

# Forecasting VAR

Alexander Haider<sup>\*1</sup>

<sup>1</sup>The New School for Social Research, New York, NY.

October 2019

## Abstract

Abstract...

*Keywords:* Bayesian vector autoregression, Forecasting, Non-linear Modelling, Stochastic Volatility

*JEL Classification:* C53, E37, E44, E50

---

<sup>\*</sup>[haidera@newschool.edu](mailto:haidera@newschool.edu)

# 1 Introduction

More than ten years after the onset of the Global Financial Crisis the effects on the U.S. economy are still ubiquitous. In macroeconomic research the crash of 2007-2009 led to a reexamination of the interaction between financial markets and the real side of the economy. The role of financial markets has been neglected for a long time in many macroeconomic models. In the predominant DSGE tradition financial markets were often not explicitly modelled. In cases where they were integrated into the model, the locally magnifying financial accelerator model of Bernanke and Gertler (1989) and Bernanke et al. (1999) dominated.

At the same time authors such as Aliber and Kindleberger (2015) and Minsky (1982, 2008) stressed the instability of the financial system and the possibility of economic crisis; theories of speculation, credit expansion and leverage, finally cumulating in financial distress and crisis ‘survived’ outside standard macroeconomics for a long time.<sup>1</sup> These destabilizing effects of financial manias are nowadays even accounted for in more standard approaches; see for example Brunnermeier and Sannikov (2014). Empirical evidence clearly suggests that the U.S. economy is distinguished by two financial regimes, which may be described as a tranquil regime and a stress regime, as pointed out by He and Krishnamurthy (2014) and Mitnik and Semmler (2013).<sup>2</sup>

Non-linear linkages and financial market regime switching are now firmly established in macroeconomic modelling. However, they are rarely accounted for in forecasting. In fact, forecasting with multivariate non-linear autoregressive models remains a neglected research area (Hubrich and Teräsvirta, 2013): most studies focus on structural analysis with non-linear models, while only few researchers evaluate their forecasting performance.

The lack of forecasting studies with financial variables and non-linear models is not very surprising. Financial market indicators are often seen as too noisy, making signal extraction for macroeconomic forecasting difficult. Limited success of such a strategy is reported, for example, by Stock and Watson (2003). Regarding forecasts based on non-linear models Teräsvirta (2006) observes that small sample sizes may affect non-linear models more strongly than linear models. As a result forecasts of non-linear models will be of low quality, even if the non-linear model is the more accurate description of the data generating process. This result is not surprising given the stronger parameterization of most non-linear

---

<sup>1</sup>Of course, the role of ‘overtrading’ and its linkages with financial distress has already been formulated in the classical tradition as pointed out by Aliber and Kindleberger (2015).

<sup>2</sup>Evidence on these strong non-linear effects of financial regimes are not confined to the U.S. economy. These linkages can also be found for European economies as described in Mitnik and Semmler (2013) and Schleer and Semmler (2015).

models compared with linear models. Clearly this result constitutes a serious problem in macroeconomic research given the short to moderate sample size of most macroeconomic time series. Furthermore, problems of overfitting, the misclassification of observations, and the explanation of infrequent events may prohibit competitive forecasts by non-linear models.

Still, a small number of studies focusing on forecasting with multivariate non-linear models exists. Examples are given by Alessandri and Mumtaz (2017) and Galvão (2006), but results remain inconclusive. Galvão (2006) reports positive results on forecasting the 2001 recession based on a structural break threshold vector autoregression. On the other hand, Alessandri and Mumtaz (2017) find that their threshold VAR (*TVAR*) only improves forecasts over the recession period following the Global Financial Crisis compared with a VAR. At the same time, a VAR with stochastic volatility improves forecasting performance over the entire sample period and would have been the preferred forecasting device at the beginning of the crisis.<sup>3</sup>

From a methodological point of view this study is very closely related to the approach taken by Alessandri and Mumtaz (2017). I build on a Bayesian framework to compare forecasting performances of VARs, threshold VARs, and VARs with stochastic volatility. The prior distribution is based on the work of Bańbura et al. (2010). I assess point forecasts for all three models for a forecast horizon up to twelve months. In addition, I take uncertainty of the forecasts into account as well by evaluating density forecasts. In contrast to Alessandri and Mumtaz (2017) I build on a larger models with 6 variables. Bańbura et al. (2010) point out that conditioning on a larger dataset influences forecast performances and impulse response dynamics. Moreover, I evaluate density forecasts based on multivariate scoring rules and analyze pooling forecasts to improve forecast performance.

My focus on VARs, TVARs and VARs with stochastic volatility is twofold. First of all, the choice of these three models is ‘natural’ in the sense that a baseline model is needed. The baseline model is given by the VAR with a constant covariance matrix. The first contender model is given by the TVAR. The TVAR model takes the non-linear linkages discussed above explicitly into account. However, the threshold model is considerably more complex than the linear model and therefore more costly to deploy. Thus for the TVAR to become a candidate for forecasting it has to ‘outperform’ the VAR. This is not an easy task. It is well known in the forecasting literature that simple models work well and this result holds for VARs as well (Karlsson, 2013). The third model, the VAR with stochastic volatility, is chosen because

---

<sup>3</sup>It should be noted here that Galvão (2006) works with real time data while Alessandri and Mumtaz (2017) do not.

time-varying volatility is seen as a pervasive feature of macroeconomic data for the U.S. (see, for example, Stock and Watson, 2002). VARs with stochastic volatility have therefore been widely applied in structural analysis and forecasting and their excellent forecasting performance is documented by Alessandri and Mumtaz (2017), Clark and Ravazzolo (2015), and Ravazzolo and Vahey (2014). The VAR with stochastic volatility is therefore used as an even tougher competitor for the TVAR.

Secondly, a focus on these three models allows for a comparison with the results obtained in Alessandri and Mumtaz (2017). In contrast to their results, I show ...

The remainder of the paper is organized as follows. The next section introduces the prior for the VAR parameters used for all three models. Section 3 discusses the three models, additional model specific priors and estimation strategy. Forecast evaluation metrics are introduced in Section 4. The dataset and the results are discussed in Section 5. Finally, Section 6 concludes.

## 2 Prior Choice

I build on the prior for medium and large Bayesian VARs suggested by Bańbura et al. (2010). They use a natural conjugate prior based on the Minnesota prior (Doan et al., 1984; Litterman, 1986). The Minnesota prior represents a shrinkage prior to account for over-parameterization in VAR models, which can affect even small VARs with only a few variables. The autoregressive parameters are shrunk towards a random walk for persistent variables and towards white noise for non-persistent variables. The prior of Bańbura et al. (2010) controls for this shrinkage via a single scalar hyperparameter,  $\lambda$ . Setting this parameter is crucial and discussed in Section 5 in detail. In brief, a small VAR is estimated via OLS on a training sample and  $\lambda$  is adjusted such that the in-sample fit of the forecasting model approaches the in-sample fit of the small VAR.

Before discussing the prior in detail it is instructive to describe the basic structure of a VAR. A VAR can be written as

$$Y = XB + E, \tag{1}$$

with  $Y$  being a  $T \times N$  matrix. The rows of matrix  $Y$  represent  $T$  observations of a time series vector,  $Y'_t$ , of size  $N$ . Matrix  $X$  is the regressor matrix and is of dimension  $T \times K$  with  $K = N \times P + 1$  where  $P$  is the lag length. Row  $t$  of  $X$  is given by  $X_t = (Y'_{t-1}, \dots, Y'_{t-P}, 1)$ .

