Forecasting US GDP with Bayesian VARs

Seminar in Econometrics
Prof. Kai Carstensen

Wirtschafts- und Sozialwissenschaftliche Fakultät
Christian-Albrechts-Universität

05.01.2025

Luca Preuße

Matrikelnummer 1182372
Quantitative Economics

Abstract

In this paper I inspect the out-of-sample fit of forecasts with large Bayesian Minnesota-type hierarchical VARs as introduced in Chan (2021). Differences and extensions to that paper are the introduction of sampling the shrinkage factors associated with each lag, considering dogmatically shrinking all contemporaneous relationships to 0, using GARCH instead of stochastic volatilities and formalizing differences between using a gamma and inverse-gamma prior for the shrinkage factors. When testing for individual forecast gains the forecast of interest is GDP-growth. I find that supressing contemporaneous relationships helps with forecasting of individual variables but severely hurts the joint forecast due to ignoring any contemporaneous correlations and that the data supports the lag-shrinkage imposed by the Minnesota prior.

1 Introduction

VARs are a workhorse model in applied macroeconometrics both for forecasting and structural analysis. Key features are their simple linear structure and accurately modelling multivariate macroeconomic variables as being endogenous. Linearity is known to perform well in out-of sample forecasts compared to overfitting nonlinear structures. Modelling multiple variables allows to model the correlation between many economic variables of interest. Endogeneity is appealing as it allows to compute multi-step forecasts by recursively iterating the system into the future outperforming direct forecasts in relevant data sets (Marcellino et al., 2006).
However, VARs suffer from overparameterization as the number of parameters grows quadratically with the dimension of the VAR. Therefore Bayesian estimation using informative priors to prevent overfitting seems appealing. An early application and common benchmark model is the Minnesota prior (see for example Litterman (1986)) using empirical wisdom to set priors and shrinking a VAR to a parsimonious model. Such empirical wisdom is that

for the equation of one variable higher lags are less relevant and the lags of other variables are less relevant than the own lags.

With computation power increasing over time the possibilities of Bayesian applications increased and high dimensional Bayesian VARs have been shown to work well for forecasting (Banbura et al., 2010).

High dimensional Bayesian VARs have also become more flexible with adaptive hierarchical hyperparameters to allow for more flexible estimation (for example Huber and Feldkircher (2019)). However, Chan (2021) show that the general structure from the Minnesota prior still holds up well and has empirical relevance and attempt to integrate this structure into such hierarchical priors.

Another recent contribution quickly becoming standard for forecasting with Bayesian VARs was the introduction of persistent time-varying volatility, commonly modelled by stochastic volatility (SV) (Carriero et al., 2019).

This paper builds on the Minnesota-type hierarchical prior framework in Chan (2021) with applications to forecasting and mainly forecasting of GDP-growth in the US. I extend this framework by also introducing a hierarchical prior for the lag-specific shrinkage whereas Chan (2021) uses the Minnesota suggestion. Further I capture the features of persistent time-varying volatility with a GARCH(1,1) model instead of the more commonly used SV model.

2 Adaptive Minnesota-type hierarchical prior

2.1 Model and Likelihood

The M-dimensional VAR(p) is written as

$$\begin{aligned}&A_0 y_t = a + A_1 y_{t-1} + A_2 y_{t-2} ... + A_p y_{t-p} + u_t \\ &u_t \sim N(0, \Sigma_t)\end{aligned} \tag{1}$$

Let α be a vector containing all elements of $A_0, a, A_1 ... A_p$.

Σ_t is diagonal with elements $\sigma^2_{i,t}$ which follow GARCH(1,1) processes

$$\sigma^2_{i,t} = a_{i,0} + a_{i,1} u^2_{i,t-1} + b_{i,1} \sigma^2_{i,t-1} \tag{2}$$

Write $\theta_i = (a_{i,0}, a_{i,1}, b_{i,1})$ and $\theta = (\theta_1, \theta_2, ... \theta_p)$. For a known starting value $\sigma_{i,0}$ (which we assume to be known and leave out in the following to simplify notation), conditional on θ and information up until period t-1, Σ_t is known with this specification.

A_0 is assumed to be upper triangular with ones on the main diagonal following Chan (2021) allowing for a recursive representation of the system.

For known starting values the likelihood is given by
$p(y_{1:T}|\alpha, \theta) = p(y_T|y_{0:T-1}, \alpha, \theta)p(y_{T-1}|y_{0:T-2}, \alpha, \theta)...p(y_1|y_0, \alpha, \theta)$
Further we can write
$p(y_t|y_{0:t-1}, \alpha, \theta) = p(y_{1,t}|y_{2:M,t}, y_{0:t-1}, \alpha, \theta)p(y_{2,t}|y_{3:M,t}, y_{0:t-1}, \alpha, \theta)...p(y_{M,t}|y_{0:t-1}, \alpha, \theta)$
Due to the upper triangular structure of A_0 these are all simple normal distributions in this framework. In this expression θ_i and the elements of the ith row of A are redundant for all conditional densities except for $y_{i,t}$ which allows us to treat the VAR as M independent regressions allowing for efficient equation by equation sampling as in Chan (2021). Let us write each equation i as
$y_{i,t} = X_{i,t}\beta_i + u_{i,t} \; X_{i,t} = (1, y'_{t-1}, ... y'_{t-p}, y'_{-i,t})$

so that the likelihood of observations $y_{i,1:T}$ is given as

$p(y_{i,1:T}|\beta_i, \theta_i) \propto |H_i|^{0.5} exp(-0.5(Y_i - X_i\beta_i)' H_i (Y_i - X_i\beta_i))$

H_i is diagonal and stacks $\sigma^{-2}_{i,1}, \sigma^{-2}_{i,2}, ... \sigma^{-2}_{i,T}$ and Y_i and X_i stack $y_{i,t}$ and $X_{i,t}$ rowwise.

The full sample likelihood is then

$$p(y_{1:T}|\alpha,\theta) = \prod_{i=1}^{M} p(y_{i,1:T}|\beta_i,\theta_i) \tag{3}$$

2.2 Priors

As priors we build a hierarchical Minnesota-type prior. $\alpha \sim N(0, V)$ (assuming all variables are stationary, it can alternatively also be centered around 1 for first own lags of persistent variables). V is diagonal and

$$V_{k,k} = \begin{cases} \lambda_{int} s_i^2, & \text{if } \alpha_k = a_i \\ \lambda_1 \lambda_l \lambda_k, & \text{if } \alpha_k = A_l(i,i) \\ \lambda_1 \lambda_l \lambda_2 \lambda_k \frac{s_i^2}{s_j^2}, & \text{if } \alpha_k = A_l(i,j), i \neq j \\ \lambda_{cont} \frac{s_i^2}{s_j^2}, & \text{if } \alpha_k = A_0(i,j), i \neq j \end{cases} \tag{4}$$

λ_{int} is the shrinkage factor for the intercepts. We will set it to a constant high value (10,000) for the remainder of this paper so that the VAR is shrunken towards the unconditional mean.

