned in order to obtain a reliable approximation. While implementing hard-constraints can also be complicated and time-consuming for the user, this paper already provides a solution for a large class of problems.

To ensure a maximally fair comparison as many factors were kept the same as possible, modifications were made that allowed the soft-constraint version to produce the best results that I was able to obtain (outlined in Section 1.3), and the soft-constraint version was even allowed to train significantly longer (see Figure 1.6 in the appendix for details). The purpose of these results is not to argue that the soft-constraint method does not work at all, but rather to demonstrate that the proposed hard-constraint method is more accurate and converges faster, and is therefore a better alternative.

Type		All Hard Constraints	All Soft Constraints	Hard Agg Constraints	Hard Idio Constraints
	Total Loss	3.60e-04	1.13e-02	2.09e-03	7.07e-02
Optimality	Euler Equation Loss	2.94e-04	5.64e-03	6.51e-04	4.89e-03
	Phillips Curve Loss	1.17e-07	5.14e-03	1.09e-03	4.06e-02
	Labour Supply Loss	6.64e-05	1.10e-32	3.52e-04	2.50e-02
Constraints	KKT Loss	3.25e-35	2.74e-04	1.87e-06	1.33e-34
	Output Constraint Loss	3.96e-32	1.03e-04	5.25e-32	6.62e-05
	Net Supply of Bonds	1.36e-32	1.56e-04	1.38e-32	1.25e-04

Table 1.4: Average total loss and its constituent components evaluated over the last 50 iterations of the fitting procedure for the models fit with only aggregate constraints hard and only idiosyncratic constraints hard. Previous results are repeated for ease of comparison.

1.6.2 Intermediate Cases

In order to demonstrate the effectiveness of the hard-constraint framework, this section will also compare the approximated policy functions generated by two alternative, intermediate methods. The first, as outlined in Section 1.5.3 is an intermediate approach, in which rescaling ensures that market-clearing is satisfied by construction, and the budget constraint is implemented by the FB penalty. This is henceforth referred to as the aggregate-constraint version. The second, as in L. Maliar and Maliar (2020), is the converse intermediate approach, where the borrowing constraint is satisfied by construction by clipping consumption choices at their feasible maximums, and market-clearing is enforced through a penalty. This is referred to as the idiosyncratic-constraint version.

Table 1.4 shows the overall loss, and components thereof generated by the intermediate approaches. Given that the most salient issues with the soft-constraint model seemed to be related to the budget constraint, one might expect to see that the idiosyncratic-constraint model was the best intermediate approach. However, interestingly, we see that the aggregate-constraint and the soft-constraint models actually out-perform the idiosyncratic-constraint model. Despite these small losses with regard to constraints, the aggregate policies generated by the idiosyncratic-constraint method are highly implausible, for example, we can see in Figure 1.11 in the appendix that this model produces an average output of only 0.94 (the true equilibrium value, which is the centre of the distribution generated by the hard-constraint and soft-constraint models is 1), and therefore, strikes the ZLB far too often.

The aggregate-constraint solution, on the other hand, performed much better, and did manage to out-perform the soft-constraint version, although its overall loss was still an order of magnitude larger than for the hard-constraint method. In this case it is interesting to note that the KKT loss was also significantly lower than in the soft-constraint version. This is likely because enforcing aggregate constraints creates a virtuous feedback cycle in which improvements to the approximation of aggregate variables and prices subsequently improves in the approximation of idiosyncratic variables, since the states being fed into the neural-network are closer to the true equilibrium. However, the aggregate-constraint model, even with a KKT Loss of the order $1e-6$ still does not

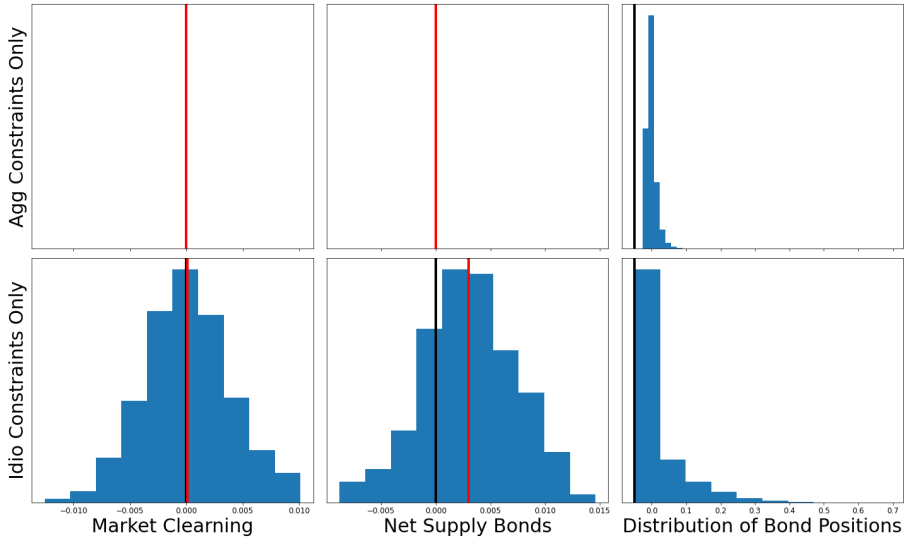


Figure 1.7: Cross-Sectional distribution of output net of consumption (leftmost column), net supply of bonds (middle column), and individual bond positions (rightmost column) for the implementation where only the aggregate constraints are hard (top row), and implementation where only the idiosyncratic constraint is hard (bottom row).

capture aspects of the cross-sectional distribution as well as the hard-constraint model. This can be seen in the budget distributions in Figure 1.7.

As expected, the aggregate-constraint version produces imperceptible errors in the aggregate constraints, while the idiosyncratic-constraint version succeeds in generating a budget distribution where the budget constraint binds frequently but is never violated. However, in contrast to the soft-constraint version, where the budget constraint was often violated, the aggregate-constraint version produces outputs that never touch the budget constraint. This model has the opposite problem, which is that the budget constraint never binds, although we know that this is an equilibrium outcome. This reflects a tendency that arises when using the FB penalty: in cases where the penalty for breaking the constraints is large relative to other loss components, the solution may produce outputs strictly inside the constraint. This is related to the tendency of the FB function to under-punish interior solutions, as highlighted by Chen et al. (2000.). On the other hand, in cases where the penalty from breaking constraints is relatively small, the model may generate outputs that ignore the constraints entirely.

Figure 1.8 shows the consumption functions and MPCs generated by the intermediate constraint models. Since these fundamentally depict an idiosyncratic behaviour, it is not surprising that the aggregate-constraint version fails to capture important features such as the kinked consumption function and very high MPC for constrained agents. However, here MPC is decreasing in wealth, which is a feature that the soft-constraint model failed to capture. The takeaway here is that even though the aggregate-constraint version manages to perform relatively well quantitatively, it still fails to capture some of the key qualitative features of the model. The idiosyncratic-constraint

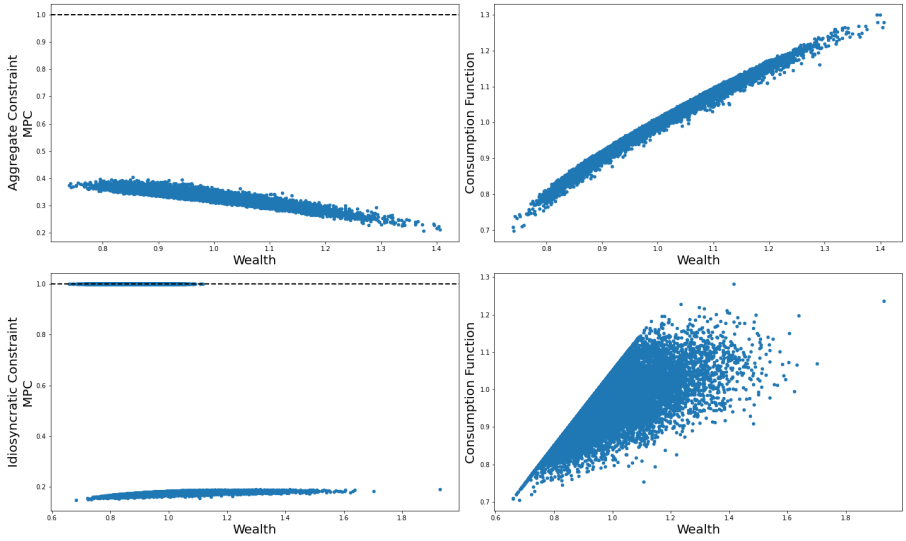


Figure 1.8: Immediate MPCs (left column) and consumption functions (right column) for the aggregate-constraint method (top row) and idiosyncratic-constraint method (bottom row), across all agents in a sample of states drawn from the ergodic distribution.

version (bottom row) does exhibit most of the features we expect to see, however, it is clear to see comparing to the hard-constraint version in Figure 1.5 that the scale is incorrect. The mean wealth is substantially lower than in the hard-constraint version, as a result of total output being below its equilibrium level. As a result, too many agents hit their budget constraint, and therefore the consumption function is more aggressively kinked compared to the hard-constraint results.

Figure 1.9 shows the generalised IRFs for the aggregate-constraint and idiosyncratic-constraint models. For ease of comparison, the baseline hard-constraint impulse responses are reproduced. As expected given its lower error, the hard-aggregate-constraint version comes significantly closer to reproducing the behaviour of the baseline model, not only for aggregate observables, but for cross-sectional ones as well. As with the pure soft-constraint version, the idiosyncratic-constraint model displays erratic behaviour for some cross-sectional variables such as the dispersion of consumption.

In practice, it is usually the case that while making one of the aggregate or idiosyncratic constraints of a heterogeneous agent model *hard* is fairly trivial, applying both at the same time can be difficult. The results of this section seem to suggest that if all of the constraints cannot be made hard at the same time, then the largest benefit relative to the soft-constraint approach comes from satisfying the aggregate constraints by construction. However, these results also show that the hard-constraint approach is the only one that approximates the model to a satisfactory degree of precision and captures all of the important qualitative features of the economic model. Therefore, when solving DSGE models using neural-networks every effort should be made to strongly enforce as many constraints of the model as possible.

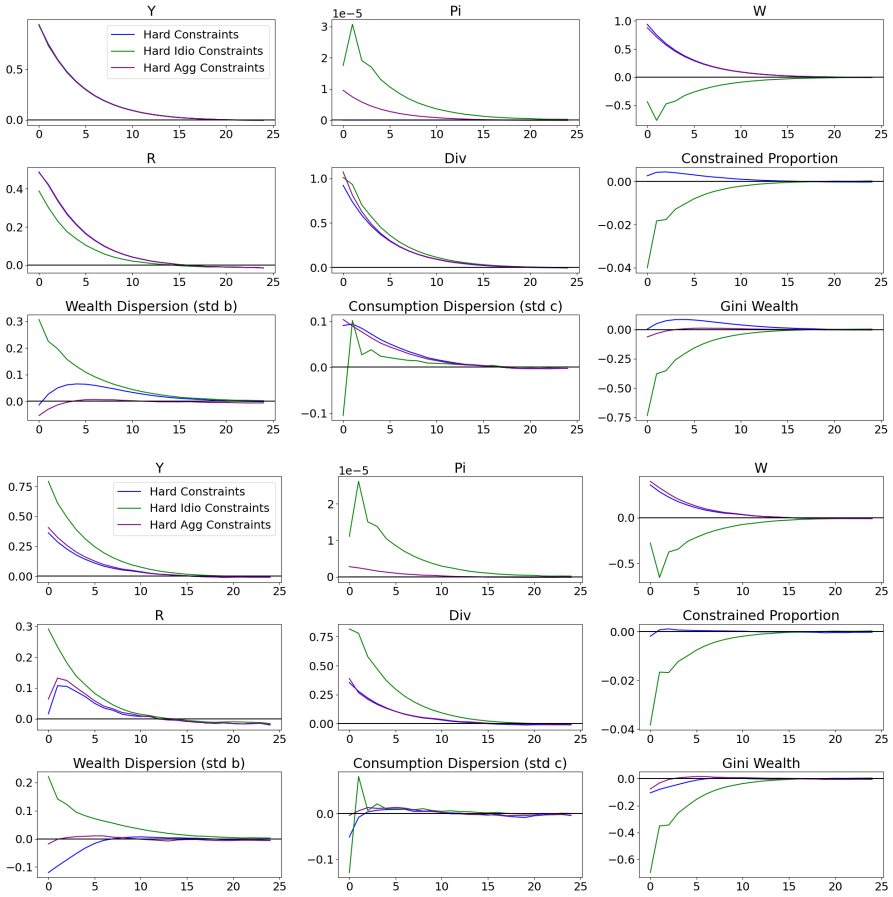


Figure 1.9: Impulse Responses to a 2 standard deviation expansionary TFP shock (ϵ_t^A) generated from the hard-idiosyncratic-constraint and hard-aggregate-constraint models. Results in the top three rows are an overall average over all states, and the bottom three rows show only ZLB states (where $R_{t-1} = 0$).

1.7 Conclusion

This paper has contributed to the understanding of how to deal with constraints when solving heterogeneous agent DSGE models using deep neural-networks. The results show that relatively small differences in implementation, which may seem to be equivalent in theory, may actually have significantly different implications in practice. This is primarily because every approximation technique has some degree of error, and when these errors relate to the constraints of a model they are more prone to propagation. Practitioners should be wary of how their treatment of constraints will impact the results obtained, and since there is still no disciplined way when working with neural-networks to determine how small the error generated by a solution has to be in order to be accepted, they should test their solutions carefully against expected quantitative and qualitative features of the model. A broad implication of these results is that while the neural-network offers a global solution, any potential method that can be used to reduce and constrain the search space to one that is closer to the true equilibrium solution is likely to speed up convergence and thus improve the quality of the final results obtained.

While no solution method is perfect, this paper has argued that the most appropriate approach is to give lexicographic priority to constraints — that is to attempt to restrict the output space of the approximating function to the set of feasible outputs, and only then attempt to optimise relative to the first-order conditions. The approach commonly suggested by the existing literature, which is to use penalties to enforce constraints, is theoretically sound and can work in some applications, but has some fundamental drawbacks that make it appealing to consider alternatives. This paper has provided a novel methodology that allows for all constraints to be satisfied by construction when using a neural-network as an approximating function in the important case of heterogeneous agents models with incomplete markets. The results demonstrate that this method is superior to alternative approaches that leave at least some of the constraints to be implemented via a penalty. Furthermore, the results consider intermediate cases in order to understand where this improvement originates from. These results seem to indicate that leaving any type of constraint to be soft, while simplifying the implementation, also decreases the quality of approximation substantially. Leaving aggregate constraints soft and idiosyncratic constraints hard makes it possible to capture the correct behaviour in a neighbourhood of the idiosyncratic constraint, but makes it difficult to correctly approximate aggregate variables and prices. Conversely, the opposite was observed in the case of hard aggregate constraints and soft idiosyncratic constraints.

There are, however, some shortcomings of my suggested method that have already been identified. Therefore, there is a need for further research in order to improve these methods. In particular, as mentioned in Section 1.3 there are corner cases where the rescaling can fail to produce outputs inside the constraints. Hopefully further research into similar types of rescaling will yield an equally effective method which is robust to these limitations. Furthermore, while I argue that the method I provide is applicable in a wide range of interesting applications, there is still a

need to develop a more general solution method that can deal with general constraints.

This paper is part of a growing literature on machine-learning solution methods for DSGE models, and as these methods improve, it will unlock the potential to consider increasingly complex and realistic models that can, for example, more accurately capture what happens when large shocks or structural breaks hit the economy, and how this interacts with inequality.

Bibliography

Aarts, L. P., & Van Der Veer, P. (2001). Neural network method for solving partial differential equations. *Neural Processing Letters*, *14*(3), 261–271.

Acharya, S., Challe, E., & Dogra, K. (2020). Optimal monetary policy according to hank.

Achdou, Y., Han, J., Lasry, J.-M., Lions, P.-L., & Moll, B. (2022). Income and wealth distribution in macroeconomics: A continuous-time approach. *The review of economic studies*, *89*(1), 45–86.

Adcock, B., Brugiapaglia, S., Dexter, N., & Moraga, S. (2020). Deep neural networks are effective at learning high-dimensional hilbert-valued functions from limited data. *arXiv preprint arXiv:2012.06081*.

Ahn, S., Kaplan, G., Moll, B., Winberry, T., & Wolf, C. (2018). When inequality matters for macro and macro matters for inequality. *NBER macroeconomics annual*, *32*(1), 1–75.

Auclert, A., Bardóczy, B., Rognlie, M., & Straub, L. (2021). Using the sequence-space jacobian to solve and estimate heterogeneous-agent models. *Econometrica*, *89*(5), 2375–2408.

Azinovic, M., Gaegauf, L., & Scheidegger, S. (2022). Deep equilibrium nets. *International Economic Review*.

Azinovic, M., & Žemlička, J. (2023). Economics-inspired neural networks with stabilizing homotopies. https://arxiv.org/abs/2303.14802

Bach, F. (2017). Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, *18*(1), 629–681.

Bayer, C., & Luetticke, R. (2018). Solving heterogeneous agent models in discrete time with many idiosyncratic states by perturbation methods.

Bellman, R. (1957). Dynamic programming, princeton univ. *Press Princeton, New Jersey*.

Carroll, C. D. (2006). The method of endogenous gridpoints for solving dynamic stochastic optimization problems. *Economics letters*, *91*(3), 312–320.

Chen, B., Chen, X., & Kanzow, C. (2000). A penalized fischer-burmeister ncp-function. *Mathematical Programming*, *88*(1), 211–216.

Dissanayake, M., & Phan-Thien, N. (1994). Neural-network-based approximations for solving partial differential equations. *communications in Numerical Methods in Engineering*, *10*(3), 195–201.

Duarte, V. (2018). Machine learning for continuous-time economics. *Available at SSRN 3012602*.

Fernández-Villaverde, J., & Rubio-Ramírez, J. F. (2007). Estimating macroeconomic models: A likelihood approach. *The Review of Economic Studies*, *74*(4), 1059–1087.

Fernández-Villaverde, J., Nuño, G., Sorg-Langhans, G., & Vogler, M. (2020). Solving high-dimensional dynamic programming problems using deep learning. *Unpublished working paper*.

Fischer, A. (1992). A special newton-type optimization method. *Optimization*, *24*(3-4), 269–284.

Galí, J., López-Salido, J. D., & Vallés, J. (2007). Understanding the effects of government spending on consumption. *Journal of the european economic association*, *5*(1), 227–270.

Greenwood, J., Hercowitz, Z., & Huffman, G. W. (1988). Investment, capacity utilization, and the real business cycle. *The American Economic Review*, 402–417.

Han, J., & Yang, Y. (2021). Deepham: A global solution method for heterogeneous agent models with aggregate shocks. *arXiv preprint arXiv:2112.14377*.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359–366.

Judd, K. L., Maliar, L., Maliar, S., & Valero, R. (2014). Smolyak method for solving dynamic economic models: Lagrange interpolation, anisotropic grid and adaptive domain. *Journal of Economic Dynamics and Control*, *44*, 92–123.

Kaplan, G., Moll, B., & Violante, G. L. (2018). Monetary policy according to hank. *American Economic Review*, *108*(3), 697–743.

Kase, H., Melosi, L., & Rottner, M. (2022). Estimating nonlinear heterogeneous agents models with neural networks.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koop, G., Pesaran, M. H., & Potter, S. M. (1996). Impulse response analysis in nonlinear multivariate models. *Journal of econometrics*, *74*(1), 119–147.

Krusell, P., & Smith, A. A., Jr. (1998). Income and wealth heterogeneity in the macroeconomy. *Journal of Political Economy*, *106*(5), 867–896.

Maliar, L., & Maliar, S. (2020). Deep learning: Solving hanc and hank models in the absence of krusell-smith aggregation. *Available at SSRN 3758315*.

Maliar, L., Maliar, S., & Winant, P. (2021). Deep learning for solving dynamic economic models. *Journal of Monetary Economics*, *122*, 76–101.

Maliar, S., Maliar, L., & Judd, K. (2011). Solving the multi-country real business cycle model using ergodic set methods. *Journal of Economic Dynamics and Control*, *35*(2), 207–228.

Reiter, M. (2009). Solving heterogeneous-agent models by projection and perturbation. *Journal of Economic Dynamics and Control*, *33*(3), 649–665.

Rotemberg, J. J. (1987). The new keynesian microfoundations. *NBER macroeconomics annual*, *2*, 69–104.

Smith, M., et al. (2011). Estimating nonlinear economic models using surrogate transitions. *Federal Reserve Board, manuscript.*

Appendix

A Neural-network Background

A fully-connected, feed-forward neural-network is defined by *layers* made by applying non-linear transformations known as *activation functions* σ_i to linear combinations of inputs (Bengio et al., 2017). The linear combinations are defined by multiplicative parameter matrices W_i known as *weights* and additive parameter matrices b_i known as biases. In the case of the first layer the inputs are data, but otherwise the inputs are the outputs of the previous layer. Layers are then stacked a given number of times, until the output of the last layer is deemed to be the output of the neural-network. The dimensions of the output layer parameter matrix are chosen to match the desired output shape of the neural-network, and activation functions can in some cases be chosen in order to limit outputs to a certain range, depending on the application. For example, the activation function $\sigma(x) = exp(x)$ may be chosen if strictly positive outputs are desired.

A neural-network with N layers consists of a set $\theta = \{W_i, b_i\}_{i=1}^{N}$ of *learnable parameters*. These are called *learnable* to emphasise the difference with *hyper-parameters*, such as the step size or learning rate. These hyper-parameters are usually chosen before the estimation procedure, and remain fixed throughout. The learnable parameters are usually initialised randomly, and then updated iteratively via SGD by a learning algorithm:

$$\theta_{t+1} = \theta_t - \alpha L_\theta(X; \theta_t) \tag{1.37}$$

where α is the *learning rate*. In baseline SGD α is fixed, however, more advanced learning algorithms such as ADAM (Kingma & Ba, 2014), which is applied in this paper, can be used to dynamically update the learning rate. Performing SGD requires calculating the gradient of the loss with respect to the learnable parameters. This can be done very efficiently with modern machine-learning software thanks to *backpropogation*. Due to the nested nature of the neural-network, application of the chain rule implies continually recalculating the gradient of intermediate layers. By caching these intermediate gradients, this seemingly expensive computational operation can be performed rather quickly. The gradient can efficiently be calculated in this way without any user code by most modern machine-learning software, such as JAX, which was used in this project.

B Aggregate States and Prices Generated by Different Solution Methods

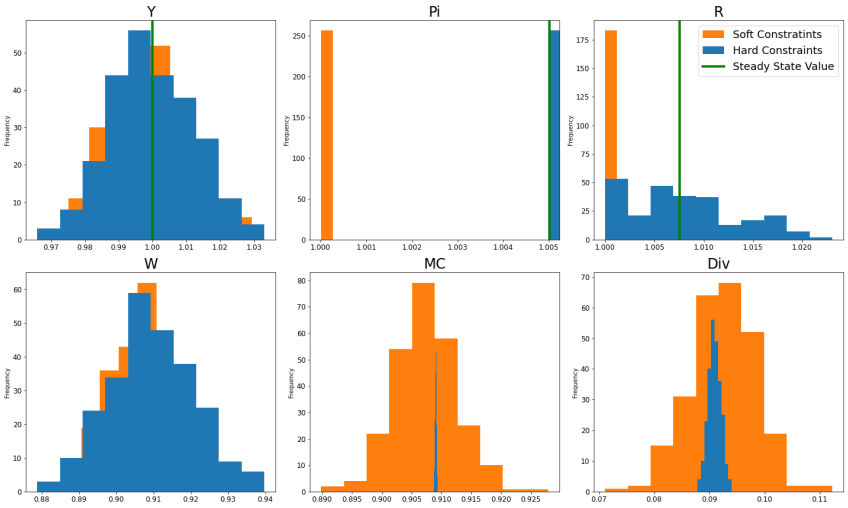


Figure 1.10: Distribution of aggregate variables and prices from a sample drawn from the ergodic distribution of sets for the hard-constraint method (Blue) and soft-constraint method (Orange). Steady-state values, where known, are marked as green vertical lines.

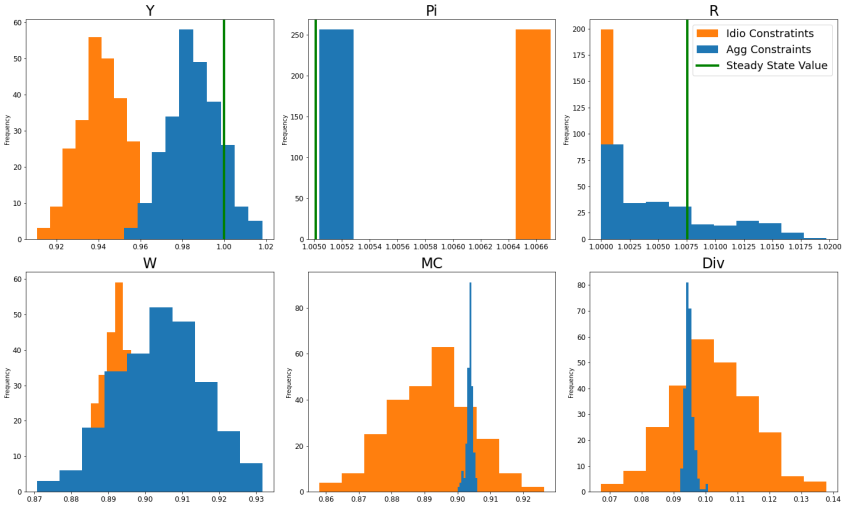


Figure 1.11: Distribution of aggregate variables and prices from a sample drawn from the ergodic distribution of sets for the aggregate-constraint method (Blue) and idiosyncratic-constraint method (Orange). Steady-state values, where known, are marked as green vertical lines.

C Neural-network Fitting Algorithm

Algorithm 3: Neural-network Fitting Algorithm

Input: $\theta_0 \in \mathbb{R}^M, X_0 \in \mathbb{R}^{mb \times k}, I \in \mathbb{N}, N \in \mathbb{N}, tol_0 \in \mathbb{N}, tol_1 \in \mathbb{N}, tol_2 \in \mathbb{N}, \alpha \in \mathbb{R}_+$
Output: $\theta_t \in \mathbb{R}^M, \theta_t \approx \operatorname{argmin}_\theta \mathbb{E}_{X,\Gamma} L(X, \theta; \Gamma)$

$X \leftarrow$ drawInitState();
$\Gamma \leftarrow$ drawStructParams();
$\theta \leftarrow \theta_0$;
$itsAtCurrentN \leftarrow 0$;
$lastLoss \leftarrow \infty$;
$n \leftarrow 1$;
$i \leftarrow 0$;
while $lastLoss > tol_0$ ***and*** $i < I$ **do**
  $loss \leftarrow \frac{1}{mb} \sum_{j=1}^{mb} L(X, \theta; \Gamma)$;
  $\theta \leftarrow \theta - \alpha \frac{1}{mb} \sum_{j=1}^{mb} \nabla_\theta L(X, \theta; \Gamma)$;
  if isnan($loss$) ***or*** $loss > tol_1$ **then**
    if $n > 1$ **then**
      $n \leftarrow n - 1$;
    $X \leftarrow$ drawInitState();
    $itsAtCurrentN \leftarrow 0$;
  if $loss < tol_1$ ***and*** $\neg$ isnan($loss$) ***and*** $itsAtCurrentN > tol_2$ ***and*** $n < N$ **then**
    $n \leftarrow n + 1$;
    $itsAtCurrentN \leftarrow 0$;
  $\Gamma \leftarrow$ drawStructParams();
  for $k \in \{1, \ldots, n\}$ **do**
    $X \leftarrow S(X, \theta; \Gamma)$;
  $itsAtCurrentN \leftarrow itsAtCurrentN + 1$;
  $lastLoss \leftarrow loss$;
  $i \leftarrow i + 1$;

D Soft-constraint Results for Different Penalty Weights

		Penalty Weight		
Type		**1e1**	**1e2**	**1e4**
	Total Loss	2.46e-02	1.13e-02	3.39e-02
Optimality	Euler Equation Loss	2.39e-03	5.64e-03	3.21e-03
	Phillips Curve Loss	2.09e-02	5.14e-03	3.06e-02
	Labour Supply Loss	1.09e-32	1.10e-32	1.09e-32
Constraints	KKT Loss	1.93e-04	2.74e-04	4.56e-06
	Output Constraint Loss	1.26e-04	1.03e-04	5.06e-05
	Net Supply of Bonds	9.24e-04	1.56e-04	6.05e-05

Table 1.5: Average total loss and its constituent components evaluated over the last 50 iterations of the fitting procedure for the models fit using the soft-constraint version with various weights applied to the constraint-related penalties. All runs were warm-started with parameters from the baseline configuration presented for the soft-constraint methodology Section 1.6.

