Deep Vector Autoregression for Macroeconomic Data

Marc Agusti (marc.agusti_i_torres@ecb.europa.eu)
Patrick Altmeyer (p.altmeyer@tudelft.nl)
Ignacio Vidal-Quadras Costa (ignacio.vidal-quadras_costa@ecb.europa.eu)

November, 2021

Abstract

Vector Autoregression (VAR) models are a popular choice for forecasting time series data. Due to their simplicity and success at modelling monetary economic indicators VARs have become a standard tool for central bankers to construct economic forecasts. Impulse response functions can be readily retrieved from the conventional VAR and used for inference purposes. They are typically employed to investigate various interactions between variables that form part of the monetary transmission mechanism. A crucial assumption underlying conventional VARs is that these interactions between variables through time can be modelled linearly. We propose a novel approach towards VARs that relaxes this assumption. In particular, we offer a simple way to integrate deep learning into VARs without deviating too much from the trusted and established framework. By fitting each equation of the VAR system with a deep neural network, the Deep VAR outperforms its conventional benchmark in terms of in-sample fit, out-of-sample fit and point forecasting accuracy. In particular, we find that the Deep VAR is able to better capture the structural economic changes during periods of uncertainty and recession.[1]

1 Introduction

As stated by the European Central Bank, the monetary transmission mechanism is the process through which monetary policy decisions affect the economy in general and the price level in particular. Uncertainty with respect to this transmission is generally huge, given that it is characterized by long, variable and uncertain dependencies through time and variables. Hence, it is typically challenging to predict how changes in monetary policy actions affect real economic outcomes. It is therefore of foremost importance for policy-makers to use adequate tools to model the underlying mechanisms.

With this in mind, a lot of research on the forecasting of time series has been developed to assess the effect of current policy decisions on future economic variables. Thanks to this,

[1]The authors would like to thank Christian Brownlees for being a helpful supervisor. We would also like to thank Eddie Gerba, research manager of the Bank of England's Markets Directorate, for helpful comments and fruitful discussions.

over the last decades policy makers have had more information when taking decisions. This information usually comes in the form of point estimates and interval forecasts. To come up with these estimates, several methodologies have been developed and applied in the time series forecasting literature.

At the time of writing, one the most common methodologies to produce these estimates is the so-called Vector Autoregression (VAR). This framework, which belongs to the traditional toolkit of econometric forecasting techniques, has been shown to provide policy-makers with fairly good and consistent point and interval estimates. It has therefore been used extensively in the monetary policy divisions of central banks.

Simultaneously, with the recent advancements in computational power, and more importantly, the development of advanced machine learning algorithms and deep learning, interesting novel tools have become available that may be useful for forecasting time series. Whereas the good performance of techniques such as VAR is well established, it is still uncertain whether deep learning algorithms can be applied successfully to macroeconomic data.

To this end, this paper contributes a new and ground-breaking methodology that combines the VAR equation-by-equation structure with deep learning. We provide evidence that this improves the model's capacity to capture potentially highly non-linear relationships in the underlying data generating process. The primary objective of this paper is to develop a methodology that produces improved modelling outcomes while deviating as little as possible from the established VAR framework, thereby keeping things straight-forward and familiar to economists. We show that the existing VAR methodology can be easily extended to the broader class of Deep VAR models and provide solid empirical evidence that Deep VARs consistently outperform the conventional approach.

To the best of our knowledge, this is the first paper to propose a Deep VAR framework of this structure, namely, to fit a deep neural network for each equation of the VAR process. Although previous work has explored the use of deep learning to forecast macroeconomic time series, previous proposed methodologies deviate more from the conventional VAR framework. For example, Verstyuk (2020) chooses to model the whole system through one unified deep neural network. We find that the equation-by-equation approach not only helps to maintain interpretability and simplicity, but also appears to produce better modelling outcomes. To enable researchers and practitioners to easily implement our proposed methodology, we have developed a unified framework for estimating Deep VARs in R and plan to continue its development going forward.

We find that the Deep VAR methodology outperforms the traditional VAR framework in terms of in-sample and out-of-sample fit as well as with respect to forecasting accuracy. In particular, the Deep VAR appears to be better at capturing non-linear dynamics underlying the time series process. It therefore leads to consistently lower modelling errors than the VAR, especially during periods of economic downturn and uncertainty.