λ_1 is a global shrinkage factor associated with all lags in the VAR.

λ_2 is a shrinkage factor for lags of jth variables in the ith equation (cross-variables) to impose that once a variable is conditioned on its own past other variables are less important.

λ_l is a shrinkage factor for the lth lag to impose different shrinkage on different lags. Without loss of generality $\lambda_l = 1$ for l=1. The Minnesota prior sets this typically to $1/l^2$ assuming higher lags to be less relevant.

λ_k is an individual shrinkage factor for each lagged variable allowing for more flexibility of the Minnesota prior to pick out important variables for forecasting that may be in conflict with the general structure of cross-variables and higher lags being less relevant.

λ_{cont} is a shrinkage factor for contemporaneous variables.

s_i^2 is a scaling factor to account for different scaling of the variables. We will

obtain it by fitting AR(1)s and estimating their variances.
$\lambda_{\{k\}}$ and $\lambda_{\{l\}}$ denote the set of all M^2p different λ_k and p different λ_l.

In a hierarchical structure we choose a prior for each $\lambda_n \in \{\lambda_1, \lambda_2, \lambda_{\{k\}}, \lambda_{\{l\}}, \lambda_{cont}\}$. The priors for each λ_n are independent. For each λ_k we choose a gamma prior $\lambda_k \sim G(1, \nu_k)$ (gamma definition in A.1). ν_k is the same for each λ_k. For the other shrinkage parameters both gamma (Chan, 2021) and inverse-gamma priors (Huber and Feldkircher, 2019) have been used in the literature. We will consider both so either $\lambda_n^{-1} \sim G(S_n^2, \nu_n)$ or $\lambda_n \sim G(S_n^2, \nu_n)$. Following Chan (2021) we will also choose a prior for $\nu_k \sim G(0.5, 2)$ which is centered around the Bayesian Lasso (Park and Casella, 2008).

The prior for the GARCH parameters θ_i is a diffuse uniform prior (truncated at the stationary region) such that the posterior conditional on everything else is proportional to the likelihood for stationary parameters and 0 else.

2.3 Posterior and MCMC Sampling

We are now interested in the posterior
$p(\alpha, \lambda_1, \lambda_2, \lambda_{\{k\}}, \lambda_{\{l\}}, \lambda_{cont}, \nu_k, \theta | y_{1:T})$
which due to the independent and hierarchical prior choices is proportional to

$$\begin{aligned} &p(y_{1:T}|\alpha, \theta)p(\alpha|\lambda_1, \lambda_2, \lambda_{\{k\}}, \lambda_{\{l\}}, \lambda_{cont})* \\ &p(\lambda_1)p(\lambda_2)p(\lambda_{\{l\}})p(\lambda_{cont})p(\lambda_{\{k\}}|\nu_k)p(\nu_k)p(\theta) \end{aligned} \tag{5}$$

To sample from this we set up a Gibbs sampler drawing from the conditional posteriors sequentially to generate a Markov Chain that converges to the posterior. To denote conditioning on every variable that is not the random variable in the conditional distribution we write $p(.|Z)$. The conditional posteriors are derived in (A.1). Here we state them simply.

For the inverse-gamma priors of the shrinkage coefficients the posterior is
$\lambda_n^{-1}|Z,y \sim G(\tilde{S}_n^2,\tilde{\nu}_n)$
with $\tilde{\nu}_n = \sum_k I(\lambda_n \in V_{k,k}) + \nu_n$
and $\tilde{S}_n^2 = (\sum_k I(\lambda_n \in V_{k,k})\alpha_k^2(\frac{V_{k,k}}{\lambda_n})^{-1} + \nu_n S_n^2)/\tilde{\nu}_n$
Note that since the indicator function sets all elements of V that do not have λ_n to 0 by dividing $V_{k,k}$ by λ_n, $\tilde{S}_n^2$ does not depend on λ_n.

For the gamma priors of the shrinkage coefficients the posterior is
$\lambda_n|Z,y \sim GiG(a,b,p)$
with $a = \nu_n S_n^2$, $b = \sum_k I(\lambda_n \in V_{k,k})\alpha_k^2(\frac{V_{k,k}}{\lambda_n})^{-1}$, $p = 0.5(\nu_n - \sum_k I(\lambda_n \in V_{k,k}))$
GiG is the generalized inverse Gaussian distribution and is given in (A.1).[1]

The elements of α can be sampled equation by equation using the notation in section 2 and $\beta_i|Z,y \sim N(\tilde{\beta}_i,\tilde{V}_i)$ with $\tilde{V}_i = (X_i'H_iX_i + V_i^{-1})^{-1}$
and $\tilde{\beta}_i = \tilde{V}_i(X_i'H_iY_i)$
V_i are the diagonal elements of the priors that are associated with β_i.

$p(\nu_k|Z,y) \propto p(\lambda_{\{k\}}|\nu_k)p(\nu_k)$ as well as $p(\theta_i|Z,y) \propto p(y_{i,1:T}|\beta_i,\theta_i)$ have no known distribution and can be sampled with an adaptive Metropolis Hastings algorithm.[2] I design it equivalently for both so let us consider the general problem of sampling of a d-dimensional distribution p(X) which is computable up to an integrating constant. The conditional posteriors have positive parameter space. For $X > 0$ it can be more convenient to sample from Y=log(X) instead to not generate draws outside of the parameter space.[3]

[1] Sampling from the GiG was faster than sampling from the gamma distribution making use of the algorithm in Devroye (2014) which is why I use this prior for λ_k as this takes up most of the time in the Gibbs block. In unreported results I also tested using an inverse gamma prior for λ_k but this did not change a lot.

[2] The kernels are given in A.1.

[3] The script GammaSampling.m performs the MH algorithm to sample from a gamma distribution and shows in Monte Carlo simulations that sampling from the log is more efficient than sampling from the level in terms of approximating mean and variance (distances measured by MSE and Itakura Saito distance).

Using change of variables $f(Y) \propto p(exp(Y)) \prod_i^d exp(Y_i) = p(X) \prod_i^d X_i$. The algorithm generates a candidate draw $log(X_s^*) \sim N(log(X_{s-1}), \Sigma_s)$ and accepts X_s^* with probability $\frac{p(X_s^*) \prod_i^d X_{i,s}^*}{p(X_{s-1}) \prod_i^d X_{i,s-1}}$, else X_s is set to X_{s-1}.[4] Details on the adaptive scheme to generate Σ_s are in (A.2). In short Σ_s is adapted so that the acceptance rate of the draws converges to a predefined target which is 0.25 - close to commonly used rule of thumbs. This prevents acceptance rates that are too low or too high which can cause this algorithm to be inefficient although the efficient acceptance rate is generally unknown.