Thus a constant is the only exogenous variable in the VAR considered here. Matrix  $B$  represents the coefficient matrix of the VAR,  $B = [B_1, \dots, B_p, c]'$  with  $B_i$  containing the autoregressive coefficients for lag  $i$ , and  $c$  being the intercept. Finally, matrix  $E$  is of size  $T \times N$  and the residuals for period  $t$  are given by  $E_t$ . In the most basic setting it is assumed that  $E_t \sim \mathcal{N}(0, \Sigma)$ . The covariance matrix is of dimension  $N \times N$  and positive definite. The natural conjugate prior is normal-inverse Wishart with

$$p(\beta|\Sigma) \sim \mathcal{N}(\beta_0, \Sigma \otimes \Omega_0), \quad (2)$$

$$p(\Sigma) \sim \mathcal{IW}(S_0, a_0), \quad (3)$$

with  $\beta = \text{vec}(B)$ ,  $\beta_0 = \text{vec}(B_0)$  and  $\text{vec}(\cdot)$  representing the vectorization operator. The prior hyperparameters  $B_0, \Omega_0, S_0, a_0$  are data-based hyperparameters and set similar to Bańbura et al. (2010). Koop (2013) points out that the natural conjugate prior can be integrated in the dataset as arising from a fictitious sample of prior observations. Typically these observations reflect specific assumptions regarding the behavior of macroeconomic data to be discussed below. The dummy (or pseudo) observations are given by

$$Y_d = \begin{pmatrix} \text{diag}(\delta_1\sigma_1, \dots, \delta_n\sigma_n)/\lambda \\ 0_{N(P-1) \times N} \\ \text{diag}(\sigma_1, \dots, \sigma_N) \\ 0_{1 \times N} \end{pmatrix}, \quad X_d = \begin{pmatrix} J_p \otimes \text{diag}(\sigma_1, \dots, \sigma_N)/\lambda & 0_{NP \times 1} \\ 0_{N \times NP} & 0_{N \times 1} \\ 0_{1 \times NP} & \varepsilon \end{pmatrix}, \quad (4)$$

with  $J_p = \text{diag}(1, \dots, P)$ . Parameter  $\delta_i$  is set equal to 0 for non-persistent variables and equal to 1 for persistent variables. Furthermore,  $\sigma_i$  is equal to the standard deviation of the residuals of an  $AR(P)$  regression for variable  $i$ . Lastly,  $\varepsilon$  is set to 0.0001, representing a diffuse prior on the constants.

$Y_d$  and  $X_d$  represent  $T_d$  dummy observations which are appended to  $Y$  and  $X$  given in equation (1). Adding these observations reflects implementing the priors described in equations (2) and (3) with  $B_0 = (X_d'X_d)X_d'Y_d$ ,  $\Omega_0 = (X_d'X_d)^{-1}$ ,  $S_0 = (Y_d - X_dB_0)'(Y_d - X_dB_0)$  and  $\alpha_0 = T_d - K$ .<sup>4</sup>

Moreover, following Alessandri and Mumtaz (2017) and Bańbura et al. (2010) I also impose additional priors on the sum of coefficients to improve forecasting performance. These additional dummy observations reflect the belief that most economic variables can be ade-

---

<sup>4</sup>See, for example Miranda-Agrippino and Ricco (2018) for a derivation of this result.

quately represented by unit root processes with weak cross-sectional linkages, or, in case of variables in first differences, as white noise processes. The additional pseudo observations are given by:

$$Y_{dd} = \text{diag}(\delta_1\mu_1, \dots, \delta_n\mu_n)/\tau, \quad X_{dd} = ((1_{1 \times P}) \otimes \text{diag}(\delta_1\mu_1, \dots, \delta_n\mu_n)/\tau \quad 0_{N \times 1}). \quad (5)$$

Additional hyperparameters  $\mu_i$  and  $\tau$  are introduced in equation (5). Hyperparameter  $\tau$  represents another shrinkage parameter and is set to  $\tau = 10\lambda$  as in Bańbura et al. (2010), implying that stronger shrinkage on the VAR parameters also leads to ‘more exact’ differences on the lags. Parameters  $\mu_i$  reflect the average value of variable  $y_{i,t}$  and is set equal to the sample average of the variables. Finally, appending both sets of dummy observations to the observed data results in  $Y^* = (Y', Y_d', Y_{dd}')'$  and  $X^* = (X', X_d', X_{dd}')'$ .

Regarding the posterior distribution, I define  $\tilde{B}$  as  $\tilde{B} = (X^{*'}X^*)^{-1}X^{*'}Y^*$  and  $\tilde{\Sigma} = (Y^* - X^*\tilde{B})(Y^* - X^*\tilde{B})'$ . The posterior is then defined as

$$p(\beta|\Sigma, Y) \sim \mathcal{N}(\text{vec}(\tilde{B}), \Sigma \otimes (X^{*'}X^*)^{-1}), \quad (6)$$

$$p(\Sigma|Y) \sim \mathcal{IW}(\tilde{\Sigma}, T_d + T). \quad (7)$$

In the context of this paper, using the dummy prior entails certain advantages and disadvantages; see Bańbura et al. (2010) and Koop (2013) for a more detailed discussion. A first advantage is given by the simple implementation of the prior. Appending the fictitious dataset given by the matrices  $X_d, X_{dd}$  and  $Y_d, Y_{dd}$  allows for efficient computation of the conditional posterior distribution  $p(\beta|\Sigma, Y)$  and the marginal posterior of the covariance matrix  $p(\Sigma|Y)$ . As can be seen from equation (6), the expected value of the posterior is the OLS estimate on the appended dataset, while the covariance matrix of  $p(\beta|\Sigma, Y)$  requires only the inversion of the  $K \times K$  matrix  $X^{*'}X^*$ , while ‘standard’ implementations of the Minnesota prior necessitate the inversion of a  $NK \times NK$  matrix. The advantage of inverting a smaller matrix can already be substantial in a relatively small model, given that

the matrix has to be inverted in each Gibbs iteration for the TVAR.<sup>5</sup>

On the other hand, the dummy observation prior entails certain limitations (Koop, 2013): while the original Minnesota prior allows for imposing additional shrinkage on lagged values of variable  $j \neq i$  in the equation for variable  $i$ , the dummy observation prior sets only one shrinkage parameter for own lags and lags of other variables. In addition, as can be seen from equation (3), the prior variance of the coefficients on the same explanatory variable in any two equations will be proportional due to the Kronecker product.

In the end, considerations concerning computation time dominate. As this study necessitates the estimation of ?? models I acknowledge the limitations of the dummy prior but nevertheless opt for it for gains in computational speed.

### 3 Forecasting Models

As stated in the introduction, I use three multivariate time series models to forecast GDP, inflation growth rates, interest rates and the 10-year/2-year spread: a VAR, a threshold VAR and a VAR with stochastic volatility model. All three models build on the dummy observation prior described in the last section. In this section I describe the implementation of the Gibbs sampling algorithms used in estimating the models.

#### 3.1 VAR

Equation (8) restates the VAR introduced above for a single observation,  $t$ , where  $Y_t$  represents a  $N \times 1$  vector containing the endogenous variables,  $c$  is the constant and  $P$  represents the lag length with  $P = 4$  for all models. Matrix  $B_i$  holds the autoregressive coefficients for lag  $i$  and  $\Sigma$  is the covariance matrix.

$$Y_t = c + \sum_{i=1}^P B_i Y_{t-i} + v_t, \quad v_t \sim \mathcal{N}(0, \Sigma). \quad (8)$$

---

<sup>5</sup>In the case of Bańbura et al. (2010) the main advantage of using the dummy observation prior is that it allows for an analytical solution which does not rely on simulation methods. In addition, the marginal posterior of the autoregressive parameters,  $B$ , can be recovered without simulation methods and the predictive density for the one period ahead forecast is available in analytical form as well (see also Koop, 2013). This won't be the case for the threshold VAR and the stochastic volatility model considered here. I therefore use simulation method for the VAR forecasts as well. Analytical results for the VAR are only used when finding the shrinkage parameter,  $\lambda$ , for the benchmark VAR.