Arguably policy makers are not only interested in the forecasting accuracy of the model, but are typically also concerned with inference. For example, central banks are often interested in knowing to what extent interest rates granger cause other variables within the monetary transmission mechanism. Another aspect policy makers and researchers care about is how the

variables of the system evolve through time in response to innovations. This information is typically recovered using Impulse Response Functions (IRFs). The linear additive modelling assumption underlying the conventional VAR makes inference straight-forward. In the case of Deep VARs inference is arguably more complicated, though promising avenues have recently been explored (Verstyuk 2020). We believe that the methodology proposed in this paper can be relatively easily augmented to the inference realm in future work.

The remainder of the paper is structured as follows: in section **??** we present a literature review of prior research on the methodologies used to provide forecasts and on the monetary transmission mechanism in general. Section **??** provides a detailed description of the data we use for our empirical exercises. In section **??** we present the traditional VAR methodology and develop our proposed Deep VAR model. Sections 2 and **??** present our empirical findings and possible extensions and caveats, respectively. Finally, section 3 concludes.

2 Empirical results

We now proceed to benchmark the proposed Deep VAR model against the conventional VAR using out macroeconomic time series data. To begin with, we compare both models in terms of their in-sample fit. For this part of the analysis the models will be strictly run under the same framing conditions. Due to the RNN's capacity to essentially model any possible function $f_i(\cdot)$ the Deep VAR dominates the VAR in this realm. We investigate during what time periods the outperformance of the Deep VAR is particularly striking to gain a better understanding of when and why it pays off to relax the linearity constraint.

These findings with respect to in-sample performance provide some initial evidence in favor of the Deep VAR. But since a reduction in modelling bias is typically associated with an increase in variance, we are particularly interested in benchmarking the models with respect to their out-of-sample performance. To this end we split our sample into train and test subsamples. We then firstly benchmark the models in terms of their pseudo out-of-sample fit. Finally we also look at model performance with respect to n-step ahead pseudo out-of-sample forecasts.

The final part of this section relaxes the constraint on the framing conditions. In particular, we investigate how hyperparameter tuning with respect to the neural network architecture and lag length p can improve the performance of the Deep VAR.

2.1 In-sample fit

For this first empirical exercise both models are trained on the full sample. We have decided to include the post-Covid sample period despite the associated structural break, since it serves as interesting point of comparison. The optimal lag order as determined by the Akaike Information Criterium is $p = 6$, where we used a maximum possible lag of $p_{\max} = 12$ corresponding to one year. A look at the eigenvalues of the companion matrix showed that the VAR(6) is stable. The LSTMs underlying the Deep VAR model are composed of $H = 2$ that count $N = 100$ hidden units each. The dropout rate is set to $p = 0.5$.

To assess the fit of our models we use the root mean squared error (RMSE) as our preferred loss function. Figure 1 shows the cumulative RMSE of both the VAR model and Deep VAR model for each of the time series over the whole sample period. The first thing we can observe is that the RMSE of the Deep VAR is consistently flatter than the RMSE of the VAR. With respect to in-sample performance, the Deep VAR the VAR throughout the entire time period of the experimental analysis and for all of the considered variables. This empirical observation seems to confirm our expectation that the vector autoregressive process is characterized by important non-linear dependencies across time and variables that the conventional VAR fails to capture.

