Economics Informed Neural Networks for New Keynesian Models

Bachelorarbeit

zur Erlangung des akademischen Grades “Bachelor of Science” (B.Sc.)

eingereicht beim Prüfungsausschuss für den Bachelorstudiengang

Economics (Politische Ökonomik)

der

Fakultät für Wirtschafts- und Sozialwissenschaften der

Ruprecht-Karls-Universität Heidelberg

2024

Malte Tölle

geboren am 28.02.1996 in Hildesheim

Matrikelnummer: 3723292

Datum der Abgabe: 5. September 2024

Betreuer: Jun. Prof. Joep Lustenhouwer

Lehrstuhl für Makroökonomie

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig und ohne unerlaubte fremde Hilfe verfasst habe und dass alle wörtlich oder sinngemäß aus Veröffentlichungen entnommenen Stellen dieser Arbeit unter Quellenangabe einzeln kenntlich gemacht sind.

Heidelberg, den 5. September 2024

Malte Tölle

Economics Informed Neural Networks for New Keynesian Models

Malte Tölle

September 5, 2024

Contents

List of abbreviations

BP Backpropagation

CRRA Constant Relative Risk Aversion

DSGE Dynamic Stochastic General Equilibrium

L-BFGS Limited-memory Broyden–Fletcher–Goldfarb–Shanno

LM Levenberg-Marquardt

NKM New-Keynesian Model

NKPC New-Keynesian Phillips Curve

SW Smets-Wouters

VAR Vector Autoregression

List of Figures

List of Tables

1 Introduction

In the perfect world of flexible macroeconomics the central bank's job is somewhat superfluous, money does not affect real variables (technology, output, employment) but only price levels. However, real evidence suggest the opposite (Romer and Romer, 2004; Mankiw and Reis, 2018). Empirical work has found that money movements are followed by output movements in the same direction, i.e. a monetary contraction leads to a decline in output. The so called new Keynesian school argues with real world imperfections that counteract the implications of perfect competition and information as well as zero transaction costs"(Galí, 2018). Prices are assumed to only adjust sluggishly to the equilibrium predicted from the classical economics.

While Keynes (1936) discussed his ideas from a classical high macroeconomic standpoint that markets do not by themselves find back to an equilibrium state without interventions from the state, New Keynesian economists argue with a model that has its foundations in microeconomics. In a famous paper that produced the Lucas' critique the adjustment of expectations have been found to ruin the Keynes' ideas (Lucas, 1976). Inflation has been found to move anti-proportional to unemployment, low unemployment would come in line with high inflation rates due to expansionary (governmental) spending. However, as expectations adjust and the anti-cyclical spending of governments in times of economic crises are anticipated this effect disappears. Consumers expect the rising inflation and have already priced in the higher price level in their consumption decisions.

By basing their arguments on microfounded evidence New Keynesian economists produced a model that revives Keynes ideas with a solid mathematical evidence. Its strength lies in its explainability of real world macroeconomic phenomena by employing rational expectation models that account for the expectation of consumers of future values for e.g. inflation (Sargent and Wallace, 1976). However, with their explainability comes the downside of a worse fit to data and prediction performance compared to correlational regression models such as variational autoregression (VAR)"(Stock and Watson, 2001). Artificial intelligence has shown impressive results on many tasks most recently in natural language processing (e.g. ChatGPT[1] or Llama3[2]). Their flexible form promises a possible approximation of arbitrary functions (Hornik et al., 1989). However, while economists and especially New-Keynesian models (NKM) strive for maximal explainability AI models have an inherent black-box nature due to their non-linearity and manifold of computational operations (Molnar, 2022). But recent work has shown promising performance of NNs on forecasting macroeconomic timeseries (Tänzer, 2024). This work aims at additional to a pure data regression task condition a model on the predictions of a explainable macroeconomic NKM. This represents a compromise between explainability and predictive performance respectively. In this work an neural network (NN) is used for enhancing the estimation of a NKM, more precisely the famous Smets and Wouters (2007) (SW) model. While the broad direction of the model is determined by Bayesian estimation, the model is fine-tuned with an neural network that increases the predictive forcasting performance.

[1]`https://chatgpt.com/`
[2]`https://ai.meta.com/blog/meta-llama-3/`

2 Background

2.1 New Keynesian Models

Although the complexity of the final, estimated model is higher for the sake of understanding we will start by introducing a baseline NKM. The model is inspired by the lecture *Computational Macroeconomics* from Jun. Prof. Joep Lustenhouwer at Heidelberg University[3].

Goods Market. In this model households consume a basket of goods based on their utility maximization. Intermediate goods producer produce a basket of goods under *staggered Calvo pricing* (see next Section). Final goods producer take prices of intermediate goods as given and sell a basket of goods to households under perfect competition.

Labor Market. The intermediate firms hire labor to maximize their profits under monopolistic competition. The labor is supplied by households based on their utility function. Wages are assumed to be perfectly flexible, but households have some power in their negotiations due to e.g. unions.

Financial Markets For now financial markets are modeled with the investment of households into a one-period riskless asset.

[3]`https://www.awi.uni-heidelberg.de/en/chairs/economic-policy/team/junprof-dr-joep-lustenhouwer`

Sticky Prices

In New-Keynesian economics the prices are believed to adjust to a new level in a staggered manner, which is backed by a body of empirical work (Alvarez et al., 2006). Thus, they are subject to some form of inflexibility i.e. stickiness.

Calvo Pricing. Calvo (1983) assume a state-independent probability θ for each firm *for being stuck* with its price for the upcoming period. Importantly, this probability is independent from the time the firm changed prices the last time (time-dependent stickiness as opposed to state-dependent stickiness).

Households

Households maximize their expected utility according to

$$E \sum_{t=0}^{\infty} \beta^t U(C_t, N_t) \quad ,0 \quad 0 < \beta < 1 \quad , \tag{1}$$

where C_t denotes the basket of goods produced from the final producer, N_t the hours worked. The factor β_t weights future utility, making it less valuable as time progresses, according to the household's time preference. The nominal and real budget constraints are

$$\begin{aligned} P_t C_t + P_t B_{t+1} &\leq (1 + i_{t-1}) P_{t-1} B_t + P_t W_t N_t + P_t \Xi_t \\ C_t + B_{t+1} &\leq \frac{1 + i_{t-1}}{1 + \pi_t} B_t + W_t N_t + \Xi_t \ . \end{aligned} \tag{2}$$

Households then choose optimal paths for C_t, N_t, and B_{t+1} with taking prices P_t, $P_t W_t$, and i_t as given. A No-Ponzi game constraint is used to rule out explosive debt by conditioning the value of debt to converge to 0 in the long run

$$\lim_{T \to \infty} \frac{B_T}{(1 + i_0)(1 + i_1) \cdots (1 + i_{T-1})} = 0 \ . \tag{3}$$

For this introductory example a constant relative risk aversion (CRRA) utility function is chosen:

$$U(C_t, N_t) = \frac{C_t^{1-\sigma}}{1 - \sigma} - \frac{N_t^{1+\eta}}{1 + \eta} \quad , \tag{4}$$

which implies concave utility out of consumption and convex dis-utility out of hours worked. After setting up the Langragian with the real budget constraint, taking derivatives w.r.t. C_t, N_t, and B_{t+1} we obtain the consumption Euler and labor-leisure equation:

$$\begin{aligned} C_t^{-\sigma} &= \beta E_t C_{t+1}^{-\sigma} \frac{1+i_t}{1+E_t\pi_{t+1}} \\ \hat{C}_t &= E_t\hat{C}_{t+1} - \sigma^{-1}\left[\hat{i}_t - E_t\hat{\pi}_{t+1}\right] \quad \text{(log-linearized)} \\ N_t^{\eta} &= C_t^{-\sigma}W_t \\ \hat{N}_t &= \frac{1}{\eta}\left[\hat{W}_t - \sigma\hat{C}_t\right] \quad \text{(log-linearized)} \quad . \end{aligned} \tag{5}$$

The Euler consumption equation summarizes the trade-off between consumption today and tomorrow indicating an inverse relationship between expected interest rate and consumption today. The labor-leisure equation states that the marginal gain from working must equal the marginal disutility from working. Capital is excluded from the model as otherwise a small change in interest rate would result in a large response of capital. This can be mitigated in larger models by introducing capital adjustment costs.