As discussed in Section 2 the posterior of model (8) can be recovered analytically and the same holds for the one-step ahead forecast. This result won't hold for the TVAR and the stochastic volatility VAR where simulation methods for posterior distribution and forecasts have to be used. Therefore, I use simulation methods for the forecasts of the VAR as well. After evaluating  $\tilde{B}$ ,  $(X^{*'}X^*)^{-1}$  and  $T_d + T$  for a given sample, values for  $\beta$  and  $\Sigma$  are drawn via Gibbs sampling. The Gibbs sampler has been implemented such that unstable draws for the VAR are rejected. This is accomplished by building the companion matrix of the autoregressive coefficients for a  $\beta$  sample and evaluating its eigenvalues (Lütkepohl, 2005, page 15f).

A draw for  $B$  is stable if the modulus of the largest eigenvalue of the companion matrix is smaller than one.<sup>6</sup> The Gibbs sampler starts with a random draw for  $\Sigma$  drawn from the posterior distribution given by equation (7) and samples in turn from  $p(B^m|\Sigma^{m-1})$  and  $p(\Sigma^m|B^m)$  for  $m = 1, \dots, M$  with  $M$  being the number of replications.<sup>7</sup>

After the burn-in phase is completed, forecasts of the VAR are simulated for a given sample for  $B$  and  $\Sigma$  by forward iteration starting from the last observed values. Point and density forecasts and their evaluation – discussed in the next section – are based on these recursive forecasts. The predictive distribution,  $p(Y_{t+k}|Y_t)$ , is given by

$$p(Y_{t+k}|Y_t) = \int \int p(Y_{t+k}, B, \Sigma|Y_t) dB d\Sigma = \int \int p(Y_{t+k}|Y_t, B, \Sigma) p(B|Y_t, \Sigma) p(\Sigma|Y_t) dB d\Sigma, \quad (9)$$

with  $p(B|Y_t, \Sigma)p(\Sigma|Y_t)$  being the conditional posterior distribution. The predictive distribution can therefore be recovered from the saved Gibbs draws for  $h = 1, \dots, 4$  with  $h$  being the forecast horizon.

---

<sup>6</sup>This is implemented in the code by allowing for up to 10,000 drawing attempts of the coefficients. If stability is not accomplished after 10,000 draws the last stable  $B$  draw is retained and forecasts are not evaluated. In general, allowing stable draws only is not an uncontroversial issue, especially in structural analysis; see Cogley and Sargent (2005) for more details. I regard unstable draws – leading to explosive behavior – as irrelevant in a forecasting scenario. Therefore I disregard them.

<sup>7</sup>More details on the settings of the Gibbs sampler are provided in Section 5.



### 3.2 Threshold VAR

The *threshold VAR (TVAR)* with two regimes is given by

$$Y_t = \begin{cases} c_1 + \sum_{i=1}^P B_{1,i} Y_{t-i} + v_{1,t} = X_1 B_1 + v_{1,t} & \text{if } z_{t-d} \leq z^*, \\ c_2 + \sum_{i=1}^P B_{2,i} Y_{t-i} + v_{2,t} = X_2 B_2 + v_{2,t} & \text{if } z_{t-d} > z^*, \end{cases} \quad (10)$$

with  $c_1$  and  $B_{1,i}$  being the coefficients of the first regime and  $c_2$  and  $B_{2,i}$  representing the coefficients of the second regime. The model allows for changes in dynamic behavior and shock transmission. Variable  $z_{t-d}$  is an endogenous variable, belonging to vector  $Y_{t-d}$ . In the model employed here  $z_{t-d}$  is given by the *Chicago Fed National Financial Conditions Index*, as discussed in more detail below. Parameter  $d$  represents the delay parameters which will be estimated via the Gibbs sampling approach described below. The threshold value,  $z^*$ , is estimated as well. I use a flat prior over  $1, \dots, \max_d$  for the delay parameter with  $\max_d = P$ . A rather non-informative normal prior,  $\mathcal{N}(\bar{z}, 10)$ , is employed for the threshold value with  $\bar{z}$  being the sample mean of the threshold variable for a given sample. Furthermore,  $v_{1,t} \sim \mathcal{N}(0, \Sigma_1)$  and  $v_{2,t} \sim \mathcal{N}(0, \Sigma_2)$  such that the TVAR allows for changes in volatility between the regimes as well. Matrices  $B_1$  and  $B_2$  are the coefficient matrices and matrices  $X_1$  and  $X_2$  are the regressors, as before.

I focus on a threshold model with two regimes for two reasons. Foremost, there is ample evidence of two financial regimes in the U.S. macroeconomy as already mentioned in the introduction. In addition, focusing on a TVAR with only two regimes reduces complexity of the estimation procedure. Chen and Lee (1995) show for the univariate threshold model that a Metropolis-within-Gibbs algorithm allows for estimating all parameters of interest in a straightforward way (see also Alessandri and Mumtaz, 2017). Below,  $p(\Phi|\Xi)$  denotes the full conditional posterior distribution for parameter  $\Phi$  conditional on all the other parameters of the model. The Gibbs sampling procedure works as follows:

1. starting values are initialized for  $d$  and  $z^*$ . For  $d$  a random starting value between 1 and  $\max_d$  is chosen. Defining  $d$  also implies choosing a random starting vector  $z_{t-d}$ . From the random starting vector,  $z_{t-d}$ , a random starting value for  $z^*$  is chosen such that both regimes hold at least  $K$  observations. In doing so, I split the sample in two starting regimes with  $Y_1$  containing all  $Y_t$  vectors with the corresponding  $z_{t-d}$  being smaller or equal to  $z^*$ . The corresponding right-hand side variable for  $Y_1$  are collected in matrix  $X_1$  where the time series structure of the observations is maintained. In a

similar way,  $Y_2$  and  $X_2$  collect the left hand side and right-hand side observations for  $z_{t-d} > z^*$ .

2. Given  $Y_1, X_1$  and  $Y_2, X_2$  the model becomes linear in the two regimes and the dummy prior observations from Section 2 are appended to both regimes. As pointed out by Alessandri and Mumtaz (2017), using the same prior structure in both regimes might seem counter-intuitive, but given the strong parameterization of the TVAR model shrinkage is crucial. Furthermore, the dummy prior does not impose strong beliefs on the distinct dynamics of the model which leads to data determined behavior in the two regimes.
3. After estimating the  $B$  matrices for both regimes via a simple OLS regression on the appended dataset, I draw starting value for  $\Sigma_1$  and  $\Sigma_2$  based on equation (7) with  $\tilde{S}_1$  and  $\tilde{S}_2$  being the sum of squared residuals of the two linear models described before. The degrees of freedom for the inverse Wishart distributions are given by  $T_d + T_1$  for the first regime and  $T_d + T_2$  for the second regime. Drawing starting values for  $\Sigma_1$  and  $\Sigma_2$  is repeated until positive definite matrices are obtained. Based on these starting values the Gibbs sampler is started.
4. **Drawing  $p(B_1|\Xi), p(\Sigma_1|\Xi), p(B_2|\Xi), p(\Sigma_2|\Xi)$  in turn:** observations are split up into the two regimes and arranged in  $Y_1, X_1, Y_2, X_2$ . The dummy observations are appended for both regimes. Then  $B_1$  and  $B_2$  are drawn the same way as in the VAR, including the test for stability. The same holds for the  $\Sigma_1$  and  $\Sigma_2$  samples; they are drawn in the same way as in the linear model with the degrees of freedom adjusting for the number of observations in each regime.
5. **Drawing  $p(z^*|\Xi)$ :** the threshold value is drawn via a random walk Metropolis step within the Gibbs sampler. The candidate value,  $z_c$ , is drawn as  $z_c \sim \mathcal{N}(z^*, \psi_1)$  with  $\psi_1$  being the scale parameter. Tuning  $\psi_1$  is accomplished via a *Delayed Rejection Adaptive Metropolis (DRAM)* step (Haario et al., 2006) where I impose again the restriction that at least  $K$  observations have to be in each regime. To the best of my knowledge this procedure has not been applied for the TVAR model so far, but it was tested and worked well during estimation. DRAM is the combination of a *Delayed Rejection (DR)*; Mira, 2001) step with an *adaptive Metropolis (AM)*; Haario et al., 2001) adjustment.

