Interpretability of Deep Neural Networks

Modelling with Big Data and Machine Learning

George Kapetanios and Felix Kempf

King's Business School
King's College London

1. Key Summary
2. Introduction
3. Motivation
4. Methodology
   - Neural Network
   - Variable Importance
   - Impulse Responses
5. Empirical Study
6. Conclusion

Key Summary

We provide an alternative *toolbox* to interpret deep neural networks in the context of macroeconomic modelling. In this presentation, we

- propose the use of partial derivates to enhance model interpretability.
- introduce a non-linear and time-varying impulse response analysis.
- provide a first and preliminary empirical example (US economy).

One of the key criticisms of (deep) neural networks is the limited scope for model interpretability. With this project, we aim to shed some light on this *black-box*.

In particular, first preliminary results look quite interesting, so that we pursue this idea further. For example, we find that

- for certain dependent variables, input variable influence only changes during specific points in time (e.g.times of increased volatility) while it remains constant otherwise. For other dependent variables, this variable influence seems to change more frequently.
- non-linear impulse responses can be supported by economic theory.

Interpretability in AI or ML in general (1/2)

No clear definition for interpretability, but can be summarised as: *the degree to which a human can understand the cause of a decision*, see [Miller, 2018].

Why do we care at all about interpretability if accuracy is competitive? Single metric such as OOS-MSE is an incomplete description, e.g. see [Doshi-Velez and Kim, 2017].

There are, however, many other reasons, why interpretability is relevant (e.g. see [Molnar, 2019]), including:

- scientific research
- safety measure
- detecting biases
- debugging

Interpretability in AI or ML in general (2/2)

There already exist different approaches to generally enhance model interpretability in ML, examples include

- Partial Dependence, e.g. see [Friedman, 2001].
- Individual Conditional Expectation, e.g. see [Goldstein et al., 2015].
- Accumulated Local Effect, e.g. see [Apley, 2016].
- Shapley Values, e.g. see [Joseph, 2019].

In this project, we turn our focus explicitly to neural networks in macroeconomics.

Motivation

The central motivation for this research is two-fold:

1. **Universal approximation**:
   - We know that multilayer feed-forward networks are universal approximators, e.g. see [Hornik et al., 1989]
2. **Time-varying effects**:
   - Through their inherent non-linearity, neural networks can display time-varying effects, e.g. see [Kapetanios, 2007]

We therefore hope to make meaningful contributions to current debates by offering time-varying analysis tools.

Neural Network Setup (1/3)

In its most fundamental form, we describe an arbitrary economic process as

$$y_t = \mathbb{E}(y_t|x_{t-1}) + \epsilon_t, \tag{1}$$

where

$$\mathbb{E}(y_t|x_{t-1}) = g(x_{i,t-1}, \theta), \tag{2}$$

is a (non-)linear approximation for the true but unknown DGP. Note the lag in x which is supported by the fact of publication delays but also by idiosyncratic persistence. Moreover, we allow x to include lags of y. All model parameters including the weights and biases as well as hyper-parameters are denoted by θ.

What is $g(\cdot)$?

Neural Network Setup (2/3)

We investigate the case where we approximate $g(\cdot)$ with a neural network.

In general, for feedforward networks, $g(\cdot)$ takes the form

$$g(\mathbf{X}, \mathbf{\Theta}) = \sigma_L(...\sigma_2(\mathbf{W_2}^T \sigma_1(\mathbf{W_1}^T \mathbf{X} + \mathbf{b}_1) + \mathbf{b}_2)), \tag{3}$$

where σ are activation functions, $l = 1, 2, ..., L$ denotes the number of layers. While the weights $\mathbf{W}$ and biases $\mathbf{b}$ are summarised in $\mathbf{\Theta}$, all model parameters, including weights, biases and all other hyper-parameters, are denoted by θ, e.g. see [Goodfellow et al., 2016]. The considered loss function is

$$\tilde{\mathcal{L}}(\mathbf{Y}, \mathbf{X}, \mathbf{\Theta}) = \mathcal{L}(\mathbf{Y}, \mathbf{X}, \mathbf{\Theta}) + \lambda \Omega(\mathbf{\Theta}), \tag{4}$$

with

$$\mathcal{L}(\mathbf{Y}, \mathbf{X}, \mathbf{\Theta}) = (g(\mathbf{X}, \mathbf{\Theta}) - \mathbf{Y})^T (g(\mathbf{X}, \mathbf{\Theta}) - \mathbf{Y}). \tag{5}$$