Firms and Production

Firms are modeled under monopolitic competition meaning that each firm has some power over setting its prices as their products are non-perfect substitutes. It is assumed that intermediate good producers produce differentiated goods that are sold under monopolistic competition and price stickiness to final goods producers. These final goods producers sell a final consumption good i.e. a basket of intermediate goods to households under perfect competition.

The profit maximization problem of final good producers is

$$\max_{Y_t(j)} = P_t \left(\int_0^1 Y_t(j)^{\frac{\epsilon-1}{\epsilon}} \right) - \int_0^1 P_t(j)Y_t(j)dj \quad , \tag{6}$$

with $Y_t(j)$ being differentiated goods and Y_t the consumption basket. This results in the demand of differentiated goods of:

$$Y_t(j) = \left(\frac{P_t(j)}{P_t}\right)^{-\epsilon} Y_t \quad , \tag{7}$$

which depends on the relative price of good j.

As the baseline NKM abstracts from capital the production function and nominal profits can be written as:

$$Y_t(j) = A_t N_t(j) \quad , \tag{8}$$

$$P_t \Xi_t = [P_t(j)Y_t(j) - P_t W_t Y_t(j)] \quad , \tag{9}$$

where A_t denotes the technology level, which is usually set to one.

The first order condition of optimizing the nominal profits with taking the ability into account for not being ablt to reset its price θ can be written as:

$$\begin{aligned} &\max_{P_t(j)^*} \frac{\epsilon}{\epsilon - 1} E_t \sum_{s=0}^{\infty} \theta^s \beta^s \left(\frac{P_{t+1}}{P_t} \right) Y_{t+s} W_{t+s} \quad , \\ &\tilde{p}_t(j) = E_t(1 - \theta\beta) \sum_{s=0}^{\infty} (\theta\beta)^s \left[\hat{W}_{t+s} + \hat{P}_{t+s} - \hat{P}_t \right] \quad \text{(log-linearized)} \quad , \end{aligned} \tag{10}$$

with $\tilde{p}_t(j)$ denoting the percentage deviation from the steady-state price. Rewriting the first order condition recursively and integrating over all firms yields:

$$p_t = (1 - \theta\beta)\hat{W}_t + \theta\beta E_t[\tilde{p}_{t+1} + \hat{\pi}_{t+1}] \quad . \tag{11}$$

With assuming aggregated inflation is dependend on the amount of firms that can reset their price $\hat{\pi}_t = \frac{1-\theta}{\theta}\tilde{p}_t$ a measure for inflation, the New Keynesian Phillips Curve (NKPC), is derived:

$$\hat{\pi}_t = \frac{(1-\theta)(1-\theta\beta)}{\theta}\hat{W}_t + \beta E_t \hat{\pi}_{t+1} \quad . \tag{12}$$

While the traditional Philipps curve has empirically connected inflation to output and unemployment the NKPC is derived from microfoundations and is, thus, immune to the Lucas critique.

With assuming $\hat{Y}_t = \hat{C}_t$ and $Y_t = N_t$ the labor-leisure equation can be rewritten as

$$\hat{W}_t = (\sigma + \eta)\hat{Y}_t \ . \tag{13}$$

Inserting this into the NKPC we obtain the first equation of a simple NKM:

$$\hat{\pi}_t = \kappa\hat{Y}_t + \beta E_t\hat{\pi}_{t+1} \quad \text{with} \quad \kappa = \frac{(1-\theta)(1-\theta\beta)(\eta+\sigma)}{\theta} \quad . \tag{14}$$

Using the same substitutes for the consumotion Euler equation the second equation is obtained:

$$\hat{Y}_t = E_t\hat{Y}_{t+1} - \sigma^{-1}[\hat{i}_t - E_t\hat{\pi}_{t+1}] \quad . \tag{15}$$

Monetary Policy

The only missing part that is a rule for monetary policy i.e. setting the interest rate, which is usually set ad hoc with a Taylor rule:

$$i_t = \phi_1\pi_t + \phi_2 y_t \quad . \tag{16}$$

2.2 DSGE Models

DSGE models are based on a system of equations that are derived from economic theory i.e. microfoundations such as the NKM described above (Christiano et al., 2018). With these equations actions from participants in the economy can be linked to counterfactual question answering and analysis of shock influences. This possibility of answering *what if* questions makes them a popular tool among policy makers.

Strengths and Weaknesses

DSGE models claim superiority over completely stochastic models such as VARs by their microfounded nature and their resulting interpretability for policy design. However, for the sake of data fit a large number of ad hoc economic mechanisms are introduced to fit persistence properties of the data rather than because of theoretical explainability of the shocks. This results in a large number of unexplained shocks, which are often highly persistent. Modern behavioral economics has shown that the assumption of pure rational expectation agents is flawed (Stiglitz, 2018). However, multiple aspects that are not modeled in VARs are featured in DSGE models such as the imposition of budget constraints, a consistent story for how agents behave and a coherent handling of expectations. These enable DSGE models to answer *what if* questions in policy design and forecasting with a sound theoretical foundation better than VARs.

2.3 The Smets-Wouters Model

One very popular DSGE model is the Smets and Wouters (2007) model. The model's equations are introduced in the following. All variables are log-linearized around their steady-state balanced growth path. Generally, starred variables denote steady-state values.

In summary, the model introduces adjustment costs, utilisation costs and habit persistence to make variables more sluggish by giving random shocks a more long-lasting effect. It is criticized whether these newly introduced shocks are really microfounded or rather still ad-hoc and thus hardly immune to the Lucas critique.

Supply Side

The aggregate production function is given by

$$y_t = \phi_p(\alpha k_t^s + (1-\alpha)t_t + \epsilon_t^a) \ , \tag{17}$$

where y_t denotes the gross domestic product (GDP), k_t^s capital in use, l_t labor input, and ϵ_t^a total factor productivity. ϕ_p is one plus the share of fixed costs in production, reflecting the presence of fixed costs in production. The capital in use is determined by the level of capital investment from the previous period (k_{t-1}) and a capacity utilisation variable (z_t):

$$k_t^s = k_{t-1} + z_t \ , \tag{18}$$

s.t. newly installed capital becomes only effective with a one-quarter lag. Adjusting the amount of capital in use is associated with a cost. So, optimisation conditions for producers mean the rate of capacity utilisation is linked to the marginal productivity of capital:

$$z_t = z_1 r_t^k \ , \tag{19}$$

where the marginal productivity of capital is a function of the capital labor ratio and the real wage:

$$r_t^k = -(k_t - l_t) + w_t \ . \tag{20}$$

The total factor productivity ϵ_t^a follows a first-order autoregressive process

$$\epsilon_t^a = \rho_a \epsilon_{t-1}^a + \eta_t^a \ . \tag{21}$$

Demand Side

The expenditure formulation of the aggregate ressource constraint is given by

$$y_t = c_y c_t + i_y i_t + z_y z_t + \epsilon_t^g \ , \tag{22}$$

where c_t denotes consumption, i_t investment and ϵ_t^g exogenous spending. $c_y = 1 - g_y - i_y$ is the steady-state share of consumption with g_y and i_y being respectively the steady-state exogenous spending-output ration and investment-output ratio. The steady-state investment-output ratio is given by $i_y = (\gamma - 1 + \delta)k_y$, where γ is the steady-state growth rate, δ denotes the depreciation rate of capital, and k_y stands for the steady-state capital-output ratio. It is assumed there are costs associated with having high rates of capacity utilisation, s.t. z_t is included in the formulation, which is multiplied by $z_y = R_*^k k_y$ with R_*^k denoting the steady-state rental rate of capital. Exogenous spending follows a first-order autoregressive process and is also affected by the productivity shock:

$$\epsilon_t^g = \rho_g \epsilon_{t-1}^g + \eta_t^g + \rho_{ga} \eta_t^a \ , \tag{23}$$

which is motivated by the empirical finding that exogenous spending also includes net exports, which can be affected by domestic productivity developments.

Consumption

Consumption is determined by the Euler equation with a backward looking element, which represents habit formation:

$$c_t = c_1 c_{t-1}(1 - c_1)E_t c_{t+1} + c_2(l_t - E_t l_{t+1}) - c_3(r_t - E_t \pi_{t+1} + \epsilon_t^b) \ , \tag{24}$$

where c_1, c_2, c_3[4] are constant parameters, r_t denotes the interest rate on a one-period safe bond and ϵ_t^b is a risk premium shock ($\epsilon_t^b = \rho_b \epsilon_{t-1}^b + \eta_t^b$). This shock determines the willingness of households to hold the one-period bond, which can also be described by a preference shock that influences the short-tem consumption saving decision.