E Settings Used for Generating Results

Parameter	Value (Soft-Constraint and Idio-Constraint)	Value (Hard-Constraint and Agg-Constraint)
Training Iterations	200000	100000
Batch Size	256	256
Max Forward Sims per Update	20	20
Learning Rate	1e-6	1e-4
Precision	x64	x64
Scale of Initial Parameters	1e-2	1e-2
eps (parameter for ADAM optimiser)	1e-12	1e-12

Table 1.6: Hyper-parameters used while training models to generate results presented. The soft-constraint version was allowed to train longer at a lower learning rate compared to the hard-constraint version because it requires significantly more precision in parameter updates to converge around the constraints. The hard-constraint version on the other hand can be used with a larger learning rate (and thus, less precise parameter updates), which allows it to converge more quickly while maintaining its inherently high level of precision.

Chapter 2

HANK and the Minimum Wage

Abstract

This paper employs a novel machine-learning based solution technique to solve a single asset Heterogeneous Agent New-Keynesian (HANK) model, with the addition of a real minimum wage policy, which is calibrated to the range observed in OECD countries. The application of a non-linear solution and a rich cross-sectional distribution allows for the development of new insights on this classical topic. In particular, although an increase in the minimum wage can cause a strong demand side impulse as it increases the income primarily of poorer agents, who have high Marginal Propensity to Consume (MPC), this is in all cases studied dominated by a contractionary general equilibrium effect resulting from the distortion of the labour market caused by the minimum wage. The overall result is a contraction in output of $\sim 0.1\%$ after a shock that increases the proportion of workers who earn the minimum wage by $\sim 20\%$ (relative), in the baseline calibration. The overall contraction is worse when the starting minimum wage is higher or there is more dispersion in productivity, even though the demand side effect is growing more positive in both. This contractionary effect is also worse when the economy is in a deep recession that has driven it to the Zero Lower Bound (ZLB) on nominal interest rates. Increases in the minimum wage may also have the presumably unintended side effect of having the highest cost for agents in the centre of the productivity distribution, who sell their savings primarily to richer workers to self-insure against the adverse shock to their income.

2.1 Introduction

The minimum wage is a classical and fiercely debated topic in economics research. Empirical research into the topic has studied the effect of the statutory minimum wage on employment (Card & Krueger, 2000; Clemens & Strain, 2020), inequality (Autor et al., 2016; Lee, 1999), and productivity (Ku, 2022; Riley & Bondibene, 2017). However, most empirical research focuses on local or partial equilibrium effects. This is perhaps because the effects at the macro level are unclear in the data. Figure 2.1 shows a scatter plot with the relationship between real GDP growth at PPP for OECD countries in the years 2000-2024. Given the salience of the topic, the correlation seen here is remarkably close to zero.

However, as pointed out by Sabia (2015), these aggregates may obscure interesting cross-sectional and dynamic effects. In order to get an understanding for what these effects and underlying mechanisms might be, it would make sense to employ a theoretical modelling approach. Yet, from the theoretical side, it has been difficult to model the general equilibrium effect of both a rich cross-sectional distribution of wealth and productivity and the minimum wage at the same time, as this introduces non-linearity and high-dimensionality to the problem. This is where this paper aims to contribute to the literature.

In order to do so, this paper will consider a non-linear HANK model featuring a rich cross-sectional productivity distribution and a stochastic and time-varying real minimum wage. The results of this paper, notwithstanding its limitations, which will be discussed, show that the overall static and dynamic effects of the minimum wage on aggregate variables is in nearly all cases adverse. Indeed, the effect on inequality can even be ambiguous, as the policy simply changes the balance of winners and losers rather than causing a uniform reduction in inequality. Therefore, the results of this paper can actually be seen as a confirmation of the traditional wisdom on the topic of the minimum wage: despite modelling a rich cross-sectional distribution and enabling HANK style demand side effects, the policy causes large distortions that are adverse on the aggregate level, and has consequences that are presumably unintended for the cross-sectional income distribution.

The primary reason for this result is that despite pricing friction in the market for their outputs, firms are able to pass along a large portion of cost increases from the increase in the minimum wage in the form of real wage cuts for more productive workers. This in turn strongly disincentivises labour supply, in particular for the most productive workers, whose decreased hours have a relatively large impact on aggregate output. On the other hand, the workers who benefit from the minimum wage tend to not increase their labour supply particularly strongly in response to increased wages because they are already working a relatively high amount of hours in order to avoid the budget constraint. Even if they did, this would still have a relatively small effect on aggregate output due to their low productivity. Although there is a substantial (previously outlined) demand-side *redistributive* effect working against the supply-side *distortionary* effect, which comes from the increased demand of minimum wage workers, quantitatively the former is almost always

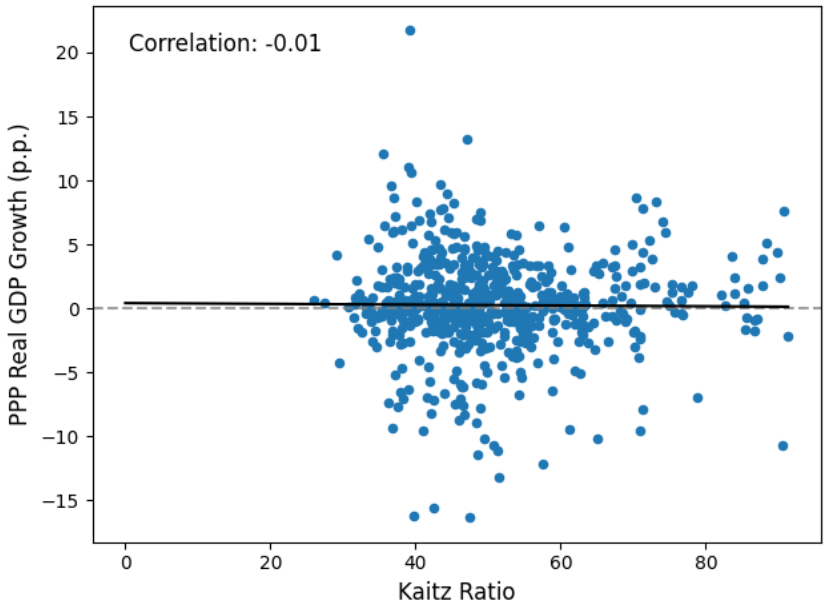


Figure 2.1: Relationship between real GDP at PPP (y-axis) and Kaitz ratio (ratio of minimum to median wage) across OECD countries in the years 2000-2024.

dominated by the latter, resulting in (increases in) the minimum wage being contractionary. In other words, while the model captures a strong positive partial equilibrium effect of the minimum wage, this is more than offset in general equilibrium.

Furthermore, increases to the minimum wage tend to cause a *hollowing of the middle* or "U-shaped" effect, wherein the poorest and richest are better off whereas those in the middle bear most of the cost of the policy. The reason that the richest agents benefit is that labour income makes up a smaller portion of their overall income, so they are relatively sheltered from the effect of wage cuts, while they profit from middle-income households selling off their savings to self-insure against the adverse shock to their wages. The most productive agents also collect the largest proportion of the firms' profits, which actually tend to increase following an increase to the minimum wage (although these make up a relatively small portion of the most richest agents' incomes), which to some extent may amplify this effect.

There is a common thread that flows through the argumentation of many recent and prominent HANK papers, wherein the explicit modelling of a rich cross-sectional distribution of agents allows for conclusions that contradict the traditional wisdom on a topic, which is itself usually based on models involving perfect markets and in particular representative agents. Examples include G. Kaplan et al. (2018), Bayer et al. (2023), and Kwicklis (2025). The minimum wage at first glance seems to be a likely candidate for a similar type of analysis: classical theories suggest that the minimum wage is a typical example of well-intentioned policy that actually backfires because the distortions it causes outweigh any potential benefits. This is, for example, the argument of Hicks (1932), based on a neoclassical model with perfect markets.

Yet there is good reason to believe that in a HANK model this classical reasoning may be overturned. Since the root source of inequality in HANK models is the uneven distribution of productivity across agents, those with low productivity, and thus low wages, are therefore most likely to be affected by the minimum wage also tend to be those who are less wealthy and closer to the borrowing constraint. These agents are the key to all deviations of implications of heterogeneous agent models from their representative agent counterparts, as the high MPC of agents at or close to the budget constraint causes so called "hand-to-mouth" consumption behaviour resulting in implications that deviate from unconstrained forward-looking Ricardian behaviour implied by standard representative agent models. The high MPC of (nearly) constrained agents means that policies designed to increase their income tend to cause disproportionately large increases in aggregate demand, which in a (New) Keynesian framework then leads to increases in output (G. Kaplan et al., 2018). The minimum wage therefore seems like a perfect candidate to invoke such a response. However, as discussed, while such an effect can be observed in the model, it is dominated by the distortions caused by the minimum wage, resulting in adverse overall effects. Furthermore, it is a policy that only makes sense to consider within the context of a model featuring a rich unequal distribution over productivity, which is possible to model with a significant level of detail and flexibility in the framework employed in this paper.

As outlined in Hall-Hoffarth (2023), this high-dimensional and non-linear setting is an example of a situation where the machine-learning solution method outlined there is particularly helpful vis-à-vis alternative methods. Therefore, this paper aims to leverage these novel machine-learning techniques to shed new light on this topic. Furthermore, the exact implementation in this paper combines many contributions made in recent years by various authors, in particular Pascal (2024), in a way that has to my knowledge not been done in any other paper, representing the current *state-of-the-art* for the neural-network DSGE solution method. This model and global non-linear solution method will allow for the generation of new perspectives on how the minimum wage interacts with inequality and the state of the economy. In particular, it will be possible to consider whether the effects of an exogenous increase in the real minimum wage are state-dependent, and whether any of these effects interact with the Zero Lower Bound (ZLB) on nominal interest rates. The results show that the adverse effects of an increase to the minimum wage are simply amplified by the ZLB.

Furthermore, I employ the extended neural-network methodology of Kase et al. (2022) to approximate the model over a range of parameters simultaneously, meaning that once the model is trained, inferences can be queried from the model for various combinations of parameters without having to re-solve the model. In particular, this allows for the comparison of simulations and impulse responses over different levels of the minimum wage and values for the variance of the productivity distribution. The results show that when either of these are higher, both the redistributive and distortionary effects grow, but the latter grows more quickly resulting in a larger

overall negative effect.

The remainder of this paper is organised as follows: Section 2.2 contains a review of the relevant empirical and theoretical literature. Section 2.3 introduces the model to be analysed, including the specific implementation of the minimum wage. Section 2.4 discusses details of the deep-learning solution method specific to this paper. Section 2.5 then presents and analyses the implications of the model. Section 2.6 discusses some limitations of this paper and how they could be addressed in future research. Finally, Section 2.7 concludes.

2.2 Literature Review

2.2.1 Minimum Wage

The minimum wage is a classical topic in economics, and there is a long-standing debate around when and if it may be an effective policy tool. The (neo)-classical view is that the labour market is predominantly efficient, and that the minimum wage is therefore a market inefficiency which will increase unemployment, decrease output, and decrease total welfare. However, more recent empirical and theoretical research has cast doubt on this view, as summarised by Dube and Lindner (2024). The labour market, especially for relatively low-wage workers, may be subject to a number of frictions that potentially make it inefficient. For example, firms may have substantial market power in low-wage sectors; workers on low wages may have very inelastic labour-supply due to a subsistence minimum and be particularly immobile as a result of moving costs, allowing their employers to suppress wages below the efficient level. In these cases, a minimum wage, even if compliance is imperfect, can create upwards pressure on wages that pushes the equilibrium towards the second-best outcome. However, a view to the contrary is given by Neumark and Wascher (2008), who in their review conclude that the macroeconomic effects of the minimum wage are limited and that it has minimal effect on reducing poverty. Similarly, Caliendo et al. (2019) study the effect of the introduction of a minimum wage in Germany in 2015. They focus on the short-run implications of the policy, especially on low-wage workers. They find that although there is a meaningful increase in wages for workers at the bottom of the income distribution, and relatively small dis-employment effects, there was also substantial non-compliance, and a tenancy among these workers to reduce their hours, which diminished the effect of the policy on reducing poverty.

Interestingly, although one might expect these effects to be non-linear, and for the distortion to be stronger at higher levels, some empirical research shows small employment effects even for relatively large increases in the minimum wage (Dube & Zipperer, 2024). However, it is important to note that many of these empirical studies following the seminal work of Card and Krueger (2000) have considered localised effects, such as industry-specific price effects, and therefore may miss out on spillovers that only become relevant in the macroeconomic scale. For example Aaronson et

al. (2008) find that minimum wage increases are passed on and result in inflationary pressure, however their analysis is limited to the restaurant sector. Amongst papers that consider broader, macroeconomic effects, Sabia (2015) finds in a panel-data analysis of OECD countries that while there is no statistically significant aggregate effect of the minimum wage on output, higher minimum wages do result in a reallocation from low-productivity to high-productivity sectors, suggesting that the policy backfires in its attempt to aid low-wage workers. Harasztosi and Lindner (2019) find that in the case of a large increase in the minimum wage in Hungary the result was a relatively mild reduction in employment and profit margins, and that therefore the largest effect was an increase in prices born by consumers.

2.2.2 Theoretical Models

Relatively few papers have attempted to approach the topic of the minimum wage from a macroeconomic modelling or DSGE perspective, considering the importance of the topic within the broader discipline of economics. Perhaps the most rigorous and detailed of these is Drechsel-Grau (2023), who estimates a DSGE model of the German economy with search and matching frictions in the labour market in the tradition of Mortensen and Pissarides (1994), and of course a minimum wage policy. They find that the minimum wage can be an effective policy, increasing output and having only minor negative effects on employment up to a relatively high Kaitz ratio (minimum wage relative to median wage) of 0.7. The primary mechanism that they identify for this is a reallocation towards more productive worker-firm matches. Similarly, in their search-and-matching model, Bauducco and Janiak (2018) find that the minimum wage can be an effective policy tool: up to moderate levels employment effects are small, and even at higher levels it is possible that increases in productivity and capital accumulation outweigh decreases in employment leading to overall higher output. Conversely, Braun et al. (2020) also attempt to model the effect of the introduction of the minimum wage in Germany, however, in their model output decreases, despite increased productivity, after the introduction of the minimum wage.

Other papers using DSGE models to evaluate the minimum wage include Šauer (2018) and Antonova (2018). Both of these papers use a Two-Agent New Keynesian (TANK) model to model the effect of a minimum wage policy. Both of these papers find a more limited role for the minimum wage: it is usually inflationary, and has an adverse effect on aggregate output, except for in the case that the proportion of hand-to-mouth or non-Riccardian households is high. Finally, Alege et al. (2021) also study the minimum wage in the context of a calibrated model of the Nigerian economy, and overall find adverse aggregate effects, especially to large increases. Relative to these papers, which are based on TANK models, I contribute a more nuanced HANK model, which allows for a continuous distribution of agent productivity and wealth, and allows for the distribution of these to change dynamically and endogenously to the policy. This in particular allows the model to capture the fact that increasing the minimum wage will not only increase the income of the agents

who were previously at the minimum wage, but also increase the proportion of workers for whom the minimum wage is binding.

2.3 Model

The modelled economy includes heterogeneous households, heterogeneous but symmetrical intermediate firms, a final goods firm, and a policy block containing a monetary and fiscal authority (government).

2.3.1 Households

L households indexed by i,[1] each with their own idiosyncratic and (highly) persistent $AR(1)$ labour productivity sequence s_t^i in each period choose their hours h_t^i and consumption c_t^i in order to solve the following optimisation problem subject to a borrowing constraint $\underline{B}$:

$$\underset{\{c_t^i, h_t^i\}}{max} \sum_{t=0}^{\infty} \mathbb{E}_0 \left[\beta \Psi_t \left(\frac{{\nu_t^i}^{1-\sigma}}{1-\sigma} - \chi \frac{{h_t^i}^{1+\eta}}{1+\eta} \right) \right] \tag{2.1}$$

$s.t.$

$$\nu_t^i = c_t^i - hC_t \tag{2.2}$$

$$c_t^i + b_t^i = \frac{R_{t-1}}{\Pi_t} b_{t-1}^i + (1-\tau)\, w_t^i h_t^i + Div_t^i \tag{2.3}$$

$$b_t^i \geq \underline{B} \tag{2.4}$$

s.t.

$$s_t^i = exp(\rho_s log(s_{t-1}^i) + \sigma_s \epsilon_t^{i,s}) \tag{2.5}$$

$$\Psi_t = exp(\rho_\Psi log(\Psi_{t-1}) + \sigma_\Psi \epsilon_t^{\Psi}) \tag{2.6}$$

Where Ψ_t is a persistent aggregate demand shock, used, for example, to induce a recession in the modelled economy. Each households' savings (b_t^i) result from the budget constraint. The households' optimisation problem results in two first-order conditions: an Euler equation and a labour supply equation:

$$1 - \mu_t^i = \beta R_t \mathbb{E}_t \left[\frac{1}{\Pi_{t+1}} \frac{\Psi_{t+1}}{\Psi_t} \left(\frac{\nu_{t+1}^i}{\nu_t^i} \right)^{-\sigma} \right] \tag{2.7}$$

$$h_t^i = \left(\frac{w_t^i (1-\tau)}{\chi {\nu_t^i}^{\sigma}} \right)^{\frac{1}{\eta}} \equiv \kappa {w_t^i}^{\frac{1}{\eta}} {\nu_t^i}^{-\frac{\sigma}{\eta}} \tag{2.8}$$

[1]This can be thought of as a discretisation of an underlying model with a continuous productivity distribution, however, for simplicity I always refer to a discrete number of agents.

For each combination of parameters used, χ is to the value for the labour-supply equation that implies that $\bar{h^i} = \bar{Y} = 1$ in the Deterministic Steady State (DSS).

2.3.2 Firms

The firm sector contains both monopolistically-competitive intermediate firms and a competitive final-goods firm, as is standard in NK models. All intermediate firms are assumed to be symmetrical and produce with a linear production technology using labour as the only input, therefore:

$$Y_t = N_t \equiv \frac{1}{L} \sum_{i=1}^{L} s_t^i h_t^i \tag{2.9}$$

The key assumption in this model is that the intermediate firms do not hire individual workers, but rather demand labour from the households in bulk, and only internalise the average wage paid in their cost function. As a result, firms are unable to exclude the hiring of workers whose productivity is less than their marginal product of labour (MPL) due to the minimum wage. This is a strong assumption, but it might be justified by for example considering firm cost planning only on a high level or the existence of some (unmodelled) complementarity between high and low productivity workers, which implies that firms will always demand some low-productivity labour. The minimum wage is implemented as a floor on the real wage per actual hour worked h_t^i, relative to the prevailing wage per actual hour worked ($w_t^{i\star}$) paid to unconstrained, and therefore efficiently remunerated households:

$$w_t^{i\star} = s_t^i W_t \tag{2.10}$$

$$w_t^i = \max\left\{w_t^{i\star}, W_t \underline{W}_t\right\} = W_t \max\left\{s_t^i, \underline{W}_t\right\} \equiv W_t p_t^i \tag{2.11}$$

In the absence of a minimum wage, since the final goods firm is competitive, workers are paid for each hour worked (h_t^i) proportional to the marginal product of that hour $w_t^{i\star} = W_t s_t^i$, which implies that aggregate wages paid are equal to marginal cost $W_t = MC_t$ because $\frac{1}{L}\sum_{i=1}^{L} s_t^i = 1$. The minimum wage is then some proportion $\underline{W}_t$ of the prevailing wage per productive hour of labour W_t, modified by the stochastic $AR(1)$ process g_t. Rearranging this shows that the effect of this type of minimum wage is to pay workers as if their productivity were no lower than $\underline{W} g_t$ per actual hour worked. This "paid-productivity" is denoted by p_t^i.[2]

[2]Another straightforward approach would be to treat $\underline{W}$ as an absolute rather than relative minimum wage, where $w_t^i = \max\left\{W_t s^i t, \underline{W}\frac{\epsilon-1}{\epsilon} g_t\right\}$. In this case the parameter can instead be interpreted as setting the minimum wage to a proportion of the steady-state mean wage, in other words the Kaitz ratio. The key difference between these alternative implementations is that in the previously introduced relative minimum wage formulation aggregate shocks do not interact directly with the minimum wage — it is only the households' productivity s_t^i that determines if they earn above the minimum wage or not. The idea of this formulation is to isolate and only consider the distributional impact of the minimum wage, ignoring any aggregate impact that it could have if it were set to an absolute level. This has the advantage of making it possible to separate the distortionary and redistributive channels cleanly, however, it has the disadvantage

The low-productivity households that benefit from this tend to be less wealthy overall (because the productivity process is highly persistent), and therefore have high MPC. Therefore, the redistribution of income towards these agents tends to increase overall aggregate demand and thus have an expansionary impact on inflation and output. This is a fundamental mechanism in most HANK models (vis-à-vis their representative-agent counterparts). I will refer to this effect as the *redistributive channel*. However, this is not a costless windfall for unproductive households, as it causes a distortion to the marginal cost of the firm:

$$MC_t = \frac{1}{L}\sum_{i=1}^{L}\frac{w_t^i}{s_t^i} = W_t\frac{1}{L}\sum_{i=1}^{L}\frac{p_t^i}{s_t^i} \equiv W_t P_t \tag{2.12}$$

An increase in the minimum wage (weakly) increases P_t, which denotes the aggregate distortion directly caused by firms paying workers as if their productivity were higher than it actually is. Therefore, all else equal, the minimum wage causes upwards pressure on the marginal cost of the intermediate firms. Given this, another way to understand the labour market equilibrium is as follows. First firms select aggregate wage level W_t. Then each worker's wage is obtained by multiplying this by that workers paid-productivity p_t^i. Given this the labour supply of each agent is determined by equation (2.8), and aggregation of this implies total labour supply and thus output. In general equilibrium, firms will realise that due to the minimum wage, each unit of effective productivity will cost them more than W_t, and will lower W_t accordingly. In particular, this will then disincentivise labour supply from high-productivity households, resulting in downward pressure on output. I will refer to this effect as the *distortionary channel*. In this sense, a minimum wage increase functions similarly to a cost-push shock.

Therefore, the overall effect of the minimum wage on aggregate variables is in general ambiguous in this model; higher pay to low productivity households will lead to increased demand through the *redistributive channel*, however, it has an adverse supply-side effect through the *distortionary channel*. Which of these effects dominates will determine the aggregate effect, however, it is already visible here that it is potentially state-dependent. If aggregate demand is depressed, as in a demand-led recession the redistributive channel is stronger, and therefore more likely to dominate. This mirrors the classical Keynesian logic of demand-side intervention. On the other hand, although the model does not capture the dynamics of unemployment, it does more neatly capture the intuition that a minimum wage increases friction and thus decreases efficiency in the labour market, and represents this in a closed-form, while allowing for these effects to be state and parameter dependent. The focus of this paper is not the realism of the labour market structure, but rather the state-dependant and non-linear nature of these effects.

The intermediate firms are as in the standard New-Keynesian model with Rotemberg (1982)

that it is not possible to consider how the minimum wage can function as an *automatic stabiliser*, as, without any other intervention, the proportion at the minimum wage does not increase automatically in any recession generated by an aggregate shock. Instead, we will consider the effects of a minimum wage shock at the ZLB versus from the steady-state in Section 2.5.3.

adjustment costs, aside from their marginal cost. Since firms optimise over price taking marginal cost as given, the derivation of the New-Keynesian Phillips Curve (NKPC) is standard (see Bianchi et al. (2021)) and as a result the NKPC (in levels) can be written as:

$$\phi\left(\frac{\Pi_t}{\Pi}-1\right)\frac{\Pi_t}{\Pi}=(1-\epsilon)+\epsilon MC_t+\beta\phi\mathbb{E}_t\left[\frac{\Lambda_{t+1}}{\Lambda_t}\left(\frac{\Pi_{t+1}}{\Pi}-1\right)\frac{\Pi_{t+1}}{\Pi}\frac{Y_{t+1}}{Y_t}\right] \tag{2.13}$$

2.3.2.1 Proportion at minimum wage

For a constant minimum wage $\underline{W}$, the proportion of workers at the minimum wage depends implicitly on the assumed log-normal productivity distribution with parameters σ_s and ρ^s. Specifically, the long run distribution of the $AR(1)$ process generating s_t^i is $s_t^i \sim LogNorm\left(0, \frac{{\sigma_s}^2}{1-{\rho_s}^2}\right)$. In order to disentangle the proportion of minimum wage workers from the productivity distribution and facilitate calibration and interpretation, the minimum wage process in each period is passed through the quantile function of the distribution of s_t^i such that:

$$\underline{W}_t = F_{\{\sigma_s,\rho_s\}}(\underline{W}g_t) \tag{2.14}$$

Where g_t is itself also an $AR(1)$ process with parameters σ_g and ρ_g that generates variation in the minimum wage. This means that the parameter $\underline{W}$ can directly be interpreted as the average proportion of agents at the minimum wage. In a stochastic simulation this amount will vary based on the particular realisations of the shocks s_t^i.

2.3.2.2 Disentangling channels

In order to disentangle these the *redistributive* and *distortionary* channels, a parameter $\iota \in \{0, 1\}$ is implemented that governs whether the intermediate firms internalise the effect of the minimum wage, and the neural-network is trained simultaneously over both values.

$$MC_t = W_t\left(\iota P_t + (1-\iota)\right) \tag{2.15}$$

When $\iota = 0$ the intermediate firm behaves "as-if" the minimum wage did not affect their marginal cost, so the outcomes are affected solely through the redistributive channel. Therefore, the difference between policies calculated with $\iota = 1$ and $\iota = 0$ is the distortionary channel, and $\iota = 1$ is the total effect.

2.3.3 Monetary and Fiscal Authority

The government sector consists of a monetary authority and a fiscal authority. The monetary authority sets the nominal interest rate according to a dual-mandate Taylor rule without persistence and subject to the ZLB. I also implement the asymmetric policy rule of Bianchi et al. (2021), where the monetary authority reacts more strongly to deviations below the target than above it. The

effect of allowing some longer periods of high inflation is to offset the deflationary bias caused by risk parameters σ_Ψ and σ_s that occurs in a global solution. In particular, this is essential in order to be able to use a high enough value of σ_s to match data on the proportion of workers at the minimum wage and the value of the minimum wage relative to the median wage, without causing so much deflationary bias that the model finds no equilibrium due to a deflationary spiral. This deflationary bias is shown in Figure 2.17 in the appendix.

$$R_t = \max\left\{R\left[\mathbf{1}\left\{\Pi_t > \Pi\right\}\left(\frac{\Pi_t}{\Pi}\right)^{\overline{\theta^\Pi}} + \mathbf{1}\left\{\Pi_t \leq \Pi\right\}\left(\frac{\Pi_t}{\Pi}\right)^{\underline{\theta^\Pi}}\right]\left(\frac{Y_t}{Y}\right)^{\theta^Y}, 1\right\} \tag{2.16}$$

The fiscal authority collects revenue from a fixed proportional tax on labour income τ and uses this to service debt D_t, spending G_t, and transfers T_t. In the default specification T_t is fixed to its steady-state value, and G_t adjusts to satisfy the government's budget constraint.

$$D_t = D \tag{2.17}$$

$$T_t = T \tag{2.18}$$

$$G_t = D_t - \frac{R_{t-1}}{\Pi_t}D_{t-1} + \tau N_t - T_t \tag{2.19}$$

2.3.4 Equilibrium

As in Azinovic et al. (2022), the neural-network solves for a *Functional Rational Expectations Equilibrium* (FREE). The states of the model consist of the current values of shock processes (exogenous states) $X_t^e = \left\{\Psi_t, g_t, mp_t, s_t^i\right\}$ and the endogenous states $X_t^n = \left\{R_{t-1}, D_{t-1}, b_{t-1}^i\right\}$ such that $X_t = \{X_t^e, X_t^n\}$ with domain $\mathscr{X}$. The minimal policies (from which all other time-t variables can be calculated) are $Y_t = \left\{\Pi_t, MC_t, h_t^i, c_t^i, \mu_t^i\right\}$. The FREE is a mapping $X_t \to Y_t$ such that equations (2.7) – (2.19) hold. This mapping is parametrised as a neural-network $\hat{Y}_t = \hat{f}(X_t)$ and the process of approximating the equilibrium consists of finding the parameters for the neural-network that minimise the errors in the equilibrium conditions. More details are provided in Section 2.4. Note that although in principle the mapping is valid for any $X_t \in \mathscr{X}$, in practice, especially in high-dimensional state-spaces only a vanishingly small subset of states in $\mathscr{X}$ are actually visited in equilibrium (S. Maliar et al., 2011), known as the ergodic set of states. Using the current estimate of $\hat{f}(X_t)$, along with the state-transition rule of the model to periodically simulate new states ensures that the sampled states over which the neural-network is trained are maximally likely to lie in the ergodic set. This greatly improves the computational feasibility of an otherwise potentially intractable computational problem: the number of grid points required to have a sufficiently dense grid over hundreds of dimensions would be computationally infeasible, regardless of the sampling method used.