Moreover, with L_1, respectively L_2 regularisation, the loss function becomes

$$\tilde{\mathcal{L}}(\mathbf{Y}, \mathbf{X}, \mathbf{\Theta}) = (g(\mathbf{X}, \mathbf{\Theta}) - \mathbf{Y})^T (g(\mathbf{X}, \mathbf{\Theta}) - \mathbf{Y}) + \lambda ||\mathbf{\Theta}||_1 \tag{6}$$

$$\tilde{\mathcal{L}}(\mathbf{Y}, \mathbf{X}, \mathbf{\Theta}) = (g(\mathbf{X}, \mathbf{\Theta}) - \mathbf{Y})^T (g(\mathbf{X}, \mathbf{\Theta}) - \mathbf{Y}) + \frac{1}{2} \lambda \mathbf{\Theta}^T \mathbf{\Theta} \tag{7}$$

Neural Network Setup (3/3)

In this first preliminary draft, we consider:

- one-layer networks
- non-linear activation functions (e.g. tanh, relu)
- stochastic gradient descent
- 90%-10% training-test split, with 20% validation split
- fixed window (other alternatives such as rolling or expanding are considered for future drafts)
- random grid search for hyperparameter tuning
- L_1 and/or L_2 regularisation for the weights

Many other specifications are imaginable!

Variable Influence (1/2)

We propose the usage of partial derivatives at each point in time to evaluate variable influence over time.

$$I_{ij,t} = \frac{\partial g_j(\mathbf{X}, \theta)}{\partial x_{i,t-1}} \tag{8}$$

Note that in this preliminary draft, each target variable j ($j = 1, 2, ..., N$) has its own $g(\cdot)$. Alternatives are also imaginable, where the network output could be multidimensional.

Due to the inherent non-linearity of the neural network we expect the derivative to vary over time.

The motivation for using the partial derivative is that the derivative can be interpreted as the marginal influence each input variable has on the process at each point in time. It can therefore be understood as a sensitivity to changes in the input variables, where we would assign relatively greater influence to larger absolute values of the partial derivative and relatively smaller influence if the absolute value is also smaller.

Variable Influence (2/2)

Moreover, we propose the usage of confidence bands in extension to equation (8). In particular, we propose using the moving-block bootstrap methodology, where we draw $T - b + 1$ overlapping blocks from the original data, where T is the total number of observations per variable and b denotes the block-length. From these blocks, T/b blocks are drawn at random with replacement and aligned following the order they were drawn with to build the bootstrapped observations, e.g. see [Kunsch, 1989].

For each bootstrapped dataset, we fit a neural network as described before. We then calculate the partial derivatives, but with respect to the original input data:

$$I_{ijB,t} = \frac{\partial g_{j,B}(\cdot)}{\partial x_{i,t-1}}, \tag{9}$$

where B indicates the respective bootstrap.

We report the 95% confidence interval of the *centred* partial derivatives.

Non-linear impulse response functions

Similarly to the methodology applied in a linear setting, we propose the usage of impulse response functions (IRF) in the context of neural networks. Generally, we apply the framework

$$IRF(h, \nu, \Omega) = E[y_{t+h}|\nu_t, \Omega_{t-1}] - E[y_{t+h}|\Omega_{t-1}], \tag{10}$$

where ν_t denote structural shocks at time t, e.g. see [Pesaran and Shin, 1998]. Given

$$y_t = g(x_{i,t-1}, \theta) + \epsilon_t, \tag{11}$$

with

$$E[\epsilon_t \epsilon_t'] = \Sigma \tag{12}$$

being a symmetric and positive definite covariance matrix whose off-diagonals are non-zero. We apply a Cholesky Decomposition of the covariance matrix such that $\Sigma = PP'$, where P is a lower-triangular matrix. It follows

$$\nu_t = Pu_t \tag{13}$$

with u_t being reduced form residuals, e.g. see [Sims, 1980].

US Economy – Overview

We consider the US economy as an empirical example with GDP, inflation, unemployment, export prices, S&P500 returns and Fed Fund rates as our dependent, with their lagged values as explanatory variables. The dataset ranges from Q4 1984 to Q2 2019.