In AM  $\psi_1$  is adapted during the MCMC procedure to find an optimal acceptance rate for the Metropolis step. Starting from an initial value of 0.001, the scale parameter

is adapted according to  $\psi_1 = (\text{Cov}(\text{chain}_{\psi_1, 1:i}) + \delta_{AM})s$  with a periodicity (adaptation frequency) of 50, starting at iteration 100. Thus,  $\psi_1$  depends on the covariance of the chain and parameter  $\delta_{AM}$  which is set to a small number,  $1e-8$ , to prevent  $\psi_1$  from becoming zero. Finally, parameter  $s$  represents the scaling factor and is given by  $s = 2.4^2$ . Although the adaption procedure tunes  $\psi_1$  well, it has to be noted here that it destroys the Markovian property of the chain.<sup>8</sup> The AM step is therefore ‘turned off’ before the burn-in phase is completed. I use the (arbitrary) rule that adaptation ends as soon as the iteration number exceeds  $\min(\text{replications} \times 0.75, \text{burn-in} \times 0.95)$ . The *DR* procedure, on the other hand, starts immediately with the first iteration. It is activated as soon as a candidate value is rejected. Instead of advancing with the next iteration and retaining the old position, a second candidate is proposed. Its visiting distribution is given by the tighter distribution  $\mathcal{N}(z^*, \psi_2)$  with  $\psi_2 = \psi_1/3$ . Using the DR step has the advantage of an asymptotically more efficient chain than a straightforward Metropolis-Hastings chain (Mira, 2001). However, the acceptance probability has to be adapted for the second proposal such that reversibility of Markov chain is not destroyed. In general, it is possible to draw more than one additional candidate value in case the second proposal is rejected as well. To keep the sampling procedure relatively simple I only work with one *DR* step which results in one (in case the first value is already accepted) or two proposals in each iteration. The first *DR* step follows the standard acceptance rule for a random walk Metropolis algorithm,  $\alpha_1(z^*, z_c) = \min(1, p(z_c|\Xi, Y)/p(z^*|\Xi, Y))$  with  $p(z_c|\Xi, Y)$  being the posterior based on the candidate and  $p(z^*|\Xi, Y)$  representing the posterior based on  $z^*$ . The second *DR* step, if activated, computes the new acceptance probability for candidate  $z_{cc}$  as

$$\begin{aligned}\alpha_2(z^*, z_c, z_{cc}) &= \min(1, w_1/w_2), \\ w_1 &= p(z_{cc}|\Xi, Y) \times q_1(z_{cc}, z_c) \times q_2(z_{cc}, z_c, z^*) \times [1 - \alpha_1(z_{cc}, z_c)], \\ w_2 &= p(z^*|\Xi, Y) \times q_1(z^*, z_c) \times q_2(z^*, z_c, z_{cc}) \times [1 - \alpha_1(z^*, z_c)].\end{aligned}\tag{11}$$

Here  $q_1(x, y)$  represents drawing  $y$  when the current position is given by  $x$  under  $\mathcal{N}(x, \psi_1)$ , while  $q_2(x, y, z)$  represents drawing  $y$  followed by  $z$  when the current po-

---

<sup>8</sup>Ergodicity of the chain, on the other hand, is maintained by the AM step.

sition is given by  $x$  under  $\mathcal{N}(x, \psi_2)$  and then  $N(y, \psi_2)$ .<sup>9</sup> It is important to note here that the DR step retains the Markovian property and reversibility of the chain. Thus, there is no need of ‘turning it off’ after the burn-in phase.

6. **Drawing  $p(d|\Xi)$ :** given a flat prior on  $d$ , the conditional posterior for  $d$  follows a multinomial distribution with probability  $\mathcal{L}(Y|d_i, \Xi) / \sum_{j=1}^{max_d} \mathcal{L}(Y|d_j, \Xi)$  for  $d_i$ , where  $\mathcal{L}(\cdot)$  is the likelihood function.

Finally, forecasts are simulated recursively as in Section 3.1. Here I have to integrate out  $B_1, B_2, \Sigma_1, \Sigma_2, z^*$  and  $d$ . The simulation procedure therefore becomes slightly more involved as I have to account for draws of  $z^*, d$  and regime switches during forecasts.

### 3.3 VAR with Stochastic Volatility

The *VAR with stochastic volatility (SVOL)* can be written as:

$$\begin{aligned} Y_t &= c + \sum_{i=1}^P B_i Y_{t-i} + v_t = XB + v_t, \\ v_t &= A^{-1} H_t^{1/2} \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I), \\ H_t &= \text{diag}(h_{1,t}, \dots, h_{N,t}), \\ \log(h_{i,t}) &= \log(h_{i,t-1}) + u_{i,t}, \quad u_{i,t} \sim \mathcal{N}(0, g_i). \end{aligned} \tag{12}$$

The first equation in system (12) represents the linear autoregressive structure with  $B$  once again collecting the intercept and autoregressive coefficients and  $X$  containing all regressors. The second equation depicts the time-varying covariance structure. The covariance matrix is implicitly given by  $\Sigma_t = A^{-1} H_t A^{-1'}$  with  $A$  being lower triangular, constant, and with ones on the diagonal. The elements of  $A$  below the diagonal are referred to as the ‘free elements’ of  $A$ . As can be seen from the third equation in system (12),  $H_t$  is diagonal and time-varying. The last equation represents the stochastic volatility process for the orthogonal errors. The logarithm of the squared volatility terms follow a random walk.

Additional priors are needed for the free elements of  $A$  and the variances of the stochastic volatility processes,  $g_i$ . Following Alessandri and Mumtaz (2017) I use uninformative priors for both of them:

---

<sup>9</sup>Note that equation (11) simplifies substantially here because I am working with a Gaussian random walk visiting distribution. Consequently the  $q_2(\cdot, \cdot, \cdot)$  terms in  $w_1$  and  $w_2$  simplify and cancel out eventually.

$$p(a_{ij}) \sim \mathcal{N}(0, 1000) \quad \text{for } i > j,$$

$$p(g_i) \sim \mathcal{IG}(1, 0.0001) \quad \text{for } i = 1, \dots, N,$$

with  $a_{ij}$  being the free elements of  $A$ . Moreover,  $\mathcal{IG}(\alpha_{IG}, \beta_{IG})$  is the inverse-gamma distribution with shape parameter  $\alpha_{IG}$  and scale parameter  $\beta_{IG}$ . Lastly, the  $h_{i,t}$  terms are drawn via the independence Metropolis-Hastings algorithm presented in Jacquier et al. (1994). The algorithm is designed for an univariate series,  $h = (h_1, \dots, h_T)'$ , and draws one stochastic volatility term at a time (single-move algorithm). Jacquier et al. (1994) show that the draw for period  $t$  depends on draws from period  $t - 1$  and  $t + 1$ . Thus a starting value,  $h_0$ , is needed for drawing  $h_1$ , while  $h_T$  only depends on  $h_{T-1}$ . The prior for  $h_0$  is given by  $\log h_0 \sim \mathcal{N}(\mu_h, \sigma_h^2)$ . The mean of the distribution,  $\mu_h$  is set equal to the log of the variance of an  $AR(1)$  regression on a training sample of size  $??$ . The training sample is excluded from the Gibbs sampler. Given the uncertainty concerning the stochastic volatility terms,  $\sigma_h^2$  is set equal to 100 to form an uninformative prior. The posterior distribution for  $h_1, \dots, h_T$  can be found in Jacquier et al. (1994).