Parameter	Min Value	Max Value	Note / Source
β	0.9975	0.9975	1% Annual Real Interest
Π	1.005	1.005	2% Annual Inflation
σ	1	1	Standard Value
η	1	1	Standard Value
ϕ	100	100	Fernández-Villaverde et al. (2023) (rounded)
$\overline{\theta^{\Pi}}$	1.	2.5	Bianchi et al. (2021)
$\underline{\theta^{\Pi}}$	2.5	2.5	Bianchi et al. (2021)
$\overline{\theta^{Y}}$	0.	0.5	Kase et al. (2022) (Extended down to zero)
ϵ	7.67	7.67	Fernández-Villaverde et al. (2023)
D	0.25	0.25	Fernández-Villaverde et al. (2023)
$\underline{B}$	-0.5	-0.1	Kase et al. (2022) (Exclude 0. from range)
$\underline{W}$	0	0.25	Somewhat larger than the range in (Eurostat, 2025)
ρ_s	0.9	0.9	Kase et al. (2022)
ρ_Ψ	0.7	0.7	Kase et al. (2022)
ρ_g	0.7	0.7	Standard Value
σ_s	0	0.08	Kase et al. (2022)
σ_Ψ	0	0.03	Kase et al. (2022)
σ_g	0	0.15	Kase et al. (2022)

2.3.5 Calibration

The calibration used is based on related papers by Fernández-Villaverde et al. (2023) and Kase et al. (2022). I use the technique of the latter to solve the model simultaneously over a range of certain key parameters whose effect I would like to study, specifically σ_s and $\underline{W}$, but leave most other parameters constant because I find that allowing more parameters than necessary to vary can slow down convergence and harm the accuracy of the final solution.

The key parameters of interest in this paper are $\underline{W}$ and the income distribution parameters σ_s and ρ_s. These together determine the level of the minimum wage as a proportion of the median wage (Kaitz ratio). In European countries, the Kaitz ratio ranges approximately from 40% to 65%, and the proportion of workers at the minimum wage ranges approximately from 1% to 15% (Eurostat, 2025). The latter is straightforward to calibrate, as this is exactly what the parameter $\underline{W}$ represents, but the former is somewhat more difficult, and reaching the low end of this range requires a problematically high σ_s of over 0.1, which generates a deflationary spiral due to deflationary bias. Therefore, the baseline calibration with $\sigma_s = 0.04$ and $\underline{W} = 0.1$ generates a relatively high Kaitz ratio of 0.89. The most probable reason for this discrepancy is that in reality the income and productivity distribution is likely significantly more skewed than the standard log-AR(1) distribution which is assumed here. Furthermore, in the one-asset model bonds are the only recourse for risk averse agents, which results in particularly strong downward pressure on real interest rates (deflationary bias). The relationship between these two parameters over the considered range and the Kaitz Ratio in steady-state is shown in Figure 2.2.

The range for the parameter σ_g seems at first glance particularly large, however, due to the quantile function applied to g_t high values are required to produce large but plausible moves in the minimum wage. For example, at the top end of the range (σ_g) a 2 standard-deviation shock

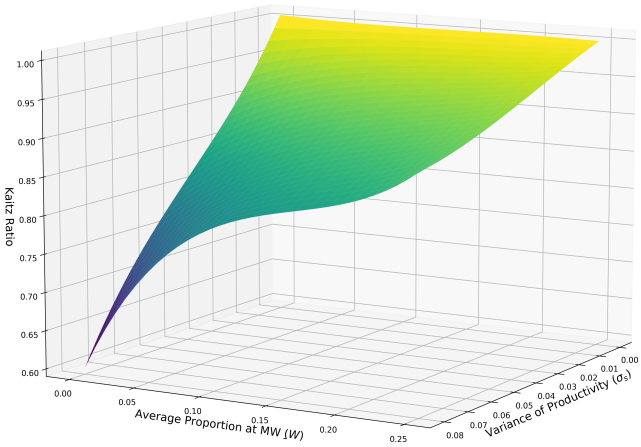


Figure 2.2: Relationship between the ratio of mean wage to minimum wage (Kaitz ratio, vertical axis) and variance of the idiosyncratic productivity process σ_s, and average proportion of workers at the minimum wage $\underline{W}$.

implies an approximately 50% relative increase in the proportion of agents at the minimum wage, so for example from 10% to approximately 15%.

2.4 Methods

In order to solve this model, numerous improvements to the original deep learning solution method of L. Maliar et al. (2021), which have since been suggested by other papers, have been combined and implemented here in a way that to my knowledge no previous paper has done. Since the background and basic implementation of this methodology were already discussed in Hall-Hoffarth (2023), this section will simply outline improvements made relative to that paper, some of which were suggested by other authors in new papers since that paper was written, and some of which are my own innovations. I will discuss these improvements in that order.

Firstly, I incorporate the approach of Han and Yang (2021) and Kahou et al. (2021) by adding a separate neural-network that takes all idiosyncratic-states as an input and outputs a set of moments summarising the distribution. These moments are then used instead of the entire idiosyncratic distribution for the downstream neural-networks that produce aggregate and idiosyncratic policies. Note that these downstream neural-networks are themselves also perfectly capable of recognising and exploiting the symmetry of the problem, so the benefit of this step is primarily computational: since in each step it is necessary to calculate the idiosyncratic policies for every agent independently, implying L forward steps per policy evaluation, reducing the input shape of

this neural-network significantly reduces the overall computational cost, despite the up-front cost of calculating moments.[3] This "moment-generating" (in the spirit of Krusell and Smith (1998)) neural-network then forms part of the computational graph, and it is trained in the same way as the rest of the model.

In order to evaluate the loss of the current model against first-order equations such as the Euler equation which involve expectations over future variables, it is necessary to integrate the error over the distribution of future shocks. In order to obtain an unbiased estimator of this expectation, L. Maliar et al. (2021) propose the "all-on-one" integration operator, whereby the total loss is taken to be the product of the FoC error generated by two independently drawn (sets of) shocks. However, this estimator is very noisy (high variance) due to the low number of draws. In order to address this, I implement the Bias-Corrected Monte-Carlo (BCMC) expectations integration proposed by Pascal (2024) to help obtain the most precise results possible. This approach consists of calculating the FoC error for $n = 128$ independently drawn sets of shocks, and then taking the loss to be the average pairwise product of the $\frac{k(k-1)}{2}$ unique pairs of errors. In particular, this feature will increase the precision of the first-order conditions that are related to inter-temporal optimisation, and therefore, operate through the expectations channel, in this case the Euler equation and NKPC. This is essential, because the expectations channel is the main driver of responses to the primary aggregate shock Ψ_t and of behaviour at the ZLB, since households act under the expectation of higher inflation than in the no ZLB counterfactual.

In order to deal with the constraints and market clearing conditions of the model I apply the technique suggested by Azinovic and Žemličkа (2023) and Hall-Hoffarth (2023). In this approach, I rescale the outputs of the idiosyncratic policy neural-network in order to enforce the budget constraint of the household and the aggregate market clearing conditions simultaneously and by construction. This has the effect of greatly reducing the search space of the neural-network by preventing the simulated states from following a divergent path where constraints are not fulfilled, leading to faster convergence, and higher precision (in terms of loss).

In order to improve the speed of convergence and ultimately the precision of the results obtained a residual learning architecture is implemented in the neural-network (He et al., 2016). Specifically, in the forward step the output of each hidden layer is accumulated and added to the final output layer. This results in the gradient of each layer being available directly to the optimiser rather than being nested below the gradients of other layers due to the chain rule. This allows for the parameters of the earlier layers in the neural-network to be learned more quickly. The neural-network fitting moments to idiosyncratic states has a depth of 3 and 128 nodes in each hidden layer, whereas both the aggregate and idiosyncratic policy functions have 5 layers with 256 nodes in each hidden layer.

This paper also contains a few methodological contributions of its own. One broad takeaway

[3]Another approach would be to have the model output the idiosyncratic policies of all agents at the same time, in a single step. However, in my experimentation, this approach was much less accurate

from Hall-Hoffarth (2023) was that the performance of the neural-network is improved when as many restrictions as possible are placed on the output. One key restriction that can be added in this case is that the representative-agent steady-state of the model, which is obtained by setting the standard deviation of all shocks — including the idiosyncratic labour productivity shock — to zero, is known. It is possible to enforce that the model's solution passes through this point by parametrising deviations from the steady-state as an increasing function of the standard deviation of each shock, that passes through zero. For example, inflation is parametrised as:

$$\Pi_t = \Pi \left(1 + \left(\sigma_i + \sigma_a\right) \phi_2(\cdot)\right) \tag{2.20}$$

Where σ_i is total idiosyncratic volatility, and σ_a is total aggregate volatility. This has the effect of "tethering" the outputs, so that they cannot drift too far from reasonable values (especially early in training), and restricting the problem further, such that the neural-network is now only expected to learn how much shocks cause policies to deviate from the (deterministic) steady-state.

In addition to the states and structural parameters, I also found that adding some important moments that can be derived from the states to the set of neural-network inputs also helped improve the speed of convergence. These moments included:

$$m_t^0 = \mathbf{1}\left\{R_{t-1} \leq 1\right\} \text{ (At ZLB)} \tag{2.21}$$

$$m_t^1 = \mathbf{1}\left\{R_{t-1} \leq 1\right\} \times \mathbf{1}\left\{zlb = 1\right\} \text{ (At ZLB and ZLB is enforced)} \tag{2.22}$$

$$m_t^2 = \frac{1}{L} \sum_{i=1}^{L} \frac{p_t^i}{s_t^i} \text{ (Average wage distortion)} \tag{2.23}$$

$$m_t^3 = \frac{1}{L} \sum_{i=1}^{L} \mathbf{1}\left\{p_t^i > s_t^i\right\} \text{ (Proportion of agents at min wage)} \tag{2.24}$$

$$m_t^{4,i} = p_t^i \text{ ("paid productivity" of each agent)} \tag{2.25}$$

It is helpful in many deep-learning problems to insert additional inputs in this way because on the one hand the additional dimensions and collinearity induced have a low cost, and on the other hand it is sometimes possible to get a large boost by doing some of the learning for the model in advance of training. For example m_t^2 can be calculated using the inputs alone ($\underline{W}$, s_t^i, and g_t), so in theory it can also be represented internally by the neural-network. However, the involved variables are quite noisy and this representation may take a large amount of training to learn. However, once this transformation of the inputs is known, its application is very clear: it makes it straightforward to calculate $MC_t = \hat{W}_t m_t^2$, which again, without knowledge of m_t^2 would be a very noisy object.

Finally, while most structural parameters are sampled uniformly, I apply a Beta distribution to the sampling of shock variances. Although it is useful to be able to turn the shocks off and check

that the model is actually at the Deterministic Steady State (DSS), when shocks are small, so are the deviations from the steady-state and therefore, there is relatively little for the model to learn. The more interesting but also more difficult cases are those where the shocks are relatively large. Since learning the solution for larger shocks will likely require more training, they are oversampled in order to ensure a precise solution across all parameter combinations.

The model and solution method were implemented in PyTorch.

2.4.1 Solution Accuracy

The accuracy of the solution obtained is measured primarily through the loss function, which in turn represents the relative error in the first-order conditions of the model. In this paper and other similar papers using a machine-learning solution method it tends to be possible to get errors of the magnitude of $1e{-}5$ or even $1e{-}6$. This is objectively quite low, but it can also be helpful to visualise this error in order to gain some intuition about how accurate the solution is. In particular, this can be done by asking how close the predicted outputs are to the optimal implied value, obtained by rearranging the FoC to solve for each policy and plugging all of the other relevant policy functions and states into the relevant model equations. Considering the relationship between these optimal and predicted policies makes it possible to demonstrate how small these FoC errors are relative to the extent to which the respective policies vary in equilibrium. This is straightforward for model equations that only involve time-t variables. For example, given the labour supply equation (2.8) plug in any given w_t^i and ν_t^i to generate an optimal $h_t^{i\star}$, and then compare this to the predicted h_t^i from the neural-network. For forward looking equations, in this case the Euler equation and NKPC, it is necessary to evaluate expectations, which can be done by simulating and aggregating over a large number of draws of shocks (n=512). In this way values for ${\Pi_t}^\star$ and $c_t^{i\star}$ are obtained. Results of this are shown in Figure 2.3, which visualises the accuracy of the obtained solution in a uniform sample over all parameters and states. For each plot the R^2 is also shown, which in each case is over 99%. In particular, note that due to the lack of linearisation, the solution manages to stay accurate in the extreme outlier cases, which are likely caused by large shocks or initial states far from the steady-state of the model.

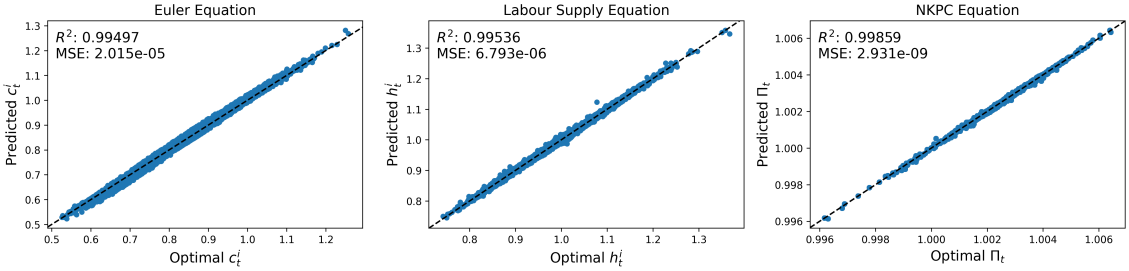


Figure 2.3: Actual versus optimal policies implied by model first-order conditions evaluated over 1024 states, each with structural parameters drawn randomly from their respective ranges. Expectations (for the Euler equation and NKPC) are evaluated over 512 draws of shocks for each future state.

2.5 Results and Discussion

2.5.1 Channel Decomposition

Combining the labour supply equation (2.8), wage equation (2.11), and the production function (2.9) implies:

$$Y_t = \kappa \underbrace{W_t^{\frac{1}{\eta}}}_{\text{Distortionary Effect}} \underbrace{\frac{1}{L}\sum_{i=1}^{L} s_t^i p_t^{i\,\frac{1}{\eta}} \nu_t^{i\,-\frac{\sigma}{\eta}}}_{RedistributiveEffect} \tag{2.26}$$

This equation cleanly expresses the previously discussed decomposition into a distortionary and redistributive channel. Analysing this makes it possible to discern the cases in which an increase to the minimum wage is more likely to be expansionary or contractionary. The change in efficient wages W_t is weakly negative, so any positive effect is maximised when the change in this term is minimised. Equations (2.12) and (2.13) show that this happens whenever the firms are unable to pass on these costs and marginal costs increase strongly in response to the minimum wage. To a limited extent, this in turn happens when prices are more sticky (ϕ is higher).

Far more relevant in this model however, is the ability of the firm to counteract the upwards pressure on their marginal cost by reducing the wages of workers who earn above the minimum wage. In the absence of any direct friction on wages, this power is indeed quite strong. Figure 2.4 shows this most clearly. This figure shows how the direct distortion term P_t, wages W_t, and marginal cost MC_t vary as a function of $\underline{W}$ in the steady-state of the model, as a percent relative to their DSS values. The left panel shows that when P_t increases W_t decreases almost as strongly, resulting in only a muted increase in MC_t. The pass-through of P_t into wages can be quantified by $\frac{MC_t}{P_t}$, which is shown in the right panel. Here it is shown that for low levels of $\underline{W}$ almost none of the distortion is absorbed through an increase in marginal costs (or in other words almost all of the distortion is passed through to the workers via lower wages) and even for large distortions only a maximum of about 25% is absorbed. So, due to their lack of any particular market power, workers above the minimum wage can be forced to absorb a quantitatively large proportion of

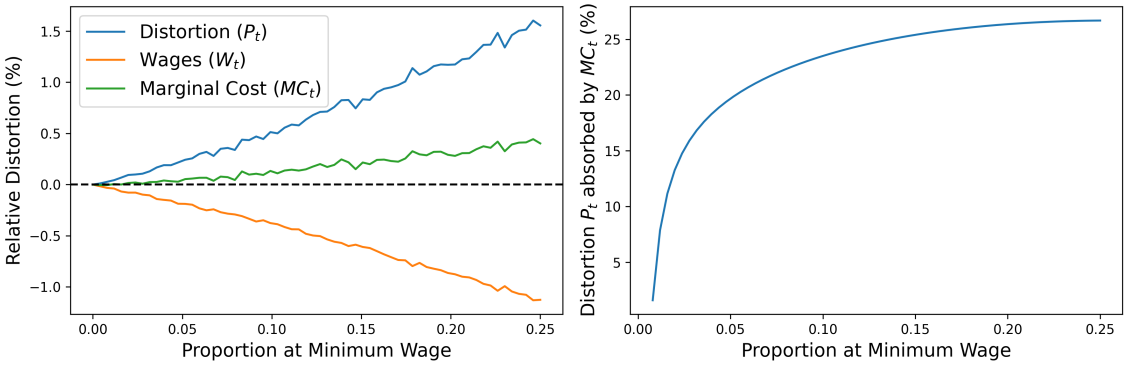


Figure 2.4: Average change in direct distortion P_t, wages W_t, and marginal costs MC_t, as a percentage of their DSS values (left panel), and percent of increase in P_t that is "absorbed" by MC_t ($\frac{MC_t}{P_t}$), both as a function of the proportion of agents at the minimum wage $\underline{W}$, from a large number of samples from the steady-state of the model.

the labour market distortion caused by the minimum wage, by being paid lower wages by firms, whose margins remain relatively unaffected. Indeed, as will be discussed later, firms profits may even increase. These relatively more productive workers will respond to these large wage cuts by strongly reducing their labour supply, which is in part why the distortionary channel is so strong in this model.

The second term in equation (2.26) (inside the summation) captures the distributional impact of a shock on output. An increase to the minimum wage is a (weakly) positive shock to p_t^i by construction. More precisely, an increase in the minimum wage changes this term either for households already at the minimum wage or for the marginal households who were previously above the lower minimum wage, but for whom the new minimum wage now binds. It has no impact on the other households, whose earnings are still above the new minimum wage. The effect is larger when either the productivity s_t^i of the affected households is higher, which is the case when the base level of the minimum wage $\underline{W}$ is higher, or when they have a larger wealth effect through ν_t^i, which is the case when the affected households are poorer and thus have lower consumption. The first component implies somewhat counterfactually that minimum wage increases are more effective when starting from a high level, however, the results will show that it is insufficient to generate an overall positive effect for high levels of the minimum wage. The second component relates to the wealth effect. Since agents at the minimum wage by definition have low(er) productivity, and this productivity is highly autocorrelated, these same agents are likely to have low wealth and consumption. This correlation is depicted for the cross section of agents across many states sampled from the steady-state of the model in Figure 2.19 in the Appendix. Therefore, in particular, this channel is stronger when σ is higher or η is lower, and agents respond more strongly to changes in their current income. Concretely, this component suggests that there might be more positive effects from increasing the minimum wage from a low starting level, where the agents affected experience a large wealth effect. This mirrors Antonova (2018), who finds that the effect of the

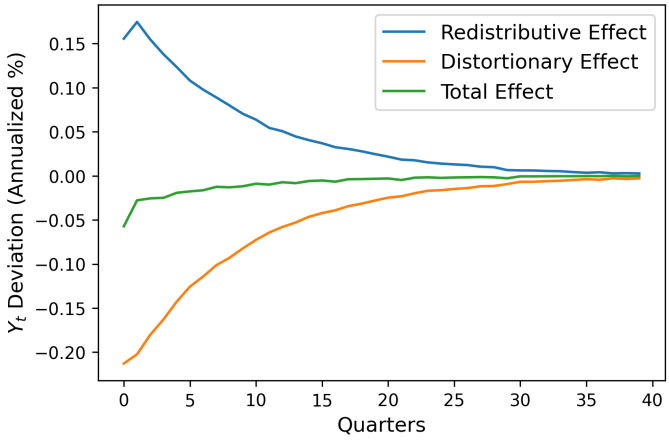


Figure 2.5: Example of the breakdown of the impulse response of output to an increase in the minimum wage for a single set of parameters, in particular, $\sigma_s = 0.04$ and $\underline{W} = 0.1$, into the distortionary and redistributive channels.

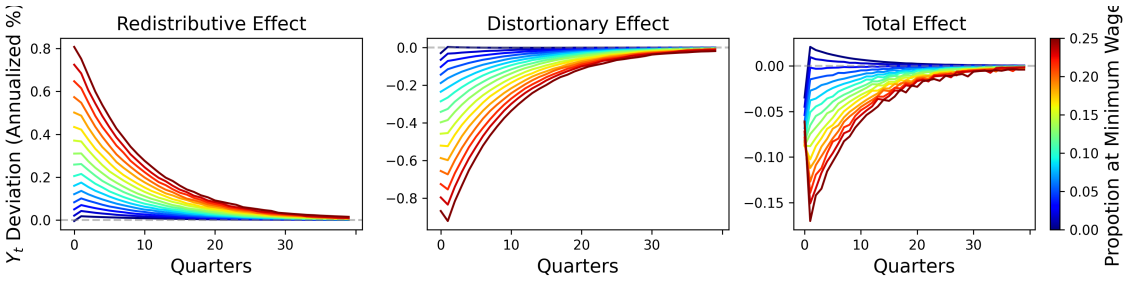


Figure 2.6: Impulse response of output to a temporary increase in the minimum wage, for various levels of the initial minimum wage $\underline{W}$.

minimum wage is larger when more agents are hand-to-mouth and thus have a high propensity to consume out of current income.

2.5.2 Impulse Responses

It is possible to quantify the effect of each of these channels by simulating from the model. As shown in equation (2.15), the breakdown can be generated by comparing policies for the same state, changing only the value of ι. Furthermore, there are multiple ways to embody and express these channels, either with impulse responses or by observing how the Stochastic Steady State (SSS) changes as a function of parameters. Figure 2.5 shows an example impulse response of output to an increase in the minimum wage, broken down into its constituent channels. In this particular example, the overall effect is mildly negative, although the *redistributive* and *distortionary* effects are quite large in magnitude relative to the overall effect.

Figure 2.6 contains three impulse responses to a two standard deviation shock to ϵ_t^g, for different levels of $\underline{W}$, and thus different initial proportions of agents at the minimum wage.[4] The leftmost

[4]Note that unlike in Hall-Hoffarth (2023), these are impulse responses from the steady-state rather than GIRFs. Aggregate shocks other than the minimum wage shock in question are therefore turned off, but

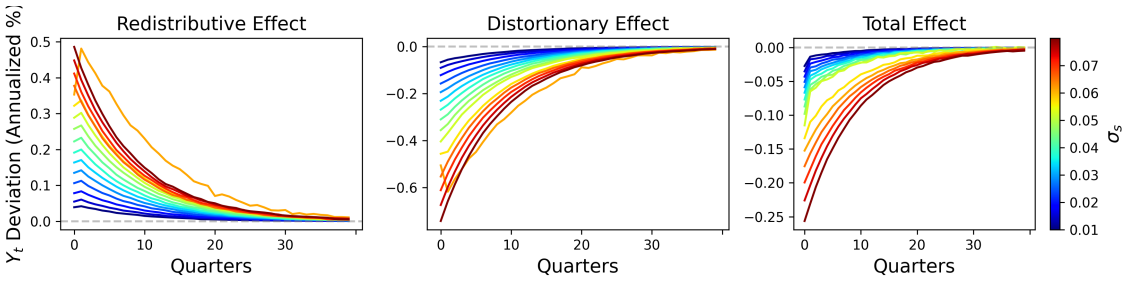


Figure 2.7: Impulse response of output to a temporary increase in the minimum wage, for various levels of the variance of the idiosyncratic productivity distribution σ_s.

diagram shows the redistributive effect of the shock, the middle diagram the distortionary effect, and the rightmost diagram the total effect on output. The contraction grows as the initial level of the minimum wage increases, because even though the redistributive effect becomes significantly stronger, the distortionary effect grows even more. So in this sense the model matches the intuition that any benefits of increasing the minimum wage are likely to come when the minimum wage starts at a low level, as this is when distortions are likely to be relatively small. However, for the calibrations considered here, the effect is almost always negative, sometimes strongly so, and certainly in the range of realistic values for the minimum wage. Similarly, Figure 2.7 shows this decomposition over a range of values for σ_s. This shows again that, all else equal, the distortion caused increases strongly as dispersion in productivity increases, resulting in larger contractions in response to an increase in the minimum wage. Another way of summarising this information is displayed in Figure 2.20 in the appendix, which shows how the initial impact of a minimum wage increase varies as a function of σ_s.

Figures 2.8 and 2.9 show impulse responses for all policies for a low, medium, and high value of $\underline{W}$ and σ_s, respectively. In general, the pattern is that of a contractionary shock, causing inflation, wages, and consumption to fall, and profits to rise, except in the case where the minimum wage starts at a high level, where further increases actually cause an increase in inflation, a fall in output, and a fall in profits.

However, as mentioned, it is also possible to consider how the minimum wage affects the steady-state of the model.[5] Figure 2.10 shows how steady-state output varies as a function of $\underline{W}$ and σ_s. Note how when the minimum wage is turned off ($\underline{W} = 0$), output is an increasing function of σ_s. This happens because the increasingly strong precautionary savings motive of households causes them to increase labour supply above the DSS level. However, for higher levels of σ_s, output is a *decreasing* function of $\underline{W}$. This is because, as previously discussed, the distortionary channel

many states with different draws of $\epsilon_t^{i,s}$ are still averaged over.

[5]Here it is important to disambiguate between multiple concepts of the steady-state. There is the deterministic steady-state where all exogenous shocks are turned off, the stochastic steady-state which is the long-run mean of aggregate variables under all shocks, and the stationary equilibrium, which is the long-run mean of aggregate variables, when only idiosyncratic shocks are turned on. Here I refer to this third type of steady-state.

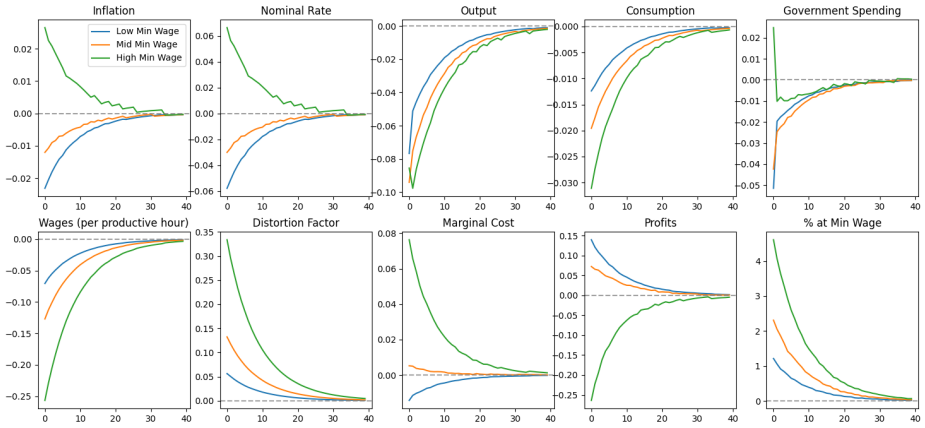


Figure 2.8: Impulse responses for various levels of $\underline{W}$. Calibration used: $\sigma_s = 0.04$, and $\sigma_g = 0.01$ Inflation, nominal rate and output are shown in annualised percentage points. All other variables are percentage deviations from their steady-state values, except % at minimum wage which is also in percentage points.