The 1-step forecast density can be written as
$p(y_{T+1}|y_{1:T}) = \int \int p(y_{T+1}|\alpha, \theta, y_{1:T}) p(\alpha, \theta|y_{1:T}) d\alpha d\theta$
which suggests that a forecast from this density can be generated by simulating (1) one step forward.[5] The 2-step forecast density is
$p(y_{T+2}|y_{1:T}) = \int \int \int p(y_{T+2}|y_{1:T+1}, \alpha, \theta) p(y_{T+1}|\alpha, \theta, y_{1:T}) p(\alpha, \theta|y_{1:T}) d\alpha d\theta dy_{T+1}$
which suggests that the general h-step forecast from the forecast distribution can be drawn with the drawn parameters by recursively simulating (1) and (2) forward. Note that the variance at t+h from the GARCH process is known given information up until t+h-1.

Initial values are always given by the centering of the prior and by an (almost) constant variance equal to the scaling factor for the GARCH parameters, meaning the autoregressive coefficients of the variance are close to 0. The chain length used in this paper is 11,000 with 1,000 draws being used as a burn-in.

[4] To avoid numerical 0/0 or ∞/∞ expressions it is better to take the exponential of the log-difference between numerator and denominator.

[5] As in the derivation of (3) $p(y_{T+1}|.)$ can be factorized into conditionals and marginals of $p(y_{i,T+1}|.)$ allowing for recursive simulation across equations.

3 Forecast Experiment

3.1 Data

3.1.1 Transformation

The dataset is constructed from the FRED-QD database at the Federal Reserve Bank of St. Louis containing 248 macroeconomic and financial variables for the US (McCracken and Ng, 2020). I use the same 23 variables as Chan (2021) in the VAR. The data spans from to 1959Q1 to 2024Q1. The variables are transformed to achieve stationarity according to the transformation codes that is provided in the dataset. An exception is that I treat log-price variables as I(1) instead of I(2) as in Medeiros et al. (2021) which led to better forecast performance in Goulet Coulombe et al. (2022).

3.1.2 Missing Data

The dataset has missing observations although for the variables in our VAR this affects only the first 4 observations of New Private Housing Units. To restore the missing data we apply the EM algorithm suggested in Stock and Watson (2002). A factor model is created using a PCA estimate with all variables without missing observations. Then for each variable that has missing observations that variable is regressed on the factors and the missing values are replaced with their fitted value from the estimated factor and loading for that variable. Then the now balanced panel is used to re-estimate the factors and loadings and the missing values are again replaced by their fitted values from this factor model. This process is repeated until convergence. The number of factors is chosen with the information criterion from Bai and Ng (2002).

3.1.3 COVID Outliers

In contrast to Chan (2021) our dataset includes data after the COVID-19 pandemic which brings in many extreme observations, especially for GDP-growth, which strongly affects estimates and forecasts of VARs despite such outliers not being caused by standard economic relationships and hence such forecasts seem implausible (Carriero et al., 2024). In standard stochastic volatility and also GARCH-models changes in volatility are highly persistent while such extreme observations are better represented by jumps/spikes in volatility. In response Carriero et al. (2024) propose models that allow for such jumps in volatility with decent forecast performance. However, they also find that treating outliers as a missing data problem performs about as well as their models for forecasting non-outliers which is what I adopt here. Instead of screening for outliers I simply treat the observations between 2020Q1 and 2020Q3 as missing data. For the forecast-evaluation those periods are left out. If those periods are not left out and the untreated data is used, the outliers can dominate any forecast-statistic so that for example the best model for forecasting GDP-growth in terms of MSE is just the one that by chance forecasted the lowest GDP-growth during COVID (even though no model should be able to predict that). To replace the missing data I assume for each variable that it follows a 2-state model where we have an underlying AR(1) state variable with additional COVID shocks that are zero during non-COVID periods. Details are given in (A.3). In the Baseline model in our forecast experiment this increased the 1-step ahead out-of-sample forecast-accuracy of GDP-growth by about 18 percentage points in terms of MSE which was driven only by the few observations after COVID.

3.2 Statistics

To evaluate forecast performance I apply an expanding window approach with estimation starting using data up until 1982Q4. We use data up to period t to sample from the 1-h-step forecast distribution and compare them to the actual observations $y_{t+1:t+h}$. I look at forecast horizons up to h=8 periods giving us 156 out-of-sample observations.

As in Chan (2021) statistics of interest are the average out-of-sample log-predictive likelihood (ALPL) and root-mean-square error (RMSE). ALPL tells us how well the forecast fits into the fitted forecast distribution accounting not only for point forecasting but also an accurate measure of forecast uncertainty. This is defined as $ALPL = \frac{1}{T-h-t0-1} \sum_{t=t_0}^{T-h} log(p(y_{i,t+h}|y_{1:t}))$ evaluated at the actual observation.

We are also interested in the ALPL of the vector-forecast (VEC-ALPL), simply by removing the i-index, which represents the fit of the joint forecast of all variables in the VAR. The forecast distribution $p(y_{i,t+h}|y_{1:t})$ integrates out the parameters and is generally unknown but we generate a sample from it as a byproduct of our Gibbs sampling.

The predictive density here is calculated similar to Adolfson et al. (2007) by using simulated forecasts and fitting a parametric distribution. Different to them however I fit a multivariate t distribution instead of a normal.[6] This takes into account mean and covariance information as well as fat-tailedness which are the main features of the forecast samples which can be seen in A.4.

Alternatively one can get an estimate of $p(y_{i,t+h}|y_{1:t})$ by averaging conditional densities which are known normals over the Gibbs sampler, see for example Geweke and Amisano (2011). For highly multivariate densities this may however suffer from the problem of the conditional densities being numerically indistinguishable from 0 so that one would end up taking logs of 0s and taking logs and averaging is not interchangeable. By fitting a multivariate density to the forecast sample the log can be applied directly to simplify the density avoiding numerical 0s.

RMSE is simply defined as $RMSE = \sqrt{\frac{1}{T-h-t_0-1} \sum_{t=t_0}^{T-h} (E(y_{i,t+h}|y_{1:t}) - y_{i,t+h})^2}$. The expectation is estimated with the mean of the forecast-sample.

Relative gains in RMSE between two models are given by their ratio, gains in ALPL by their difference.

[6] I do this by matching mean and then iterating between maximizing the likelihood with respect to the degree of freedom and matching the covariance until convergence.