Cogley and Sargent (2005) describe a Metropolis-within-Gibbs algorithm which allows for estimating the model. The sampling procedure for the VAR with stochastic volatility works as follows:

1. obtain starting values for the Kalman filter: starting values for the mean of  $B$ ,  $b_{00}$ , are obtained by a linear regression of the dummy observations  $Y_d$  on  $X_d$ . The starting value for the variance of  $B$ ,  $p_{00}$  is set to  $p_{00} = \Sigma_d \otimes (X_d' X_d)^{-1}$  with  $\Sigma_d$  being a diagonal matrix with the covariances of the residuals from the dummy  $AR(P)$  model on the diagonal.
2. Starting values for the VAR residuals are obtained by a VAR estimated via OLS after removing a training sample of size  $??$  which will be used in the next step. Starting values for  $h_{i,t}, i = 1, \dots, N, t = 0, \dots, T$  are set equal to the first difference of  $Y$  squared. Two additional values are added to adjust for size differences and a small value,  $\varepsilon = 0.0001$ , is added to each observation. This is done to avoid zero values for the stochastic volatility terms which would cause problems for the sampling algorithm.
3. **Drawing  $p(g_i|\Xi)$ :** given the prior distribution described above, they are distributed as  $IG((u'_{i,t} u_{i,t} + 0.0001)/2, (T + 1)/2)$  with  $u_{i,t}$  being the residuals from the stochastic

volatility process:  $u_{i,t} = \log(h_{i,t}) - \log(h_{i,t-1})$ .

4. **Drawing**  $p(a_{ij}|\Xi)$ : defining  $Av_t = e_t$  results in  $e_t \sim \mathcal{N}(0, H_t)$  by definition. Rearranging yields  $N - 1$  linear regressions of the form

$$v_{i,t} = - \sum_{j=1}^{i-1} a_{i,j} v_{j,t} + e_{i,t} \text{ for } i = 2, \dots, N. \quad (13)$$

These equations exhibit a known form of heteroskedasticity which is given by the last draw of  $h_{i,t}$  available from the Gibbs sampler. Heteroskedasticity can be removed by dividing equation  $i$  by  $\sqrt{h_{i,t}}$  which yields  $\tilde{v}_{i,t} = v_{i,t}/\sqrt{h_{i,t}}$  and  $\tilde{v}_{-i,t} = \sum_{j=1}^{i-1} v_{j,t}/\sqrt{h_{i,t}}$ . Now a linear regression with homoscedastic errors,  $e_{i,t} \sim \mathcal{N}(0, 1)$ , can be estimated. Given the prior on the free elements described above, this allows for drawing posterior values for the terms below the diagonal of  $A$ :

$$\begin{aligned} a_{i,j} &\sim \mathcal{N}(M^*, V^*), \\ V^* &= (\text{diag}(1/1000) + \tilde{v}'_{-i,t} \tilde{v}_{-i,t})^{-1}, \\ M^* &= V^* (\tilde{v}'_{-i,t} \tilde{v}_{i,t}). \end{aligned} \quad (14)$$

5. **Drawing**  $p(B|\Xi)$ : given a time-varying covariance matrix, drawing  $B$  is done via the Kalman filter with  $p(B|\Xi, Y) \sim \mathcal{N}(B_{T|T}, P_{T|T})$ . Period  $T$  stands for the last iteration of the Kalman filter. The implementation of the Kalman filter builds on a state-space model with a time-varying variance in the observation equation,  $\Sigma_t$ , but a transition equation without an error term ( $B$  is assumed to be time-invariant). As in Section 3.1 I test for stability of the draw via the eigenvalues of the companion matrix. In a last step the errors,  $v_t$ , are computed based on the last accepted draw of  $B$ .
6. **Drawing**  $p(h_{i,t}|\Xi)$ : Using the current draw for  $A$  and  $v_t$ , I recalculate  $e_t = Av_t$ . As already mentioned above,  $v_t \sim \mathcal{N}(0, H_t)$ , with  $H_t$  being diagonal. This implies that the  $e_t$  is contemporaneously uncorrelated and each element of  $e_t$ ,  $e_{i,t}$ , can be written as  $e_{i,t} = \xi_t \sqrt{h_{i,t}}$  with  $\xi_t \sim \mathcal{N}(0, 1)$ . Thus the algorithm of Jacquier et al. (1994) can be applied for each  $e_{i,t}$  to obtain estimates for the stochastic volatility terms.

Forecasting with the stochastic volatility VAR is similar to forecasting with the VAR of Section 3.1. Taking another look at equation (9), it is easy to see that I only have to

to account for the additional parameters – given by the variance of the stochastic volatility process and the free elements of  $A$  – and the latent states,  $h$ . All other steps remain unchanged.

## 4 Forecast Evaluation Metrics

Following, amongst others, Alessandri and Mumtaz (2017) and Koop (2013) I evaluate point forecasts as well as density forecasts for my out-of-sample test set. Weiss (1996) points out that for a forecaster with a quadratic loss function the mean square error ( $MSE$ ) is the appropriate measure for model choice. The average root mean square error,  $RMSE$ , for model  $m$  with  $m = \{\text{VAR}, \text{TVAR}, \text{SVOL}\}$ , variable  $i$  and forecast horizon  $h$  is defined as:

$$RMSE_{i,h}^m = \frac{1}{T_1 - H - T_0 + 1} \sum_{t=T_0+H-h}^{T_1-h} \sqrt{(y_{i,t}^{obs} - \bar{y}_{i,t}^m)^2}, \quad (15)$$

where  $y_{i,t}^{obs}$  represents the observed value and  $\bar{y}_{i,t}^m$  is the posterior predictive mean. Furthermore,  $H$  stands for the maximum forecast horizon,  $H = 4$ .  $T_0$  represents the starting point of the first forecast for forecast horizon  $h = H$  and  $T_1$  is the sample endpoint. Equation (15) guarantees that all forecasts start in the same time period,  $T_0 + H$  and end in  $T_1$ .

However, the the Gibbs samplers of Section 3 do not only provide point estimates. They also produce predictive samples. These samples can be used for evaluating predictive distributions based on equation (9). Gneiting et al. (2007) argue that probabilistic forecasts, that is forecasts based on a full probability distributions over future events, should aim for maximizing the *sharpness* of the predictive distribution subject to *calibration*. Calibration demands that observations should be indistinguishable from a random sample from the predictive distribution, while sharpness describes the concentration of the forecasts. A sharper predictive distribution is preferred.

Various approaches for evaluating calibration and sharpness of predictive distributions exist. Here I focus on forecast comparison via *scoring rules* which assess calibration and sharpness simultaneously in such a setting. A scoring rule assigns a numerical penalty based on the forecast distribution and the observed value which allows for ranking competing forecasts (Gneiting and Raftery, 2007; Gneiting and Katzfuss, 2014).