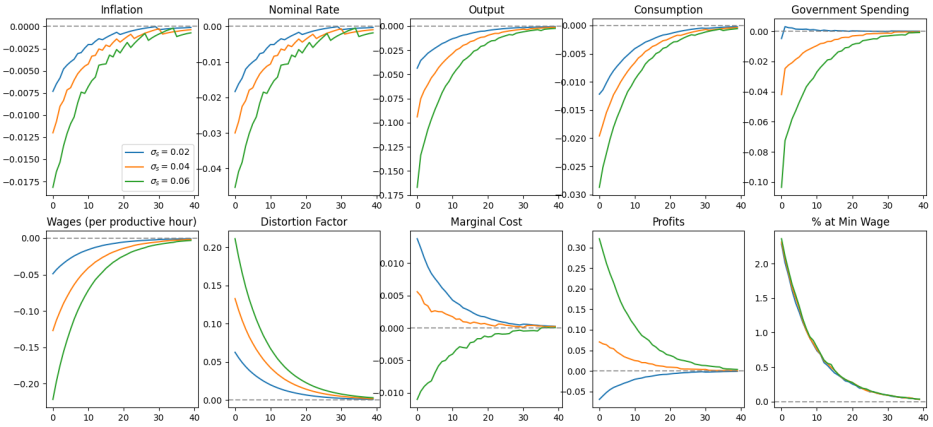


Figure 2.9: Impulse responses for various levels of σ_s. Calibration used: $\underline{W} = 0.1$, and $\sigma_g = 0.01$ Inflation, nominal rate and output are shown in annualised percentage points. All other variables are percentage deviations from their steady-state values, except % at minimum wage which is also in percentage points.

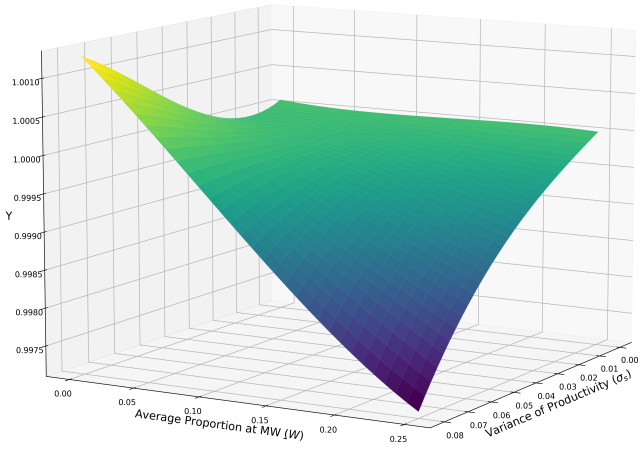


Figure 2.10: Relationship between output (vertical axis) and variance of the idiosyncratic productivity process σ_s, and average proportion of workers at the minimum wage $\underline{W}$, in the steady-state of the model.

dominates the redistributive channel. For combinations of high $\underline{W}$ and σ_s, which is the case where the minimum wage has the most benefit for those directly affected, output is even below the DSS level, as a result of the strong distortions. Thus, overall higher minimum wages decrease equilibrium output in this model.

2.5.3 Interaction with the ZLB

The non-linear global solution method also allows for the computation of state-contingent impulse responses. One important case is when the economy is in a recession that pushes the nominal interest rate down to the ZLB. A demand-side recession in the model is induced by applying a sufficiently large negative demand shock (ϵ_t^{Ψ}). The impulse response to this shock is shown in Figure 2.18 in the Appendix. Three counterfactuals can then be compared: the minimum wage shock occurs in the steady-state, the minimum wage shock occurs during a recession (but there is no ZLB), compared to the counterfactual where the minimum wage is not increased, and finally, the minimum wage shock occurs at the ZLB, compared to the counterfactual where the minimum wage is not increased.

In this case, one might expect to see that with wages already depressed, the nominal interest rate pinned at the lower bound, and any inflationary side-effects actually welcomed, that there would be a greater possibility for the minimum wage to be expansionary. However, as shown in Figure 2.11 the opposite is actually the case in this model. For this calibration, the shock, which

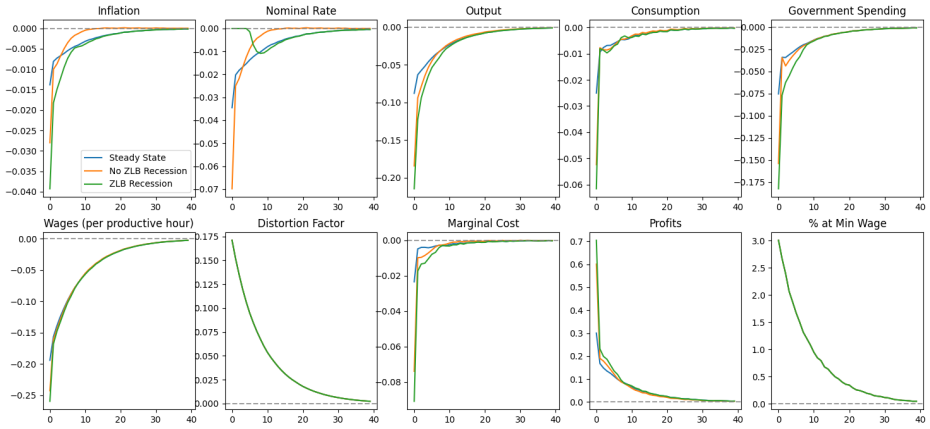


Figure 2.11: Impulse responses to a 2-standard deviation shock to the minimum wage ϵ_t^g in the steady-state of the model, and in a demand-led recession where the ZLB binds, and where it does not bind (i.e. nominal interest rates are allowed to be negative). Calibration used: $\sigma_s = 0.04$, $\underline{W} = 0.1$ and $\sigma_g = 0.01$ implying 10% of workers at the minimum wage in steady-state, rising initially by 13% after the shock. Inflation, nominal rate and output are shown in annualised percentage points. All other variables are percentage deviations from their steady-state values, except % at minimum wage which is also in percentage points.

causes an initial contraction of output of about 0.1% from the steady-state, is amplified in a ZLB recession to cause a contraction approximately twice as large. Furthermore, the shock worsens the deflationary environment, by exacerbating declines in inflation and wages.

2.5.4 A Counterproductive Policy?

While it may be possible to find scenarios in which the redistributive channel dominates and there is an overall benefit to the minimum wage policy, under all calibrations of the model considered here, and indeed under the calibrations that are most realistic, an increase in the minimum wage either has no effect or is contractionary. Indeed, for high levels, it can be strongly contractionary, as in these cases the distortionary effects intensify dramatically.

In some sense, it may seem counter-intuitive that such a model does not predict large and widespread benefits from the minimum wage. After all, the most obvious drawback of such a policy is the effect that is has on unemployment, which is a feature that is notably missing from this model. However, there are still some major drawbacks to the policy, even in this setting. This outcome is influenced primarily by how the budget constraint interacts with the minimum wage. As discussed in Section 2.5.1, because productivity is highly autocorrelated, agents who are at the minimum wage are mostly also close to the borrowing constraint because, on average, they have had low productivity and income for many sequential periods. As a result, there is a kink in a labour supply function, and these constrained or nearly constrained agents tend to have high labour supply already, relative to that of comparable, unconstrained households with

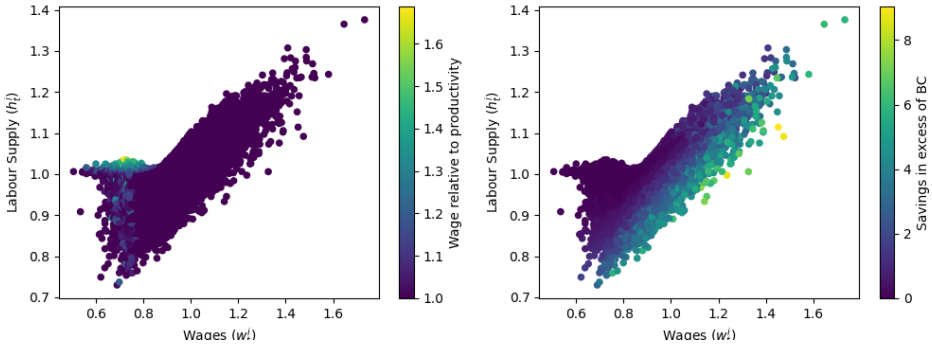


Figure 2.12: Labour supply h_t^i against wages w_t^i, coloured by distance to the budget constraint (right panel) and whether or not the agent is at the minimum wage (left panel), for all 100 agents, in a sample of 1024 states from the steady state of the model.

the same income. This can be seen in Figure 2.12. Therefore, these households are likely to have relatively low own-wage labour elasticity. This intuition will be quantitatively verified in Section 2.5.6. Furthermore, even if these households did increase their labour supply significantly, they are still especially unproductive, which further dampens any potential increase in output.

In general, any overall expansionary effect of an increase in the minimum wage could only occur when the minimum wage starts out affecting a small proportion of households, and when dispersion in productivity is low. This is primarily because when the tails of the distribution are longer the minimum wage causes more distortion, all else equal. While the demand-side benefits grow at higher levels of inequality and of the minimum wage itself, this benefit tends to quickly be outweighed by the larger distortion that it also causes. This also means that any potential expansionary effect would be small.

2.5.5 Effect on Inequality

Aside from any *aggregate* effects of the minimum wage, the policy might also be justified by its *distributional* effects, specifically in reducing inequality. However, as mentioned in the introduction, in this model the effect in this regard is not as unambiguous as it might seem. Indeed, as seen in Figure 2.13, which shows the impulse response of various quantiles of the respective total income and savings distributions to a temporary increase in the minimum wage, the beneficiaries are not only the agents with the lowest incomes, who are also overwhelmingly those at the minimum wage, but also those in the top 10%. On the other hand, the workers disadvantaged by this policy in terms of income are those in the middle of the income distribution. Thus, the policy has a "U-shaped" effect on the income distribution.

There are a couple of potential explanations for this outcome. Firstly, since profits are distributed proportionally to productivity as in Fernández-Villaverde et al. (2023), any increase in profits benefits the most productive and therefore wealthiest agents relatively more. This, in turn, occurs when the distortionary channel dominates and wages fall by more than marginal cost, re-

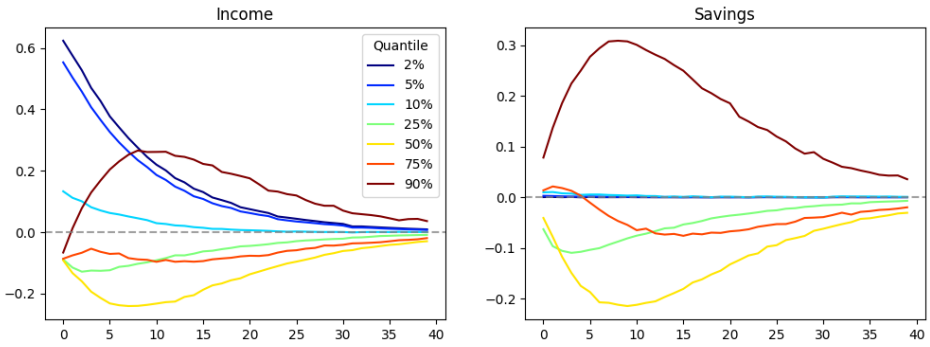


Figure 2.13: Impulse response of various cross-sectional quartiles of total income ω_t^i and savings b_t^i in response to a temporary shock increasing the minimum wage. Units are percentage-points of steady-state values.

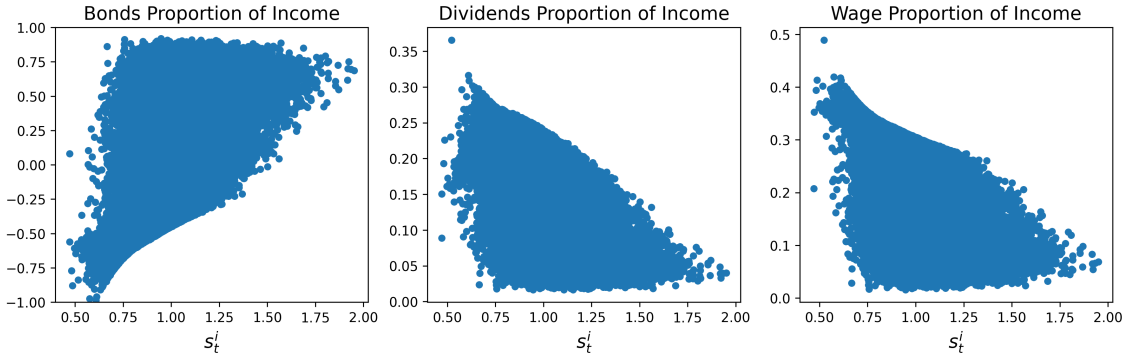


Figure 2.14: Proportion of total income ω_t^i derived from each of its three constituent components (returns on savings $\frac{R_{t-1}}{\Pi_t} b_{t-1}^i$, shares of firm profits $s_t^i Div_t$, and labour income $w_t^i h_t^i$) for all 100 agents across 1024 *iid* draws from the steady-state of the model.

sulting in decreased output and increased firm profits. This has already been shown to be the case, especially when the variance of productivity is high, as shown in Figure 2.9. However, this effect by itself, while consistent in direction with the "U-shaped" effect, is quantitatively small. This can be seen in Figure 2.14, which shows the proportion of total income drawn from each of its three constituent parts: returns on bonds, shares of profit, and labour income, across agents and a large number of states drawn from the models' steady-state. In particular, note that despite being allocated a larger proportion of the firms' profits, wealthier agents only draw a small proportion of their income from this source, and indeed a similarly small proportion of their income from labour. Instead, the overwhelming majority of their income comes from returns on their savings.

Turning back to the quantile impulse responses in Figure 2.13, the main conclusion is that the effect on savings is the main source of the "U-shaped" effect. The agents in the lowest quantiles are at the budget constraint, and thus their savings do not change, however, the savings of those with moderate amounts of savings falls. These agents are also those who tend to be agents with moderate productivity, just above the minimum wage, who still derive a large proportion of their income from labour, and are therefore particularly vulnerable to shocks which adversely impact

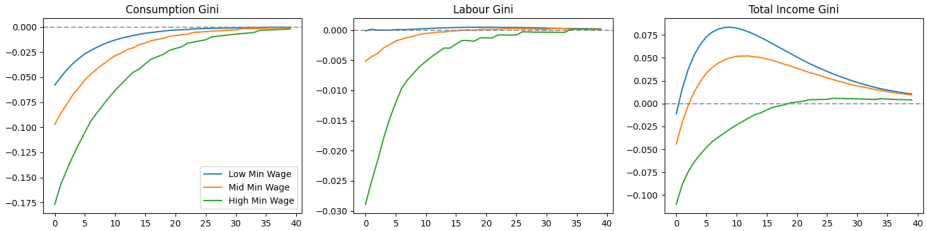


Figure 2.15: Impulse responses of the cross-sectional Gini coefficient of consumption c_t^i, hours h_t^i and total income ω_t^i in response to a temporary shock increasing the minimum wage for 3 values of the average proportion of agents at the minimum wage parameter $\underline{W}$. Units are percent deviations relative to steady-state.

their labour income. These agents sell off their savings in order to smooth over this shock, and in particular to the wealthier agents, who due to deriving only a small portion of their income from labour, are relatively unaffected. This accumulation of assets by the wealthiest agents leads to the hump-shaped increase seen in the income quantiles shown in Figure 2.13.

Another way to quantify the effect on inequality caused by a temporary increase in the minimum wage is to look at the impulse response of the Gini coefficient of idiosyncratic variables. This is shown for consumption, labour, and total income in Figure 2.15. In the right panel, note the hump-shaped increase in income inequality, particularly for lower levels of $\underline{W}$. However, there is an unambiguous decrease in the Gini of consumption and that of labour supply, which is particularly strong when $\underline{W}$ is high. Taken together, these results imply that the inequality effects of a temporary increase in the minimum wage are actually more pronounced when the initial level of the minimum wage is *lower*.

2.5.6 Marginal Responses

Since the neural-network is a differentiable functional mapping from states to polices, it is straightforward to calculate what the derivative of any policy is with respect to any state (including idiosyncratic ones). Therefore, not only is it possible to calculate the aggregate response to a shock, but also how individual agents respond to any given shock. In this way, it is possible, for example, to calculate each agents' MPC. Figure 2.16 shows the derivative of consumption and labour to a minimum wage shock for a large sample of states drawn from the steady-state and all agents. Note here that, as expected, the agents at the minimum wage have a stronger consumption and labour supply response to the minimum wage shock. Another interesting point, which matches the conclusions of the previous subsection is that the labour supply of agents with moderate wages is most negatively impacted, whereas the mean response for agents with the highest wages is closer to zero.

Table 2.1 shows the breakdown of the mean derivative of consumption and labour from Figure 2.16, into four groups depending on whether the agent is at the minimum wage or the budget constraint. What can be seen here in particular is that the agents who are at the minimum wage

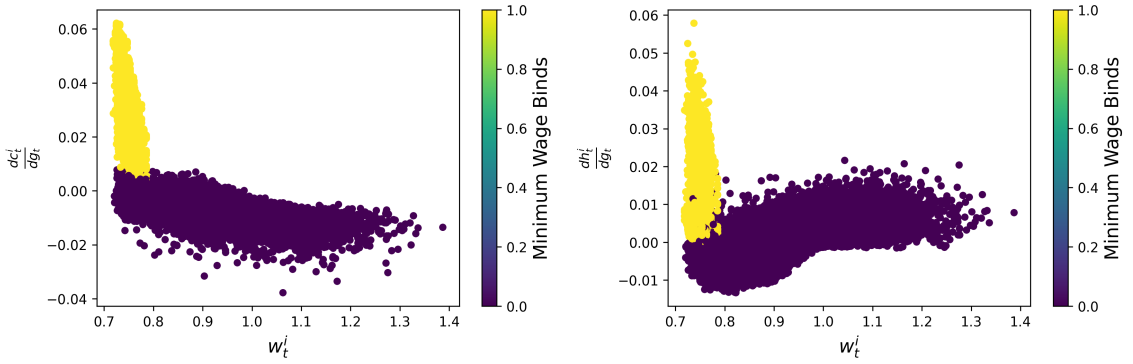


Figure 2.16: Partial derivative of consumption $\frac{\partial c_t^i}{\partial g_t}$ (y-axis, left panel) and labour $\frac{\partial h_t^i}{\partial g_t}$ (y-axis, right panel) to a temporary shock increasing the minimum wage g_t for all 100 agents across 1024 *iid* draws from the stochastic steady-state of the model, plotted against each agents total income ω_t^i (x-axis). Agents already at the minimum wage are highlighted in yellow.

but *not* at the budget constraint have a substantially higher propensity to increase their labour supply in response to an increase in the minimum wage (0.015 vs 0.003). This is quantitative confirmation of the previously discussed intuition that the agents at the budget constraint (who are also more likely to be at the minimum wage) are already working more than they would like to in absence of the budget constraint, and are therefore not likely to increase their labour supply substantially when the minimum wage increases. On the other hand, note that agents who earn the minimum wage respond by increasing their consumption more than their labour supply, and even more strongly so if they are also at the budget constraint. This is the demand-side aspect of the redistributive channel.

Min Wage Binds?	Budget Constraint Binds?	$\frac{\partial c_t^i}{\partial g_t}$	$\frac{\partial h_t^i}{\partial g_t}$	Proportion of Agents
No	No	-0.005	-0.002	0.78
No	Yes	-0.006	-0.003	0.11
Yes	No	0.023	0.015	0.03
Yes	Yes	0.037	0.003	0.08

Table 2.1: Average response of idiosyncratic labour supply and consumption to a positive minimum wage shock, over four groups of agent, defined by whether they earn the minimum wage and whether they are at the budget constraint, from a large number of states ($n = 1024$) in the SSS of the model.

2.6 Future Research

The implications of the model in this paper, as with any model, depend critically on the assumptions used to design the model. In this paper I have presented an initial attempt at considering the issue of how the minimum wage interacts with a HANK model, by making simple assumptions and deviating minimally from the baseline case in order to maintain tractability. Despite this, the model was still able to generate a number of insights. Nevertheless, in this section, I would like

to suggest some extensions to this model that should be considered in future research in order to gain a deeper understanding of the topic, and what their likely implications would be.

The first point is that the model contains no notion of unemployment. This is actually not (yet) a standard feature of HANK models, although recent papers by Consolo and Hänsel (2024) and Payne et al. (2024) make progress in this regard, the latter also employing a machine-learning methodology. This is of course a major drawback for any model considering the minimum wage, as in empirical studies employment effects are a major focus, and this is often named as the most likely drawback of increasing the minimum wage. While more detailed modelling of the labour market would be a welcome improvement to this model, the fact that the minimum wage is likely to cause distortion in the labour market is still captured by the model, and this distortion was indeed shown to be very strong. Therefore, modelling the labour market in more detail is unlikely to change the sign of the overall result, although it may affect the magnitude.

Secondly, the main channel through which this distortion manifests is through a reduction of wages for higher earners. This is at least somewhat counterfactual, as the sum of empirical evidence suggests that the most relevant macroeconomic effect is an increase in inflation (see for example Harasztosi and Lindner (2019)). At face value, it seems unlikely that a CEO's pay would be cut strongly as a result of their least paid employees being paid slightly more. On the other hand, while it may seem much more likely that firms respond to an increased minimum wage by decreasing the amount of low wage workers that they employ, rather than cutting (in real terms) the wages of workers higher up the hierarchy, studies by Dube et al. (2007) and Hirsch et al. (2015) both find that increases in the minimum wage result in no measurable change in employment whereas there is a measurable decrease in wage dispersion within firms. However, these papers only consider localised effects on the restaurant industry, and it is not clear that this generalises to the macro scale. In any case, the implementation of real wage rigidities in the style of Erceg et al. (2000) and Uhlig (2007) would likely dampen the distortionary channel significantly, as wages for productive workers would fall much less in response to a minimum wage hike, and this could cause the aggregate effect thereof to be positive, particularly if the variance of the productivity distribution is low.

Finally, the model also lacks capital used in production. If this were to be included in the model, then an increase in the minimum wage would lead to firms substituting in favour of capital, which, assuming that this is also distributed unevenly, may exacerbate the "U-shaped" effect already identified here in the one-asset case. The addition of this "outlet-valve" for the inefficiency of labour caused by the minimum wage may result in an attenuation of the negative effects on output. This is a major finding of Bauducco and Janiak (2018).

2.7 Conclusion

This paper has extended a basic one-asset HANK model to consider the effects of a minimum wage policy. Using a novel machine-learning solution method, a global and non-linear solution for this model was calculated. Analysis of the model concluded that the policy is counterproductive on both the aggregate and distributional levels. In the aggregate sense higher minimum wages are associated with lower output and higher inflation, and are thus contractionary. In the cross-section the minimum wage fails to uniformly reduce inequality as it burdens middle class households to the benefit of the poorest and wealthiest.

However, these conclusions are dependent on the particular modelling assumptions made here. The primary channel for these adverse effects is a reduction in wages for workers earning above the minimum wage. This could be better represented in the model through the addition of (downwards) price stickiness for higher earning workers, which would force firms to either accept lower margins or to pass on their increased wage costs in the form of higher prices, in a ratio that depends on the market power of the intermediate firms. Furthermore, the addition of capital to the model would allow for substitution of inputs that is not possible in this model, which, while potentially still to the detriment of workers, might reduce the overall negative impact of the minimum wage on output and inflation.

There are also some drawbacks of the methodology employed that should be mentioned, so that future researchers can avoid these pitfalls. Although as shown previously it is possible to obtain a highly accurate and global solution with this method, these can take a long time to finish training. To fit the model used to generate the results in this paper it took almost a week of training on two RTX 3060 GPUs. This training time greatly hinders the ability of the researcher to iterate on new models and variation of models. Therefore, researchers should not believe that machine-learning can be used as a substitute for a deep understanding of the model they aim to solve, it is instead a compliment, that allows for deeper insights to be obtained. Although with infinite computing resources the model could in theory be trained until the loss is arbitrarily small, in practice, due to the neural scaling laws J. Kaplan et al. (2020), improvements decay exponentially in additional training time, and the ultimately obtained loss will be non-zero. As a result, and similarly to the previous point, researchers must apply some degree of pragmatism and ask if the solution obtained matches sensible economic intuitions about the model in question. Tests like the one shown in Section 2.4.1 can also help quantitatively diagnose whether the deep-learning model is sufficiently well trained. Researchers should be aware that insufficiently training the model can lead to misleading conclusions.

Bibliography

Aaronson, D., French, E., & MacDonald, J. (2008). The minimum wage, restaurant prices, and labor market structure. *Journal of Human Resources*, *43*(3), 688–720.

Alege, P., Oye, Q., Ogundipe, A., & Adu, O. (2021). Macroeconomic effect of minimum wage increase in nigeria: A dsge approach. *Nigerian Journal of Economic and Social Studies*, *63*(2), 271–299.

Antonova, A. (2018). Macroeconomic effects of minimum wage increases in an economy with wage underreporting. *Visnyk of the National Bank of Ukraine*, (246), 10–33.

Autor, D. H., Manning, A., & Smith, C. L. (2016). The contribution of the minimum wage to us wage inequality over three decades: A reassessment. *American Economic Journal: Applied Economics*, *8*(1), 58–99.

Azinovic, M., Gaegauf, L., & Scheidegger, S. (2022). Deep equilibrium nets. *International Economic Review*.

Azinovic, M., & Žemlička, J. (2023). Economics-inspired neural networks with stabilizing homotopies. https://arxiv.org/abs/2303.14802

Bauducco, S., & Janiak, A. (2018). The macroeconomic consequences of raising the minimum wage: Capital accumulation, employment and the wage distribution. *European Economic Review*, *101*, 57–76.

Bayer, C., Born, B., & Luetticke, R. (2023). The liquidity channel of fiscal policy. *Journal of Monetary Economics*, *134*, 86–117.

Bianchi, F., Melosi, L., & Rottner, M. (2021). Hitting the elusive inflation target. *Journal of Monetary Economics*, *124*, 107–122.

Braun, H., Döhrn, R., Krause, M., Micheli, M., & Schmidt, T. (2020). Macroeconomic long-run effects of the german minimum wage when labor markets are frictional. *Jahrbücher für Nationalökonomie und Statistik*, *240*(2-3), 351–386.

Caliendo, M., Wittbrodt, L., & Schröder, C. (2019). The causal effects of the minimum wage introduction in germany–an overview. *German Economic Review*, *20*(3), 257–292.

Card, D., & Krueger, A. B. (2000). Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania: Reply. *American Economic Review*, *90*(5), 1397–1420.

Clemens, J., & Strain, M. R. (2020). *Minimum wage analysis using a pre-committed research design: Evidence through 2018* (tech. rep.). IZA Discussion Papers.

Consolo, A., & Hänsel, M. (2024). Hank faces unemployment. *Available at SSRN 4902309*.

Drechsel-Grau, M. (2023). *Employment and reallocation effects of higher minimum wages* (tech. rep.). CESifo Working Paper.

Dube, A., & Lindner, A. (2024). Minimum wages in the 21st century. *Handbook of Labor Economics*, *5*, 261–383.

Dube, A., Naidu, S., & Reich, M. (2007). The economic effects of a citywide minimum wage. *ILR Review*, *60*(4), 522–543.

Dube, A., & Zipperer, B. (2024). *Own-wage elasticity: Quantifying the impact of minimum wages on employment* (tech. rep.). National Bureau of Economic Research.

Erceg, C. J., Henderson, D. W., & Levin, A. T. (2000). Optimal monetary policy with staggered wage and price contracts. *Journal of monetary Economics*, *46*(2), 281–313.

Eurostat. (2025). Minimum wage statistics. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Minimum_wage_statistics#Minimum_wage_levels_in_relation_to_median_gross_earnings

Fernández-Villaverde, J., Marbet, J., Nuño, G., & Rachedi, O. (2023). *Inequality and the zero lower bound* (tech. rep.). National Bureau of Economic Research.

Hall-Hoffarth, E. (2023). Non-linear approximations of dsge models with neural-networks and hard-constraints. *arXiv preprint arXiv:2310.13436*.

Han, J., & Yang, Y. (2021). Deepham: A global solution method for heterogeneous agent models with aggregate shocks. *arXiv preprint arXiv:2112.14377*.

Harasztosi, P., & Lindner, A. (2019). Who pays for the minimum wage? *American Economic Review*, *109*(8), 2693–2727.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hicks, J. R. (1932). The theory of wages.

Hirsch, B. T., Kaufman, B. E., & Zelenska, T. (2015). Minimum wage channels of adjustment. *Industrial Relations: A Journal of Economy and Society*, *54*(2), 199–239.

Kahou, M. E., Fernández-Villaverde, J., Perla, J., & Sood, A. (2021). *Exploiting symmetry in high-dimensional dynamic programming* (tech. rep.). National Bureau of Economic Research.

Kaplan, G., Moll, B., & Violante, G. L. (2018). Monetary policy according to hank. *American Economic Review*, *108*(3), 697–743.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kase, H., Melosi, L., & Rottner, M. (2022). Estimating nonlinear heterogeneous agents models with neural networks.

Krusell, P., & Smith, A. A., Jr. (1998). Income and wealth heterogeneity in the macroeconomy. *Journal of Political Economy*, *106*(5), 867–896.