Investment

The investment Euler equation is given by

$$i_t = i_1 i_{t-1} + (1 - i_1)E_t i_{t+1} + i_2 q_t + \epsilon_t^i \ , \tag{25}$$

where i_1, i_2[5] are constant parameters again and ϵ_t^i represents a disturbance to the technology process and follow a first-order autoregressive process: $\epsilon_t^i = \rho_i \epsilon_{t-1}^i + \eta_t^i$. The real value of the capital stock with its corresponding arbitrage equation is given by:

$$q_t = q_1 E_t q_{t+1} + (1 - q_1)E_t r_{t+1}^k - (r_t - \pi_{t+1} + \epsilon_t^b) \ .^6 \tag{26}$$

q_t is the main driving force behind investment and depends positively on expected future marginal productivity of capital and negatively on future real interest rates. A shock to investment also affects the capital stock. Finally, the accumulation of installed capital is

[4] $c_1 = \frac{\lambda/\gamma}{1+\lambda/\gamma}$, $c_2 = \frac{(\sigma_c - 1)(W_*^h L_*/C_*)}{\sigma_c(1+\lambda/\gamma)}$, $c_3 = \frac{1-\lambda/\gamma}{(1+\lambda/\gamma)\sigma_c}$

[5] $i_1 = \frac{1}{1+\beta\gamma^{(1-\sigma_c)}}$, $i_2 = \frac{1}{(1+\beta\gamma^{1-\sigma_c}\sigma^2\phi}$ with ϕ being the steady-state elasticity of the capital adjustment cost function.

[6] $q_1 = \beta\gamma^{-\sigma_c}(1-\delta) = \frac{1-\delta}{R_*^k + (1-\delta)}$

a function of the flow of investment and the relative efficiency of these investments:

$$k_t = k_1 k_{t-1} + (1 - k_1) i_t + k_2 \epsilon_t^i \ .^7 \tag{27}$$

Prices

The mark-up of prices over marginal costs is given by:

$$\mu_t^p = \alpha(k_t - l_t) + \epsilon_t^a - w_t \ , \tag{28}$$

which is dependent on diminishing marginal productivity of capital, the productivity shock on costs, and the real wage. Following, inflation is determined by a New-Keynesian Philips curve with lagged inflation:

$$\pi_t = \pi_1 \pi_{t-1} + \pi_2 E_t \pi_{t+1} - \pi_3 \mu_t^p + \epsilon_t^p \ , \tag{29}$$

which is modelled by the assumption that most firms index their prices to past inflation. with ϵ_t^p representing a price mark-up disturbance ($\epsilon_t^b = \rho^p \epsilon_{t-1}^p + \eta_t^p - \mu_p \eta_{t-1}^p$).

Wages

Wages are modelled similar to prices with gradual adjustments to marginal costs determined by the marginal rate of substitution between working and consumption. The gap is defined as wage mark-up:

$$\mu_t^w = w_t - mrs_t = w_t - \left(\sigma t_t - \frac{1}{1 - \lambda/\gamma}(c_t - \lambda c_{t-1}) \right) \quad . \tag{30}$$

Wages are determined by:

$$w_t = w_1 w_{t-1} + (1 - w_1) E_t(w_{t+1} + \pi_{t+1}) - w_2 \pi_t + w_3 \pi_{t-1} - w_t \mu_t^w + \epsilon_t^w \ , \tag{31}$$

with $\epsilon_t^w = \rho^w \epsilon_{t-1}^w + \eta_t^w - \mu_w \eta_{t-1}^w$.

[7] $k_1 = (1 - \delta)/\gamma, k_2 = (1 - (1 - \delta)/\gamma)(1 + \beta \gamma^{(1-\sigma_c)} \gamma^2 \phi$

Monetary Policy

The final missing element is monetary policy. The central bank sets short-term interest rates according to

$$r_t = \rho r_{t-1} + (1-\rho)(r_\pi \pi_t + r_y(y_t - y_t^p)) + r_{\Delta y}\left[(y_t - y_t^p) - (y_{t-1} - y_{t-1}^p)\right] + \epsilon_t^r \ , \qquad (32)$$

with $\epsilon_t^r = \rho^r \epsilon_{t-1}^r + \eta_t^r$. The interest rate adjust gradually to the target interest while also depending on last period's realization. Potential output is defined as the level that would arise if prices and wages were fully flexible.

2.4 Vector Autoregressions

In contrast to DSGE models Vector Autoregressions (VAR) in their simplest form do not make as many a-priori assumptions. They rather model the data as straightforward estimation of their own lagged values, while each variable can be used to predict the others (Stock and Watson, 2001). The simplest VAR that models two feature with one lag in the reduced form is given by:

$$\begin{aligned} y_{1t} &= a_{11}y_{1,t-1} + a_{12}y_{2,t-1} + e_{1t} \\ y_{2t} &= a_{21}y_{1,t-1} + a_{22}y_{2,t-1} + e_{2t} \\ Y_t &= AY_{t-1} + e_t \end{aligned} \qquad (33)$$

$$\text{with} \quad Y_t = \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} \ , \quad A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \ , \quad e_t = \begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} \ .$$

Unbiased forecasts are straightforward with the matrix formulation as $EY_{t+1} = AY_t$. The shock term can be discarded as $Ee_{t+1} = 0$. The model can be adjusted to allow for more lags. This reduced-form VAR is a purely econometric model without any theoretical element. From a theory perspective the reduced-form shocks e_1 and e_2 can be interpreted as separate shocks to e.g. output and inflation. But if they rather represent an aggregate demand or supply shock the interpretation is different. Further, contemporaneous interactions between variables are likely, in this case one speaks of structural shocks. The model is then adjusted to:

$$\text{Structural Form}: \ AY_t = BY_{t-1} + \epsilon_t \quad \text{with} \quad A = \begin{pmatrix} 1 & -a_{12} \\ -a_{21} & 1 \end{pmatrix} \tag{34}$$
$$\text{Reduced Form}: Y_t = DY_{t-1} + e_t \quad \text{with} \quad D = A^{-1}B \quad , \quad e_t = A^{-1}\epsilon_t \quad .$$

While the forecasting remains straightforward with the reduced form VAR, the extension with modelling structural interactions between variables enables answering more theoretical sound questions such as the influence of an isolated shock that is uncorraleted with others.

2.5 Deep Learning

Deep learning, a subset of machine learning, has made tremendous progress in the past years, achieving super-human performance in a variety of fields (Sarker, 2021). Most often the term deep learning refers to the usage of neural networks, which due to their non-linearity can approximate any function according to the universal approximation theorem (Hornik et al., 1989). An NN is a mathematical framework inspired by the biological nervous system. A neuron or perceptron computes a weighted sum of its inputs multiplied with a non-linearity such as e.g. ReLU ($ReLU(x) = max(0, x)$). The ReLU activation function imitates the firing of human neurons when the electric potential reaches a threshold. Many stacked perceptrons comprise one (hidden) layer, while many sequential hidden layers make up an NN. Thus, each hidden layer receives information from the previous layer and multiplies it with a weight and adds bias. A forward pass for one layer in a network with L hidden layers can be mathematically formulated as:

$$z^{(l+1)} = \phi\left(\sum_i \theta_i^{(l+1)} x_i^l + b^{(l+1)}\right) \quad \Longleftrightarrow \quad \mathbf{z}^{(l+1)} = \phi\left(\boldsymbol{\theta}^{(l+1)}\mathbf{x}^l\right) \quad , \tag{35}$$

where $l \in \{1, \ldots, L\}$ indexes the hidden layer, θ^l and b^l denote the weights and biases at layer l, and ϕ is a non-linear activation function (Bishop, 2006). Generally, the more hidden layer a model has, the more accurate its predictions will be, but overfitting can be a negative side effect. Further, due to its non-linearity the wider (more neurons per hidden layer) and deeper (more hidden layers) the NN is, the more it loses explainability. NNs are therefore often referred to as black box models, which regularly prevents economists from using the models in practice.

