A peek inside the 'Black Box' - interpreting neural networks

Marc Agustí (marc.agusti@barcelonagse.eu)
Patrick Altmeyer (patrick.altmeyer@barcelonagse.eu)
Ignacio Vidal (ignacio.vidalquadrascosta@barcelonagse.eu)

February, 2021

Economists and policy makers recognise the undeniable potential of emerging AI technologies and Data Science, but they are rightly hesitant to adopt novel models and tools that they cannot trust. The ability to explain one's actions and decisions is especially crucial for institutions such as central banks whose policies affect the livelihoods of millions of people. It is therefore not surprising that the body of machine learning literature concerned with interpretability has recently gained considerable momentum.

Some advancements have emerged from that literature, among those post-hoc methods which aim to extract feature importance. Shapley values which come from cooperative game theory are often used, but come with a prohibitive computational burden. A more novel approach recently proposed by Ish-Horowicz et al. (2019) provides an intuitive entropy measure for variable importance, but is only applicable in the Bayesian setting (see here for a summary).

Another recent paper take more of an ad-hoc approach to interpretability, that is it aims at keeping the model perse interpretable: Bussmann, Nys, and Latré (2020) propose a Neural Additive Vector Autoregression (NAVAR) model for causal discovery in time series data. Let the equation below denote the standard linear VAR

$$\mathbf{X}_t{}^{(j)} = \beta^j + \sum_{i=1}^{N}\sum_{k=1}^{K}[A_k]_{ij}\mathbf{X}_{t-k}{}^{(i)} + \eta_t^j \tag{1}$$

where each variable in the system depends linearly on its own lags and those of its covariates. Then the NAVAR model is then denoted as

$$\mathbf{X}_t{}^{(j)} = \beta^j + \sum_{i=1}^{N} f_{ij}\left(\mathbf{X}_{t-K:t-1}{}^{(i)}\right) + \eta_t^j \tag{2}$$

instead allows for non-linear interactions between covariates where f_{ij} is the i-th output from a deep neural network that maps from all of j-th past lags (up to K) to all covariates. Notice that if f is linear we are just back to the simple VAR case.

In our paper we aim to compare outputs from a standard VAR model for the monetary transmission mechanism, to outputs from a NAVAR model. In light of the discussion above, our focus would primarily be on model interpretability, with forecasting performance taking only a secondary role (we keenly want to avoid a "horse race" type of analysis).

Notice that the neural networks run for each variable at each step t still suffer from the inherent Black Box problem. Should we find that an understanding of the step-wise interactions between all covariates also need to be well understood, we may try to use either Shapley values or the novel measure mentioned above.

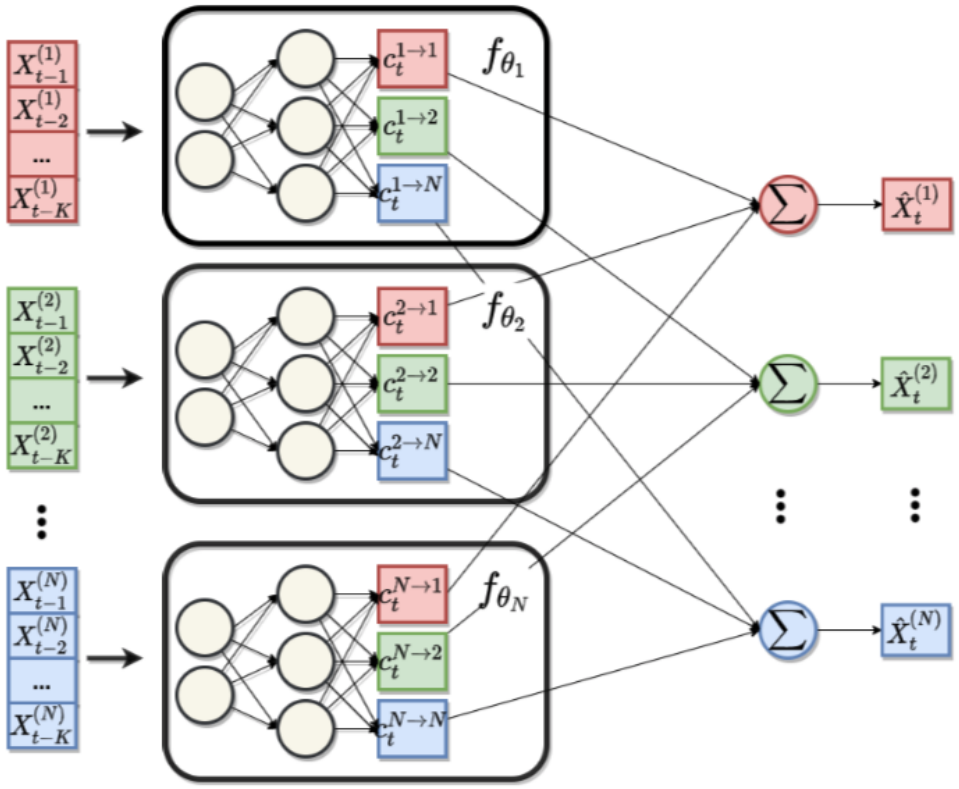

Figure 1: Graphical illustration of NAVAR model with MLPs. Source: Bussmann, Nys, and Latré (2020)

References

Bussmann, Bart, Jannes Nys, and Steven Latré. 2020. "Neural Additive Vector Autoregression Models for Causal Discovery in Time Series Data." *arXiv Preprint arXiv:2010.09429*.

Ish-Horowicz, Jonathan, Dana Udwin, Seth Flaxman, Sarah Filippi, and Lorin Crawford. 2019. "Interpreting Deep Neural Networks Through Variable Importance." *arXiv Preprint arXiv:1901.09839*.