3.3 Prior Settings

The Baseline specification is given using p=4 lags, using inverse-gamma priors with $S_1^2 = 0.04$ $S_2^2 = 0.04$ with $\nu_{1,2} = 4$ which is centered around the same values as in Carriero et al. (2015). $S_{cont}^2 = 1/10000$ with $\nu_{cont} = 1000000$ so we basically impose strict 0s on the contemporaneous relationships. $S_l^2 = 1/l^2$ with $\nu_l = 4$ so we center the prior for lag-shrinkage around the Minnesota suggestion. A degree of freedom parameter of $\nu = 4$ implies finite first but infinite second moment of the implied inverse-gamma distribution for λ_n. Table 1 summarizes all alternative prior settings that I test:

Prior	Differences to Baseline
1	$\nu_l = 1000000$
2	$S_l^2 = 1$
3	Gamma instead of inverse-gamma priors with $\nu_{1,2,l} = 4$ $S_{1,2}^2 = 1/0.04$ and $S_l^2 = l^2$
4	$\nu_{cont} = 4$ $S_{cont}^2 = 1$
5	$\lambda_{\{k\}} = 1$
6	$\nu_{2,l} = 1000000$ $S_2^2 = 1$ $S_l^2 = 1$

Table 1: Prior Settings

Prior 1 is using the Minnesota lag-shrinkage factors dogmatically. Prior 2 tests a prior belief that assumes all lags are equally important. Prior 3 tests the gamma instead of the inverse-gamma priors. Prior 4 includes sampling of contemporaneous effects with adaptive data-driven shrinkage. Prior 5 tests the Minnesota structure without local shrinkage factors. Prior 6 is a global-local prior as in Huber and Feldkircher (2019) which does not impose any structure from Minnesota.

4 Results

4.1 Out-of-Sample

Table 2 shows the results of the forecast experiment. As a benchmark I use a very simple model assuming all variables to follow independent AR(1)s with constant variance. In terms of RMSE this benchmark is not beaten at longer forecast horizons but there are strong gains over the benchmark in terms of predictive density, specifically in the multivariate, and also in the 1-step ahead RMSE.

For GDP-growth the Baseline performs well and Prior 2 (in the short run) and Prior 4, 5 and 6 perform worse so the general Minnesota structure seems to do well for GDP forecasting and deviating from it with Prior 2 or Prior 6 hurts forecasting performance. Prior 5 hurts forecasting performance so adding local shrinkage components on top of the Minnesota structure to become more flexible seems to be recommended for GDP forecasting.

To evaluate general forecast performance beyond just GDP one should focus on the Median statistics in parentheses and on VEC-ALPL.

When comparing the Baseline with Prior 1 we see a slight improvement by using the Minnesota lag-shrinkage factors dogmatically for the 1-step forecast. Over longer horizons this difference disappears. Prior 2 shows generally a worse performance than the Baseline implying that prior information that assumes higher lags being less relevant does help for overall forecasting performance. The Minnesota specification does generally seem to do well and allowing the data to deviate from it does not lead to improvements.

When comparing the Baseline with Prior 3 we do not see severe differences implying robustness towards the shape of the prior being gammas or inverse-gammas.

When comparing the Baseline to Prior 4 we see that Prior 4 does much better in terms of VEC-ALPL because it accounts for the correlation across variables leading to a better joint forecast density. The ALPL and also RMSE

h	Baseline	Prior 1	Prior 2	Prior 3	Prior 4	Prior 5	Prior 6
	RMSE						
1	12.9 (4.7)	12.7 **(5.4)**	11.7 (4.7)	12.8 (4.5)	9.7 (3.7)	11.7 (4.7)	9.4 (4.7)
2	3.5 (3.3)	3.5 (3.2)	2.7 (2.4)	**3.5** (3.2)	1.9 (-0.5)	2.7 (2.6)	1.6 (-0.5)
3	-1.4 (-0.3)	-1.7 (-0.8)	-2.2 (-1.6)	-1.1 (-0.4)	-1.8 (-3.1)	-2.7 (-0.8)	-2.1 (-2.1)
4	-6.0 (-2.5)	-6.2 (-2.9)	-6.7 (-3.2)	-5.6 (-2.4)	-5.4 (-5.4)	-7.7 (-3.1)	-5.7 (-3.9)
5	-9.3 (-4.2)	-9.8 (-4.4)	-10.3 (-4.6)	-9.3 (-4.3)	-8.2 (-5.5)	-11.3 (-4.8)	-9.1 (-4.3)
6	-8.3 (-4.1)	-8.6 (-4.0)	-8.9 (-5.7)	-8.2 (-4.0)	-4.6 (-4.6)	-9.8 (-4.7)	-7.7 (-4.3)
7	-9.1 (-4.4)	-9.3 (-4.5)	-9.7 (-4.8)	-9.2 (-4.5)	-3.9 (-4.4)	-10.8 (-5.4)	-8.7 (-4.2)
8	-10.5 (-3.8)	-10.7 (-4.0)	-11.0 (-4.3)	-10.2 (-4.0)	-4.5 (-4.4)	-12.0 (-4.9)	-10.1 (-4.5)
	ALPL						
1	21.3 (16.1)	**21.3 (18.8)**	20.0 **(16.5)**	**21.7 (18.0)**	11.4 (11.2)	20.4 **(17.8)**	17.9 **(17.4)**
2	19.3 (18.7)	19.2 (18.5)	18.2 (18.2)	19.3 (16.5)	10.1 (10.1)	17.8 (16.3)	17.6 (14.8)
3	14.9 (14.9)	14.7 **(14.9)**	14.2 **(15.1)**	14.9 (14.7)	6.3 (8.6)	14.0 (14.7)	14.3 (14.3)
4	10.2 (10.1)	**10.7** (9.2)	**10.6** (9.5)	**10.5** (9.1)	2.7 (5.7)	8.7 (9.3)	**10.8** (9.4)
5	7.0 (6.1)	**7.4** (5.4)	**7.3** (4.9)	**7.4** (4.8)	-0.1 (3.6)	5.6 (5.7)	**7.4 (6.9)**
6	6.4 (5.3)	**6.9 (5.6)**	**6.9** (5.0)	**6.8 (6.5)**	-0.1 (1.7)	5.8 (5.2)	**6.6 (5.8)**
7	5.7 (5.1)	**5.7** (4.5)	**5.8** (4.7)	5.6 (4.9)	-0.5 (0.4)	4.7 (4.7)	**5.7** (4.8)
8	4.2 (4.2)	**4.3** (4.1)	**4.3 (4.3)**	**4.6 (4.6)**	-1.4 (-0.4)	3.6 (3.6)	**4.3 (4.3)**
	VEC-ALPL						
1	497.3	**509.7**	489.4	488.8	**1225.3**	489.6	473.8
2	650.7	648.4	641.4	648.8	**1283.9**	636.1	620.3
3	670.2	666.9	660.1	**670.2**	**1291.2**	654.5	646.8
4	641.0	639.0	633.0	**643.5**	**1280.6**	628.5	621.8
5	616.5	611.9	606.1	**618.3**	**1256.5**	597.5	593.0
6	602.1	595.6	585.6	601.1	**1229.8**	580.0	578.8
7	580.8	580.6	570.1	**581.1**	**1204.9**	564.3	567.1
8	569.7	566.6	561.4	**571.9**	**1189.3**	555.4	554.5

Table 2: Out-of-sample statistics

Table shows the percentage gains of a h-step forecast in RMSE and ALPL for GDP-growth and for the full vector of variables relative to an AR(1) benchmark. Median across all variables in parentheses. Bold if better than Baseline and benchmark.

at short horizons however suggest that this comes at a cost of forecasting individual variables.