I focus on two proper<sup>10</sup> score scoring rules here. The most commonly used scoring rule is given by the *Logarithmic Score* (*LogS*; Good, 1952). Equation (16) defines the average logarithmic score for a given sample running from  $T_0 + H$  until  $T_1$ :

$$\text{LogS}(f_{i,h}^m) = -\frac{1}{T_1 - H - T_0 + 1} \sum_{t=T_0+H-h}^{T_1-h} \log f_{i,h}^m(y_{i,t}^{obs}), \quad (16)$$

where  $f_{i,h}^m$  denotes the predictive density for variable  $i$ , model  $m$ , and forecast horizon  $h$ . Note that *LogS* is defined as the negative log likelihood. It is therefore negatively oriented and constitutes a penalty term as demanded above. In my case the predictive density for computing the logarithmic score is only given indirectly via the MCMC samples of the three models. Thus I use kernel density estimation to approximate the predictive densities for each model. The density estimates,  $\hat{f}_{i,h}^m$ , are based on a Gaussian kernel with the rule-of-thumb bandwidth selection used in Silverman (1986), given by  $h_{bw} = 1.06\hat{\sigma}_{\mathcal{M}}\mathcal{M}^{-1/5}$  with  $\mathcal{M}$  representing the number of forecast samples and  $\hat{\sigma}_{\mathcal{M}}$  being the standard deviation of a given MCMC forecast sample.

However, density forecast evaluation via *LogS* has been criticized for being too sensitive to extreme events (Gneiting and Raftery, 2007). In case of MCMC output and kernel density estimation the score can become highly sensitive to bandwith choice if the observed value falls in the tail of the simulated forecast distribution (Krüger et al., 2019). Instead, these studies argue in favor of the *Continuous Ranked Probability Score* (*CRPS*; Matheson and Winkler, 1976) in evaluating density forecasts which rewards predictive distributions with mass concentrated around the outcome (‘sensitivity to distance’; see also Clark and Ravazzolo, 2015).<sup>11</sup> The average CRPS is defined as:

---

<sup>10</sup>A scoring rule is proper if reporting the true distribution as the forecasting distribution constiutes the optimal strategy in expectation (Gneiting and Katzfuss, 2014). That is, if observation  $y \sim \mathcal{G}$ , then a proper scoring rule will always have the property  $\mathbb{E}_{Y \sim \mathcal{G}} S(\mathcal{G}, Y) \leq \mathbb{E}_{Y \sim \mathcal{G}} S(\mathcal{F}, Y)$  for all  $\mathcal{F}, \mathcal{G}$ . The scoring rule is *strictly proper* if and only if the equality only holds for  $\mathcal{F} = \mathcal{G}$ .

<sup>11</sup>On the other hand, some authors argue in favor of the logarithmic score due its ‘locality’ property. A local scoring rule only pays attention to the density value of a realized outcome. Ultimately the choice between locality and sensitivity to distance is subjective. However, the problem of density estimation remains.



$$\begin{aligned}
CRPS(F_{i,h}^m, y_{i,h}^{obs}) &= \frac{1}{T_1 - H - T_0 + 1} \sum_{t=T_0+H-h}^{T_1-h} \int_{-\infty}^{\infty} \left( F_{i,h}^m(z) - \mathbb{1}\{y_{i,h}^{obs} \leq z\} \right)^2 dz \\
&= \frac{1}{T_1 - H - T_0 + 1} \sum_{t=T_0+H-h}^{T_1-h} \mathbb{E}_F |y_{i,t}^m - y_{i,h}^{obs}| - \frac{1}{2} \mathbb{E}_{FF} |y_{i,t}^m - y_{i,t}^{m,2}|
\end{aligned} \tag{17}$$

with  $\mathbb{1}\{\cdot\}$  being the indicator function and  $F(\cdot)$  representing the predictive CDF. Furthermore,  $y_{i,t}^m$  and  $y_{i,t}^{m,2}$  represent two independent draws from the posterior predictive distribution. As can be seen from the first line of equation (17), the CRPS measures the area between predicted and the realized cumulative distribution (see also Ravazzolo and Vahey, 2014) and is negatively oriented. The alternative formulation of the CRPS, in the second line of equation (17) – due to Gneiting and Raftery (2007) – allows for a straightforward application of the CRPS for a simulated sample as shown in equation (18). The CRPS based on the ECDF is implemented efficiently<sup>12</sup> for a single observation as:

$$CRPS(\hat{F}_{i,h}^{\mathcal{M}}, y^{obs}) = \frac{2}{\mathcal{M}^2} \sum_{i=1}^{\mathcal{M}} (Y_{(i)} - y^{obs}) \left( \mathcal{M} \mathbb{1}\{y^{obs} \leq Y_{(i)}\} - i + \frac{1}{2} \right) \tag{18}$$

with  $Y_{(1)}, \dots, Y_{(\mathcal{M})}$  representing the order statistic for  $y_{i,T_0+H-h}^m, \dots, y_{i,T_1-h}^m$ . The *CRPS*, like *LogS*, can be generalized for multivariate observations. Here I only consider the multivariate extension of the CRPS, the *Energy Score (ES)* as described in Gneiting et al. (2008). I do not include multivariate logarithmic scores here due to the density estimation problems described above and the ‘curse of dimensionality’ of multivariate density estimation. The implementation of the ES for a *single* observation is given by:

$$ES(\hat{F}_{i,h}^{\mathcal{M}}, y^{obs}) = \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} \|\mathbf{Y}_i - \mathbf{y}^{obs}\| - \frac{1}{2\mathcal{M}^2} \sum_{i=1}^{\mathcal{M}} \sum_{j=1}^{\mathcal{M}} \|\mathbf{Y}_i - \mathbf{Y}_j\| \tag{19}$$

with  $\mathbf{y}^{obs}$  and  $\mathbf{Y}_i$  now being  $d$  dimensional vectors and  $\|\cdot\|$  representing the Euclidean norm. Again, the average value over all samples will be used for forecast evaluation.

---

<sup>12</sup>The implementation is efficient in the sense that its computational complexity is  $\mathcal{O}(\mathcal{M} \log \mathcal{M})$  while a more straightforward implementation of equation (17) would result in a  $\mathcal{O}(\mathcal{M}^2)$  algorithm; see Krüger et al. (2019) for more details.

## 5 Results

I use monthly data for taken from the FED St. Louis *FRED*. All three models are based on six variables: total non-farm employment (PAYEMS), the consumer price index (CPIAUCSL), the effective Federal Funds rate (FEDFUNDS), the unemployment rate (UNRATE), housing starts (HOUST), and the *Chicago Fed National Financial Conditions Index* (NFCI). The data used in this study starts in January 1971 and runs until July 2019 which results in 583 observations in total.

As in Bańbura et al. (2010) I use total non-farm employment as a proxy for real economic activity.<sup>13</sup> Another proxy for real activity is given by the unemployment rate, while housing starts is a well-known leading indicator which also captures changes in investment behavior. Finally, the NFCI measures financial conditions in money markets, debt and equity markets. It takes traditional and “shadow” banking systems into account. Brave and Butters (2012) show that the NFCI is an accurate leading indicator for financial stress which makes a natural choice for the threshold variable in the TVAR. In addition, the NFCI effectively transforms the models of Section 3 into factor models as pointed out by Alessandri and Mumtaz (2017).