Ku, H. (2022). Does minimum wage increase labor productivity? evidence from piece rate workers. *Journal of Labor Economics*, *40*(2), 325–359.

Kwicklis, N. (2025). Active vs. passive policy and the trade-off between output and inflation in hank. *Journal of Monetary Economics*, 103732.

Lee, D. S. (1999). Wage inequality in the united states during the 1980s: Rising dispersion or falling minimum wage? *The quarterly journal of economics*, *114*(3), 977–1023.

Maliar, L., Maliar, S., & Winant, P. (2021). Deep learning for solving dynamic economic models. *Journal of Monetary Economics*, *122*, 76–101.

Maliar, S., Maliar, L., & Judd, K. (2011). Solving the multi-country real business cycle model using ergodic set methods. *Journal of Economic Dynamics and Control*, *35*(2), 207–228.

Mortensen, D. T., & Pissarides, C. A. (1994). Job creation and job destruction in the theory of unemployment. *The review of economic studies*, *61*(3), 397–415.

Neumark, D., & Wascher, W. L. (2008). *Minimum wages*. MIT press.

Pascal, J. (2024). Artificial neural networks to solve dynamic programming problems: A bias-corrected monte carlo operator. *Journal of Economic Dynamics and Control*, *162*, 104853.

Payne, J., Rebei, A., & Yang, Y. (2024). Deep learning for search and matching models.

Riley, R., & Bondibene, C. R. (2017). Raising the standard: Minimum wages and firm productivity. *Labour Economics*, *44*, 27–50.

Rotemberg, J. J. (1982). Sticky prices in the united states. *Journal of political economy*, *90*(6), 1187–1211.

Sabia, J. J. (2015). Minimum wages and gross domestic product. *Contemporary Economic Policy*, *33*(4), 587–605.

Šauer, R. (2018). The macroeconomics of the minimum wage. *Journal of Macroeconomics*, *56*, 89–112.

Uhlig, H. (2007). Explaining asset prices with external habits and wage rigidities in a dsge model. *American Economic Review*, *97*(2), 239–243.

Appendix

A Deflationary Bias

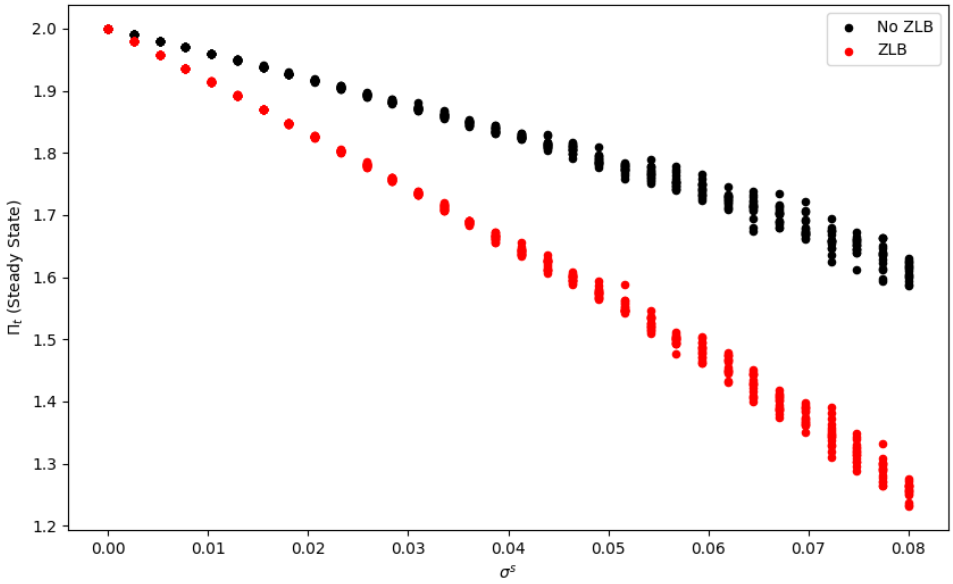


Figure 2.17: Deflationary bias in the SSS caused by increasing σ_s, due to the precautionary savings motive of the households.

B ZLB Episode

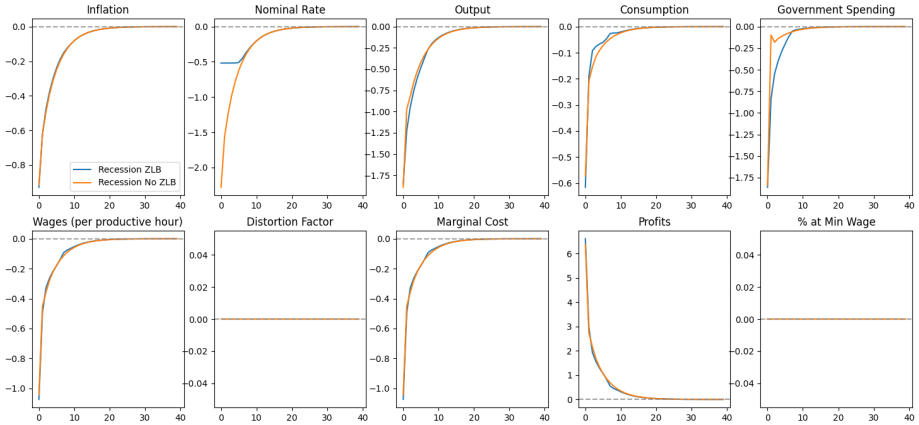


Figure 2.18: Impulse responses to a 3-std demand shock, which brings the economy into a deep recession, as discussed in Section 2.5.3

C Productivity and Wealth Distribution

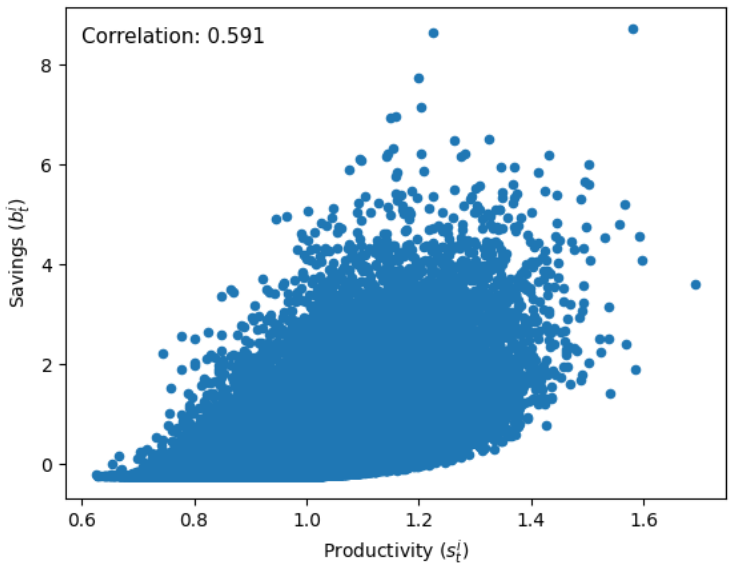


Figure 2.19: Cross-section of savings b_t^i and productivity s_t^i of individual agents from a large number of draws from the steady-state of the model under the default parametrisation.

D Output Response Breakdown

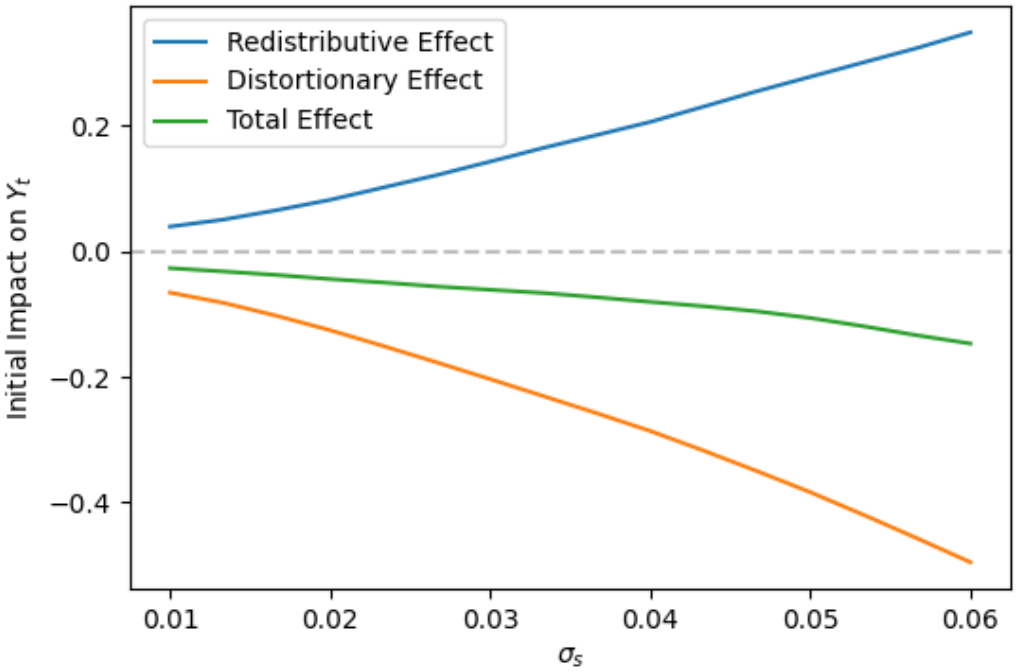


Figure 2.20: Effect on impact of a shock that increases the proportion of agents for whom it binds from approximately 10% to approximately 12% on output Y_t over a range of values for the variance of the productivity distribution σ_s. The total effect is broken down into a supply-side *distortionary* effect resulting from firms reducing wages in response to the shock, and a demand-side *redistributive* effect resulting from higher labour supply and consumption demand from the directly affected households.

Chapter 3

Causal Discovery of Macroeconomic State-Space Models

Abstract

This paper presents a set of tests and an algorithm for agnostic, data-driven selection among macroeconomic Dynamic Stochastic General Equilibrium (DSGE) models inspired by structure learning methods for Directed Acyclical Graphs (DAGs). As the log-linear state-space solution to any DSGE model is also a DAG, it is possible to use associated concepts to identify a unique ground-truth state-space model that is compatible with an underlying Data Generating Process (DGP), based on the conditional independence relationships that are present in that DGP. In order to operationalise search for this ground-truth model, the algorithm tests feasible analogues of these conditional independence criteria against the set of combinatorially possible state-space models over observed variables. This process is consistent in large samples. In small samples, however, the result may not be unique, so conditional independence tests can be combined with likelihood maximisation in order to select a single optimal model. The efficacy of this algorithm is demonstrated for simulated data, and results for real data are also provided and discussed.

> *"... the most important issue holding back the DAGs is the lack of convincing empirical applications. History suggests that those are what is driving the adoption of new methodologies in economics and other social sciences, not the mathematical elegance or rhetoric."*
>
> – Guido Imbens, *Journal of Economic Literature, 2020*

3.1 Introduction

In the machine-learning literature, causal discovery is generally defined as the act of inferring causal relationships from observational data (Huang et al., 2020). This however also exactly describes the goal of much of empirical economic research, and therefore, in this context it is most reasonable to append to this definition that which is taken for granted in machine-learning — that this inference is done *algorithmically*. The field of (algorithmic) causal discovery has benefited from intense development in recent years, however, it is hardly a new discipline. Work along these lines started in the 1980s with early contributions from Judea Pearl, Thomas Verma, and Peter Spirtes, among others. Indeed, there has been considerable work done in the field of economics regarding algorithmic model selection, in particular the general-to-specific model selection of Krolzig and Hendry (2001).

While there are many approaches to causal discovery, this paper focuses on the inference of a DAG, sometimes also (somewhat misleadingly) referred to as a Bayesian Network.[1] These are a type of *graphical model* which can be used to illustrate, and, with the aid of some associated statistical techniques, to infer causal relationships between variables. While the use of these models as a descriptive tool has been fiercely debated (Pearl & Mackenzie, 2018), what is perhaps more exciting for the field of economics is the fact that numerous algorithms exist which, under relatively mild conditions, can identify a DAG, and thus a causal model, directly from observational data. Despite the clear potential of these methods, to my knowledge, relatively little research has been done in the macroeconomic literature about potential applications of causal discovery algorithms. This paper therefore aims to contribute to the literature by evaluating how these methods might be used by considering a particularly well-suited and important application: learning the classification of states among macroeconomic variables, assuming that these can be explained by a DSGE model.

DSGE models such as the *Real Business Cycle* (RBC) model first popularised by Kydland and Prescott (1982), and subsequent *New Keynesian* models were formulated primarily as a response to the *Lucas critique*; that reduced form macroeconomic time-series models such as Vector Auto-Regressions (VARs) are unsuitable for inferring the causal effects of changes to microeconomic or structural parameters or of truly exogenous (uncorrelated) shocks (Lucas et al., 1976). The key feature of DSGE models is that they are based on *micro-foundations* — that is, they explicitly model the optimal behaviour of economic agents in order to derive equilibrium conditions among observed macroeconomic variables. However, these optimisation problems are still subject to assumptions about the nature of constraints faced by agents, the information available to them, and in some cases even their degree of rationality. For example, do agents form expectations in a purely forward-looking fashion, or do they employ some form of indexing to past values? In the relevant literature these assumptions are generally justified either with microeconomic evidence or com-

[1] This is somewhat misleading because Bayesian Networks do not require the application of Bayes rule, although one could choose to estimate the associated parameters in this way.

paring the *impulse response functions* generated by the model to those estimated by econometric models (Christiano et al., 2018).

Different assumptions about micro-foundations or frictions will sometimes, but not always, imply different state-space models. For example, in a basic log-utility DSGE model consumption is a control variable, however, if habits in consumption are assumed as in J. C. Fuhrer (2000) (i.e. if $U(\cdot) = ln(c_t - hc_{t-1})$) it becomes a state variable, such that the past value of consumption becomes relevant in determining the current value of other variables in the model. Other prominent examples include inflation, which is a control variable under rational expectations, but becomes a state variable if it is persistent due to firms that update their prices in a *rule-of-thumb* manner (Christiano et al., 2005; J. Fuhrer & Moore, 1995; Galí & Gertler, 1999), or persistence in monetary policy as in Sack and Wieland (2000). Although it is not the explicit justification for most modelling decisions, getting these classifications right can have significant implications for understanding underlying mechanisms and making policy recommendations. For example, whether or not inflation is *backward-looking* can have significantly different implications for the optimal conduct of monetary policy, as shown in Steinsson (2003).

In these cases, the test and algorithm presented in this paper can be seen as another tool that can be used by empirically-minded researchers to evaluate whether the assumptions made in their proposed model are consistent with the data. This evidence is particularly valuable because it is obtained in a *maximally agnostic* way that makes no assumptions about any of the particular observables (e.g. inflation, interest rate, output), only about the nature of the relationships between them.[2] In other words, the algorithm regards any observables as ex-ante equally likely to be either state variables or controls, so any conclusions drawn in this way solely reflect the data to the greatest extent possible. Furthermore, as shown in Section 3.5.4.1, selecting states by using the conditional independence tests introduced in this paper is much more reliable than picking the states that have the most predictive power (i.e. maximise the (posterior) likelihood). This latter approach is the basis of most standard model-comparison techniques such as marginal likelihood maximisation and Bayesian model averaging (Geweke & Amisano, 2012). This is because in this context the goal is to discover the causal structure of the data (in a particular sense that will be defined more precisely later), and the model with the most predictive power is not necessarily the model with the most credible causal structure. However, despite the advantages of the method presented in this paper, what this paper does not do is present a solution to the problem of *micro-economic dissonance* (Levin et al., 2008). In cases where the linear state-space model implied by DSGE models are equivalent, this procedure cannot determine which set of micro-foundations are more reasonable.

In order to test the ability of various algorithms in practice I generate random observations from well-known DSGE models and then test the ability of various algorithms to identify the ground-

[2] I assume linearity and Gaussian shocks for simplicity, although in principle these too could be relaxed.

truth, which in this context is known. Despite considerable promise, and theoretical guarantees of asymptotic consistency, in these simulation experiments existing structure learning algorithms for DAGs performed poorly at identifying the correct ground-truth state-space model. This is likely due to the fact that these algorithms search over the set of all possible DAGs, of which those that are also state-space models are only a small subset. This is compounded by the fact that in macroeconomics sample sizes available are usually small relative to the number of observables. Conversely, the algorithm I propose explicitly assumes that the solution is within the subset of DAGs which are also state-space models. It is asymptotically consistent, and thanks to the smaller search space, simulation evidence demonstrates that it is also much more successful than existing structure learning algorithms at identifying the ground-truth state-space model, even given realistic (small) sample sizes.

While there is considerable potential for the application of such a tool in economics, thus far relatively little work in this vein has taken place. Indeed, Imbens (2020) considers the value of DAGs for empirical economics and concludes that the reason this framework has not caught on is precisely because few useful applications have been demonstrated. Notable exceptions include the work of Demiralp and Hoover (2003), who consider structure learning algorithms for DAGs in the context of Structural Vector Autoregressions (SVARs), and Bazinas and Nielsen (2015), who utilise concepts of conditional (in)dependence closely related to those used in DAGs to develop the notion of *causal transmission*. Notwithstanding these, this paper aims to provide a substantive contribution to the literature by presenting an application of DAGs to macroeconomic DSGE models. Specifically, I show that a DSGE model's log-linear state-space solution can be represented as a particular form of DAG. Furthermore, I develop a new algorithm for finding DAGs of this particular form using conditional-independence tests. This algorithm, given its substantially restricted problem space and domain-specific adaptations is able to handily out-perform out-of-the-box causal discovery algorithms such as that used in Demiralp and Hoover (2003) at the task of learning a state-space, and thus a *valid* set of DSGE models from observational data alone.

The remainder of this paper is organised as follows. Section 3.2 covers background information on both DAGs and DSGE models. Section 3.3 introduces the proposed structure learning tests and algorithm. Section 3.4 briefly introduces the simulated and real world data which will be used for empirical validation. Section 3.5 provides and discusses the results of the proposed algorithm, as well as some existing alternatives on those data sets. Section 3.6 includes some closing remarks.

3.2 Literature Review

Before continuing, it is important to clearly define certain concepts and conditions related to causal discovery and DSGE models, which will be used later to derive some of the key results in this paper. This is provided because many of the terms, especially related to the causal discovery literature, seem to be used under various different meanings, and it is therefore essential to clarify which

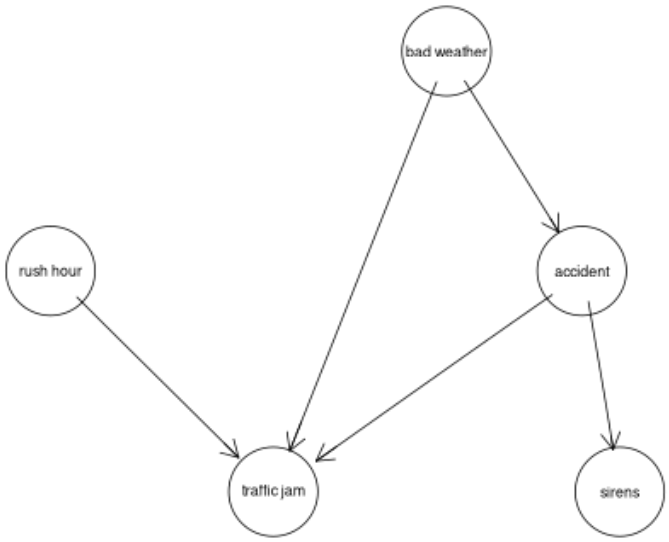


Figure 3.1: A simple example of a DAG (Liszka, 2013)

definition I use to avoid any ambiguity. The most relevant concepts are therefore summarised as succinctly as possible in this section.

3.2.1 DAGs

3.2.1.1 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 $k \times k$ matrix of *edges* or *arcs* (Kalisch & Bühlmann, 2007). $(x, y) \in E$ indicates the presence of a directed edge from node x to node y. As the name DAG 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 $\{p_1, p_2, ..., p_k | p_i \in E\}$ containing a directed path starting and ending at the same node. Figure 3.1 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 it is possible to more specifically define V as a $(k \times 1)$ vector and E as a $k \times k$ adjacency 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 as before, and the acyclic property is now equivalent to the statement that E^n has zeros on its diagonal for $\forall n > 0$. In the spirit of traditional econometric SVARs, the model will now also include a $k \times 1$ vector ϵ containing mutually independent Gaussian shocks, one for each node.

The set of nodes from which an arc into a node x originates in a graph G are known as the *parents* of x ($pa_G(x)$), and the set of nodes that have an incoming arc from x are known as

the *children* of x ($ch_G(x)$) (Pearl, 2009). The set of all nodes from which a directed path into x originates are known as the *ancestors* of x ($ans_G(x)$) and the set of all nodes that have an incoming path from x are known as the *decedents* of x ($des_G(x)$).

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

Definition 1. *Stability Let $f(\mathbf{w};\theta)$ represent some DGP defined over observable variables $\mathbf{w}$ with true parameters θ, and $I(f(\mathbf{w};\theta))$ be all the conditional independence relationships that exist in $f(\mathbf{w};\theta)$ between variables in $\mathbf{w}$. Then f is **stable** if $I(f(\mathbf{w};\theta)) = I(f(\mathbf{w};\theta'))\ \forall\ \theta'$ in a neighbourhood of θ. (Pearl, 2009, p.48)*

Definition 2. *Faithfulness A DAG G is said to be **faithful** to $f(\mathbf{w};\theta)$ if $f(\mathbf{w};\theta)$ is stable and G satisfies $I(G) = I(f(\mathbf{w};\theta))$ (Spirtes et al., 2000, p.31)*

Other than the optional assumption of linearity and Gaussian errors that are made here for simplicity, and the assumed lack of unobserved confounders, *stability* is the primary assumption necessary for the identification of a DAG that represents a true DGP. This is the untestable component of the perhaps more commonly referenced concept of *faithfulness*.[3] Faithfulness as defined here also consists of the (testable) assumption that the graph G captures all the conditional independence relationships in the DGP f.[4] Stability is the assumption that these relationships are invariant to small perturbations in the parameters of the true DGP. Intuitively, if one wishes to use conditional independence relationships to identify a model then it is necessary to assume that the observed conditional independence relationships do not belie the underlying relationships between variables. This assumption is violated only 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 the parameter restriction $\beta_{yx} = -\beta_{yz}\beta_{zx}$ is imposed 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. In this case the ground truth graph cannot be learned from observational data, as the causal dependence that exists between x and y is not captured by the conditional independence relationships in the data generated by this model. It is also necessary to assume that the set of observed variables $\mathbf{w}$ are complete, in the sense that any other variables that are unobserved are orthogonal to the ones that are. While this is a strong assumption, it is also a fundamental one that underlies practically any model of a dynamic system. In any case, under these assumptions it is possible to use conditional independence tests in the following way to evaluate whether a DAG G is consistent with (or more precisely *faithful* to) f.

[3]Note that this definition of faithfulness includes an equivalence relationship and therefore encompasses what is sometimes referred to separately as the *Causal Markov Condition* which states that $I(G) \subseteq I(f(\mathbf{w};\theta))$ (Spirtes & Zhang, 2016).

[4]What it means for variables to be (conditionally) independent in a graph will be covered very shortly.

Definition 3. *D-Separation A path P starting at node x and ending at node y in a DAG G is said to be* ***d-separated*** *or* ***blocked*** *by a set of variables* $\mathbf{z}$ *if and only if the following two conditions hold:*

1. If P contains a chain $x \to m \to y$ or fork $x \leftarrow m \to y$ then $m \in \mathbf{z}$ ***and***

2. If P contains a collider $x \to m \leftarrow y$ then $\{m \cup des(m)\} \cap \mathbf{z} = \emptyset$

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)*

D-separation is sometimes also referred to in terms of *backdoor* and *frontdoor* paths (Pearl, 2009). A backdoor path is a path that links two nodes going back against the direction of an arrow from at least one node. Conversely, a frontdoor path is a path that travels down in the direction of the arrow from every node. Then a set of variables $\mathbf{z}$ can equivalently be said to D-separate nodes x and y if and only if $\mathbf{z}$ blocks all the backdoor paths from x to y (1.) and none of the frontdoor paths (2.). D-separation is closely related to the econometric concepts of *orthogonality* and *exogeneity*. In OLS consistent estimates can be obtained only if the residuals are uncorrelated with regressors. This in turn occurs when there are no unobserved confounders, and thus no omitted variable bias (in DAG language, there are no unblocked backdoor paths), and there are no *bad controls* (Angrist & Pischke, 2014) (in DAG language, there are no blocked frontdoor paths).

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 $f(\mathbf{w};\theta)$ of x and y, then x and y are independent conditional on* $\mathbf{z}$*. (Pearl, 2009, p.18)*

Corollary 1. *Test of Faithfulness If x and y are d-separated in G by* $\mathbf{z}$ *but x and y are not independent conditional on* $\mathbf{z}$ *in the true DGP $f(\mathbf{w};\theta)$, and $f(\mathbf{w};\theta)$ is stable then G is not faithful to $f(\mathbf{w};\theta)$.*

The corollary is simply the negation of Theorem 1, and it shows how the faithfulness of some DAG G is falsifiable as long as f is known to be *stable*. This result is essential for defining the constraint-based tests in Section 3.3.2. In particular, it implies the following result that will be leveraged:

Definition 4. *Parental Markov Condition Given some DAG G, a node x in G is d-separated from and therefore independent of all its non-decedents by its parents. (Pearl, 2009, p.16, p.19)*

Corollary 2. *If G is faithful to the DGP f over a set of observable variables* $\mathbf{w}$ *then f admits the following factorisation:*

$$f(\mathbf{w};\theta) = \prod_{i=1}^{k} f(w_i | pa_G(w_i);\theta) \tag{3.1}$$

(Pearl, 2009, p.16)

This means that if a DAG G is faithful for f, then the correct conditioning set that assures

Figure 3.2: A DAG before structure learning

orthogonality in a regression for each variable is the set of its parents in G. If a variable has no parents then it is exogenous relative to the system.

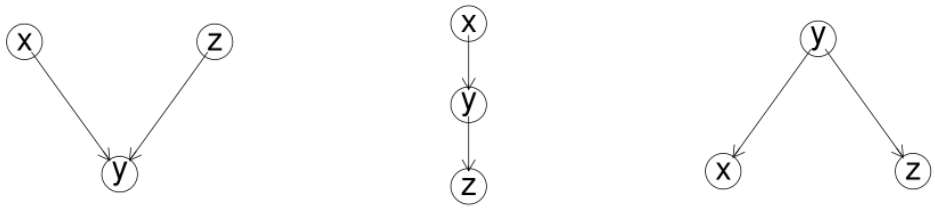
3.2.1.2 Estimation

There are two fundamental problems to solve when estimating a DAG. The first is known as *parameter learning*, and the other *structure learning* (Ermon, 2017). Given a DAG as in Figure 3.2, the first task is simply to estimate the parameters of the network, such as the parameter matrices **A**, **B**, **C**, **D**, and **E** in equations (3.2) - (3.4) in Section 3.2.2. This is usually done via maximum likelihood or perhaps with Bayesian techniques.

The second and more onerous task, as demonstrated by Figure 3.2 is that, starting with some observational data and no model, it is not obvious which edges between nodes need to be estimated in the first place. One way to do this is for the researcher to specify explicitly which edges should be present in the graph, and simply fit the parameters of that graph. As discussed in Section 3.3, this is straightforward to do for DSGE models, assuming the true state variables are known. However, doing so in this context would achieve little. This is equivalent to specifying a system of linear equations (VAR with some parameters restricted to zero) to be estimated, probably based on some economic model that was developed by other means. While this is then automatically encapsulated in a convenient, easily interpreted representation of the underlying assumptions, this approach does not offer anything particularly novel.

Instead, a more promising approach is to algorithmically learn the structure of the graph, that is to learn a causal model, directly from observed data. One *brute force* method to solve this problem is to compute the posterior likelihood of every possible network. However, this number is super-exponential in the number of variables and, therefore it becomes very computationally expensive, very quickly (Chickering, 1996). As a response to this, many heuristic approximation techniques have been developed. These can be broadly grouped into two categories: constraint-based and score-based structure learning algorithms (Spirtes & Glymour, 1991) (Verma & Pearl, 1991), which I will now briefly discuss in that order.

Constraint-based algorithms rely on the fact that changing the direction of an arc changes the conditional independence relationships 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 3.3. Consider a graph with three nodes, such that no one node is completely independent of the other two (as this would make the graph trivial, and this case



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

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

could be ruled out with an (unconditional) 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 3.3. These are known as 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 it is easy to test for the presence of conditional and unconditional independence under the assumption of joint-normality using a t-test or F-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. Since this only separates one case from the other two, for every set of three variables the network is only partially identifiable, however, full identification can (but will not always) be achieved when more variables are observed. This is done 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 independence relationships. There are many algorithms that have been implemented using this general approach, the most popular of which is the PC algorithm first developed by Spirtes and Glymour (1991). 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 and Gaussian conditional probability functions, stability, lack of unobserved confounders, and structural complexity that does not grow too quickly relative to n (Kalisch & Bühlmann, 2007).