Prior 5 generally tends to perform worse than the Baseline so adding local shrinkage components seems to be recommended for general forecasting purposes. Prior 6 does generally worse than the Baseline mainly noticable for VEC-ALPL confirming again the benefits of the Minnesota structure whose lacking cannot be entirely compensated by only local and a global shrinkage component.

4.2 Full Sample

Parameter	10,000 draws	100,000 draws
λ_1	0.1192 (0.0281)	0.1126 (0.0302)
λ_2	0.0467 (0.0117)	0.0499 (0.0136)
$\lambda_{\{l=(2,3,4)\}}$	(0.1473 0.1067 0.0537) ((0.0403) (0.0338) (0.0198))	(0.1494 0.1133 0.051) ((0.0408) (0.0354) (0.0208))
ν_k	1.7724 (0.9862)	1.961 (0.9159)

Table 3: Posterior Moments Full Sample

Posterior means and standard deviatons in parentheses of Baseline specification for 10,000 posterior draws with 1,000 burn-in draws and 100,000 draws with 10,000 burn-in draws.

Table 3 shows some results of the sampler using the full dataset with the Baseline specification. Moments do not change a lot by increasing the chain with exception for ν_k which hints at convergence issues for this parameter. I apply the diagnostic in Geweke (1991) to all simulated forecast distributions with 10,000 draws and it rejects convergence 7.61% out of 184 times at the 5% level implying that the convergence issues do not appear to affect the convergence of the forecast distribution which is ultimately our statistic of interest in the forecast experiment.

A visual inspection of the trace plots in (A.4) supports this. There are obvious convergence issues for ν_k with the smaller chain. The forecasts however also seem well converged visually. The other parameters seem to have converged well with the smaller chain and ν_k also appears to have converged with the longer chain. Increasing the number of draws in the forecast experiment is beyond the scope of this Seminar due to limited computation power but due to the convergence of the forecast distribution I do not expect this to affect the experiment a lot except possibly for longer forecast horizons which have higher variance and heavier tails and therefore require generally longer chains to pick this up reliably. The convergence results using the smallest sample considered in our forecast experiment are similar.

The degree of freedoms of the fitted forecast-t-distributions differ a lot between variables. They tend to get smaller as the horizon increases (implying heavier tails) and are always far away from the boundary of $\nu \leq 2$ which is ruled out to prevent infinite variances. For GDP-growth the 1-step forecast density has a degree of freedom larger than 30 so is approximated by a normal distribution and then goes down from the 2-step to the 8-step forecast from 19.3 to 11.8. For the multivariate t-density that is fit from the forecasts the degree of freedoms only get slightly smaller than 30 for the latest forecast horizons implying that the multivariate forecast is fairly well approximated by a normal.

The posterior means of $\lambda_{\{k\}}$ range from 0.378 to 6.181. The acceptance rates of the MH sampler range from 0.2029 to 0.2845 for the GARCH parameters and is 0.2598 for ν_k so the adaptive MCMC seems work as intended bringing the acceptance rates close to the target of 0.25.

The posterior means of $\lambda_{\{l\}}$ are close to the Minnesota specification except for the second lag for which the data suggests stronger shrinkage than Minnesota. When applying Prior 2, so a prior belief that all lag-shrinkage parameters are the same, the posterior means of lags larger than 1 become (0.2777, 0.2409, 0.1901) so the data supports the idea of shrinking higher lags stronger even without imposing it in the prior.

5 Further Issues and Research Directions

5.1 Recursiveness Assumption

A problem arises from the assumption that A_0 in (1) is upper triangular. This assumption is adopted from the literature (see for example Carriero et al. (2019), Chan (2021)) and is not problematic beyond structural analysis for VARs with constant variance as such an assumption is required for identification and therefore the likelihood is invariant to the ordering of the variables when assuming a triangular structure. However, nontrivial time-varying volatility does provide identification in SVARs (see for example Lütkepohl and Milunovich (2016) for GARCH or Carriero et al. (2021) for SV). Therefore a recursiveness assumption in this framework provides overidentifying restrictions and the order of variables clearly matters even asymptotically beyond structural analysis.[7]

Chan et al. (2024) discuss this in further detail and show also how this affects forecasting. The triangular assumption is convenient to write the model as M regressions and sample equation by equation. If one wants to generalize this, conditional on A_0 the posterior and sampling scheme is still simple but sampling of the elements of A_0 requires a more sophisticated method like a Metropolis Hastings sampler (see for example Baumeister and Hamilton (2015)) which is potentially problematic in high dimensions. As our Baseline model shrinks the off-diagonal elements of A_0 to 0 dogmatically most of our results here are of course not affected by this. Testing Prior 4 with the reverse ordering of the variables doesn't change the results relative to the Baseline but of course both orderings are just arbitrary and another order or an order-invariant approach as in Chan et al. (2024) could lead to better

[7] Carriero et al. (2019) also point out a sensitivity to the order of variables due to the way the prior is built which is irrelevant asymptotically.

results than the Baseline and is definitely required, if one is interested in joint forecasting or structural analysis.

5.2 Possible Extensions

An obvious extension tackling the convergence problem of ν_k and robustness of the results would be to scale up the experiment by increasing the chain length. It is also worth pointing out that our adaptive MCMC approach to sample from ν_k differs from Huber and Feldkircher (2019) who apply a Metropolis Hastings algorithm with constant proposal variance and from Chan (2021) who use an independence chain algorithm where the proposal distribution is a normal using the posterior mode and hessian of the log-posterior to compute mean and variance. A comparison of convergence speed between these algorithms would be interesting.

Another deviation in this paper is the usage of GARCH volatility instead of SV and a comparison between those choices and possibly different GARCH-type specifications would be interesting as well. Literature comparing GARCH and SV models find typically support for SV (see for example Pederzoli (2006) and Chan and Grant (2016)) but these comparisons are done for high-frequency univariate financial data which does not necessarily translate to low-frequency high-dimensional macro data and our Bayesian VAR setting.

References

M. Adolfson, J. Lindé, and M. Villani. Forecasting performance of an open economy dsge model. *Econometric Reviews*, 26(2-4):289–328, 2007.

J. Bai and S. Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

M. Banbura, D. Giannone, and L. Reichlin. Large bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92, 2010.

C. Baumeister and J. D. Hamilton. Sign restrictions, structural vector autoregressions, and useful prior information. *Econometrica*, 83(5):1963–1999, 2015.

A. Carriero, T. E. Clark, and M. Marcellino. Bayesian vars: Specification choices and forecast accuracy. *Journal of Applied Econometrics*, 30(1): 46–73, 2015.

A. Carriero, T. E. Clark, and M. Marcellino. Large bayesian vector autoregressions with stochastic volatility and non-conjugate priors. *Journal of Econometrics*, 212(1):137–154, 2019.

A. Carriero, T. E. Clark, and M. Marcellino. Using time-varying volatility for identification in vector autoregressions: An application to endogenous uncertainty. *Journal of Econometrics*, 225(1):47–73, 2021.