In particular, we include export prices as another price variable to avoid encountering price puzzle (e.g. see [Sims, 1992], [Buch et al., 2014], [Balke et al., 1994]). However, we find that an impulse response analysis (e.g. using VAR) is occasionally still displaying the price puzzle dependent on where we split the data.

We perform variable transformation to make them stationary.

In this draft, we feed the network the already transformed data. However, in next steps we will experiment with standardised and / or raw data.

Since the Cholesky Decomposition is sensitive to variable ordering we use the following ordering:
GDP $\rightarrow$ CPI $\rightarrow$ Unemployment $\rightarrow$ Export Prices $\rightarrow$ S&P 500 $\rightarrow$ Rates

US Economy – Predictions

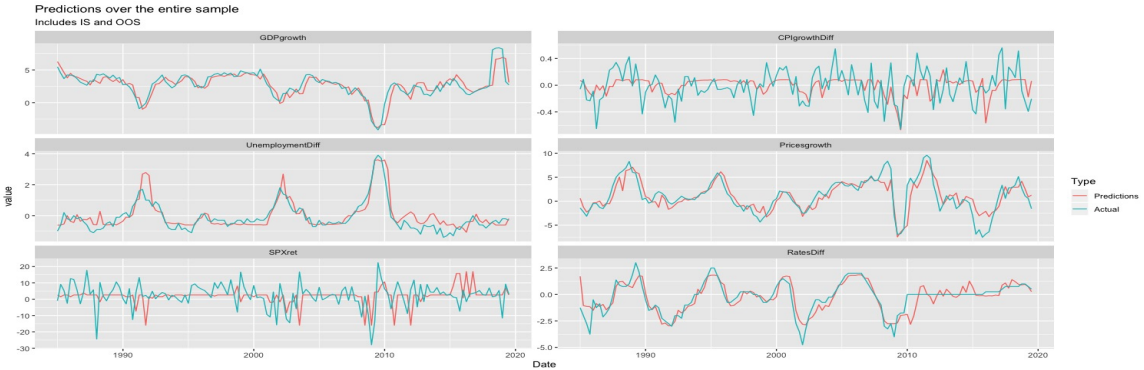


Figure: Full Sample (in- and out-of-sample) predictions

US Economy – Variable Influence (CPI)

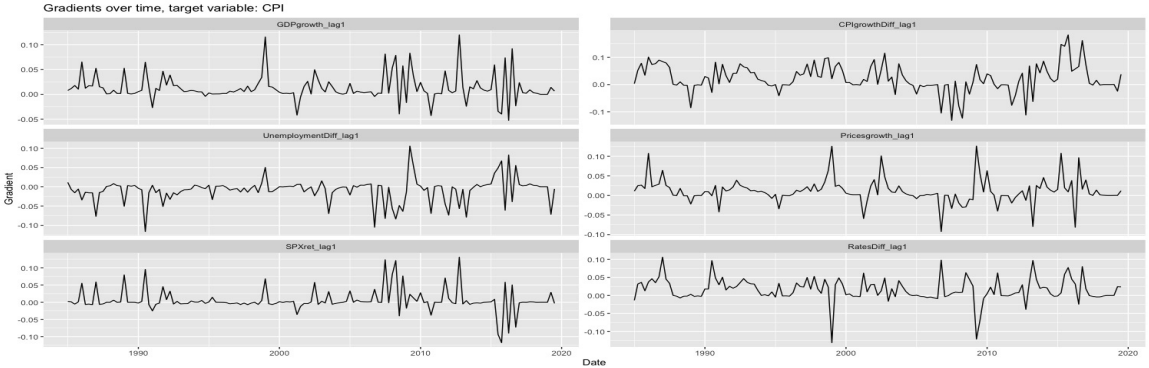


Figure: Partial Derivatives: CPI

US Economy – Variable Influence (GDP)