Score-based[5] methods assign some score to every network based on its predictive accuracy (usu-

[5]Here I use the meaning of "score" that is typical in the machine-learning literature — some function to be maximised in order to improve model fit. This should not be confused with the common definition

ally related to the likelihood of the model) and then use (stochastic) gradient-descent[6] to identify the optimal network structure. There are a number of functions and hill climbing algorithms that can be used to achieve this. In the case of continuous data the log-likelihood of the model or some penalised variant is usually used as the score function. A consistency result for the GES score-based algorithm is given in Chickering (2002). The assumptions are slightly stronger 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-based methods, which effectively maximise the predictive accuracy of the model, and there is seemingly no guarantee that the best 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 reasonable to believe that bias due to finite samples or slight deviations from stated assumptions is likely to be more prominent 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) (Spirtes & Glymour, 1991). 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 a set of DAGs that are said to be *observationally equivalent* or *Markov equivalent* (Colombo & Maathuis, 2014) is obtained. This is problematic because it is difficult or impossible to fit parameters to and thereby derive counterfactual implications from 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 in 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 the resulting graph is always directed, it does not always correctly reflect the observed v-structures because it trades off flexibly between constraint satisfaction and score maximisation (instead of giving lexicographic priority to constraint satisfaction, which is the approach that my algorithm will take). Nandy et al. (2018) give an asymptotic consistency result for a particular hybrid learning algorithm called ARGES.

of "score" in the econometrics and statistics literatures, which is the gradient of the likelihood function.

[6] In this case to be precise it is really a gradient-*ascent*.

3.2.2 DSGE Models

Suppose a DSGE model is defined over a set of k variables in a vector $\mathbf{w}_t$, for which one observation is available per time period, for example, quarterly or yearly data. The log-linear approximation to a stationary DSGE model solution can be written as a state-space model (King et al., 1988) that partitions $\mathbf{w}_t$ into three mutually exclusive vectors $\mathbf{x}_t$, $\mathbf{y}_t$, and $\mathbf{z}_t$. This state-space model is defined by equations (3.2) - (3.4):

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}\mathbf{z}_t + u_t^y \tag{3.2}$$

$$\mathbf{x}_t = \mathbf{C}\mathbf{x}_{t-1} + \mathbf{D}\mathbf{z}_t + u_t^x \tag{3.3}$$

$$\mathbf{z}_t = \mathbf{E}\mathbf{z}_{t-1} + \epsilon_t \tag{3.4}$$

Where $\mathbf{x}_t$ is a vector of endogenous state variables, $\mathbf{y}_t$ is a vector of control variables, $\mathbf{z}_t$ is a vector of exogenous state variables, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{E}$ 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 simply not necessary for the type of inference proposed in this paper. Note that while ϵ_t is unobserved, the same cannot be said for $\mathbf{z}_t$. This means, for example, that it is assumed that the Toal Factor Productivity (TFP) process is observable, although the shocks to it are not. The shocks u_t^y and u_t^x are not part of the DSGE model as such, but are policy-invariant measurement errors that are assumed to exist when $\mathbf{y}_t$ and $\mathbf{x}_t$ are actually observed, because otherwise the statistical tests of conditional independence laid out in 3.3.2.2 are not well defined, as the $\mathbf{y}_t$ and $\mathbf{x}_t$ are constant conditional on $[\mathbf{x}_{t-1}, \mathbf{z}_t]$.[7]

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 is stationary. Note that this structure implies that the exogenous states 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 in principle be generalised to allow for longer lags, if for example, it is believed that some effects may take multiple periods to play out.

In this setup, all variables can be categorised 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

[7] In practice, instead of adding measurement error noise to the simulations, the algorithm simply detects when the outcomes are constant conditional on the correct conditioning set.

variables are everything else; their past is irrelevant to the current values of the model. State variables can be further categorised as either endogenous states (the capital stock in the economy is a typical example) or exogenous states (the state of technology or productivity is a typical example) (Ravenna, 2007). As the name suggests, endogenous states are determined simultaneously (endogenously) with contemporaneous controls in the model, however, their past is relevant to the determination of the current values of the model. Exogenous states, on the other hand, are exogenous in the strongest possible sense. In this setup they are *strictly exogenous* relative to any other variable in the model, including the other exogenous states.

3.3 Methodology

Given equations (3.2) - (3.4) it is straightforward to characterise the general solution to a DSGE model as a DAG. This is illustrated by Figure 3.4. This expresses in graphical format all the relationships that exist in those equations, taking into consideration that the edges in the DAG are assumed to imply linear relationships. In particular, it captures all the conditional independence relationships in the DSGE model, therefore, if the underlying distribution is stable, then this DAG is faithful.

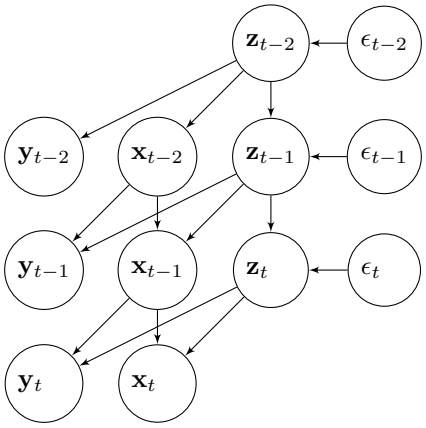


Figure 3.4: DSGE solution expressed as a DAG

Already it would seem straightforward to input random samples generated from a DSGE model into the available structure learning algorithms in order to find the correct model, given that these algorithms have well established asymptotic convergence properties. Unfortunately, results obtained this way (provided in Sections 3.5.4.2 and 3.5.4.3) are less than convincing, 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 variable. While correlation is undefined for constants, in practice when the calculation is forced the result tends to infinity, as it involves

division by a number very close to zero. This is particularly problematic because the case where the true parents are conditioned on is exactly the case in which one would like to conclude that the remaining partial correlation is in fact zero. Furthermore, these results are only asymptotic, and it seems that finite-sample bias may be important in economic applications, where in practice sample sizes are small relative to the dimension of the problem. Particularly problematic is that structure learning algorithms consider all possible DAGs given observed variables as potential candidates, whereas in this context it is assumed that the solution takes on a particular form, as in equations (3.2) - (3.4).

As a result of these limitations, I found that a more effective approach in this context involved a bespoke algorithm that takes into account the relatively stringent assumptions that can be made about DSGE solutions. For the reasons outlined in Section 3.2.1.2 this will be a hybrid algorithm. Therefore, before introducing the algorithm I will define relevant constraint and score tests in turn. But first, I will discuss the validity of the *stability* assumption (Definition 1), which is essential for any DAG estimation procedure to be justified, in the context of DSGE models. Note that although the discussion here is about stability, which is the fundamental untestable assumption under the definitions given in this paper, often in the DAG literature it is instead the faithfulness assumption that is examined. This is because when discussing the appropriateness of a specific DAG, faithfulness is the relevant concept. In order to be faithful, the DAG and DGP must satisfy both stability and have the same conditional independence relationships. However, the focus here is on whether any appropriate DAG exists, which depends only on stability. If stability holds then it is possible to search for a DAG that also satisfies the testable part of faithfulness using particular conditional independence tests that will be introduced in Section 3.3.2.

3.3.1 Validity of the Stability Assumption

A sufficient condition for stability is that the DGP parameters are jointly continuous and vary freely over the parameter space (Steel, 2006) for different populations, 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 as in the example in Section 3.2.1.1 have Lebesgue measure 0. If under the assumption that the true DGP of the macroeconomy is a 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 stability assumption. 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 stability, Steel (2006) notes that such failures or near-failures (that is near-zero statistical dependence despite clear causal pathways) are likely to 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 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 the variation caused by selection. This condition is perhaps more likely to be violated as there is considerable cross-country variation in macroeconomic conditions. However again, surviving the selection and homogeneity test does not guarantee that the stability assumption is verified.

Despite these concerns, I would argue that the stability assumption is plausible in most macroeconomic contexts. For simulations, whether or not the assumption is violated can be read straight off the model. For real data, it seems unlikely that any macroeconomic variable (even the policy rate) is determined in an entirely systematic or deterministic way. In reality, monetary authorities face a number of constraints that would prevent them from completely stabilising inflation including informational constraints, political influences, and the zero lower bound. 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, suggesting that even in the context of monetary policy where decision-making is highly rigorous, it is nonetheless neither strictly deterministic, nor determined in the same way everywhere. For other macroeconomic variables which are not determined by a centralised authority, it is even easier to believe there is significant room for free variation of parameters across populations.

3.3.2 Constraint Tests

3.3.2.1 Independence Relationships

Applying the parental Markov condition (Corollary 4) to the DAG in Figure 3.4 implies the following four independence relationships among the time t and $t-1$ variables:

$$x_t \perp\!\!\!\perp x'_t \,||\, [\mathbf{x}_{t-1}, \mathbf{z}_t] \text{ for all } x_t \neq x'_t \in [\mathbf{x}_t, \mathbf{y}_t] \tag{3.5}$$

$$x_{t-1} \perp\!\!\!\perp z_t \,||\, \mathbf{z}_{t-1} \text{ for all } x_{t-1} \in \mathbf{x}_{t-1} \text{ and } z_t \in \mathbf{z}_t \tag{3.6}$$

$$x_t \perp\!\!\!\perp z_{t-1} \,||\, [\mathbf{x}_{t-1}, \mathbf{z}_t] \text{ for all } x_t \in [\mathbf{x}_t, \mathbf{y}_t] \text{ and } z_{t-1} \in \mathbf{z}_{t-1} \tag{3.7}$$

$$z_t \perp\!\!\!\perp z'_t || \mathbf{z}_{t-1} \text{ for all } z_t \neq z'_t \in \mathbf{z}_t \tag{3.8}$$

The first condition (3.5) 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 $\mathbf{z}_t$ (I sometimes hereafter refer to these as 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 (3.6) states that the time t lagged endogenous states are independent of every exogenous state conditional only 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 (3.7) holds because the time t states d-separate the time t endogenous variables from the lagged exogenous states. If further lags were to be considered, 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 (3.8) holds that all exogenous states are mutually independent conditional on past exogenous shocks. This is a stronger condition than the other three, and depends crucially on the assumptions that $\mathbf{E}$ and Σ are diagonal.

Only (conditional) independence relationships are considered because it is usually the case in macroeconomic time-series that all observables have non-zero pairwise correlation. Therefore, the *lack* of a relationship in the form of conditional independence is more useful for identification than the presence of one. Corollary 1 can then be applied to these constraint tests in order to already provide a powerful selection criteria for empirical DSGE models, which I will refer to as *validity*:

Theorem 2. *Suppose that a log-linearised DSGE model M generates a DGP $f(\mathbf{w}_t;\theta)$ over a set of observed variables $\mathbf{w}_t$, that partitions $\mathbf{w}_t$ into three mutually exclusive vectors $\mathbf{x}_t$, $\mathbf{y}_t$, and $\mathbf{z}_t$ representing the endogenous states, controls, and exogenous states of M respectively. Further suppose that there is some DAG G which is faithful to $f(\mathbf{w}_t;\theta)$. Then G is the only faithful DAG which satisfies conditions (3.5), (3.6), and the Minimum State Variable (MSV) criterion (McCallum, 1999). This G is said to be valid.*

Proof of this theorem is provided in Appendix A. The introduction to Section 3.3 showed that under the assumption of stability a log-linear DSGE solution can be represented by at least one faithful DAG, therefore, by applying the previous theorem, the following corollary is obtained:

Corollary 3. *If an underlying DSGE model which can be expressed as in equations (3.2) - (3.4) generates a stable distribution $f(\mathbf{w}_t;\theta)$, then there is exactly one valid DAG which is faithful to $f(\mathbf{w}_t;\theta)$.*

Note that while the proof of Theorem (2) makes use of only constraints (3.5) and (3.6), (3.7) and (3.8) are still applicable (necessary conditions) because they are implied by the DAG, but they are not in the minimal set of sufficient conditions for a unique solution. To be more general, these assumptions could be dropped as long as the shocks only directly affect the exogenous states, and the other constraints would still hold and be valid tests of the model. However, these constraints (and the associated assumptions) can nonetheless be included because they are satisfied by a wide range of DSGE models including all of those considered in the empirical portion of this paper, and more importantly testing a larger number of conditions will, all else equal, give more *power* to reject incorrect models, which will prove critical in dealing with issues arising from small available sample sizes.

The MSV criterion simply states that the chosen model should have the fewest number of state variables among those which satisfy the conditional independence criteria. This criterion is necessary for the proof and it is a natural and intuitive requirement to impose. Adding state variables increases the size of the conditioning set, and therefore weakly increases the plausibility of any conditional independence relationship that may be tested.[8] As a result, if some model with m states is valid, then another model with the same states save for one control that is changed to a state variable will also trivially be valid. Yet that model would also be less parsimonious and is therefore less desirable. 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 (3.2) - (3.4). Exogenous states are involved in all three equations, endogenous states two, and controls only one. Another way to see this is in figure 3.4. Among time t and $t-1$ variables, adding an exogenous state results in the addition of edges in four places and thus eight parameters (one slope parameter and one variance parameter), an endogenous state in three places, and a control in only two. Therefore, according to this principle, models with fewer states, especially exogenous states are preferable, all else equal.

It is worth stopping now to consider what Corollary 3, and the results leading up to it, collectively imply. Consider some distribution $f(\mathbf{w}_t, \theta)$, and assume that all the conditional independence relationships implied by it ($I(f(\mathbf{w}_t, \theta))$) are known. Furthermore, assume that $f(\mathbf{w}_t, \theta)$ is *stable*, and was generated by some sort of log-linear DSGE model as in equations (3.2) - (3.4), whose exact specification is unknown. Then there is exactly one DAG, and hence state-space model that is *valid*. It is then possible to identify this state-space model by testing all possible state-space models against the validity criteria. While this does not uniquely pin down a single set of micro-foundations, it does rule out any structural model that does not generate the correct state-space in its reduced form. This shows the degree of identification that is at least theoretically obtainable using this approach. However, it will be necessary to consider many practical issues in order to operationalise this, including how to test for conditional independence, which is the issue that will be considered next.

3.3.2.2 Testing Procedure

The proof in Theorem 2 assumed that the conditional independence relationships in the true distribution of variables are known. Of course, this is not the case in practice, therefore, this section will discuss the implementation of an empirically viable strategy for testing conditions (3.5) - (3.8). In the present application, I make the assumption that observed variables are normally distributed, such that testing for conditional independence is equivalent to testing for non-correlation among partial residuals. This assumption is in general not required as it is possible to test for conditional

[8] To see this, note that $\mathbf{y}_{t-1}$ has no children, so $y_{t-1} \in \mathbf{y}_{t-1}$ cannot result in a blocked frontdoor path if it is moved into $\mathbf{x}_{t-1}$, while no new conditional independence restrictions that must be satisfied are implied.

independence non-parametrically (see Strobl et al. (2019) for a review of recent contributions in this vein), however, this assumption is made here because Gaussian assumptions are common in DSGE models and economic applications more generally, and the resulting simplifications will allow for more clear exposition of the main contributions of this paper.

Partial linear correlations can be estimated by regressing the set of target variables of interest $\mathbf{x}$ on the set of conditioning variables $\mathbf{z}$ and then estimating the correlations between the resulting estimated residuals $\hat{\mathbf{u}_x}$. Therefore, one way to implement tests for conditions (3.5) - (3.8) would be to perform a t-test on the estimated partial linear correlation implied by each of these conditions for every model, and then reject the model if any of these t-tests reject the null hypothesis at the specified significance level (after applying a Bonferroni (1936) correction). If the incorrect conditioning set is applied these residuals will be correlated. If the correct conditioning set is applied, these residuals will be consistent for the measurement errors, which are assumed to be independent. Furthermore, in principle, one could adjust these tests to allow for heteroskedasticity, however, to maintain simplicity and focus on the core contribution of this paper only the homoskedasticity case will be considered. Hereafter this is referred to as the *multiple testing approach*. As shown in Section 3.5, this approach does seem to perform well on simulated data, with higher power and lower size than the second approach which I will soon introduce. However, it has a number of significant drawbacks.

Firstly, the Bonferroni (1936) correction assumes independence of each of the tests, which is highly implausible in this case. Indeed, this explains why the empirical size of these tests is less than the specified significance level. Since the degree of correlation between tests may take on any form it is difficult or impossible to pin down important statistical properties (such as the size or power) of this procedure. Furthermore, there is the issue (which was also noted as a drawback of alternative approaches) that computation of partial correlations can be unstable if residuals are very close to or equal to zero. Indeed, in principle the residuals produced by the correct model should be exactly zero, which is a constant, and therefore pairwise correlation undefined. In practice, when simulated data is used, residuals for the ground truth model are very close to (but not equal to) zero. In this case pairwise correlation can be computed, however, it is not particularly meaningful since it only reflects floating point imprecision or rounding error in the simulation. Since this computation involves dividing two near-zero values it tends to produce an estimated correlation close to 1. This is problematic because this is exactly when we do not want to reject the null hypothesis of conditional independence. As a workaround for this the algorithm will detect small residuals below some tolerance threshold and pass the model through the test (do not reject the hypothesis of independence) if they are observed. This is a highly idiosyncratic correction that is an undesirable feature of this approach. Finally, the number of tests conducted can grow very large if there is a large number of observables resulting in implausibly large critical values (due to the Bonferroni correction), and exponentially growing computational complexity.

For these reasons, I also propose the implementation of a different test provided by Srivastava (2005). This test is for the null hypothesis that a covariance matrix is diagonal. In order to use this, I combine and slightly rearrange conditions (3.5) - (3.8) such that they have the same conditioning set, and imply a relationship of *complete partial independence* between tested variables. To do this, I roll conditions (3.5) and (3.7) back one period in time,[9] and add $\mathbf{x}_{t-2}$ to the conditioning sets in conditions (3.6) and (3.8). This latter change is justified because in both cases every backdoor path between the variables of interest have already been blocked and $\mathbf{x}_{t-2}$ is not part of any frontdoor path between them, and therefore d-separation is maintained. In other words, if the exogenous states are mutually independent, conditioning on endogenous states contains no new relevant information, so it is harmless to add these as regressors. The modified conditions are shown in (3.9) - (3.12).

$$x_{t-1} \perp\!\!\!\perp x'_{t-1} \,||\, [\mathbf{x}_{t-2}, \mathbf{z}_{t-1}] \text{ for all } x_{t-1} \neq x'_{t-1} \in [\mathbf{x}_{t-1}, \mathbf{y}_{t-1}] \tag{3.9}$$

$$x_{t-1} \perp\!\!\!\perp z_t \,||\, [\mathbf{x}_{t-2}, \mathbf{z}_{t-1}] \text{ for all } x_{t-1} \in \mathbf{x}_{t-1} \text{ and } z_t \in \mathbf{z}_t \tag{3.10}$$

$$x_{t-1} \perp\!\!\!\perp z_{t-2} \,||\, [\mathbf{x}_{t-2}, \mathbf{z}_{t-1}] \text{ for all } x_{t-1} \in [\mathbf{x}_{t-1}, \mathbf{y}_{t-1}] \text{ and } z_{t-2} \in \mathbf{z}_{t-2} \tag{3.11}$$

$$z_t \perp\!\!\!\perp z'_t || [\mathbf{x}_{t-2}, \mathbf{z}_{t-1}] \text{ for all } z_t \neq z'_t \in \mathbf{z}_t \tag{3.12}$$

Each of the conditional independence relationships (3.9) - (3.12) now relies on the same conditioning set. Furthermore, when combined these conditions imply that all the variables in the vector $[\mathbf{y}_{t-1}, \mathbf{x}_{t-1}, \mathbf{z}_t, \mathbf{z}_{t-2}]$ are completely independent, conditional on $[\mathbf{x}_{t-2}, \mathbf{z}_{t-1}]$. $\mathbf{z}_{t-2}$ can be optionally excluded from this vector[10] as it is associated with test (3.7), which is not required for a unique *valid* model as in Theorem 2. On the other hand it will be necessary to impose (3.8), which is also not required for *validity*, in order to implement this test. This condition is now necessary because without it the partial covariance matrix would have some unrestricted elements and not be strictly diagonal under the null hypothesis, which would therefore be a substantially more difficult hypothesis to test. To test whether a model is valid using this approach, first regress the vector $[\mathbf{y}_{t-1}, \mathbf{x}_{t-1}, \mathbf{z}_t]$ on $[\mathbf{x}_{t-2}, \mathbf{z}_{t-1}]$, collect estimated residuals, estimate the covariance matrix S of the $T \times k$ matrix of residuals and, perform a z-test at some specified significance level α on the test statistic $\hat{T}_3$ from Srivastava (2005), which is asymptotically normally distributed. This test statistic is defined by equations (3.13) - (3.16):

[9] Equivalently, one could roll forwards the other two conditions, but this would require data on a lead rather than two lags.

[10] In the application this will be excluded because the independence of $\mathbf{z}_{t-2}$ here depends on the strict exogeneity property of the exogenous states rather than the mutual independence of their AR processes (since the conditioning set does not include $\mathbf{z}_{t-3}$). This may lead to some false rejections of the true model because of the possibility of a spurious regression when the exogenous states are close to unit roots.

$$\hat{T}_3 = \left(\frac{n}{2}\right) \frac{(\hat{\gamma}_3 - 1)}{\left(1 - \left(\frac{1}{p}\right)\left(\frac{\hat{a}_{40}}{\hat{a}_{20}^2}\right)\right)^{\frac{1}{2}}} \tag{3.13}$$

$$\hat{\gamma}_3 = \frac{n}{n-1} \frac{tr(S^2) - \frac{1}{n}(tr(S))^2}{\sum_{i=1}^{m} s_{ii}^2} \tag{3.14}$$

$$\hat{a}_{20} = \frac{n}{p(n+2)} \sum_{i=1}^{m} s_{ii}^2 \tag{3.15}$$

$$\hat{a}_{40} = \frac{1}{p} \sum_{i=1}^{m} s_{ii}^4 \tag{3.16}$$

Note that s_{ij} is the (i, j) element of S. Also note that the denominator in $\hat{T}_3$, $\left(1 - \left(\frac{1}{p}\right)\left(\frac{\hat{a}_{40}}{\hat{a}_{20}^2}\right)\right)$ can be negative, and thus, the test statistic undefined. In order to alleviate this I take the same approach as in Wang et al. (2013) and replace this term with $1 - \sum_{i=1}^{m} s_{ii}^4 / \left(\sum_{i=1}^{m} s_{ii}^2\right)^2$ when it is negative.

This strategy alleviates a number of the drawbacks of the first approach. Since this approach utilises estimated covariance rather than correlations it avoids unstable computation around the true model, where residuals are very close to zero. Therefore, it is possible to test all models without making exceptions for special cases. Furthermore, it is much simpler to describe the properties of this test. Asymptotically, it will have exactly α type I error rate, without the need for any corrections. Estimates of the power of this test against numerous alternatives can be found in Wang et al. (2013), and are also provided over a range of scenarios in Appendix B. Finally, this approach results in exactly one test being performed regardless of the complexity of the model under consideration. While it is true that the test is somewhat more computationally intensive for larger covariance matrices, it has much better computational scaling properties than the multiple testing strategy. However, as shown in Section 3.5, it unfortunately does not seem to be as accurate at identifying the ground truth model in simulations on more complex data sets as the multiple testing strategy. This seems to be due to a lack of power against alternatives.

3.3.3 Score Tests

Notwithstanding the uniqueness proof in Theorem 2, in finite samples it is not uncommon to encounter cases where more than one model is *valid*. The results in Section 3.5 show that these models will generally be very similar to the ground truth, and represent a small minority of all considered models. At this point it could be left up to expert opinion to select the most sensible of the remaining models, however, since one of the most important benefits of this approach is agnosticism it is desirable to implement some heuristic way of selecting a single model with the algorithm. The approach that I will use to achieve this is score maximisation over valid models. Essentially, this will sort the models which are deemed to be valid by their likelihood in order

to choose a unique winning model. In principle, one could evaluate models solely on their score, however, for the reasons outlined in Section 3.2.1.2, my preferred approach is to use this only in a secondary role. For comparison simulation results for pure score-based estimation will be considered in Section 3.5.4.1.

The most basic score function for Gaussian Bayesian networks is the log-likelihood function. According to the parental Markov condition (Definition 4) if DAG G is faithful to the DGP f, then f admits factorisation of the joint probability distribution into the product of the distribution of each variable conditional on its parents:

$$f(\mathbf{w};\theta) = \prod_{i=1}^{k} f(w_i|pa_G(w_i);\theta) \tag{3.17}$$

Therefore, the log-likelihood can be calculated as:

$$\mathcal{L}(\mathbf{w},\theta) = \sum_{i=1}^{k} ln(f(w_i|pa_G(w_i);\theta)) \tag{3.18}$$

Now consider the assumptions in the current context. $\mathbf{w}$ is partitioned into $\mathbf{z}$, $\mathbf{x}$, and, $\mathbf{y}$. It is assumed that the conditional probabilities are linear functions and follow a mean-zero normal distribution, so the only parameters are the slope parameters in matrices $\mathbf{A}$ - $\mathbf{E}$ and the variance-covariance matrix $\tilde{\Sigma}$. Furthermore, the model predicts time t values *given* time $t-1$ values, so it is not necessary to consider the distribution of lags which are constant with respect to the model. Therefore, the log-likelihood of the model is given by equation (3.20):

$$\begin{aligned}
\mathcal{L}(\mathbf{w};\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D},\mathbf{E},\tilde{\Sigma}) = &\sum_{z_{i,t}\in\mathbf{z}_t}\left(\sum_{t=1}^{T} ln(\phi(z_{i,t}|z_{i,t-1}|\mathbf{E},\tilde{\Sigma}_z))\right)+ \\
&\sum_{y_{i,t}\in[\mathbf{y}_t,\mathbf{x}_t]}\left(\sum_{t=1}^{T} ln(\phi(y_{i,t}|[\mathbf{x}_{t-1},\mathbf{z}_t]|\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D},\tilde{\Sigma}_y))\right) \qquad (3.19) \\
= &\sum_{i|y_{i,t}\in\mathbf{y}_t}\left(\sum_{t=1}^{T} ln(\phi(\mathbf{a}_i\mathbf{x}_{t-1}+\mathbf{b}_i\mathbf{z}_t|\mathbf{a}_i,\mathbf{b}_i,\sigma_i^2))\right)+ \\
&\sum_{i|x_{i,t}\in\mathbf{x}_t}\left(\sum_{t=1}^{T} ln(\phi(\mathbf{c}_i\mathbf{x}_{t-1}+\mathbf{d}_i\mathbf{z}_t|\mathbf{c}_i,\mathbf{d}_i,\sigma_i^2))\right)+ \\
&\sum_{i|z_{i,t}\in\mathbf{z}_t}\left(\sum_{t=1}^{T} ln(\phi(\mathbf{e}_i z_{i,t-1}|\mathbf{e}_i,\sigma_i^2))\right) \qquad (3.20)
\end{aligned}$$

Where $\mathbf{x_i}$ is the i_{th} row of $\mathbf{X}$, σ_i^2 are the (i,i) diagonal elements of $\tilde{\Sigma}$, and ϕ is the probability density function of the normal distribution. Notice that it is possible to calculate the variances separately in each linear projection because the parental Markov condition implies that each regression equation is independent. Therefore, substituting in for the maximum likelihood estimate

of σ_i^2 for each regression and the functional form of ϕ results in a substantially simpler expression for the log-likelihood function:

$$\mathcal{L}(\mathbf{w}) = -\frac{T}{2}\left(k\left(1 + ln(2\pi)\right) + \sum_{i=1}^{k} ln(\hat{\sigma}_i^2)\right) \tag{3.21}$$

$$\hat{\sigma}_i^2 = \frac{1}{T}\sum_{t=1}^{T}(w_{i,t} - \hat{w}_{i,t})^2 \tag{3.22}$$

Where $\hat{w}_{i,t}$ are the predicted values of some w_i in $\mathbf{w}$ implied by equations (3.2) - (3.4) using maximum likelihood estimates of the coefficient matrices. Note that the log-likelihood is inversely proportional to the mean squared error (MSE). This is consistent with the interpretation that maximising the score function is equivalent to finding the model with the best predictive performance in this context. Indeed, since the rest of the terms are constant, it suffices to minimise the MSE to maximise the log-likelihood in this setup.

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*. The most common response to this is to use a penalised score function such as the Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarz et al., 1978), which will regularise the estimated model by reducing the score by some increasing function of the number of parameters estimated. Indeed, in their proof of the consistency of their score-based GES algorithm, Chickering (2002) require that the score function used adequately penalise complexity.