A. Carriero, T. E. Clark, M. Marcellino, and E. Mertens. Addressing COVID-19 Outliers in BVARs with Stochastic Volatility. *The Review of Economics and Statistics*, 106(5):1403–1417, 2024.

J. C. Chan. Minnesota-type adaptive hierarchical priors for large bayesian vars. *International Journal of Forecasting*, 37(3):1212–1226, 2021.

J. C. Chan and A. L. Grant. Modeling energy price dynamics: Garch versus stochastic volatility. *Energy Economics*, 54:182–189, 2016.

J. C. C. Chan, G. Koop, and X. Yu. Large order-invariant bayesian vars with stochastic volatility. *Journal of Business & Economic Statistics*, 42 (2):825–837, 2024.

L. Devroye. Random variate generation for the generalized inverse gaussian distribution. *Statistics and Computing*, 24:239–246, 2014.

J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Staff Report 148, Federal Reserve Bank of Minneapolis, 1991.

J. Geweke and G. Amisano. Hierarchical markov normal mixture models with applications to financial asset returns. *Journal of Applied Econometrics*, 26(1):1–29, 2011.

P. Goulet Coulombe, M. Leroux, D. Stevanovic, and S. Surprenant. How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, 37(5):920–964, 2022.

F. Huber and M. Feldkircher. Adaptive shrinkage in bayesian vector autoregressive models. *Journal of Business & Economic Statistics*, 37(1):27–39, 2019.

R. B. Litterman. Forecasting with bayesian vector autoregressions: Five years of experience. *Journal of Business & Economic Statistics*, 4(1):25–38, 1986.

T. Lux. Bayesian Estimation of Agent-Based Models via Adaptive Particle Markov Chain Monte Carlo. *Computational Economics*, 60(2):451–477, 2022.

H. Lütkepohl and G. Milunovich. Testing for identification in svar-garch models. *Journal of Economic Dynamics and Control*, 73:241–258, 2016.

M. Marcellino, J. H. Stock, and M. W. Watson. A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135(1):499–526, 2006.

M. McCracken and S. Ng. Fred-qd: A quarterly database for macroeconomic research. Working Paper 26872, National Bureau of Economic Research, 2020.

M. C. Medeiros, G. F. R. Vasconcelos, A. Veiga, and E. Zilberman. Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics*, 39(1):98–119, 2021.

T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

C. Pederzoli. Stochastic volatility and garch: a comparison based on uk stock data. *The European Journal of Finance*, 12(1):41–59, 2006.

G. O. Roberts and J. S. Rosenthal. Coupling and ergodicity of adaptive markov chain monte carlo algorithms. *Journal of Applied Probability*, 44 (2):458–475, 2007.

J. H. Stock and M. W. Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162, 2002.

A Appendix

A.1 Posterior Derivation

A.1.1 $p(\lambda_n^{-1}|Z,y)$ with Inverse-Gamma Prior

If $x \sim G(S^2,\nu)$, $p(x) = (\frac{\nu S^2}{2})^{0.5\nu}\Gamma(0.5\nu)^{-1}x^{\frac{\nu-2}{2}}exp(-0.5x\nu S^2)$
with $E(x) = S^{-2}$ and $Var(x) = 2S^{-4}/\nu$

In (5) λ_n only appears in the prior for α and its own prior.

For $\lambda_n^{-1} \sim G(S_n^2,\nu_n)$:
$p(\lambda_n^{-1}|Z,y) \propto p(\lambda_n^{-1})p(\alpha|\lambda_{\{n\}},Z) \propto (\lambda_n^{-1})^{\frac{\nu_n-2}{2}}exp(-0.5\lambda_n^{-1}\nu_n S_n^2)|V|^{-0.5}exp(-0.5\alpha' V^{-1}\alpha)$

V is defined as in (4). Because V is diagonal we can write

$= (\lambda_n^{-1})^{\frac{\nu_n-2}{2}}exp(-0.5\lambda_n^{-1}\nu_n S_n^2)(\prod_k V_{k,k})^{-0.5}exp(-0.5\sum_k \alpha_k^2 V_{k,k}^{-1})$

Then we can factorize it into terms that do not and do depend on λ_n and write

$\propto (\lambda_n^{-1})^{\frac{\nu_n-2}{2}} exp(-0.5\lambda_n^{-1}\nu_n S_n^2) *$
$(\lambda_n^{\sum_k I(\lambda_n \in V_{k,k})})^{-0.5} exp(-0.5\sum_k \alpha_k^2 V_{k,k}^{-1} I(\lambda_n \in V_{k,k}))$

$\propto (\lambda_n^{-1})^{\frac{\nu_n+\sum_k I(\lambda_n \in V_{k,k})-2}{2}} exp(-0.5\lambda_n^{-1}(\nu_n S_n^2 + \sum_k \alpha_k^2(\frac{V_{k,k}}{\lambda_n})^{-1} I(\lambda_n \in V_{k,k})))$

which is proportional to $G(\frac{\nu_n S_n^2+\sum_k \alpha_k^2(\frac{V_{k,k}}{\lambda_n})^{-1} I(\lambda_n \in V_{k,k})}{\nu_n+\sum_k I(\lambda_n \in V_{k,k})}, \nu_n + \sum_k I(\lambda_n \in V_{k,k}))$.

A.1.2 $p(\lambda_n|Z, y)$ with Gamma Prior

Equivalently we only need to change the prior for λ_n to derive for
$\lambda_n \sim G(S_n^2, \nu_n)$:

$p(\lambda_n|Z, y) \propto \lambda_n^{\frac{\nu_n-2}{2}} exp(-0.5\lambda_n \nu_n S_n^2) *$
$(\lambda_n^{\sum_k I(\lambda_n \in V_{k,k})})^{-0.5} exp(-0.5\lambda_n^{-1} \sum_k \alpha_k^2 V_{k,k}^{-1} I(\lambda_n \in V_{k,k}))$

$\propto \lambda_n^{\frac{\nu_n-\sum_k I(\lambda_n \in V_{k,k})-2}{2}} exp(-0.5\lambda_n \nu_n S_n^2) exp(-0.5\lambda_n^{-1} \sum_k \alpha_k^2(\frac{V_{k,k}}{\lambda_n})^{-1} I(\lambda_n \in V_{k,k}))$

The Generalized inverse Gaussian distribution has pdf $p(x) \propto x^{p-1} exp(-0.5ax) exp(-0.5b/x)$ which has the same kernel as $p(\lambda_n|Z, y)$ with $p = \frac{\nu_n-\sum_k I(\lambda_n \in V_{k,k})}{2}$, $a = \nu_n S_n^2$
and $b = \sum_k \alpha_k^2(\frac{V_{k,k}}{\lambda_n})^{-1} I(\lambda_n \in V_{k,k})$.