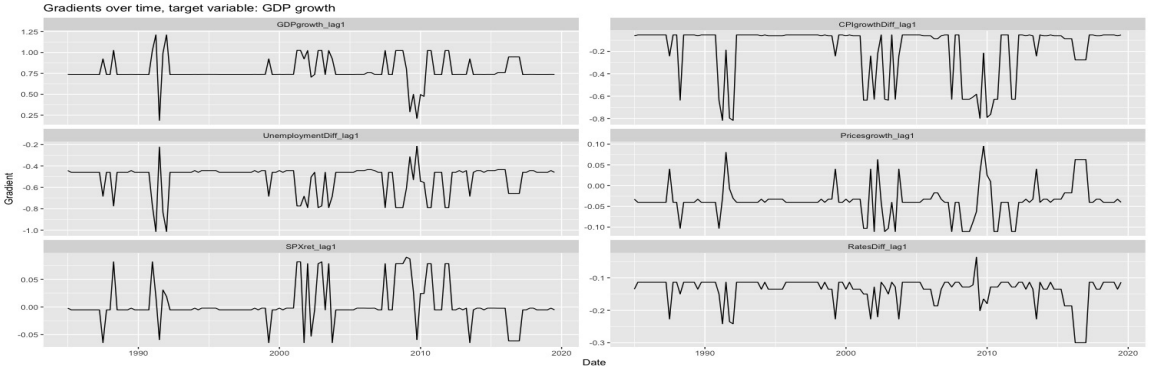


Figure: Partial Derivatives: GDP

US Economy – Variable Influence (Rates)

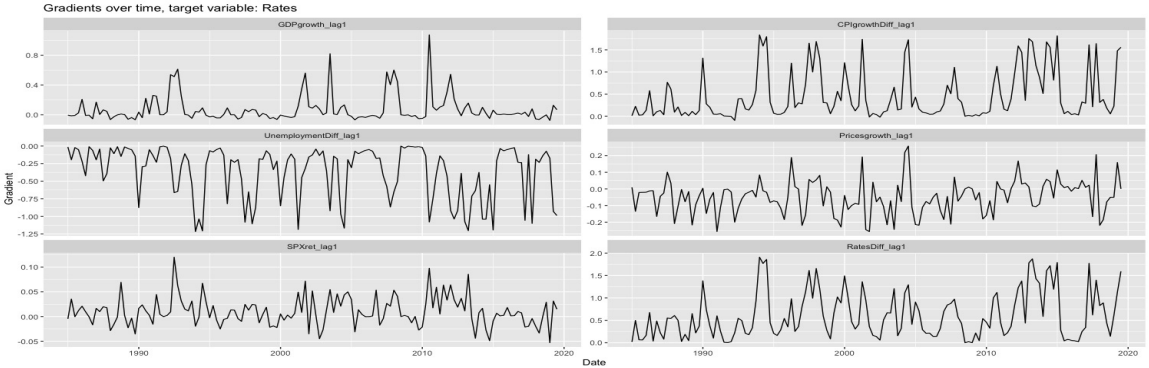
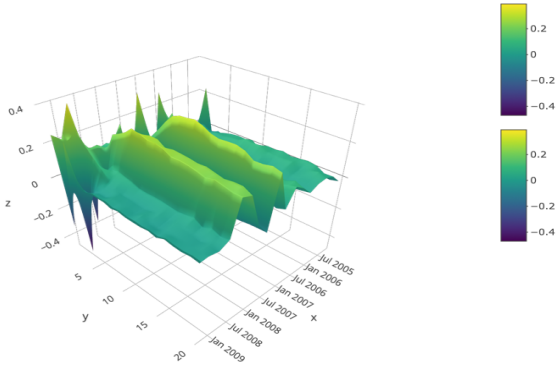


Figure: Partial Derivatives: Rates

US Economy – Variable Influence Interpretation

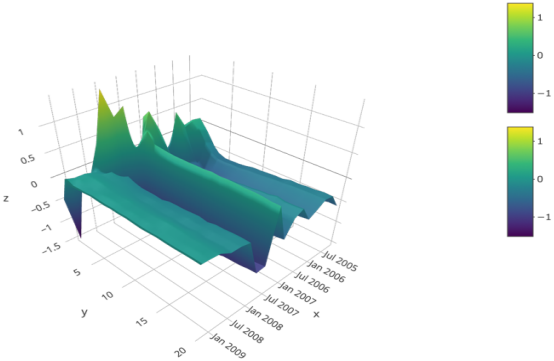
- We find, that variable influence as displayed by the partial derivative is sensitive to time. We can therefore confirm our initial hypothesis.
- In particular, we find that characteristic variable influences vary for each dependent variable.
- In the case of GDP, for example, the network is most sensitive to changes in lagged values of GDP most of the time. It is only during times of increased volatility that the network also becomes more sensitive to changes in other explanatory variables.
- For rates, on the other hand, the neural network seems to be much more sensitive to changes in almost all variables. In absolute values, CPI and Unemployment seem to have the largest influence.