When it comes to the preferred hybrid algorithm proposed here, given that stringent conditional independence criteria are already being applied, 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 refers to the tendency 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. In experimentation, I found that penalised score functions are very unlikely to overturn this bias towards exogenous states. So instead of using these, I will simply take lexicographic preference for models (among those which are valid) with more endogenous states first, and then only after this maximise the likelihood function. With this sorting, all remaining models have the same complexity so penalised scores no longer serve any purpose. Another justification for this sorting is the more general belief in macroeconomics that all observables are interrelated in some way, and therefore, the exogeneity assumptions implied by exogenous states are quite strong, and it is thus preferable to minimise them.

3.3.4 Algorithm

Having defined a number of tests for an optimal and *valid* model, we now turn our attention to developing an algorithm which will apply these tests in order to choose one from the set of all possible state-space models, which is outlined in Algorithm 4.

The algorithm is very simple and is designed to reflect a few key model selection heuristics. As previously discussed, the algorithm assumes that constraint 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 is outlined in 3.2.1.2. Essentially, unlike score functions, constraints directly rely on information about a relevant sense of causality.

The MSV criterion is imposed by the algorithm in the sense that it stops considering models with a greater number of states once some valid model is found. This is primarily because MSV is required for the uniqueness property of validity, however, the MSV criterion also allows for a potentially very large increase in the computational speed of the algorithm. Without it, it would be necessary to 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 $\sum_{r=m}^{k} 2^r \binom{k}{r}$ iterations can be skipped, which potentially reduces the search space by many orders of magnitude if $m << k$. This algorithm is nonetheless highly inefficient, however, 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 future research.

It follows almost immediately that under the assumption that the number of observables is fixed (or indeed grows sufficiently slowly relative to the number of observations n) this algorithm will consistently estimate the unique valid state-space model as $n \to \infty$. This is independent of the conditional independence test used, as long as that test has power against alternatives that is one in the limit. The test given by Srivastava (2005), and indeed the multiple testing strategy both satisfy this requirement. Therefore, since the algorithm systematically considers every one of a bounded number of possible models, it will reject every incorrect model in the asymptotic case. It will also reject the correct model in a proportion α of samples. In these cases the algorithm will (still asymptotically) yield no solution. In the rest it will yield the unique valid model.

Algorithm 4: Brute force hybrid state-space estimation algorithm

Input: *alpha*: significance level

Input: *test*: testing strategy is either 'multiple' or 'srivastava'

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, sorted by likelihood

begin

```
continue = true
n_states = 0
max_states = #observables − 2
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)
        if test = multiple:
            sig_level = alpha / length(constraint_tests)
        else:
            sig_level = alpha
        if every constraint_test .p_value > sig_level for constraint_test ∈
         constraint_tests:
            append potential_states to all_valid_states
            continue = false
sort descending all_valid_states by #endogenous_states, score_tests
return all_valid_states
```

However, in finite samples there is unfortunately no guarantee that the algorithm will yield the correct solution. Although the test from Srivastava (2005) is only asymptotically normal, in practice the type I error rate remains close to the specified significance level α for any reasonable sample size (see Appendix B), so this seems to be a reasonable approximation. On the other hand, the probability of type II error can be quite high in small samples, and this is very problematic. The algorithm will stop early if it finds some valid model with m states. However, if this is the result of a type II error, and the correct model actually has more than m states, then the algorithm will terminate before it ever even considers the correct model. Potential solutions to this problem that would improve small sample performance would be to devise a test with more power, remove the early stopping behaviour, or otherwise limit the size of the search space, although this may result in a greater reliance on sorting by score to differentiate between valid models, or a loss of

the agnosticism of the algorithm.

3.3.5 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 time-series. It is possible to draw comparisons with both Structural Vector Auto Regression (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 VAR, which was introduced by Sims (1980). This method involves regressing a vector of outcomes $\mathbf{y}_t$ on a matrix containing p lags of $\mathbf{y}_t$ in the form $\mathbf{y}_t = [\mathbf{y}_{t-1}, ..., \mathbf{y}_{t-p}]\beta + \epsilon_t$. 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 truly exogenous shock on the dynamics of observed variables. In order to address this issue the model can transformed and an assumed causal ordering imposed in the form of a Cholesky decomposition (Sims, 1980), which has the effect of making the errors of the estimated, transformed model mutually uncorrelated or in other words *structural*. Therefore, such models are known as SVARs. In particular, the kind of DAG considered in this paper also include a distributed lag, as some variables can enter into the regression of others contemporaneously. As noted by Demiralp and Hoover (2003), there is an equivalence that can be drawn between SVAR models and DAGs. Indeed, in this paper they implement the PC-algorithm (Spirtes & Glymour, 1991) to show that structure learning methods for DAGs can be used to identify the causal order of the Cholesky decomposition for an SVAR from data.

The primary advantage of DAGs is the relatively weak assumptions they require. SVAR models require the researcher to specify assumptions about the relative exogeneity of observable variables, by explicitly assuming a causal ordering of variables. These assumptions are themselves either derived from a similarly assumption-heavy model such as a DSGE model, or are in some cases 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 Jordà, (2005) criticising this type of methodology. Seen in this way, DAGs constitute a powerful new tool to choose the specification of these types of models in an agnostic and data-driven way.

3.3.6 Misspecification

As with any empirical technique, when applying this algorithm, one may worry how it behaves under misspecification. In the current context, misspecification could have two separate meanings. The first is that a log-linear DSGE model is not the correct model for a given set of data. For example, it could be the case that the data are in fact generated by a stationary process. This

type of misspecification is however applicable to the entire class of linearised DSGE models, which are widely used. The second kind of misspecification, which is likely more relevant, is that not all variables that exist in the true DGP are observed, and therefore some of them cannot be input into the algorithm. In this case, the outcome will depend on which category the missing data is in. Consider equations (3.5) - (3.8). If the variables that are unobserved are all control variables y_t, then the conditioning set in all four conditions are unaffected, and therefore, the conditional independence relationships between the variables that are observed are unaffected. Therefore, the true model among the observed variables can still be identified.

However, if the missing observables are states, then none of the models which can be constructed among the variables that are observed will be *valid* asymptotically.[11] This is because relative to the DGP, there will be unobserved confounding — or in DAG language — an unblocked backdoor path between some of the observed variables. In other words, the conditioning set will not be sufficient to ensure conditional independence because one of the relevant conditioning variables is missing. In the case of a missing exogenous state all 4 conditions (3.5) - (3.8) will fail and in the case of a missing endogenous state conditions (3.5) and (3.6) will fail. Therefore, asymptotically, the algorithm will return no valid model. In finite samples, where power is limited, some invalid models may still be accepted. Furthermore, if some states are unobserved, then unlike in the case where controls are unobserved, it is also unfortunately impossible for the algorithm to identify the correct model among the variables that are observed. Indeed, in this case, strictly speaking the true model would be rejected by all of the conditions (3.5) - (3.8). For example, condition (3.5) would fail for the true model (among variables that are observed) since the true controls would not be independent conditional on the observed states, because the presence of an unobserved state leaves open an unblocked backdoor path between the true controls.

In practice, how severe this limitation is will depend somewhat on the data available. While it is impossible to rule out misspecification, the severity of misspecification depends on the number of states that are unobserved and the strength of the causal pathways that they mediate. State variables are the variables that are most fundamental the economic system being modelled, as their behaviour propagates through the model, whereas controls are relatively more superfluous, simply being outcomes of the dynamic system. The take-away from this section should therefore be that in order to increase the credibility and generalisability of the results of the algorithm one need not simply include as many variables as possible, especially given the limited dimensionality scaling of the proposed algorithm, and should instead focus on including the most fundamental observables that likely mediate other economic relationships.

[11] This was also empirically verified in a test of the RBC model using the full dataset of 100000 synthetic observations, in which one of the exogenous states was included from the data used.

3.3.7 IRFs

One very common way of evaluating DSGE models is to compare the Impulse Response Functions (IRFs) that they imply 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. This is simply done to demonstrate that the state-space model that is estimated matches the reduced form of the original simulation. 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.

3.4 Data

In order to demonstrate the capability of the proposed algorithm empirically I will work with both simulated and real macroeconomic data. Using simulated data has a number of advantages. Firstly, since the model that generates the data is known it is possible to evaluate whether structure learning has succeeded in identifying the ground-truth DAG. Secondly, in this context it can be ensured to the greatest possible extent that the underlying assumptions of the structure learning algorithms, including linearity, normality, and observability are satisfied. In addition, since these models are standard in modern macroeconomics it provides a highly relevant controlled testing environment. Furthermore, using real data is an opportunity to demonstrate that the algorithm is also a powerful heuristic tool that can be implemented outside a rigorously controlled environment. If these results are to be believed it will allow for inferences pertaining to a number of important debates in the DSGE literature. The remainder of this section will discuss the various sources and general properties of the data used.

3.4.1 Simulations

In order to collect simulated data I consulted a GitHub repository containing Dynare code to replicate well known macroeconomic models (Pfeifer, 2020). In particular, I chose to model the baseline RBC model as a simple case and a New Keynesian model from Galí (2015) for a more difficult and complex modelling challenge. Simulations output a file containing $100,000$ observations of $i.i.d.$ draws of the exogenous shocks, and the associated observed values of the other variables in the model. This data and smaller subsets thereof, were then used as inputs for the structure learning algorithm.

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 3.1: Description of variables for the baseline RBC model.

3.4.1.1 Baseline RBC

The baseline RBC model includes 11 variables, which are summarised by Table 3.1. This model contains two exogenous state variables: technology and government spending, and one endogenous state: capital. There are two shocks in the model: one that affects only technology directly and one that affects only government spending directly. As explained in Section 3.2.2, these shocks will be dropped from the data. The shocks are Gaussian and orthogonal, and furthermore the model is taken as a first-order approximation. Therefore, all of the necessary assumptions are satisfied.

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 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 algorithm's performance when the assumption of stationarity is challenged.

3.4.1.2 Baseline New Keynesian

New Keynesian models are extremely popular in modern macroeconomics and are also considerably more complex than the baseline RBC. Therefore, this serves as a worthy challenge for this methodology. In particular, I use a model from Galí (2015) as provided by Pfeifer (2020). The variables in this model are summarised in Table 3.2.[12] This model has a total of four state variables: three exogenous states (policy rate, technology and, preferences) for which there is one $i.i.d.$ and Gaussian shock each, and one endogenous state (price level).

3.4.2 US Data

To provide an example of real macroeconomic time-series, quarterly data from the US during the period 1985-2005 were collected from FRED (2020) for 15 variables outlined in Table 3.3. All of the variables were de-trended and demeaned by taking the residuals of an estimated first-order

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

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 3.2: Description of variables for the baseline New Keynesian model.

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
z	Total Factor Productivity
u	Unemployment
l	Total Workforce
c	(Real) Personal Consumption

Table 3.3: Description of Variables for US Data

autoregression (as opposed to an HP filter). 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).

The assumption of some log-linear DSGE solution, implicitly implies the assumption that the data is generated from a stationary distribution with no structural breaks. This particular data set was chosen because during this time-frame that assumption is plausibly valid. In general, structural breaks are important to model correctly, however, at present incorporating these are left as an avenue for future research.

3.5 Results

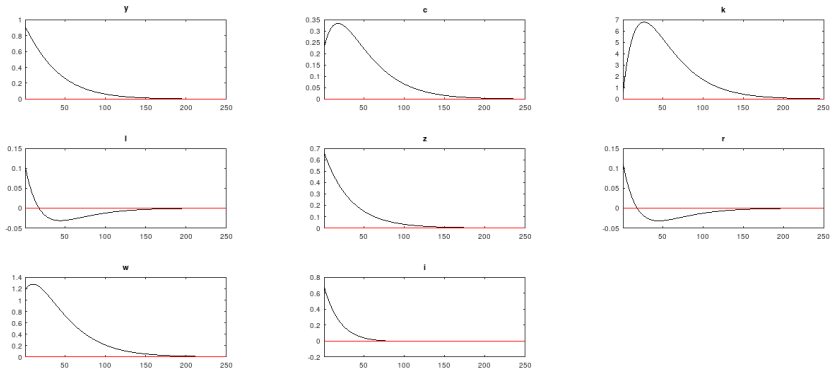
In this section many of the properties of the proposed algorithm will be thoroughly investigated. Using simulated data allows for the possibility of many experiments to test these properties in a controlled environment. In particular, for the models under consideration two scenarios will be presented. Solely to demonstrate the asymptotic consistency of the algorithm, results from the algorithm for a very large number of samples (100,000) are considered. Since at this sample size the power of the statistical tests against alternatives is close to one, one would expect to see that the true model is the only *valid* model. Conversely, to demonstrate the finite sample properties, results from a large number of runs of the algorithm (1,000) with a relatively small and realistic sample size (100) will be provided. In this case in each simulation it is recorded whether each combination of states was *valid*, and whether it was chosen as the best model (*wins*). Here one would expect to see that the true model is *valid* in $(1-\alpha)\%$ of the samples, where $\alpha = 0.05$ is the significance level of the conditional independence test, and that most of the time that the true model is *valid* it *wins*.

3.5.1 Baseline RBC

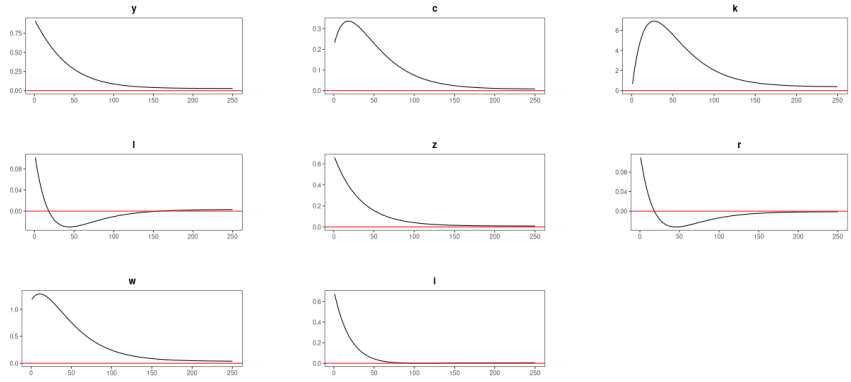
Using either testing strategy from Section 3.3.2.2 (multiple testing or Srivastava (2005)) on the entire sample of 100,000 observations for the RBC model the algorithm successfully identifies the correct states, which are exogenous states z and g, and endogenous state k. No other (incorrect) models are valid. Figure 3.5 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. Clearly, this large sample case is not empirically relevant; it is simply included to verify the asymptotic consistency of the algorithm.

Table 3.4 shows the small sample results for the algorithm using the test based on Srivastava (2005), and Table 3.5 likewise for the multiple testing strategy. I will now discuss each of these results in turn.

The results in Table 3.4 are promising for a number of reasons. The headline result is that



(a) Original Simulation



(b) Estimated State-Space Model

Figure 3.5: IRFs to a one standard deviation technology shock generated by the original simulation and estimated model.

Index	**Exogenous States**	**Endogenous States**	**Wins**	**Valid**
1	g z	k	944	944
2	g w	k	27	729
3	g y	k	27	571
4	c g	k	2	8
5	g l y		0	340
6	g r y		0	421
7	g r	k	0	576
8	g l z		0	716
9	g i r		0	781
10	g i l		0	629
11	g i	k	0	867
12	g r w		0	609
13	g r z		0	858
14	g k l		0	625
15	g l w		0	603
16	g k r		0	779
17	c g w		0	1

Table 3.4: Small-sample (n=100) simulation structure learning results for the RBC model using the Srivastava (2005) test. Algorithm was run for 1,000 iterations on different samples. Only models that were **Valid** relative to the conditional independence test in at least one iteration are displayed. **Wins** indicates the number of iterations in which that model was selected by the algorithm. The ground-truth model has **Exogenous States** g and z and **Endogenous State** k.

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 3.5: Small-sample (n=100) simulation structure learning results for the RBC model using pairwise correlation tests and a Bonferroni (1936) correction (multiple testing strategy). Algorithm was run for 1,000 iterations on different samples. Only models that were **Valid** relative to the conditional independence test in at least one iteration are displayed. **Wins** indicates the number of iterations in which that model was selected by the algorithm. The ground-truth model has **Exogenous States** g and z and **Endogenous State** k.

the ground-truth model (with exogenous states g and z and endogenous state k) is selected by the algorithm (denoted as "wins") in nearly 95% of iterations, and in every iteration where it is valid. The latter observation suggests that sorting by number of endogenous states and the likelihood function is having the intended effect. Also note that the empirical size of the test is quite close to the expected 5% significance level, as the correct model was rejected 56 times out of 1,000 iterations ($\sim 5.6\%$). Furthermore, observe that out of the 834 models that are considered in each iteration, that is, the models with less than or equal to three state variables, only 17 ($\sim 2\%$) are ever valid, and of those 17 only 4 (including the true model) are ever selected as the optimal model by the algorithm. Therefore, this testing strategy seems to have strong power to reject incorrect models in this application.

Table 3.5 mirrors the previous results in many ways, however, there are some key differences. The correct model is only rejected in 3 out of the 1,000 iterations, so the empirical size is far below the specified 5% significance level. This confirms suspicions that these pairwise correlation tests are not independent. However, this is much better than having higher than expected type I error, and this low type I error rate does not seem to have come at the cost of power, at least in comparison to the other testing strategy. Here only 8 out of the 834 models considered were ever valid, so this testing strategy actually seems to have higher power. Nonetheless, the true model does win less often using this approach (only 888 times as compared with 944), primarily because the model with exogenous states c and l wins 109 times (every time it is valid). This particular model was rejected in every iteration of the Srivastava (2005) test, despite its overall lower power. It seems likely that this particular combination of states performs so well in the multiple testing approach because these c and l are nearly collinear with the true exogenous states z and g, while being even more persistent, with very high estimated autoregressive coefficients of 0.994 and 0.972 respectively. As a result, the prediction while treating g and z as controls obtains a relatively high likelihood score. The conclusion here is that this approach may run into difficulties in small samples if there is a very high degree of multicollinearity or autocorrelation among observables.

Index	Exogenous States	Endogenous States	Log-Likelihood
1	a nu z	p	2.84×10^7
2	a nu z p		2.76×10^7

Table 3.6: Large sample (n=100,000) simulation structure learning results for one run using the New Keynesian model using pairwise correlation tests and a Bonferroni (1936) correction. The ground-truth model has **Exogenous States** a, nu and, z and **Endogenous State** p.

Index	Exogenous States	Endogenous States	Wins	Valid
1	a nu z	p	753	999
2	a mu nu	p	90	183
3	a nu w real	p	72	217
4	m real nu z	p	20	670
5	m real nu r nat	p	18	63
6	a nu r nat	p	18	966
7	nu pi y	p	14	375
8	a r real w real	p	4	6
9	a nu p z		3	1000
10	a mu nu p		2	181
11	mu nu y	p	2	240
12	a n nu	p	1	183
13	a mu r real	p	1	2
14	nu y z	p	1	26
15	i nu y	p	1	742

Table 3.7: Small-sample (n=100) simulation structure learning results for New Keynesian model using pairwise correlation tests and a Bonferroni (1936) correction. Algorithm was run for 1,000 iterations on different samples. Only models that had at least one **Win** are displayed. 55 models were **Valid** relative to the conditional independence test in at least one iteration. **Wins** indicates the number of iterations in which that model was chosen as optimal by the algorithm. The ground-truth model has **Exogenous States** a, nu, and z and **Endogenous State** p.

3.5.2 Baseline New Keynesian

We now turn our attention to the more complex baseline New Keynesian model. This model contains 17 observables, and is thus considerably more complex than the simulated RBC data. Table 3.6 shows the results for a large sample, using the multiple testing approach. The Srivastava approach are not shown, because this approach did not work in this application. This test is lacking in power (even with the full sample), such that numerous models with only two states were found to be valid, and therefore the algorithm terminated before considering the ground truth, which has four states. This unfortunately highlights one of the limitations to this approach.

On the other hand, while using the multiple testing strategy, the results are still promising. While using the full sample of $100,000$ observations, only two models are valid, and the correct model with exogenous states a, nu, and, z and endogenous state p wins both on preference for models with more endogenous states and on log-likelihood. This once again constitutes empirical validation of asymptotic properties.

Table 3.7 shows the small sample results using the multiple testing strategy. The results are not as strong as with the RBC model, but this is to be expected given the greater number of

Index	Exogenous States	Endogenous States	Log-Likelihood
1	y z u	pi rm k c	3.76×10^3
2	rm y z u	pi k c	3.74×10^3

Table 3.8: Structure learning results for the US macroeconomic data set (1985-2005) using pairwise correlation tests and a Bonferroni (1936) correction.

variables and complexity of model considered with the same sample size. I find that the ground-truth model wins in approximately 75% of iterations, while only being rejected once. Fifty-five models were valid in at least one iteration, which represents approximately 0.1% of models tested in each iteration. So despite the complexity of this problem, the realised type I and type II error rates of the multiple testing strategy in this application were actually even lower in relative terms than in the RBC setup, only not by enough to completely offset the increased complexity of the problem. Compared to that setup there were more valid models in any given iteration. As a result the task of sorting left over models by likelihood is more difficult, and this explains why the true model is not chosen as often, despite almost always being valid.

These results show that there are practical limitations to how well the algorithm and tests can perform. The tests are consistent as the sample size $n \to \infty$ with the number of observables k fixed. If k is not so small compared to the sample size then there is likely to be poor performance. This is a problem common to all high-dimensional econometric models, however, it may be particularly acute here because the number of models considered, and thus the complexity of the problem grows exponentially in the number of observables.

3.5.3 US Data

Table 3.8 shows results for structure learning on the US macroeconomic data set using the multiple testing strategy.[13] Despite the small data set of only 80 observations these tests were able to reject all but 2 of the $93,434$ models considered. Many features of this solution are consistent with what standard intuitions would imply. For example, observe that capital and the policy rate are endogenous states. Both of these are standard features of any DSGE model, and the second one reflects the well documented Taylor (1993) rule, an example of which is shown in (3.23), here also augmented with interest rate smoothing (Sack & Wieland, 2000). Note that although I assume that there is a structural shock affecting an endogenous state, this does not fundamentally violate any of the assumptions made; it simply implies that the *measurement error* has a particular interpretation in the case of this variable, which is that it is the non-systematic component of monetary policy. Also note that TFP is exogenous, which is a fairly standard assumption to make.

$$r_t = \rho_r r_{t-1} + (1 - \rho_r)(\theta^\pi \pi_t + \theta^y y_t) + \epsilon_t^{mp} \tag{3.23}$$

[13] Again, Srivastava results are not shown because, given the result from Section 3.5.2, the multiple testing strategy results seem more credible.

If these results are to be believed, then there are numerous implications for theory, at least in the context of US macroeconomic trends. First of all, the fact that consumption is an endogenous state is evidence in favour of the hypotheses of J. C. Fuhrer (2000) that DSGE models should take into account habits in consumption, thus making consumption inertial. Furthermore, observe that inflation is an endogenous state. This is evidence related to a particularly heated debate surrounding whether inflation is purely rational and forward-looking (Levin et al., 2004), and should therefore be modelled as a control variable, or whether inflation demonstrates persistence (Christiano et al., 2005) due perhaps to indexing or other forms of bounded rationality, and should therefore be modelled as a state variable. Clearly then, this evidence supports the latter hypothesis.

Perhaps more difficult to reason about is why output and unemployment enter as exogenous states. But for these too, some explanation can be suggested. Exogenous states are the only variables in the model which are directly exposed to shocks. Recall that these shocks are assumed to be structural or orthogonal. Assuming a Cobb-Douglas style production function the three determinants of output are TFP, labour input, and the capital input. Since unemployment (which is inversely proportional to the labour input) and TFP are already included as exogenous states, orthogonal shocks to output must be shocks to the capital input. Yet capital itself is included in the observables here, therefore, this is best interpreted as a shock to variable capacity utilisation (Driver, 2000). Similarly, in a model with search and matching in the labour market unemployment could be an exogenous state if a structural shock to matching efficiency were included. The fact that these variables enter as exogenous states does not suggest that the shocks to these variables are the shocks that are the most important in explaining the dynamics of the economy, but rather that these dynamics are most difficult to explain via endogenous channels in the observed data.

Figures 3.10 and 3.11 in the appendix show IRFs for shocks to the three exogenous states, as well as to the policy rate rm, which is an endogenous state in the estimated state-space model and in an unconditional VAR(1) model. The VAR IRFs are provided as a basic sanity check — so that some form of comparison can be made, but there is little reason to believe that these are necessarily a good depiction of reality. Ultimately, there is no fundamental ground-truth to compare these IRFs to, much like the choice of state variables they can only be evaluated against common heuristics and *stylised facts* in the literature. For example, consider the response to a TFP shock. All the IRFs from the state-space model match the direction of those from the VAR, except output and unemployment (which are exogenous and thus do not respond in the state-space model), and labour force, which responds negatively in the state-space IRFs. Regarding the last point however, the state-space model is probably more credible than the VAR, as declining labour input as a response to technology shocks is a well documented empirical fact (Galí & Rabanal, 2004). Now consider the IRFs generated for a (expansionary) monetary policy shock. Again, the state-space IRFs match the direction of those from the VAR for the non-exogenous variables, except for investment, which is markedly different. The state-space predicts an expansion in investment

Index	Exogenous States	Endogenous States	BIC
1	r		4.30×10^6
2		r	4.28×10^6
3	l		4.14×10^6
4	g		4.12×10^6
5		l	4.12×10^6
6		g	4.12×10^6
7	i		3.82×10^6
8	l	r	3.81×10^6
9	r	l	3.78×10^6
10		i	3.75×10^6

Table 3.9: Large sample (n=100,000) simulation structure learning results for RBC data over one run by maximising the BIC score function only. The top 10 models are shown. Since there are no conditional independence tests, all models are **Valid**. The ground-truth model has **Exogenous States** g and z and **Endogenous State** k. Using this approach the ground truth was ranked 6353 out of 16866 models.

after the monetary policy shock, whereas the VAR predicts a decrease. Again, the state-space model seems to be in agreement with empirical work in this area such as that of Christiano et al. (2005).

The purpose of this exercise is not to argue that this is the optimal model for macroeconomic behaviour in the United States, but rather, it is to demonstrate that the algorithm provides sensible results when used outside the laboratory setting provided by the simulated data used in previous sections. However, in this case there may be considerable concern about misspecification, which could never be fully assuaged. Therefore, I will forgo any deeper analysis of the IRFs produced by this model. It is entirely possible to use this approach to estimate a model that is worthy of such further discussion, perhaps even to go so far as specifying a micro-founded model that is consistent with the conclusions of the algorithm, but this is for now left as an avenue for future research.

3.5.4 Alternative Approaches

This section will briefly present some other strategies suggested by the literature, which I found to be less successful in this application, and will briefly discuss some reasons why that was the case. For ease of comparison only results for the RBC model are shown for each approach.

3.5.4.1 Score Maximisation

The main approach implements a *score function* to differentiate between models only when more than one survives the conditional independence tests. As discussed in Section 3.2.1, it is at least theoretically possible to learn the structure of a DAG using only the score function. In order to implement this, a brute force attempt is made to maximise the BIC score (Schwarz et al., 1978) and the log-likelihood over the set of all possible state-space models, which as discussed in Section 3.3 is a subset all possible DAGs. Tables 3.9 and 3.10 respectively display full sample (n=100,000) results using these two different score functions.

Index	Exogenous States	Endogenous States	Log-Likelihood
1		k g r	1.949×10^7
2	y c	k z g	1.948×10^7
3	y c z r	k g i	1.946×10^7
4	y l r i	k z g	1.945×10^7
5	i	k z g	1.945×10^7
6	c l z	k g i	1.945×10^7
7	y c l i	k z g	1.944×10^7
8	c z r w	k g i	1.942×10^7
9	y c l r	k z g	1.942×10^7
10	y c r i	k z g	1.941×10^7

Table 3.10: Large sample (n=100,000) simulation structure learning results for RBC data over one run by maximising the log-likelihood function only. The top 10 models are shown. Since there are no conditional independence tests all models are **Valid**. The ground-truth model has **Exogenous States** g and z and **Endogenous State** k. Using this approach the ground truth was ranked 10606 out of 16866 models.