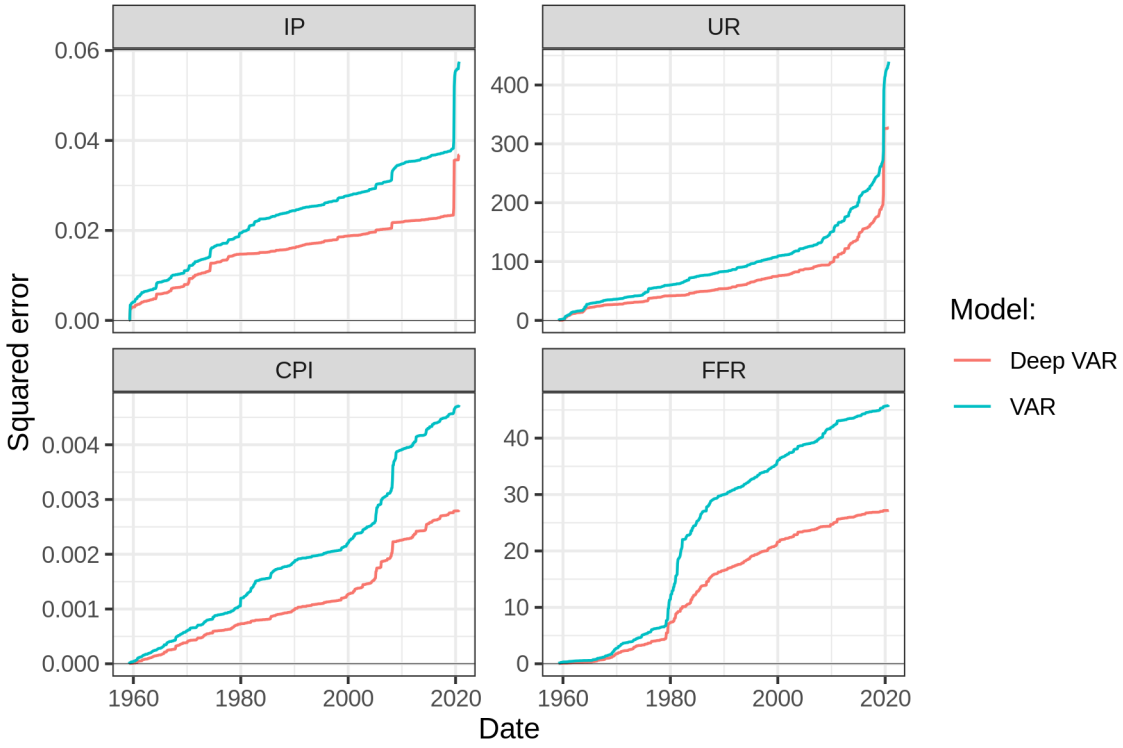

Figure 1: Comparison of cumulative loss over the entire sample period for conventional VAR and proposed Deep VAR.

Figure 1 is especially useful to asses in which specific periods the Deep VAR model achieves better modelling outcomes than the VAR model. From the very beginning and across variables, we observe that the increase in cumulative loss for the VAR model is greater than for the Deep VAR model. The US economy during 1960s was influence by John F. Kennedy's introduction of **New Economics**, which was informed by Keynesian ideas and characterized by increasing levels of inflation, a reduction in unemployment and output growth. The change in government certainly corresponded to a regime switch with respect to the economy (Perry and Tobin 2010) and in that sense it is interesting to observe that the Deep VAR appears to be doing a better job at capturing the underlying changes. The 1970s can be broadly thought of as a continuation of New Economics and loosely defined as a period of stagflation. The Deep VAR continues to outperform the VAR during that period.

The first truly interesting development we can observe in Figure 1 coincides with the onset of the Volcker disinflation period. Following years of sustained CPI growth, Paul Volcker set the Federal Reserve on course for a series of interest rate hikes as soon as he became chairperson of the central bank in August 1979. The shift in monetary policy triggered fundamental changes to the US economy and in particular the key economic indicators we are analyzing here throughout the 1980s (Goodfriend and King 2005). Despite this structural break, the increase in the cumulative RMSE of the Deep VAR remains almost constant during this decade for most variables. The performance of the VAR on the other hand is unsurprisingly poor over the same period, in particularly so for the CPI and the Fed Funds Rate, which arguably were the two variables most directly affected by the change in policy. The Deep VAR also clearly dominates the VAR with respect to the output related variables (IP) and to a lesser extent unemployment. These findings indicate that changes to the monetary transmission mechanism in response to sudden policy shifts are not well captured by a linear-additive vector autoregressive model. Instead they appear to unfold in a high-dimensional latent state space, which the Deep VAR by its very construction is designed to learn.

Following the Volcker disinflation period, Figure 1 does not reveal any clear outperformance of either of the models during the 1990s. Interestingly the dot-com bubble has little affect on either of the models, aside from a small pick-up in cumulative loss with respect to the CPI for both models. With all that noted, the Deep VAR still continuously outperforms the VAR since evidently its cumulative loss increases at a lower pace altoghether.

As the Global Financial Crisis unfolds around 2007 the pattern we observed for the Volcker disinflation reemerges, albeit to a lesser extent: there is a marked jump in the difference between the cumulative loss of the VAR and the Deep VAR, in particular so for the CPI, the Fed Funds rate and industrial production. The gap for all these variables continues to widen during the aftermath of the crisis. The Deep VAR once again does a better job at modelling the changes that the dynamical system undergoes: post-crisis US monetary policy was characterized by very low interest rates, low levels of inflation as well as the introduction of a range of non-conventional monetary policy tools including quantitative easing and forward guidance.