A.1.3 $p(\beta_i|Z, y)$

We need to derive $p(\beta_i|Z, y)$ which due to the likelihood expression in (3)

$p(\beta_i|Z, y) \propto p(y_{i,1:T}|\beta_i, \theta_i) p(\alpha|\lambda_{\{n\}})$

V_i are the diagonal elements of the priors that are associated with β_i and due to the diagonal structure of V

$\propto exp(-0.5(Y_i - X_i\beta_i)' H_i (Y_i - X_i\beta_i)) exp(-0.5(\beta_i' V_i^{-1} \beta_i))$

$\propto exp(-0.5(\beta_i'(X_i'H_iX_i + V_i^{-1})\beta_i - 2\beta_i'X_i'H_iY_i))$

$\propto N((X_i'H_iX_i + V_i^{-1})^{-1}X_i'H_iY_i, (X_i'H_iX_i + V_i^{-1})^{-1})$

A.1.4 $p(\theta_i|Z, y)$

A GARCH(1,1) process is stationary if $a_{i,1} + b_{i,1} < 1$. Hence,

$p(\theta_i|Z, y) \propto |H_i|^{0.5}exp(-0.5(Y_i - X_i\beta_i)'H_i(Y_i - X_i\beta_i))I(a_{i,1} + b_{i,1} < 1)$

$= \prod_t \sigma_{i,t}^{-1}exp(-0.5 * \sum_t(y_{i,t} - X_{i,t}\beta_i)^2\sigma_{i,t}^{-2})I(a_{i,1} + b_{i,1} < 1)$

$\sigma_{i,t}^2$ depends on θ_i, $\sigma_{i,0}$ and can be recursively computed for a given θ_i. $\sigma_{i,0}$ is fixed as the unconditional variance of a fitted AR(1).

A.1.5 $p(\nu_k|Z, y)$

$p(\nu_k|Z, y) \propto p(\lambda_{\{k\}}|\nu_k)p(\nu_k)$

$\propto \prod_k((\frac{\nu_k}{2})^{0.5\nu_k}\Gamma(0.5\nu_k)^{-1}\lambda_k^{\frac{\nu_k-2}{2}}exp(-0.5\lambda_k\nu_k))exp(-0.5\nu_k)$

$\propto (\frac{\nu_k}{2})^{M^2p0.5\nu_k}\Gamma(0.5\nu_k)^{-M^2p}(\prod_k \lambda_k)^{\frac{\nu_k}{2}}exp(-0.5(1 + \nu_k\sum_k \lambda_k))$

A.2 Adaptive MCMC

Σ_s is set adaptively similar to Lux (2022) after a burn-in period s_0

$\Sigma_s = C_s\hat{\Sigma}_{1:s-1}$, $\hat{\Sigma}_{1:s-1} = Cov(log(X_{1:s-1}))$

$log(C_s) = min(log(C_{max}), log(C_{s-1}) + \gamma_s(\zeta_s - \zeta^*))$

$\gamma_s = \frac{1}{(s-s_0)^\eta}$

With this scheme the proposal variance is proportional to the empirical covariance of the generated draws. The scaling adjusts automatically such that the scaling increases, if the empirical acceptance rate ζ_s is larger than a predefined target ζ^* and vice versa which automatically moves the acceptance rate towards the target. For such an updating algorithm to be valid it has to satisfy the diminishing adaptation condition outlined in Roberts and Rosenthal (2007). γ_s decreases over time with rate $0 < \eta < 1$ so that $\Sigma_{s+1} - \Sigma_s = O_p(s^{-\eta})$. If C_s were to get stuck at the upper bound, the diminishing adaptation assumption is satisfied as C_s does not change and $\hat{\Sigma}_{1:s}$ only changes at rate $O_p(1/s)$.

Proof:

$$\Sigma_{s+1} - \Sigma_s = exp(log(C_s) + \gamma_s(\zeta_s - \zeta^*))(\hat{\Sigma}_{1:s-1} + O_p(1/s)) - exp(log(C_s))\hat{\Sigma}_{1:s-1}$$

Since ζ_s is bounded between 0 and 1 $\gamma_s(\zeta_s - \zeta^*) = O_p(s^{-\eta})$.[8] Then due to the bound of C_s:

$$\Sigma_{s+1} - \Sigma_s = (exp(log(C_s) + O_p(s^{-\eta})) - exp(log(C_s)))\hat{\Sigma}_{1:s-1} + O_p(1/s)$$

This expression is $O_p(s^{-\eta})$ using again the bound of C_s and the local Lipschitz continuity of the exponential function

$$exp(log(C_s) + O_p(s^{-\eta})) - exp(log(C_s)) \leq K|O_p(s^{-\eta})| = O_p(s^{-\eta})$$

I use $\eta = 0.4$, $s_0 = 100$ which seemed well behaved in experiments with the full and first sample from the forecast experiment and $\zeta^* = 0.25$ which is close to commonly used rule-of thumb values.

[8] As ζ_s generally converges to ζ^* the convergence rate is in most cases even faster but $o_p(s^{-\eta})$ implies $O_p(s^{-\eta})$.

A.3 Outlier Filter

Consider the model

$$\begin{aligned}
&y_t = s_t + \epsilon_t \\
&s_t = \alpha_0 + \alpha_1 s_{t-1} + \eta_t \\
&\eta_t \sim N(0, \sigma_s^2) \\
&\begin{cases} \epsilon_t = 0, & \text{if } t < T_{break} \\ \epsilon_t \sim N(0, \sigma_b^2), & \text{if } t \geq T_{break} \end{cases}
\end{aligned} \tag{6}$$

This assumes that a variable s_t follows a simple AR(1) with accurate measurement. After a break-date T_{break} the observed variable is this underlying AR(1) and an additional shock. In context of COVID it makes sense to think of these as COVID specific shocks. As these shocks may be potentially large and influential and carry no information about actual underlying economic relationships we want to extract the underlying AR(1) process from the observations to remove these outliers. To make such an estimate feasible in real time we are interested in $E_t(s_t)$ to replace the actual observations y_t that correspond to the COVID era. The likelihood of the model is given as $p(y_{1:T}|\theta) = \prod_{t=1}^{T} p(y_t|y_{1:t-1}, \theta)$
which can of course be easily computed for periods where $y_t = s_t$. For other periods we use rules of probability (for simpler notation let us leave out conditioning on the parameters θ)

$$p(y_t|y_{1:t-1}) = \int p(y_t|s_t)p(s_t|y_{1:t-1})ds_t \tag{7}$$

where $p(y_t|s_t)$ is a simple normal distribution. Similarly we write

$$p(s_t|y_{1:t-1}) = \int p(s_t|s_{t-1})p(s_{t-1}|y_{1:t-1})ds_{t-1} \tag{8}$$

where again $p(s_t|s_{t-1})$ is a simple gaussian. Then using Bayes theorem

$$p(s_{t-1}|y_{1:t-1}) \propto p(y_{t-1}|s_{t-1})p(s_{t-1}|y_{1:t-2}) \tag{9}$$

$p(s_{t-1}|y_{1:t-2})$ is the lag of the same expression in (7) and of course a known gaussian up until the break-date. Next we will show that all these three equations are actually normal distributions, given that we start with a normal distribution, with mean and variance that depend on the last step allowing for a recursive scheme to compute the likelihood. This likelihood can then be numerically maximized and $E_t(s_t) = E(s_t|y_t, \theta_{MLE})$ can be computed for all variables replacing the outliers. Note that to mimic real time availability the maximization is only performed with data up until t, first including one outlier, then two etc.