The results are much weaker compared to the preferred hybrid approach with conditional independence testing. The ground-truth model obtains a very low rank using both scores, although the BIC seems to perform somewhat better, ranking the ground truth at 6353 which is higher than rank 10606 from the log-likelihood. Now consider the top 10 ranked models generated by each score function, as displayed by Tables 3.9 and 3.10. The bias towards complexity from maximising the log-likelihood is visible here, where all but the first ranked model have too many state variables. On the other hand, the BIC has regulated the complexity of the chosen models, but has done so too aggressively, and chosen models with too few state variables. The top 10 for the AIC score was the same as for the BIC and is therefore not shown. These results show that while sorting by score may be helpful for selection among already similar models, it does not seem to be very effective when used as the only model selection criteria in this application.

3.5.4.2 2-phase Restricted Maximisation (Hybrid Algorithm)

It is also interesting to compare the performance of some existing structure learning methods such as the hybrid rsmax2 algorithm (Scutari et al., 2014). The DAG estimated using this algorithm on a large sample of $100,000$ observations is shown in Figure 3.6. The primary limitation of this approach is that the algorithm will search over the space of all possible DAGs. This is clear when considering the estimated DAG, which does not conform to equations (3.2) - (3.4). In other words, the estimated solution must be incorrect because it is not a state-space model. That said, there is some extent to which important characteristics of the ground-truth solution to the RBC model can be seen here. For example, the exogenous states z and g depend only on their own lag. The IRFs produced by this DAG (shown in Figure 3.7) also seem to be very close to those of the original simulation. Therefore, one might conclude that although this approach did not yield the correct solution, it did recover some sense of causality from the underlying DGP.

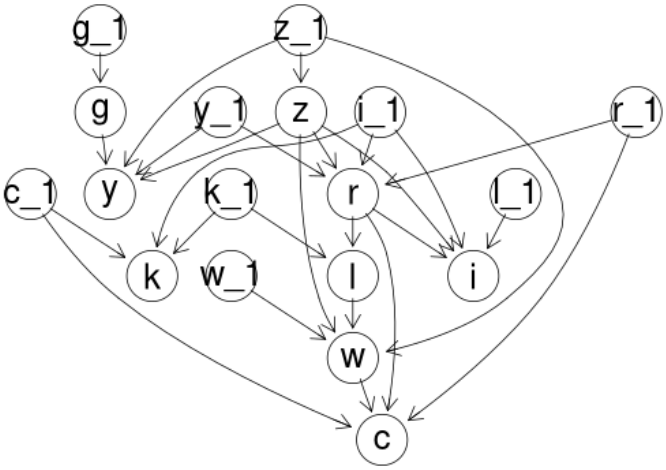


Figure 3.6: DAG fit to RBC data using rsmax2 hybrid constraint-based algorithm (Scutari et al., 2014). Additional constraint was added such that lagged values were forced to be root nodes (as they are in the ground truth).

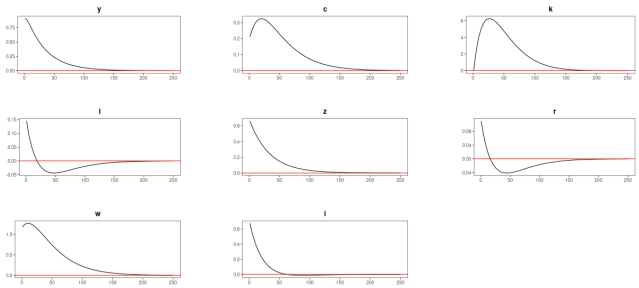


Figure 3.7: IRFs generated by DAG fit to RBC data using rsmax2 hybrid constraint-based algorithm (Scutari et al., 2014).

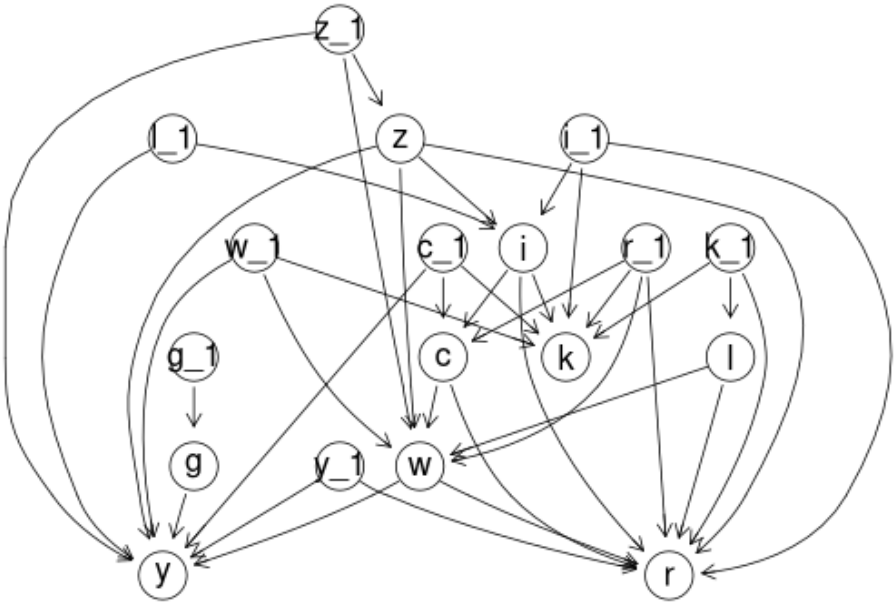


Figure 3.8: DAG fit to RBC data using PC-algorithm (Kalisch & Bühlmann, 2007). Additional constraint was added such that lagged values were forced to be root nodes (as they are in the ground truth).

3.5.4.3 PC-algorithm (Constraint-Based Algorithm)

Finally, consider the PC constraint-based structure learning algorithm (Spirtes et al., 2000) (Kalisch & Bühlmann, 2007). The large sample estimated DAG is shown in Figure 3.8. The result and conclusion here mirror those for the hybrid algorithm in many ways, however, this approach is somewhat less successful. In particular, the generated IRFs (Figure 3.9) show the time-series diverging because the estimated solution is non-stationary.

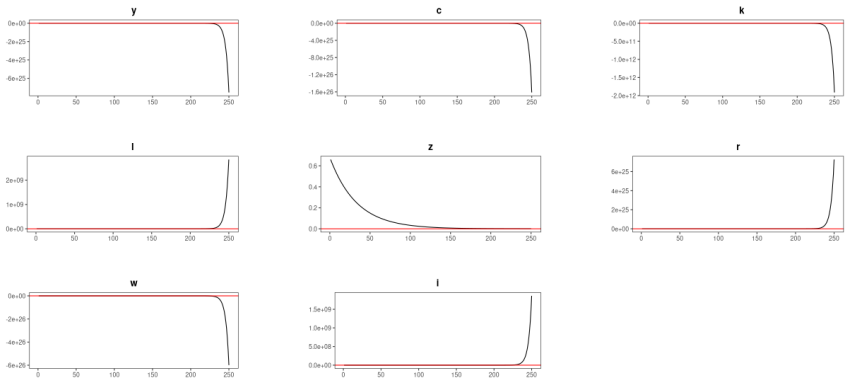


Figure 3.9: IRFs generated by DAG fit to RBC data using PC-algorithm (Kalisch & Bühlmann, 2007).

3.6 Conclusion

This paper has introduced a series of tests and an algorithm for data-driven causal discovery of macroeconomic state-space models. These tests are asymptotically consistent, and have been shown to perform well on at least relatively simple data sets given a realistic sample size. Results derived using this strategy can be used to gain insight into prominent debates in the DSGE literature. This result constitutes a concrete example of an application in which DAGs and the causal discovery toolkit more broadly can be used in empirical economics. This approach comes with a number of benefits, chief among them that it makes no assumptions about which particular relationships are present in the ground-truth DSGE model, a property that I refer to as agnosticism.

Much work remains to be done however, as this study has uncovered a number of limitations. In particular, the methodology presented here has limited scalability, in both a statistical and computational sense, to higher dimensional inputs. As more observables are provided, the number of potential state-space models increases exponentially, and as a result, so too do the statistical power and computational resources required. Since more realistic models will invariably require higher dimensional inputs, whether they come from simply adding more observables or allowing for the possibility of additional structural breaks or frictions, and due to the possibility of misspecification as discussed in Section 3.3.6 it is of paramount importance to relax these constraints. Clearly, more research is required here, but from the current vantage point it is unclear how successful this will be. Using a greedy search algorithm could possibly improve the computational performance, while using other multiple-comparison tests (see Lee and Lee (2018) for a summary of alternatives) could increase statistical power, as the Bonferroni correction is extremely conservative. Nonetheless, these adaptations may be insufficient to overcome the generally poor dimensionality scaling of this approach, which would leave it constrained to low dimensional settings where it is perhaps less useful relative to explicit model selection.

One relatively straightforward yet valuable extension to this paper would be to identify micro-founded DSGE models to match the reduced forms identified by the algorithm over some real data sets, and comparing these to the state of the art in the literature in order to see how closely they match. This exercise would not only serve as a means of validating existing macroeconomic models, but also as an example of how empirical researchers can implement the techniques introduced in this paper in applied work. Furthermore, I have attributed the relatively poor performance of existing structure learning algorithms in the context of state-space models to the large search space within which they operate. This issue seems unlikely to be unique to this particular application, suggesting that it would improve the usefulness of these structure-learning algorithms if they allowed the user to reduce the size of the search space by using domain-specific knowledge or commonly accepted rules of thumb to define more general constraints on the nature of the solutions generated.

Imbens (2020) states that DAGs are most useful, "in complex models with many variables that are not particularly popular in empirical economics." The implication is that there is limited scope

for the application of (algorithmic) causal discovery in economics. However, what this paper has shown is the opposite. DAGs and causal discovery tools are useful primarily in relatively low-dimensional cases, and the additional benefit that they provide is the ability to choose amongst models in a data-driven way. Furthermore, the converse of this could equally be true; complex models with many variables are not popular in empirical economics *because* there is a lack of tools that make these problems tractable. Particularly in the context of macroeconomics it seems that complex models such as that of Smets and Wouters (2007) are becoming increasingly popular. This paper has made an attempt at bringing data and computational power to bear on unwieldy problems to derive interpretable solutions, and further research along these veins will hopefully yield tools that help applied researchers to more easily consult data to evaluate theoretical considerations and deal with complexity.

Bibliography

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

Angrist, J. D., & Pischke, J.-S. (2014, April). *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press. https://ideas.repec.org/b/pup/pbooks/10363.html

Bazinas, V., & Nielsen, B. (2015). *Causal transmission in reduced-form models*. Nuffield College.

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. In *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.

Christiano, L. J., Eichenbaum, M., & Evans, C. L. (2005). Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of political Economy*, *113*(1), 1–45.

Christiano, L. J., Eichenbaum, M. S., & Trabandt, M. (2018). On dsge models. *Journal of Economic Perspectives*, *32*(3), 113–40. https://doi.org/10.1257/jep.32.3.113

Colombo, D., & Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, *15*(1), 3741–3782.

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

Driver, C. (2000). Capacity utilisation and excess capacity: Theory, evidence, and policy. *Review of Industrial Organization*, *16*(1), 69–87.

Ermon, S. (2017, January). Learning in directed models. https://ermongroup.github.io/cs228-notes/learning/directed/

Fernandez-Villaverde, J., Rubio-Ramirez, J. F., & Schorfheide, F. (2016). Solution and estimation methods for dsge models. In *Handbook of macroeconomics* (pp. 527–724, Vol. 2). Elsevier.

FRED. (2020). Federal reserve bank of st.louis and us. Retrieved July 12, 2020, from https://fred.stlouisfed.org/

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

Fuhrer, J., & Moore, G. (1995). Inflation persistence. *The Quarterly Journal of Economics*, *110*(1), 127–159.

Fuhrer, J. C. (2000). Habit formation in consumption and its implications for monetary-policy models. *American Economic Review*, *90*(3), 367–390.

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

Galí, J., & Gertler, M. (1999). Inflation dynamics: A structural econometric analysis. *Journal of monetary Economics*, *44*(2), 195–222.

Galí, J., & Rabanal, P. (2004). Technology shocks and aggregate fluctuations: How well does the real business cycle model fit postwar us data? *NBER macroeconomics annual*, *19*, 225–288.

Geweke, J., & Amisano, G. (2012). Prediction with misspecified models. *American Economic Review*, *102*(3), 482–486.

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

Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., & Schölkopf, B. (2020). Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, *21*(89), 1–53.

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, *58*(4), 1129–79.

Jordà, 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.

Krolzig, H.-M., & Hendry, D. F. (2001). Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics and Control*, *25*(6-7), 831–866.

Kydland, F. E., & Prescott, E. C. (1982). Time to build and aggregate fluctuations. *Econometrica: Journal of the Econometric Society*, 1345–1370.

Lee, S., & Lee, D. K. (2018). What is the proper way to apply the multiple comparison test? *Korean journal of anesthesiology*, *71*(5), 353–360.

Levin, A. T., López-Salido, J. D., Nelson, E., & Yun, T. (2008). Macroeconometric equivalence, microeconomic dissonance, and the design of monetary policy. *Journal of Monetary Economics*, *55*, S48–S62.

Levin, A. T., Natalucci, F. M., Piger, J. M., et al. (2004). The macroeconomic effects of inflation targeting. *Review-Federal Reserve Bank of Saint Louis*, *86*(4), 51–8.

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. In *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.

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

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.

Sack, B., & Wieland, V. (2000). Interest-rate smoothing and optimal monetary policy: A review of recent empirical evidence. *Journal of Economics and Business*, *52*(1-2), 205–228.

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. (2000). *Causation, prediction, and search*. MIT press.

Spirtes, P., & Zhang, K. (2016). Causal discovery and inference: Concepts and recent methodological advances. *Applied informatics*, *3*(1), 3.

Srivastava, M. S. (2005). Some tests concerning the covariance matrix in high dimensional data. *Journal of the Japan Statistical Society*, *35*(2), 251–272.

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

Steinsson, J. (2003). Optimal monetary policy in an economy with inflation persistence. *Journal of Monetary Economics*, *50*(7), 1425–1456.

Strobl, E. V., Zhang, K., & Visweswaran, S. (2019). Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, *7*(1).

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

Wang, G., Zou, C., & Wang, Z. (2013). A necessary test for complete independence in high dimensions using rank-correlations. *Journal of Multivariate Analysis*, *121*, 224–232.

Appendix

A Faithfulness Proof

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 (3.4). Therefore, the difference must be that H partitions one or more of the variables a 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 the other variable types y and z.

Consider all possible cases to see that H must produce a contradiction:

Case 1: $a \in G_y$ and $a \in H_x$

G has fewer state variables that h, which therefore does not satisfy the MSV criteria. Contradiction.

Case 2: $a \in G_y$ and $a \in H_z$

(3.6) fails because there is a direct path from $\mathbf{x}_{t-1}$ to a in G. Contradiction.

Case 3: $a \in G_x$ and $a \in H_y$

(3.5) 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 $a \in H_z$

(3.6) fails because there is a direct path from $\mathbf{x}_{t-1}$ to a in G. Contradiction.

Case 5: $a \in G_z$ and $a \in H_y$

(3.5) 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 $a \in H_x$

(3.5) fails because there is a direct path from a to any time t endogenous variable in G. Contradiction. □

B Testing Validation

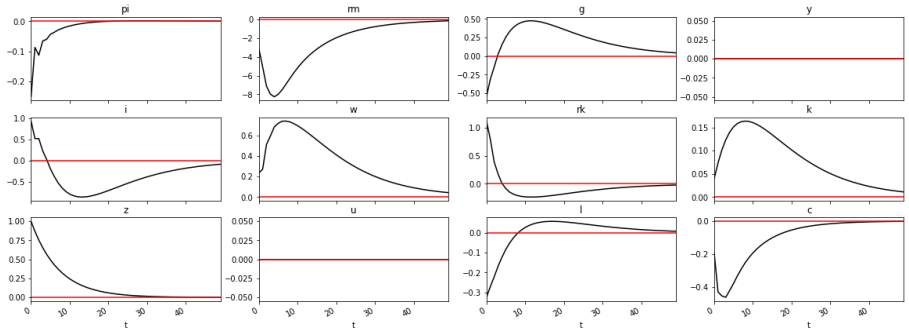
empirical Size	Alpha	Difference	n	m	Repetitions
0.041	0.010	-0.031	10	5	1000
0.104	0.050	-0.054	10	5	1000
0.015	0.010	-0.005	100	5	1000
0.052	0.050	-0.002	100	5	1000
0.021	0.010	-0.011	10000	5	1000
0.031	0.050	0.019	10000	5	1000
0.118	0.010	-0.108	10	25	1000
0.239	0.050	-0.189	10	25	1000
0.012	0.010	-0.002	100	25	1000
0.041	0.050	0.009	100	25	1000
0.019	0.010	-0.009	10000	25	1000
0.053	0.050	-0.003	10000	25	1000
0.463	0.010	-0.453	10	50	1000
0.699	0.050	-0.649	10	50	1000
0.008	0.010	0.002	100	50	1000
0.076	0.050	-0.026	100	50	1000
0.008	0.010	0.002	10000	50	1000
0.063	0.050	-0.013	10000	50	1000

Table 3.11: empirical validation of significance level of Srivastava (2005) test.

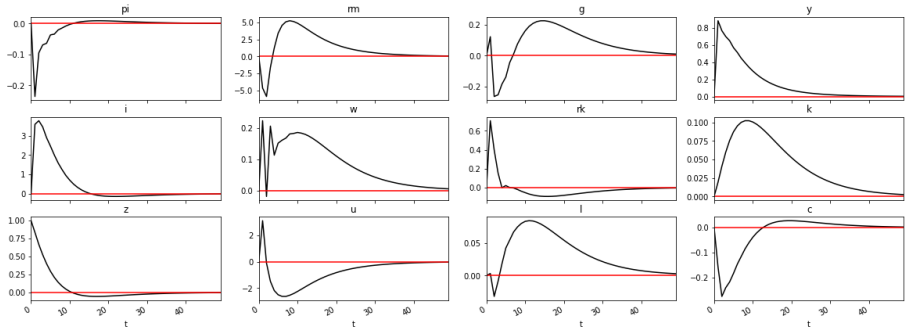
empirical Power	Alpha	n	Correlation	m	Repetitions
0.077	0.010	10	0.100	5	1000
0.110	0.050	10	0.100	5	1000
0.379	0.010	100	0.100	5	1000
0.507	0.050	100	0.100	5	1000
1.000	0.010	10000	0.100	5	1000
1.000	0.050	10000	0.100	5	1000
0.422	0.010	10	0.100	25	1000
0.553	0.050	10	0.100	25	1000
0.999	0.010	100	0.100	25	1000
1.000	0.050	100	0.100	25	1000
1.000	0.010	10000	0.100	25	1000
1.000	0.050	10000	0.100	25	1000
0.824	0.010	10	0.100	50	1000
0.909	0.050	10	0.100	50	1000
1.000	0.010	100	0.100	50	1000
1.000	0.050	100	0.100	50	1000
1.000	0.010	10000	0.100	50	1000
1.000	0.050	10000	0.100	50	1000
0.436	0.010	10	0.325	5	1000
0.518	0.050	10	0.325	5	1000
1.000	0.010	100	0.325	5	1000
1.000	0.050	100	0.325	5	1000
1.000	0.010	10000	0.325	5	1000
1.000	0.050	10000	0.325	5	1000
0.953	0.010	10	0.325	25	1000
0.969	0.050	10	0.325	25	1000
1.000	0.010	100	0.325	25	1000
1.000	0.050	100	0.325	25	1000
1.000	0.010	10000	0.325	25	1000
1.000	0.050	10000	0.325	25	1000
0.996	0.010	10	0.325	50	1000
0.999	0.050	10	0.325	50	1000
1.000	0.010	100	0.325	50	1000
1.000	0.050	100	0.325	50	1000
1.000	0.010	10000	0.325	50	1000
1.000	0.050	10000	0.325	50	1000
0.863	0.010	10	0.550	5	1000
0.871	0.050	10	0.550	5	1000
1.000	0.010	100	0.550	5	1000
1.000	0.050	100	0.550	5	1000
1.000	0.010	10000	0.550	5	1000
1.000	0.050	10000	0.550	5	1000
0.997	0.010	10	0.550	25	1000
1.000	0.050	10	0.550	25	1000
1.000	0.010	100	0.550	25	1000
1.000	0.050	100	0.550	25	1000
1.000	0.010	10000	0.550	25	1000
1.000	0.050	10000	0.550	25	1000
1.000	0.010	10	0.550	50	1000
1.000	0.050	10	0.550	50	1000
1.000	0.010	100	0.550	50	1000
1.000	0.050	100	0.550	50	1000
1.000	0.010	10000	0.550	50	1000
1.000	0.050	10000	0.550	50	1000
0.996	0.010	10	0.775	5	1000
0.997	0.050	10	0.775	5	1000
1.000	0.010	100	0.775	5	1000
1.000	0.050	100	0.775	5	1000
1.000	0.010	10000	0.775	5	1000
1.000	0.050	10000	0.775	5	1000
1.000	0.010	10	0.775	25	1000
1.000	0.050	10	0.775	25	1000
1.000	0.010	100	0.775	25	1000
1.000	0.050	100	0.775	25	1000
1.000	0.010	10000	0.775	25	1000
1.000	0.050	10000	0.775	25	1000
1.000	0.010	10	0.775	50	1000
1.000	0.050	10	0.775	50	1000
1.000	0.010	100	0.775	50	1000
1.000	0.050	100	0.775	50	1000
1.000	0.010	10000	0.775	50	1000
1.000	0.050	10000	0.775	50	1000
1.000	0.010	10	1.000	5	1000
1.000	0.050	10	1.000	5	1000
1.000	0.010	100	1.000	5	1000
1.000	0.050	100	1.000	5	1000
1.000	0.010	10000	1.000	5	1000
1.000	0.050	10000	1.000	5	1000
1.000	0.010	10	1.000	25	1000
1.000	0.050	10	1.000	25	1000
1.000	0.010	100	1.000	25	1000
1.000	0.050	100	1.000	25	1000
1.000	0.010	10000	1.000	25	1000
1.000	0.050	10000	1.000	25	1000
1.000	0.010	10	1.000	50	1000
1.000	0.050	10	1.000	50	1000
1.000	0.010	100	1.000	50	1000
1.000	0.050	100	1.000	50	1000
1.000	0.010	10000	1.000	50	1000
1.000	0.050	10000	1.000	50	1000

Table 3.12: empirical validation of power of Srivastava (2005) test against data generated from a normal distribution where the off-diagonal elements of the covariance matrix all take on the value specified by *correlation*.

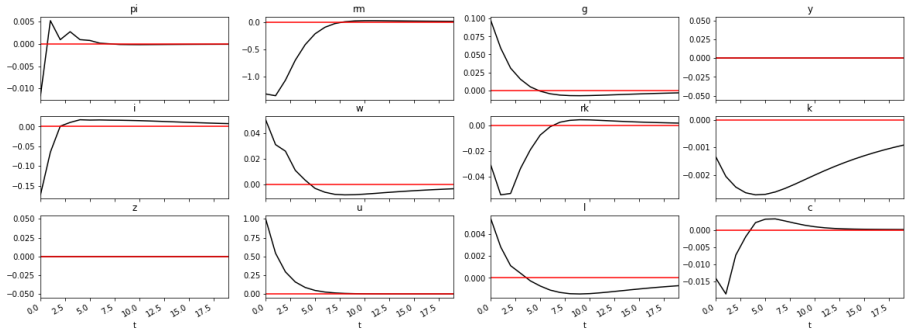
C Real Data IRFs



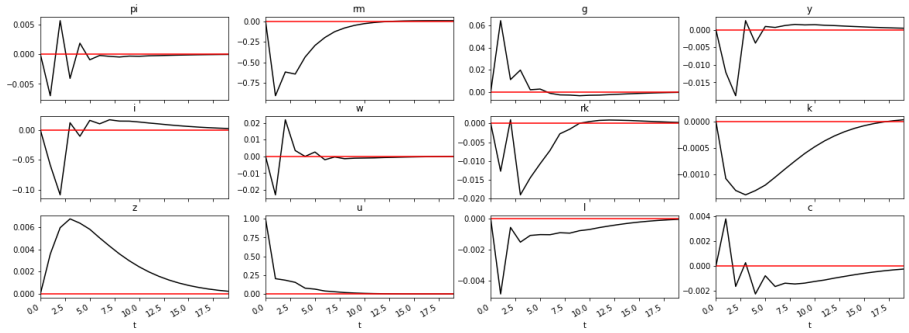
(a) TFP (State-Space)



(b) TFP (VAR)

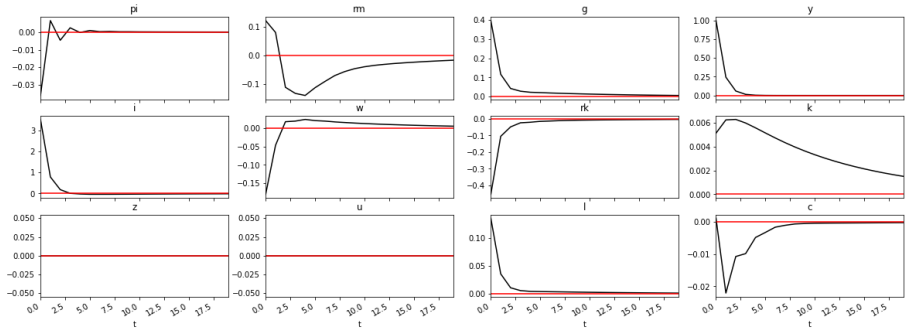


(c) Unemployment (State-Space)

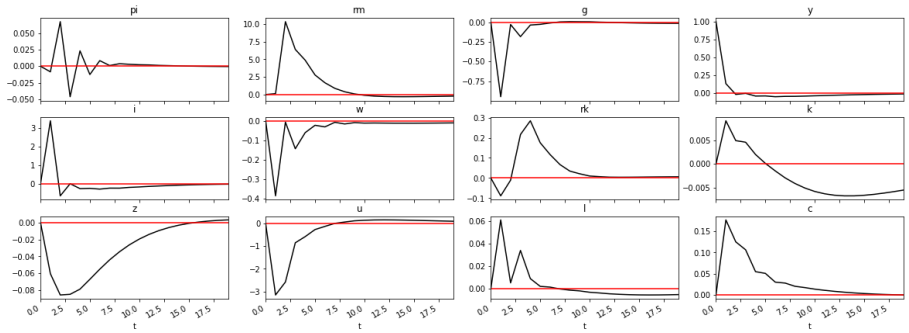


(d) Unemployment (VAR)

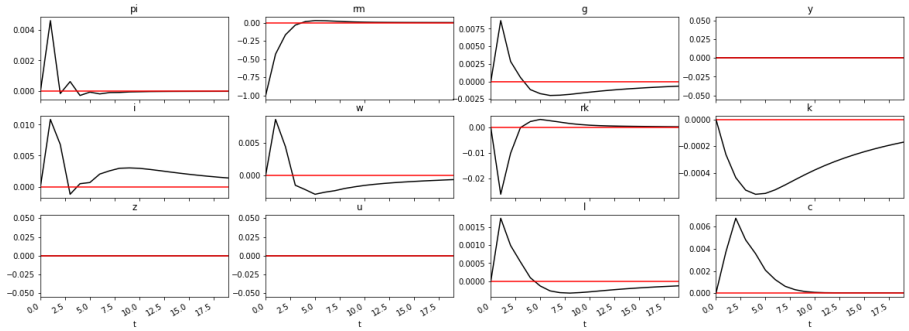
Figure 3.10: IRFs to a standard deviation shock to technology and unemployment in both the estimated state-space model and an unconditional VAR(1) fit to US macroeconomic data for the period 1985-2005.



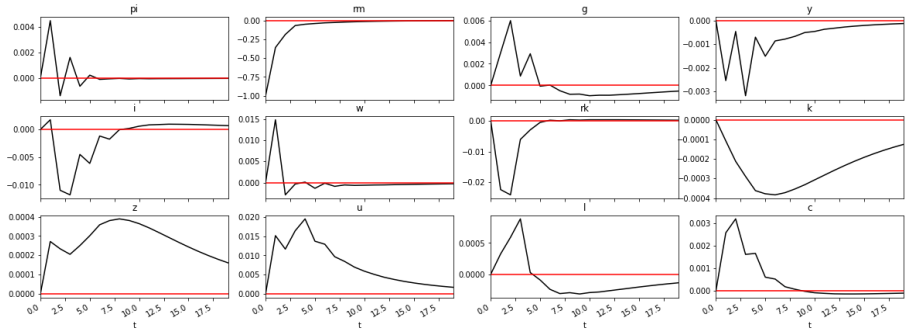
(a) Real Output (State-Space)



(b) Real Output (VAR)



(c) Federal Funds Rate (State-Space)



(d) Federal Funds Rate (VAR)

Figure 3.11: IRFs to a standard deviation shock to real output and the policy rate in both the estimated state-space model and an unconditional VAR(1) fit to US macroeconomic data for the period 1985-2005.