US Economy – Impulse Responses (Shock in Rates, Response in CPI)



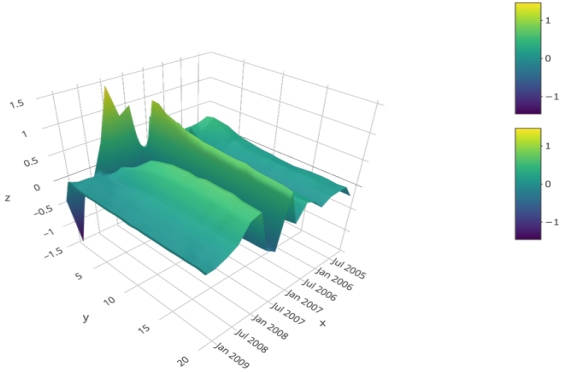
- We find that negative shocks in rates lead to positive responses in CPI as supported by economic theory.
- The magnitude of the response is time-dependent.
- During the GFC, a negative shock in rates cannot offset the effect of the crisis and CPI falls despite reduced rates.

US Economy – Impulse Responses (Shock in GDP, Response in Rates)



- We find that rates respond negatively to a negative shocks in GDP as expected by economic theory.
- We find that a shock prior the GFC leads to a lower level in rates than post the crisis.

US Economy – Impulse Responses (Shock in CPI, Response in Rates)



- We find that rates respond negatively to a negative shock in CPI.
- This is supported by economic theory.
- In particular, the effects of a shock in CPI seem to be more distinct prior the GFC.

Conclusion

We find that both partial derivatives and non-linear impulse responses can help to shed some light on economic theory.

First preliminary results look somewhat interesting, so we will pursue this project further.

There is room for further improvement, in particular with regard to model tuning and selection.

References I

[Apley, 2016] Apley, D. W. (2016).
Visualizing the effects of predictor variables in black box supervised learning models.
arXiv preprint arXiv:1612.08468.

[Balke et al., 1994] Balke, N. S., Emery, K. M., et al. (1994).
Understanding the price puzzle.
Federal Reserve Bank of Dallas Economic Review, Fourth Quarter, pages 15–26.

[Buch et al., 2014] Buch, C. M., Eickmeier, S., and Prieto, E. (2014).
Macroeconomic factors and microlevel bank behavior.
Journal of Money, Credit and Banking, 46(4):715–751.

[Doshi-Velez and Kim, 2017] Doshi-Velez, F. and Kim, B. (2017).
Towards a rigorous science of interpretable machine learning.
arXiv preprint arXiv:1702.08608.

[Friedman, 2001] Friedman, J. H. (2001).
Greedy function approximation: a gradient boosting machine.
Annals of statistics, pages 1189–1232.

References II

[Goldstein et al., 2015] Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015).
Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation.
Journal of Computational and Graphical Statistics, 24(1):44–65.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016).
Deep learning, volume 1.
MIT press Cambridge.

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

[Joseph, 2019] Joseph, A. (2019).
Shapley regressions: A framework for statistical inference on machine learning models.
arXiv preprint arXiv:1903.04209.

[Kapetanios, 2007] Kapetanios, G. (2007).
Measuring conditional persistence in nonlinear time series.
Oxford Bulletin of Economics and Statistics, 69(3):363–386.

References III

[Kunsch, 1989] Kunsch, H. R. (1989).
The jackknife and the bootstrap for general stationary observations.
The annals of Statistics, pages 1217–1241.

[Miller, 2018] Miller, T. (2018).
Explanation in artificial intelligence: Insights from the social sciences.
Artificial Intelligence.

[Molnar, 2019] Molnar, C. (2019).
Interpretable machine learning.
Lulu. com.

[Pesaran and Shin, 1998] Pesaran, H. H. and Shin, Y. (1998).
Generalized impulse response analysis in linear multivariate models.
Economics letters, 58(1):17–29.

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

References IV

[Sims, 1992] Sims, C. A. (1992).
Interpreting the macroeconomic time series facts: The effects of monetary policy.
European economic review, 36(5):975–1000.