Finally, it is also interesting to observe how both models perform in response to the unprecedented exogenous shock that Covid-19 constitutes. Both models incur huge errors with respect to both IP and UR - the two series most significantly affected by Covid. Evidently though, the magnitude of the errors is somewhat larger for the VAR than for the Deep VAR. This, once again seems to confirm our hypothesis that the Deep VAR model captures important non-linear dependencies across time and variables that the conventional VAR fails to capture.

As a sanity check we also visually inspected the distributional properties of the model residuals for the full-sample fit. The outcomes are broadly consistent across models: while for some variables residuals are clearly not Gaussian, we see no evidence of serial autocorrelation of residuals (see Figures **??** and **??** in the appendix).

Table 1: Root mean squared error (RMSE) for the two models across subsamples and variables.

Sample	Variable	DVAR	VAR	Ratio (DVAR / VAR)
test	IP	0.00608	0.01484	0.40957
test	UR	0.97022	1.65170	0.58741
test	CPI	0.00273	0.00342	0.79869
test	FFR	0.19964	0.23974	0.83271
train	IP	0.00528	0.00727	0.72610
train	UR	0.32325	0.43322	0.74615
train	CPI	0.00149	0.00232	0.64019
train	FFR	0.16115	0.25780	0.62509

2.2 Out-of-sample fit

In order to assess if the Deep VAR's outperformance is a consequence of overfitting, we now repeat the previous exercise, but this time we train the models on a subsample of our date. The training sample spans from March, 1959 to October, 2008, whereas the test data goes from November, 2008 to March, 2021. This corresponds to training the model on 80 percent of the data and retaining the remaining 20 percent for testing purposes. The optimal lag order for the training subsample is $p = 7$ where we use the same criterion and maximum lag order as before. Once again we find this VAR specification to be stable.

Tables 1 shows the Root Mean Squared Error (RMSE) for the in-sample and the out-of-sample predictions of both the VAR model and the Deep VAR model. We can see that the RMSE for the Deep VAR outperforms the one for the conventional VAR for both the training data and the test data and for all time series. The fifth column of the table shows us the ratio between the RMSEs of the Deep VAR and the VAR: the lower the ratio, the better the Deep VAR compared to the VAR. With respect to the training sample, the RMSE of the Deep VAR model is consistently less than 75% of that of the conventional VAR reflecting to some extent the results of the previous sections. Turning to the test data, there is no evidence that the Deep VAR is more prone to overfitting than the VAR. For both industrial production and unemployment, the Deep VAR yields an RMSE that is around half the size of that produced by the VAR. For inflation and interest rate predictions the outperformance on the test data is less striking, but still fairly significant.

2.3 Forecasts

Up until now we have been assessing the 1-step ahead predictions of both models. In our context these predictions can be thought of as 1-month ahead nowcasts from a practical perspective. Since real-time nowcasts have grown in popularity during recent years, the results so far should be of great interest to central bankers and other practitioners. Nonetheless, there is typically also great interest in time series forecasts at longer horizons. We therefore

Table 2: Comparison of n-step ahead pseudo out-of-sample forecasts.

Variable	VAR FRMSE	Deep-VAR FRMSE	VAR correlations	Deep-VAR correlations
IP	0.01870	0.01602	-0.30409	-0.65279
UR	0.85984	0.82785	-0.10093	0.27425
CPI	0.00946	0.00708	-0.33567	0.07823
FFR	0.52321	0.39161	-0.55935	0.01161

briefly introduce n-step ahead pseudo out-of-sample forecasts in this section and revisit them again further below.

Forecasts are produced recursively both for the VAR and the Deep VAR. Specifically, we use the models we trained on the training data to recursively predict one time period ahead, concatenate the predictions to the training data and repeat the process.[2] This way we produce one-year ahead forecasts beginning from the first date in the test sample (October, 2008).

Table 2 shows the resulting root mean squared forecast errors (RMSFE) along with correlation between forecasts and realizations. As we can see in the table, the RMSFE of the Deep VAR is consistently lower than the one for the VAR. Regarding correlations the VAR produces forecasts that are negatively correlated with actual outcomes for all time series: in other words, when the time series evolves in one direction, the VAR forecast tends to evolve in the opposite direction. For industrial production, the Deep VAR forecast also has a highly negative correlation with the actual values. For the rest of time series the Deep VAR forecasts correlate positively with actual outcome, albeit weakly. Another general observation we made with respect to these forecasts is that the forecasts from the conventional VAR are fairly volatile, while the Deep VAR forecasts swiftly revert to steady levels (see Figures **??** and **??** in the appendix).