2.5.1 Optimization

The optimal weights for a neural network are usually found with backpropagation and gradient descent, where the output of the network is compared to a ground truth and the resulting loss $\mathcal{L}$ is partially differentiated through the network and the weights updated accordingly (Rumelhart et al., 1986):

$$\theta_i^l := \theta_i^l - \eta \frac{\partial}{\partial \theta_i^l} \mathcal{L}(\boldsymbol{\theta}) \ . \tag{36}$$

Other possibilities additionally based on the second-order derivative are the Levenberg-Marquardt (LM) alogirthm (Moré, 1978) or the Limited-memory Broyden, Fletcher, Goldfarb, and Shanno (L-BFGS) optimization method (Liu and Nocedal, 1989). In LM gradient descent is accompanied by the Gauss-Newton method, which allows for updates in the steepest direction of the loss curve. Thus, LM is more likely to find the global minimum of a function independent of the initialization. L-BFGS enhances LM with an approximate method to compute the Hessian i.e. the second derivative enabling larger models to be fitted. Both methods exhibit the drawback of introducing additional computational costs, making them only suitable for small-scale models.

2.5.2 Interpretability

NNs suffer from their their black-box nature caused by the high non-linearity often resulting in a weak interpretablity. Especially in economics, where possible policy interactions need to be explained to wider audiences, interpretablity is crucial to the public acceptance of the model's implications. Shapley values are one concept that originated in cooperative game theory (Shapley, 1953). The question they originally answered was to assign a fair payout to each participant in a game. They are defined as given the current set of feature values, the contribution of a feature value to the difference between the actual prediction and the mean prediction is the estimated Shapley value (Molnar, 2022). In a linear model:

$$f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad , \tag{37}$$

the contribution ϕ_j of the j-th feature on the prediction $f(x)$ is:

$$\phi_j(f) = \beta_j - \mathbb{E}(\beta_j X_j) = \beta_j - \beta_j \mathbb{E}(X_j) \quad , \tag{38}$$

where $\mathbb{E}(\beta_j X_j)$ is the mean effect estimate for feature j. Thus, the distribution is the feature effect minus the average effect of that feature.

While the effect is easily computed for a linear model for non-linear models it becomes more difficult. The Shapley value is defined through a value function val of players i.e. features in S. The Shapley value of this feature is then defined as its contribution of the payout, weighted and summed over all possible feature combinations termed coalitions:

$$\phi_j(val) = \sum_{S \subset \{1,\ldots,p\}\{j\}} \frac{|S|!(p-|S|-1)!}{p!}(val(S \cup \{j\}) - val(S)) \quad , \tag{39}$$

where S is the subset of features used in the model, x is the vector of feature values of the instance to be explained and p is the number of features. $val_x(S)$ denotes the prediction for feature values in the set S that are marginalized over features that are not included in the set:

$$val_x(S) = \int f(x_1, \ldots, x_p) d\mathbb{P}_{x \notin S} - \mathbb{E}_X(f(X)) \quad . \tag{40}$$

Lundberg and Lee (2017) have expanded the concept to machine and deep learning by including some mathematical tricks that enable them to compute the integral for only an effective subset of all possible feature combinations.

3 Economics Informed Neural Networks

NNs exhibit a high predictive performance due to their flexibility in fitting the data, but are prone to overfitting and lack of explainability due to their black-box nature. DSGE models on the other hand enable answering of *what if* questions with their microfoundations. However, this comes at the cost of worse data fit compared to purely data-driven models. A hybrid approach leveraging the benefits of both methods is a promising approach. In addition to past observational values of economic timeseries the network is also fed with the forecasts of a DSGE model. Thus, the model is made aware of the microfounded predictions and is tasked with predicting the residual between the two.
So far, the applications of NNs in macroeconomics are limited (see e.g. Smalter Hall and Cook (2017)). Approaches include incorporating the model parameters in the NN's state prediction or loss function. Shiono (2019) use a temporal distance variational autoencoder that encodes an arbitrary number of input variables into a compressed latent space from which the prediction is derived. The transition of this latent space is conditioned on the model dynamics from a DSGE model. Another popular area of research are physics informed neural networks (PINN) that make use of the automatic differentiation ability of NNs (Raissi et al., 2019). PINNs have two loss components, one for data fit and the other incorporates the differential equations. This is used for macro finance models in contiunous time (Fan et al., 2023; Wu et al., 2024). However, solving macroeconomic in continuous time is out of scope for this work. The approach most related to ours is provied by Tänzer (2024), where a two layer network is fitting on macroeconomic time series, representing a non-linear extension of VARs. However, the model is not made aware of the economic underpinnings and no explainability measures are undertaken.

4 Data and Baseline Model

As baseline the DSGE model seminal work from Smets and Wouters (2007) is used, which is estimated with Bayesian estimation with the Metropolis-Hastings algorithm in Dynare[8] and a publicly available implementation[9].

In Smets and Wouters (2007) seven key macro-economic quarterly US time series from 1948:1 until 2004:4 as observable variables are used: the log difference of real GDP, real consumption, real investment and the real wage, log hours worked, the log difference of the GDP deflator and the federal funds rate. The corresponding measurement equation is:

$$Y_t = \begin{pmatrix} \mathrm{dlGDP}_t \\ \mathrm{dlCONS}_t \\ \mathrm{dlINV}_t \\ \mathrm{dlWAG}_t \\ \mathrm{lHOURS}_t \\ \mathrm{dlP}_t \\ \mathrm{FEDFUNDS}_t \end{pmatrix} = \begin{pmatrix} \bar{\gamma} \\ \bar{\gamma} \\ \bar{\gamma} \\ \bar{\gamma} \\ \bar{l} \\ \bar{\pi} \\ \bar{r} \end{pmatrix} + \begin{pmatrix} y_t - y_{t-1} \\ c_t - c_{t-1} \\ i_t - i_{t-1} \\ l_t \\ \pi_t \\ r_t \end{pmatrix}, \tag{41}$$

where l and dl denote log and log difference respectively, $\bar{\gamma} = 100(\gamma - 1)$ is the common quarterly trend growth rate to real GDP, consumption, investment, and wages, $\bar{\pi} = 100(\Pi_* - 1)$ is the quarterly steady-state inflation rate and $\bar{r} = 100(\beta^{-1}\gamma^{\sigma_c}\Pi_* - 1)$ is the steady-state nominal interest rate. The latter is determined by discounting the steady-state inflation rate. Last, $\bar{l}$ is steady-state hours-worked, which is normalized to be equal to zero. The parameters of the estimated SW model can be found in Tab. A.1 in the Appendix.

---

[8] `https://www.dynare.org`

[9] `https://github.com/JohannesPfeifer/DSGE_mod/tree/master/Smets_Wouters_2007`

5 Experiments

We conduct multiple experiments to test whether an approach involving a non-linear NN can enhance predictive performance of traditional models such as VARs and microfounded DSGE models. As baseline, we use an implementation of the famous Smets and Wouters (2007) DSGE model. The model is estimated in Dynare with Bayesian estimation and subsequently up to 10 periods are forecasted. Building on Tänzer (2024) we fit a non-linear NN with 40 hidden features. As activation we LeakyReLU with a negative slope of 0.01. For regularization we include a dropout layer with $p = 0.3$. As optimizer we choose L-BFGS as the commonly employed backpropagation yielded sub-par results. We hypothesize that optimizing with backpropagation resulted in the optimzation into a local optimum, while also taking the second derivative, the Hessian, into account enables the model to find solutions near the global optimum.

Due to structure with connections between every neuron between layers NNs are seen as black boxes and their results must be interpreted with care. To enhance explainability we make the model aware of the forecasts from the microfounded Smets-Wouters DSGE model. the predictions from this model for the next four periods are fed as input to the network such that it receives a hint to the possible progress. Further, similar to VARs we introduce a lag of four periods. This potentially enables the model to also learn connections between data points that are not only dependent on the last period.

To take a step towards better explainability and counteract the black-box nature of NNs we estimate the SHAP values of the resulting best model (Lundberg and Lee, 2017). SHAP values enable the attribution of feature importance also for non-linear models in reasonable time.