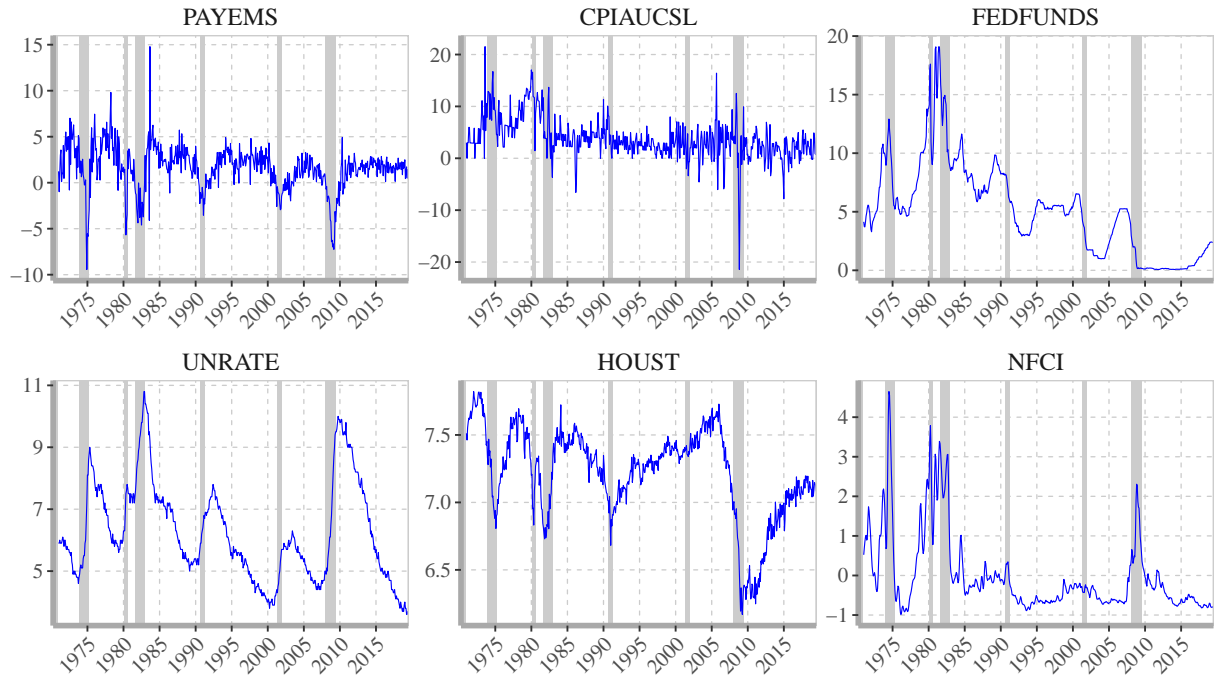
All data is provided on a monthly frequency with the exception of the NFCI which is compiled on a weekly basis. Monthly figures for the NFCI are derived by taking averages. The data is plotted in Figure 1. NBER recession dates as added as grey shaded areas. As can be seen from the figure, variables PAYEMS and CPIAUCSL have been transformed to annualized growth rates, while HOUST enters in logarithm. All other variables enter the three models without any transformation.

All estimates are obtained by recursive estimation over an expanding data window. Forecasts at horizons greater than one quarter are obtained recursively.

As discussed in Section 2, setting the shrinkage parameter at an appropriate level is crucial for obtaining reliable forecasts. I follow the standard approach of Bańbura et al. (2010) here and adjust  $\lambda$  such that its in-sample fit over the training sample running from ?? through ?? matches the in-sample fit of a small scale VAR estimated by OLS. The small scale VAR only uses data of the three main variables, PAYEMS, CPIAUCSL, and the

---

<sup>13</sup>Many authors, such as Alessandri and Mumtaz (2017), use industrial production as an indicator for real economic activity for monthly data. However, Hubrich and Tetlow (2015) point out that industrial production only represents a small share of the economy and might not be representative for real economic activity.



**Figure 1:** Variables used in all three models. PAYMES: All Employees, Total Nonfarm (annualized growth rate); CPIAUCSL: Consumer Price Index: All Items in U.S. City Average, All Urban Consumers (annualized growth rates); FEDFUNDS: Effective Federal Funds Rate; UNRATE: Unemployment Rate; HOUST: Housing Starts: Total: New Privately Owned Housing Units Started (log); NFCI: Chicago Fed National Financial Conditions Index.

Federal Funds rate. The fit of the Bayesian VAR for hyper-parameter  $\lambda$  is then defined as

$$Fit_{\lambda} = \frac{1}{3} \sum_{i=1}^3 \frac{MSE_i^{\lambda}}{MSE_i^0}, \quad (20)$$

with  $i$  representing the three main variables and  $MSE_i^{\lambda}$  being the in-sample one step ahead mean squared forecast error for the VAR.  $MSE_i^0$  is used as a normalizing constant and represents the in-sample one step ahead mean squared forecast error of the prior. The fit for the OLS VAR,  $Fit_{OLS}$  is evaluated the same way. Finally, to achieve a similar in sample fit for the Bayesian VAR and the small-scale OLS VAR,  $\lambda$  is chosen such that the absolute distance between  $Fit_{OLS}$  and  $Fit_{\lambda}$  is minimized. The optimization procedure follows a simple grid search over the interval between 0.01 and 1 with a step size of 0.01. Note that this results in the estimation of 99 Bayesian VARs to find  $\lambda$ . Implementation cost of this procedure is limited as a closed-form solution for the Bayesian VAR exists which is applied here.

The optimized value for the hyper-parameter,  $\lambda_{VAR}$  is used for the VAR and VAR with stochastic volatility. However, for the TVAR a different value for  $\lambda$  is chosen.<sup>14</sup> Hyper-parameter  $\lambda_{TVAR}$  is also found by a grid search. However, no closed-form solution exists for the TVAR. Therefore a smaller grid between ?? and ?? was used to find the optimal value. The step size remains the same.

## 6 Conclusion

In this paper I used three multivariate dynamic models to compare their forecast performances. More specifically, I tried to answer the question if a model which takes financial regime switches into account may be useful for forecasting macroeconomic data. To answer this question a TVAR had to be compared to benchmark models. I used a standard VAR and a VAR with stochastic volatility for this purpose. All three models have been estimated based on a Bayesian approach akin to the approach taken in Alessandri and Mumtaz (2017) and Bańbura et al. (2010). I used a VAR model as the baseline scenario, a standard approach in the literature (see, for example, Alessandri and Mumtaz, 2017 and Clark and Ravazzolo, 2015). Moreover, the VAR with stochastic volatility was chosen as a second benchmark, given its superb forecasting abilities. The forecasts were evaluated via point forecasts and

---

<sup>14</sup>I tried using the same  $\lambda$  value for all three models and using a distinct  $\lambda$  for the stochastic volatility VAR as well. However, the best results were obtained by using  $\lambda_{VAR}$  for the VAR and the stochastic volatility VAR and a distinct  $\lambda_{TVAR}$  for the TVAR.

density forecast metrics. In addition, I also investigated the potential of pooling the three models – based on scoring success – to improve forecasting performance.

My results show ...

All algorithms for the models described in Section 3 were written **R** and **C++** and integrated via the **Rcpp** package (Eddelbuettel and François, 2011; Eddelbuettel and Balamuta, 2017). The code is available at [https://github.com/alexhaider/Bayes\\_VARS](https://github.com/alexhaider/Bayes_VARS).<sup>15</sup>

---

<sup>15</sup>The functions are provided as-is with very little documentation so far. They are not available as a **R** package.