2.4 Varying hyperparameters

While up until now with respect to model selection we have intentionally remained strictly within the conventional VAR framework, we will now relax that constraint and vary the lag length as well as hyperparameters of the Deep VAR. In particular, we perform a grid search where we vary the number of hidden layers (1,2,5), number of hidden units per layer (50,100,150), the dropout rate (0.3,0.5,0.7) and the lag order (10, 50, 100). For each combination of parameter choices we train the two models and compute the various performance measures introduced above.[3] Our expectation is that the conventional VAR is prone to overfitting and will produce poor out-of-sample outcomes for higher lag orders. For the Deep VAR we expect to interesting variation in the outcomes for different lag order and hyperparameter choices. It is not clear ex-ante that the Deep VAR should suffer from the

[2]Note that for the Deep VAR an alternative approach would be to work with a different output dimension for the underlying neural networks.

[3]Of course, with respect to the conventional VAR only the lag order affects outcomes.

same issue of overfitting for higher lag orders. The bulk of the corresponding visualizations can be found in the appendix.

2.4.1 Tuning the Deep VAR

To begin with, we shall forget about benchmarking for a moment and focus on the outcomes for the Deep VAR as we vary parameters. Recall that a higher number of hidden layers (depth), a higher number of hidden units (width) and a smaller choice for the dropout rate all correspond to an increase in neural network complexity. Consistent with this intuition we find that the in-sample loss for the Deep VAR improve as complexity increases (Figure **??**): higher complexity leads to a reduction in bias and as we noted earlier the underlying recurrent neural networks should in principle be able to model arbitrary functions (Goodfellow, Bengio, and Courville 2016). Conversely, we observe exactly the opposite pattern for out-of-sample loss: as evident from Figure **??** a higher choice for the dropout rate and lower choices for the depth and width of the neural networks generally yields a smaller out-of-sample RMSE across variables.

Interestingly, both in- and out-of-sample loss tend to decrease significantly as the number of lags increases. In other words, the Deep VAR seems to be relatively insensitive to overfitting with respect to the lag order. With that in mind, we find that using standard lag order selection tools such as the AIC above may in fact not be appropriate for Deep VARs.

Finally, Figure **??** provides an overview of how pseudo out-of-sample forecasting errors behave as we vary the hyperparameters. As before we produce one-year ahead forecasts starting from the end of the 80% training sample. In this context, the pattern is less clear and varies across variables. As the lag order increases, for example, the forecast performance for the unemployment rate deteriorates. For inflation, forecasts are poor for the medium lag choice of $p = 50$ and much better for the low and high lag orders. The exact opposite relationship appears to hold for the Fed Funds Rate. With respect to the choices for the Deep VAR hyperparameters it is difficult to establish any clear pattern at all. The magnitude of differences in RMSFE is generally very small, so overall we conclude that to some extent the variation we do observe may be random.

In light of this evidence, we propose that for the purpose of hyperparameter tuning Deep VARs researchers should focus on the RMSE associated with the 1-step ahead fitted values. For the underlying data, a reasonable set of hyperparameter choices could be: 1 hidden layer, 50 hidden units and a dropout rate of 0.5.

2.4.2 Benchmark

Using the hyperparameter choices proposed above we now turn back to comparing the performance of the Deep VAR to the conventional VAR. Figure 2 shows the pseudo out-of-sample RMSE and RMSFE for both models across the different lag choices. For the sake of completeness we also include the performance measures we obtained when we initially ran both models in section 2.2 using the optimal lag order as determined by the AIC.

The first observation is that the Deep VAR outperforms the VAR across the board, reflecting our earlier findings. As expected, the VAR is subject to overfitting for when high lag order are chosen. This trend is observed both for the RMSE as well as the RMSFE. The fact that n-step ahead forecasts of the VAR are also subject to overfitting with respect to the lag order, while the Deep VAR appears unaffected, to some extent may reflect what we observed earlier: for the given data, Deep VAR forecasts swiftly converge to steady levels, while VAR forecasts are volatile, which may explain the relative outperformance of the Deep VAR. It appears that this effect is amplified for higher lag orders.