6 Results

Five different methods are estimated on the data from Smets and Wouters (2007): their DSGE model estimated with Bayesian optimization, a conventional VAR, and three NNs. For the NNs we distinguish between a model only fed with the past data and one which is also fed with the predictions from the Smets-Wouters model. Further, we train one with conventional backpropagation and with the L-BFGS algorithm that additionally to the Jacobian also takes the Hessian into account during optimization.
Fig. 6.1 shows the qualitative performance of the different methods for predicting the next timestep for the five different methods. While SW, VAR, and NNs with L-BFGS can estimate the overall trajectory very well, an NN trained with conventional backpropagation fails to approximate the true distrbution. We attribute this to a pathology of training with MSE loss together with a method based on the first derivative, which results in predicting the mean of all timeseries, which represents a local minimum. Using L-BFGS and by this the second derivative of the loss function enables the model to find a better optimum. Due to the sub par performance of the model trained with backpropagation we exclude it from the presentation of the remaining results. Tab. 6.1 shows the mean root mean squared error (RSME) of the different methods for the seven observational variables for the next timestep. The conventional VAR outperforms the SW DSGE model on all variables. The performance of an NN is on par or better than the VAR for the timeseries $dlGDP$, $dlCONS$, $dlINV$, and $dlWAG$. The VAR is superior for $lHOURS$, dlP, and $FEDFUNDS$. By making the NN aware of the economic microfoundations by further feeding the predictions of the SW model as input the predictive performance can be enhanced outperforming the VAR on 5 out of seven observational variables ($dlGDP$, $dlCONS$, $dlINV$, $lHOURS$, $dlWAG$) and being on par on the other two (dlP, $FEDFUNDS$). The difference between the two methods, VAR and NN with

Table 6.1: Mean RMSEs per model and observational variable over the whole time series for forecasting the next timestep. The VAR performs better than the Smets-Wouters DSGE model for all observations on average. The performance of the NN is on par with the VAR. Conditioning the NN further on the outputs of the Smets-Wouters model increases performance.

	$dlGDP$	$dlCONS$	$dlINV$	$lHOURS$	$dlWAG$	dlP	$FEDFUNDS$
SW	1.118	0.946	2.918	2.453	0.624	0.596	0.691
VAR	0.977	0.866	2.413	1.594	0.620	**0.509**	**0.527**
NN	0.996	0.837	2.410	2.396	0.603	0.616	0.644
NN SW	**0.946**	**0.824**	**2.352**	**1.527**	**0.594**	0.519	0.565

Table 6.2: p-Values from paired sample t-Test for assessing the difference in the outputs of the different methods per observational variable.

		VAR	NN	NN SW		VAR	NN	NN SW
SW	$dlGDP$	0.020^{**}	$< 0.01^{***}$	$< 0.01^{***}$	$dlCONS$	0.740	$< 0.01^{***}$	$< 0.01^{***}$
VAR		-	$< 0.01^{***}$	$< 0.01^{***}$		-	$< 0.01^{***}$	$< 0.01^{***}$
NN			-	0.040^{**}			-	0.840
SW	$dlINV$	$< 0.01^{***}$	$< 0.01^{***}$	$< 0.01^{***}$	$lHOURS$	$< 0.01^{***}$	$< 0.01^{***}$	$< 0.01^{***}$
VAR		-	$< 0.01^{***}$	$< 0.01^{***}$		-	0.019^{**}	$< 0.01^{***}$
NN			-	$< 0.01^{***}$			-	0.015^{**}
SW	$dlWAG$	0.143	0.013^{**}	$< 0.01^{***}$	dlP	0.267	0.671	0.290
VAR		-	$< 0.01^{***}$	$< 0.01^{***}$		-	0.536	0.567
NN			-	0.239			-	0.070^{*}
		VAR	NN	NN SW				
SW	$FEDFUNDS$	0.698	0.01^{***}	0.002^{**}				
VAR		-	$< 0.01^{***}$	$< 0.01^{**}$				
NN			-	0.127				

SW, is significant to the 1% level for all observational variables except observed inflation, where all methods tend to produce similar results (Tab. 6.2).

6.1 Forecasting Performance

A known drawback of DSGE models is their inferior forecasting performance for more than four periods into the future (Warne et al., 2010). The questions remains whether a hybrid approach enhancing a DSGE model with the non-linear predictive capability of an NN can enhance predictive performance for future periods. In Fig. 6.3 the RMSEs for the

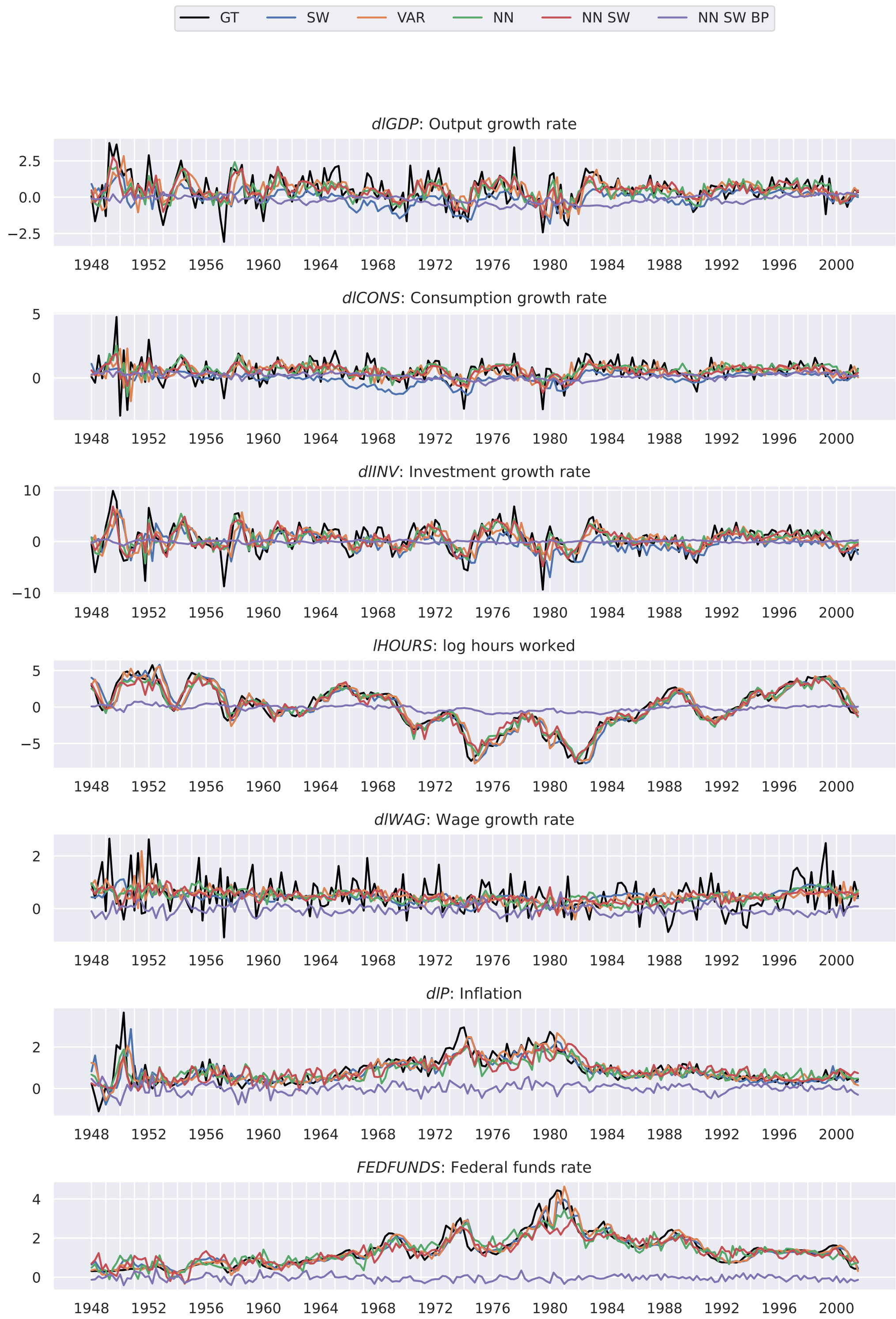

Figure 6.1: Comparison of forecasting performance for the next timestep of the different models. While the high level approach shows a good fit for almost all methods, the NN trained with conventional backpropagation fails to fit the data distribution.