## References

- Alessandri, Piergiorgio and Haroon Mumtaz (2017). “Financial conditions and density forecasts for US output and inflation”. In: *Review of Economic Dynamics* 24, pp. 66–78.
- Aliber, Robert and Charles Kindleberger (2015). *Manias, Panics, and Crashes: A History of Financial Crises*.
- Bańbura, Marta, Domenico Giannone, and Lucrezia Reichlin (2010). “Large Bayesian vector auto regressions”. In: *Journal of Applied Econometrics* 25.1, pp. 71–92.
- Bernanke, Ben and Mark Gertler (1989). “Agency Costs, Net Worth, and Business Fluctuations”. In: *American Economic Review* 79.1, pp. 14–31.
- Bernanke, Ben, Mark Gertler, and Simon Gilchrist (1999). “The financial accelerator in a quantitative business cycle framework”. In: *Handbook of Macroeconomics*. Ed. by J. B. Taylor and M. Woodford. 1st ed. Vol. 1, Part C. Elsevier. Chap. 21, pp. 1341–1393.
- Brave, Scott and R. Andrew Butters (2012). “Diagnosing the Financial System: Financial Conditions and Financial Stress”. In: *International Journal of Central Banking* 8.2, pp. 191–239.
- Brunnermeier, Markus K. and Yuliy Sannikov (2014). “A Macroeconomic Model with a Financial Sector”. In: *American Economic Review* 104.2, pp. 379–421.
- Chen, Cathy W. S. and Jack C. Lee (1995). “Bayesian Inference of Threshold Autoregressive Models”. In: *Journal of Time Series Analysis* 16.5, pp. 483–492.
- Clark, Todd E. and Francesco Ravazzolo (2015). “Macroeconomic Forecasting Performance under Alternative Specifications of Time-Varying Volatility”. In: *Journal of Applied Econometrics* 30.4, pp. 551–575.
- Cogley, Timothy and Thomas J. Sargent (2005). “Drifts and volatilities: monetary policies and outcomes in the post WWII US”. In: *Review of Economic Dynamics* 8.2. Monetary Policy and Learning, pp. 262–302.
- Doan, Thomas, Robert Litterman, and Christopher Sims (1984). “Forecasting and conditional projection using realistic prior distributions”. In: *Econometric Reviews* 3, pp. 1–100.
- Eddelbuettel, Dirk and James Joseph Balamuta (2017). “Extending R with C++: A Brief Introduction to Rcpp”. In: *PeerJ Preprints* 5, e3188v1.
- Eddelbuettel, Dirk and Romain François (2011). “Rcpp: Seamless R and C++ Integration”. In: *Journal of Statistical Software* 40.8, pp. 1–18.
- Galvão, Ana Beatriz C. (2006). “Structural break threshold VARs for predicting US recessions using the spread”. In: *Journal of Applied Econometrics* 21.4, pp. 463–487.

- Gneiting, Tilmann, Fadoua Balabdaoui, and Adrian E. Raftery (2007). “Probabilistic forecasts, calibration and sharpness”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.2, pp. 243–268.
- Gneiting, Tilmann and Matthias Katzfuss (2014). “Probabilistic Forecasting”. In: *Annual Review of Statistics and Its Application* 1.1, pp. 125–151.
- Gneiting, Tilmann and Adrian E Raftery (2007). “Strictly Proper Scoring Rules, Prediction, and Estimation”. In: *Journal of the American Statistical Association* 102.477, pp. 359–378.
- Gneiting, Tilmann, Larissa I. Stanberry, Eric P. Gneiting, Leonhard Held, and Nicholas A. Johnson (2008). “Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds”. In: *TEST* 17.2, pp. 211–235.
- Good, I. J. (1952). “Rational Decisions”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 14.1, pp. 107–114.
- Haario, Heikki, Marko Laine, Antonietta Mira, and Eero Saksman (2006). “DRAM: Efficient adaptive MCMC”. In: *Statistics and Computing* 16.4, pp. 339–354.
- Haario, Heikki, Eero Saksman, and Johanna Tamminen (2001). “An Adaptive Metropolis Algorithm”. In: *Bernoulli* 7.2, pp. 223–242.
- He, Zhiguo and Arvind Krishnamurthy (2014). *A Macroeconomic Framework for Quantifying Systemic Risk*. Working Paper 19885. National Bureau of Economic Research.
- Hubrich, Kirstin and Timo Teräsvirta (2013). “VAR Models in Macroeconomics – New Developments and Applications: Essays in Honor of Christopher A. Sims”. In: ed. by Thomas B. Fomby, Lutz Kilian, and Anthony Murphy. Vol. 32. *Advances in Econometrics*. Emerald. Chap. Thresholds and Smooth Transitions in Vector Autoregressive Models, pp. 273–326.
- Hubrich, Kirstin and Robert J. Tetlow (2015). “Financial stress and economic dynamics: The transmission of crises”. In: *Journal of Monetary Economics* 70.C, pp. 100–115.
- Jacquier, Eric, Nicholas G Polson, and Peter Rossi (1994). “Bayesian Analysis of Stochastic Volatility Models”. In: *Journal of Business & Economic Statistics* 12.4, pp. 371–89.
- Karlsson, Sune (2013). “Forecasting with Bayesian Vector Autoregression”. In: vol. 2. Elsevier. Chap. 15, pp. 791–897.
- Koop, Gary M. (2013). “Forecasting with Medium and Large Bayesian VARS”. In: *Journal of Applied Econometrics* 28.2, pp. 177–203.
- Krüger, F., S. Lerch, T.L. Thorarindottir, and T. Gneiting (2019). *Predictive inference based on Markov chain Monte Carlo output*, research rep. Heidelberg Institute for Theoretical Studies,

- Litterman, Robert (1986). “Forecasting with Bayesian vector autoregressions: five years of experience”. In: *Journal of Business and Economic Statistics* 4, pp. 25–38.
- Lütkepohl, Helmut (2005). *New Introduction to Multiple Time Series Analysis*. Berlin, Germany: Springer.
- Matheson, James E. and Robert L. Winkler (1976). “Scoring Rules for Continuous Probability Distributions”. In: *Management Science* 22.10, pp. 1087–1096.
- Minsky, H.P. (1982). *Can "it" Happen Again?: Essays on Instability and Finance*. M.E. Sharpe.
- (2008). *John Maynard Keynes*. McGraw Hill professional. McGraw-Hill Education.
- Mira, Antonietta (2001). “On Metropolis-Hastings algorithms with delayed rejection”. In: *Metron - International Journal of Statistics* 0.3-4, pp. 231–241.
- Miranda-Agrippino, Silvia and Giovanni Ricco (2018). *Bayesian Vector Autoregressions*. Research rep. 1808. Centre for Macroeconomics (CFM).
- Mittnik, Stefan and Willi Semmler (2013). “The real consequences of financial stress”. In: *Journal of Economic Dynamics and Control* 37.8, pp. 1479–1499.
- Ravazzolo, Francesco and Shaun Vahey (2014). “Forecast densities for economic aggregates from disaggregate ensembles”. In: *Studies in Nonlinear Dynamics & Econometrics* 18.4, pp. 367–381.
- Schleer, Frauke and Willi Semmler (2015). “Financial sector and output dynamics in the euro area: Non-linearities reconsidered”. In: *Journal of Macroeconomics* 46.C, pp. 235–263.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.
- Stock, James H. and Mark W. Watson (2002). *Has the Business Cycle Changed and Why?* Working Paper 9127. National Bureau of Economic Research.
- (2003). “Forecasting Output and Inflation: The Role of Asset Prices”. In: *Journal of Economic Literature* 41.3, pp. 788–829.
- Teräsvirta, Timo (2006). “Forecasting economic variables with nonlinear models”. In: *Handbook of Economic Forecasting*. Ed. by G. Elliott, C.W.J. Granger, and A. Timmermann. Vol. 1. Elsevier. Chap. 8, pp. 413–457.
- Weiss, Andrew A. (1996). “Estimating time series models using the relevant cost function”. In: *Journal of Applied Econometrics* 11.5, pp. 539–560.