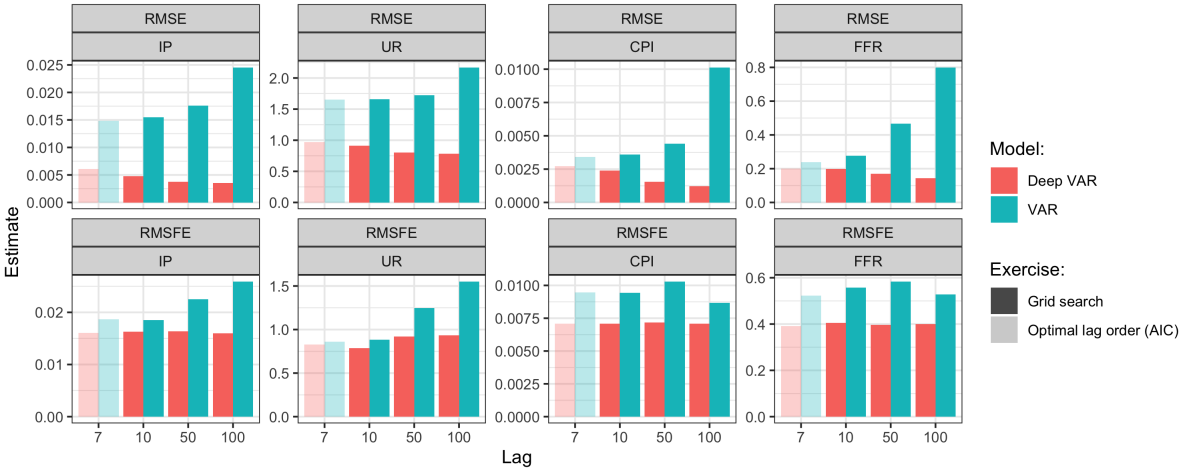

Figure 2: Pseudo out-of-sample RMSE and RMSFE for both models across the different lag choices. For the sake of completeness we also include the performance measures we obtained when we initially ran both models using the optimal lag order as determined by the AIC.

To conclude this empirical section we summarize our main findings:

1. We provide evidence that the conventional, linear VAR fails to capture important non-linear dependencies across time and variables that are typically used to model the monetary transmission mechanism.
2. Tapping into the broader class of Deep VARs leads to consistently better modelling outcomes.
3. Deep VARs appear to be relatively insensitive to very high lag orders at which conventional VARs are prone to overfitting.

3 Conclusions

Our initial motivation for this study was to see if by incorporating some of the latest developments from the machine learning and deep learning domains in the conventional VAR framework, we could attain improvements in the modelling and forecasting performance. In an effort not to deviate too much from the established framework, we only relax one single assumption to move from the conventional linear VAR to a broader class of models that we refer to as Deep VARs.

To assess the modelling performance of Deep VARs compared to linear VARs we investigate a sample of monthly US economic data in the period 1959-2021. In particular, we look at variables typically analysed in the context of the monetary transmission mechanism including output, inflation, interest rates and unemployment. Our empirical findings show a consistent and significant improvement in modelling performance associated with Deep VARs. In particular, our proposed Deep VAR produces much lower cumulative loss measures than the VAR over the entire period and for all of the analysed time series. The improvements in modelling performance are particularly striking during subsample periods of economic downturn and uncertainty. This appears to confirm or initial hypothesis that by modelling time series through Deep VARs it is possible to capture complex, non-linear dependencies that seem to characterize periods of structural economic change.

When it comes to the out-of-sample performance, a priori it may seem that the Deep VAR is prone to overfitting, since it is much less parsimonious that the conventional VAR. On the contrary, we find that by using default hyperparameters the Deep VAR clearly dominates the conventional VAR in terms of out-of-sample prediction and forecast errors. An exercise in hyperparameter tuning shows that its out-of-sample performance can be further improved by appropriate regularization through adequate dropout rates and appropriate choices for the width and depth of the neural. Interestingly, we also find that the Deep VAR actually benefits from very high lag order choices at which the conventional VAR is prone to overfitting. In summary, we provide solid evidence that the introduction of deep learning into the VAR framework can be expected to lead to a significant boost in overall modelling performance. We therefore conclude that time series econometrics as an academic discipline can draw substantial benefits from further work on introducing machine learning and deep learning into its tool kit.

We also point out a number of shortcomings of our proposed Deep VAR framework, which we believe can be alleviated through future research. In particular, policy-makers are typically concerned with uncertainty quantification, inference and overall model interpretability. Future research on Deep VARs should therefore address the estimation of confidence intervals, impulse response functions as well as variance decompositions typically analysed in the context of VAR models. We point to a number of possible avenues, most notably Monte Carlo dropout and a Bayesian approach to modelling deep neural networks.