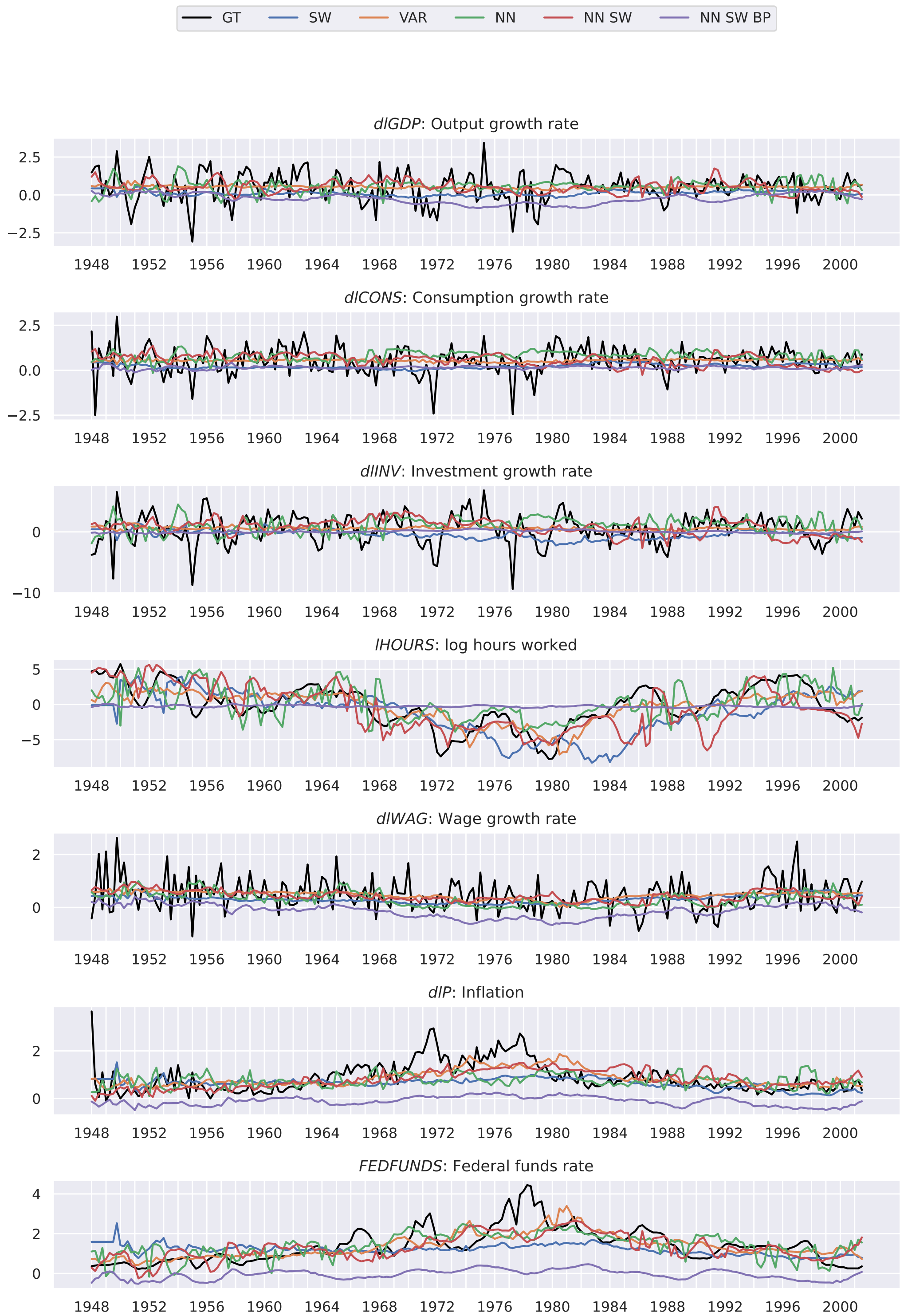

Figure 6.2: Comparison of forecasting performance for the 10th timestep of the different models. While the VAR and NN with SW are still able to capture part of the ground truth timeseries (especially for $lHOURS$), the conventional NN and SW model perform sub-par. For the timeseries $dlGDP$, $dlCONS$, and $dlINV$ also NN with SW and VAR converge to a mean prediction explaining the trajectory of the RMSE curves (Fig. 6.3).

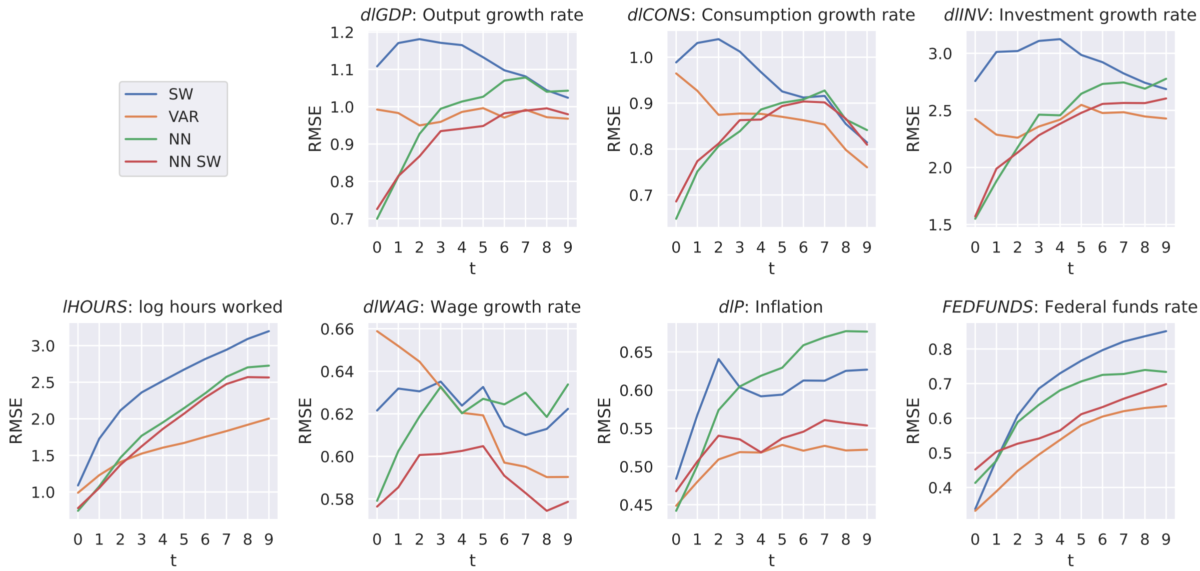

Figure 6.3: RMSE per method and observational variable. The NN conditioned on the SW model's predictions performs on average or better than the VAR model for future periods. The performance of the SW model as well as the conventional NN are inferior.

four different methods for up to ten periods for all observational variables in the future are presented. Results are similar to the previous ones: the NN additionally conditioned on the outputs of the SW model performs superior or on par with the VAR, while SW and only backwards looking NN perform worse. However, the results are mixed, e.g. for output growth rate ($dlGDP$) the NNs perform better for values that directly follow i.e. while the perfromance converges for larger forecasting periods. For e.g. the federal funds rate the forecasting performances follow a mostly parallel trend between the four methods, while NN with SW and VAR consistently show the lowest RMSE.

We also provide the plot for a far ahead forecast with tasking the models to predict ten periods into the future (Fig. 6.2). While SW model and NN with purely backwards looking inputs converge to a mean value of the timeseries, NN with SW and VAR are able to capture at least parts of the underlying true data distribution. This is especially obvious for predicting the log hours worked ($lHOURS$).

6.2 Interpretability

One drawback of NNs is their inherent black box nature due to their high non-linearity. But this feature can be mitigated by using interpretability techniques such as SHAP values (Shapley, 1953; Lundberg and Lee, 2017). The SHAP values for the NN conditioned on the SW forecasts are shown in Fig. 6.4. The mean contribution from all features lies around zero, while some features have a high positive (additive) and negative (subtractive) effect on the predicted next timestep. What is particularly interesting is the contribution of input timesteps that lie further than period in the past or future as this is not modeled with the conventional SW DSGE model. In Tab. 6.3 the five highest and five lowest contributnig features to the model's predictions according to SHAP values are presented. As already observed before, the SW model forecasts make up a substantial part of the most important features explaining the better performance.

Table 6.3: Five highest and lowest SHAP values for NN with SW. For each observational feature the five highest (95-percentile) and lowest (5-percentile) values are shown.

	pos. feature	SHAP	neg. feature	SHAP
$FEDFUNDS$	$dlP_{SW,(t+1)}$	0.590	$FEDFUNDS_{t-1}$	-0.540
	$FEDFUNDS_{t-1}$	0.423	$dlP_{SW,(t+1)}$	-0.463
	$lHOURS_{SW,(t+2)}$	0.246	dlP_{t-4}	-0.280
	$FEDFUNDS_{t-4}$	0.220	$lHOURS_{SW,(t+2)}$	-0.234
	$dlWAG_{t-1}$	0.210	$dlWAG_{SW,(t+3)}$	-0.209
$dlCONS$	$dlCONS_{SW,(t+3)}$	0.414	$dlP_{SW,(t+1)}$	-0.495
	$dlP_{SW,(t+1)}$	0.351	$dlCONS_{SW,(t+3)}$	-0.286
	$lHOURS_{SW,(t+2)}$	0.246	dlP_{t-3}	-0.212
	dlP_{t-3}	0.222	$FEDFUNDS_{SW,(t+2)}$	-0.207
	$FEDFUNDS_{SW,(t+2)}$	0.191	$FEDFUNDS_{t-4}$	-0.182
$dlGDP$	$FEDFUNDS_{t-1}$	0.380	$dlGDP_{t-4}$	-0.316
	$dlWAG_{t-2}$	0.352	$dlWAG_{t-1}$	-0.300
	$dlGDP_{t-4}$	0.309	$dlP_{SW,(t+1)}$	-0.275
	$dlWAG_{t-1}$	0.291	$dlINV_{t-2}$	-0.206
	$dlCONS_{t-4}$	0.249	$dlCONS_{t-4}$	-0.204
$dlINV$	$FEDFUNDS_{t-1}$	1.152	$dlP_{SW,(t+1)}$	-1.699
	$dlP_{SW,(t+1)}$	1.137	$dlINV_{t-3}$	-0.950
	$lHOURS_{SW,(t+2)}$	0.815	dlP_{t-4}	-0.738
	$dlINV_{t-3}$	0.735	$dlCONS_{SW,(t+1)}$	-0.602
	$dlCONS_{SW,(t+3)}$	0.629	$dlWAG_{t-1}$	-0.584
dlP	$dlWAG_{t-2}$	0.252	$FEDFUNDS_{t-1}$	-0.263
	$dlWAG_{t-1}$	0.201	$dlWAG_{t-2}$	-0.235
	dlP_{t-4}	0.194	dlP_{t-4}	-0.208
	$dlWAG_{SW,(t+3)}$	0.169	$dlINV_{t-2}$	-0.169
	$FEDFUNDS_{t-1}$	0.169	$dlWAG_{t-1}$	-0.165
$dlWAG$	$dlWAG_{t-2}$	0.195	$dlWAG_{t-2}$	-0.198
	dlP_{t-4}	0.180	dlP_{t-4}	-0.154
	$FEDFUNDS_{t-1}$	0.179	$dlP_{SW,(t+1)}$	-0.154
	$lHOURS_{SW,(t+2)}$	0.132	$FEDFUNDS_{t-1}$	-0.138
	$dlCONS_{SW,(t+3)}$	0.131	$lHOURS_{SW,(t+2)}$	-0.128
$lHOURS$	$FEDFUNDS_{t-1}$	2.478	$FEDFUNDS_{t-1}$	-3.087
	$dlWAG_{t-2}$	0.801	$dlWAG_{t-2}$	-0.940
	dlP_{t-4}	0.537	$dlCONS_{SW,(t+3)}$	-0.633
	dlP_{t-2}	0.499	dlP_{t-2}	-0.612
	$dlGDP_{t-4}$	0.466	dlP_{t-4}	-0.586

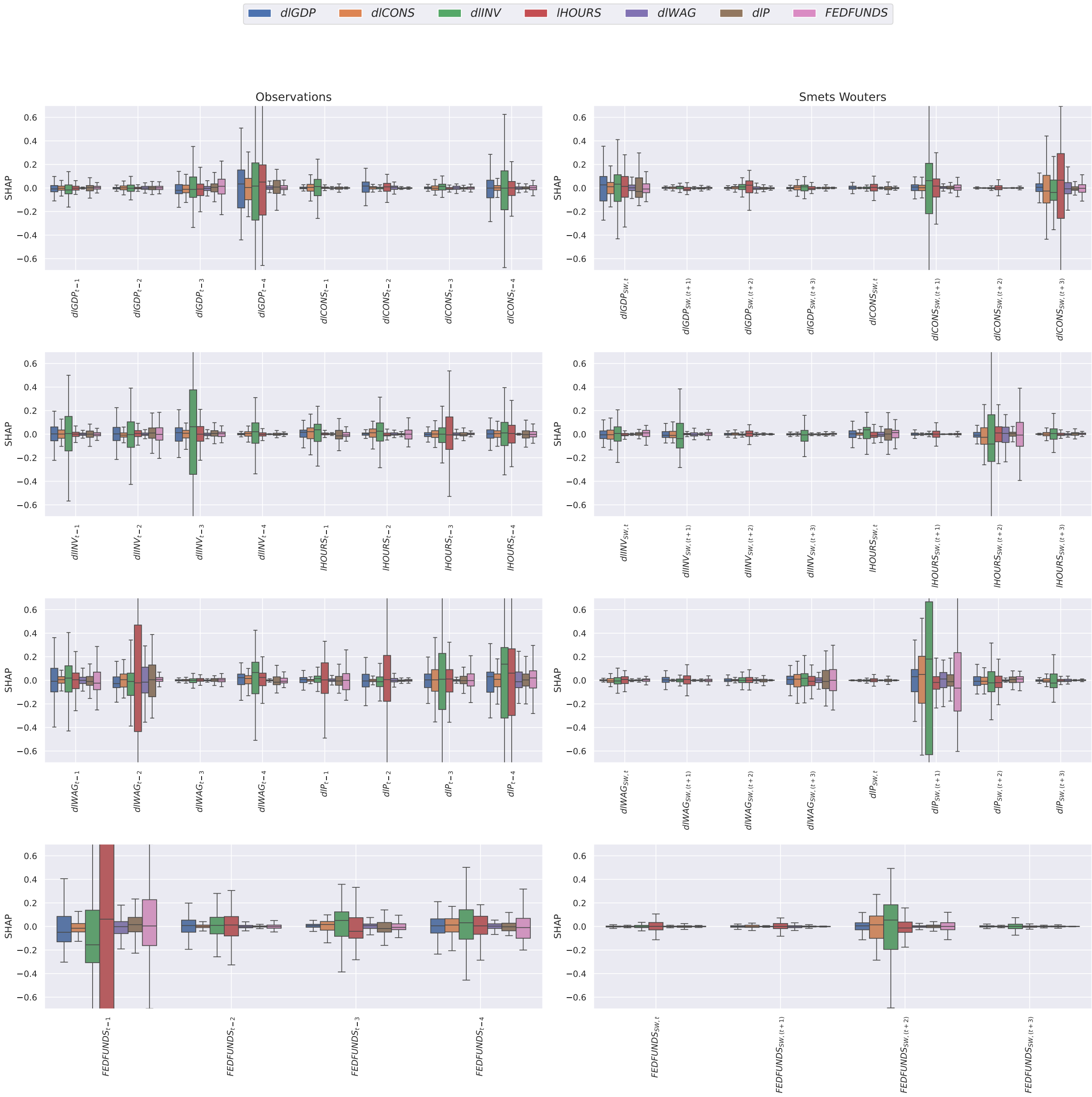

Figure 6.4: SHAP values of the input values for the non-linear NN with Smets-Wouters model predictions as input. Surprisingly, not always the variable from the period directly preceding the prediction period and the Smets-Wouters predictions from the upcoming period are important but also periods with more lags in the past and future.

7 Discussion

The famous Smets and Wouters (2007) model's fit to data coming from seven economic timeseries was compared to an NN's ability on the same task. When the NN was fitted to purely backwards looking data, i.e. like a non-linear VAR, the results exhibit a lower RMSE than the SW model but were not on par with a conventional VAR-based forecast. By providing the model with the predictions of the Smets and Wouters (2007) model as input values the fit to data could be further improved and the VAR is outperformed on five out of seven timeseries (Tab. 6.1). We attribute this to the more information the model obtains due to the incorporation of microfounded economic contexts within the SW model, which is missing in conventional VARs (Stock and Watson, 2001). Our approach resembles the prediction of residuals between SW predictions and the ground truth, as a model could also learn to simply output the prediction from the SW model, if the fit to data would represent a global optimum. We argue that the model finds a compromise between correlational forecasting and microfoundations provided by the economic underpinnings from the SW model.

This still raises the question what is missing in DSGE models such as the one from Smets and Wouters (2007). The most obvious difference is the amount of periods the models look at for their prediction. While the SW model only incorporates one lagged period and models the expectations only for the next period our model takes four periods in both directions into account.

The non-linear nature of the NN seems to outperform the linear VAR on most of the timeseries. The better performance of the non-linear model comes at the cost of less explainability. We try to shed light on the decision process of the NN by computing SHAP values for the input features (Lundberg and Lee, 2017). In addition to conditioning our NN on more economic background knowledge the step towards a better understanding of the decision making process of the network sets us apart from previous literature as e.g. Tänzer (2024).

The predictions from the SW model are among the most influential features for the individual timeseries, both with a positive and negative impact. Additionally, in a large amount of cases predictions for periods more than one in the future affect the outcome significantly. This might indicate that expectation adjustment processes might need more than period to take effect; they potentially adjust more sluggishly to real world economic circumstances. More investigations into this are needed in the future.

One obvious critique of SHAP values is their pure correlational nature. But economic interpretations of the resulting SHAP values are possible. E.g. a rising prediction $lHOURS_{SW,t+2}$ might cause a demand pull inflation due to higher wages, which in turn might cause the central bank to higher interest rates ($FEDFUNDS$). Again not the period directly succeeding the current one are most important for the model prediction, but two periods into the future. But the resulting SHAP values also have weaknesses. E.g. the $FEDFUNDS_{t-4}$ seem to affect the predicted $FEDFUNDS$ for the current period. The model might have learned a pure correlational connection as this would only make sense if also the other periods in between would affect the current $FEDFUNDS$.

When comparing the prediction performance for more than one period ahead we see a mixed picture. It seems that sometimes the RMSE is decreasing (Fig. 6.3) while the actual forecasts get worse (Fig. 6.2). In the qualitative forecasts only a line representing the mean of zero is forecasted for some of the timeseries. This emphasizes that zero represents a local minimum that minimizes the (R)MSE loss as also underlined by the optimization with backpropagation.

8 Summary and Outlook

In this work the predictive ability of a non-linear NN is used to enhance the explainable predictivity of a microfounded NKM, more precisely the famous Smets and Wouters (2007) model. It was shown that using an NN together with the predictions from the SW model even outperforms VARs on most of the time series. Further, feeding the NN the SW predictions as inputs improves its performance compared to a solely backwards looking NN. The NN that is only being fed with lagged values represents a non-linear extension of conventional linear VARs.

Lacking explainability is a known drawback of NNs due to their black box nature preventing their large scale application despite their often superior predictive performance. We are taking a step towards explaining the predictions of our NN with SHAP values. SHAP values allow for attributing influences on the outcome of a potentially black box function of specific input features (Fig. 6.4 and Tab. 6.3). Most importantly the SHAP values show that the NN does also take predictions from the SW model in account that lie more than one period in the future indicating a more sluggishly adjustment of expectations to current economic phenomena. However, more research is necessary as our measures for now are only correlational regularities and are not completely interpretable as causal relationships.

One method that might potentially enable also causal interpretations are physics informed neural networks (Raissi et al., 2019). They allow for baking known real world laws as e.g. natural physical phenomena or economic regularities into the loss function of the NN. Wu et al. (2024) have taken a step towards using physics informed NNs for estimating an economic model in continuous time.

Appendix

Table A.1: Parameter posterior modes of Smets and Wouters (2007) model.

Math Symbol	Long Name	Posterior Mode
ρ_a	Persistence productivity shock	0.9588
ρ_b	Persistence risk premium shock	0.1824
ρ_g	Persistence spending shock	0.9762
ρ_i	Persistence investment-specific technology shock	0.7096
ρ_r	Persistence monetary policy shock	0.1271
ρ_p	Persistence price markup shock	0.9038
ρ_w	Persistence wage markup shock	0.9719
μ_p	Coefficient on MA term price markup	0.7449
μ_w	Coefficient on MA term wage markup	0.8881
φ	Investment adjustment cost	5.4882
σ_c	Risk aversion	1.3952
λ	External habit degree	0.7124
ξ_w	Calvo parameter wages	0.7375
σ_l	Frisch elasticity	1.9199
ξ_p	Calvo parameter prices	0.6563
ι_w	Indexation to past wages	0.5920
ι_p	Indexation to past prices	0.2284
ψ	Capacity utilization cost	0.5472
ϕ_p	Fixed cost share	1.6150
r_π	Taylor rule inflation feedback	2.0295
ρ	Interest rate persistence	0.8153
r_y	Taylor rule output level feedback	0.0847
$r_{\Delta y}$	Taylor rule output growth feedback	0.2229
$\bar{\pi}$	Steady state inflation rate	0.8180
$100(\beta^{-1} - 1)$	Time preference rate in percent	0.1607
$\bar{l}$	Steady state hours	-0.1031
$\bar{\gamma}$	Net growth rate in percent	0.4320
ρ_{ga}	Feedback technology on exogenous spending	0.5261
α	Capital share	0.1928

Bibliography

Alvarez, L. J., Dhyne, E., Hoeberichts, M., Kwapil, C., Bihan, H. L., Lünnemann, P., Martins, F., Sabbatini, R., Stahl, H., Vermeulen, P., and Vilmunen, J. (2006). Sticky prices in the euro area: A summary of new micro-evidence. *Journal of the European Economic Association*, 4(2/3):575–584.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer, New York.

Calvo, G. A. (1983). Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics*, 12(3):383–398.

Christiano, L. J., Eichenbaum, M. S., and Trabandt, M. (2018). On DSGE Models. *Journal of Economic Perspectives*, 32(3):113–40.

Fan, B., Qiao, E., Jiao, A., Gu, Z., Li, W., and Lu, L. (2023). Deep Learning for Solving and Estimating Dynamic Macro-Finance Models. Technical report, MIT Mathematics.

Galí, J. (2018). The state of new keynesian economics: A partial assessment. *Journal of Economic Perspectives*, 32(3):87–112.

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

Keynes, J. M. (1936). *The General Theory of Employment, Interest, and Money.* Macmillan, London.

Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1:19–46.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30.

Mankiw, N. G. and Reis, R. (2018). Friedman's presidential address in the evolution of macroeconomic thought. *Journal of Economic Perspectives*, 32(1):81–96.

Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition.

Moré, J. (1978). The Levenberg-Marquardt algorithm: Implementation and theory. In *Numerical Analysis*, volume 630, pages 105–116. Springer Berlin Heidelberg.

Raissi, M., Perdikaris, P., and Karniadakis, G. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707.

Romer, C. D. and Romer, D. H. (2004). A new measure of monetary shocks: Derivation and implications. *American Economic Review*, 94(4):1055–1084.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Sargent, T. J. and Wallace, N. (1976). Rational expectations and the theory of economic policy. *Journal of Monetary Economics*, 2(2):169–183.

Sarker, I. H. (2021). Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, 2(6).

Shapley, L. S. (1953). A value for n-person games. In *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press.

Shiono, T. (2019). Combining a DSGE Model with Neural Networks for Forecasting Economic Variables. *SSRN*.

Smalter Hall, A. and Cook, T. R. (2017). Macroeconomic indicator forecasting with deep neural networks. In *Federal Reserve Bank of Kansas City Working Paper No. 17-11.*

Smets, F. and Wouters, R. (2007). Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach. *American Economic Review*, 97(3):586–606.

Stiglitz, J. E. (2018). Where modern macroeconomics went wrong. *Oxford Review of Economic Policy*, 34(1-2):70–106.

Stock, J. H. and Watson, M. W. (2001). Vector autoregressions. *Journal of Economic Perspectives*, 15(4):101–115.

Tänzer, A. (2024). Multivariate Macroeconomic Forecasting: From DSGE and BVAR to Artificial Neural Networks. In *Working Paper Series No. 205, Institute for Monetary and Financial Stability, Goethe University Frankfurt.*

Warne, A., Coenen, G., and Christoffel, K. (2010). Forecasting with DSGE models. Working Paper Series 1185, European Central Bank.

Wu, Y., Guo, J., Gopalakrishna, G., and Poulos, Z. (2024). Deep-MacroFin: Informed Equilibrium Neural Network for Continuous Time Economic Models. Technical report, University of Toronto.