For (9) we can derive this the following way assuming $s_{t-1}|y_{1:t-2} \sim N(\mu, \sigma^2)$

$p(s_{t-1}|y_{1:t-2}) \propto exp(-0.5\sigma_b^{-2}(y_{t-1} - s_{t-1})^2)exp(-0.5\sigma^{-2}(s_{t-1} - \mu)^2)$

$\propto exp(-0.5((\sigma_b^{-2} + \sigma^{-2})s_{t-1}^2 - 2s_{t-1}(\sigma_b^{-2}y_{t-1} + \sigma^{-2}\mu)))$

$\propto N((\sigma_b^{-2} + \sigma^{-2})^{-1}(\sigma_b^{-2}y_{t-1} + \sigma^{-2}\mu), (\sigma_b^{-2} + \sigma^{-2})^{-1})$

(7) and (8) follow the same structure so let us show that $p(X) = \int p(X|Y)p(Y)dY$ with $X|Y \sim N(\beta_0 + \beta_1 Y, h_1^{-1})$ and $Y \sim N(\mu, h_2^{-1})$ follows a normal distribution. The expression for the two defining moments of this normal distribution can then be applied to both equations by substituting X and Y.

$p(X) \propto \int exp(-0.5h_1(X - \beta_0 - \beta_1 Y)^2)exp(-0.5h_2(Y - \mu)^2)dY$

$\propto \int exp(-0.5h_1(X^2 + \beta_1^2 Y^2 - 2X\beta_0 + 2Y\beta_0\beta_1 - 2XY\beta_1)) *$
$exp(-0.5h_2(Y^2 - 2Y\mu))dY$

$= exp(-0.5h_1(X^2 - 2X\beta_0)) *$

$\int exp(-0.5((h_1\beta_1^2 + h_2)Y^2 - 2Y(-h_1\beta_0\beta_1 + h_1X\beta_1 + h_2\mu))dY$

The term in the integral is proportional in Y to the pdf of a normal distribution with inverse-variance $(h_1\beta_1^2 + h_2)$ and mean $(h_1\beta_1^2 + h_2)^{-1}(h_1X\beta_1 - h_1\beta_0\beta_1 + h_2\mu)$ so we can multiply and divide by terms such that it integrates to 1. Then ignoring all multiplying constants that do not depend on X we end up with

$\propto exp(-0.5h_1(X^2 - 2X\beta_0))exp(0.5(h_1\beta_1^2 + h_2)^{-1}(h_1X\beta_1 - h_1\beta_0\beta_1 + h_2\mu)^2)$

$\propto exp(-0.5h_1(X^2 - 2X\beta_0)) *$
$exp(0.5(h_1\beta_1^2 + h_2)^{-1}(h_1^2\beta_1^2X^2 - 2X(h_1^2\beta_1^2\beta_0 + h_1h_2\beta_1\mu)))$

$\propto exp(-0.5(X^2(h_1\beta_1^2 + h_2)^{-1}h_1h_2 - 2X(h_1\beta_1^2 + h_2)^{-1}h_1h_2(\beta_0 + \beta_1\mu)))$

Which is proportional to a normal distribution with inverse-variance $(h_1\beta_1^2 + h_2)^{-1}h_1h_2$ and mean $\beta_0 + \beta_1\mu$.

Figure 1 shows the data for GDP-growth with and without running this filtering approach.

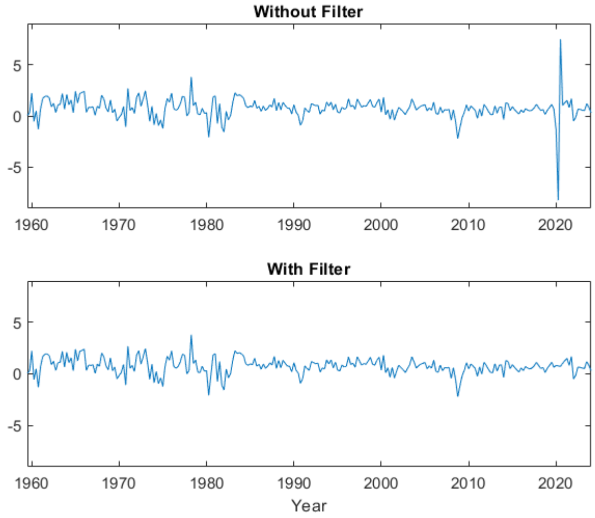

Figure 1: Quarter to Quarter GDP-growth in %

A.4 Trace Plots Full Sample

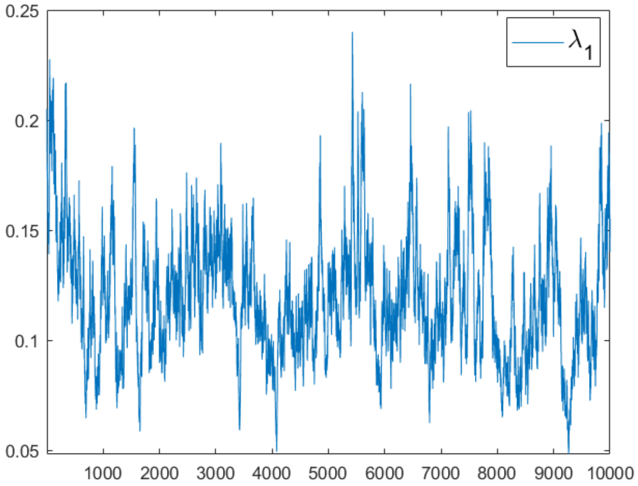

Figure 2: Posterior Draws λ_1 after burn-in

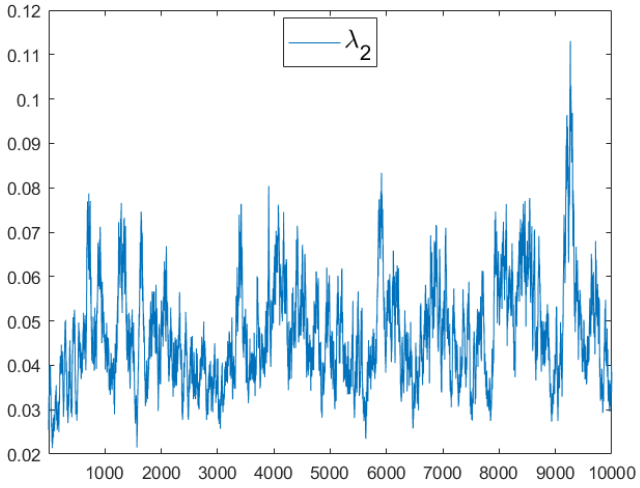

Figure 3: Posterior Draws λ_2 after burn-in

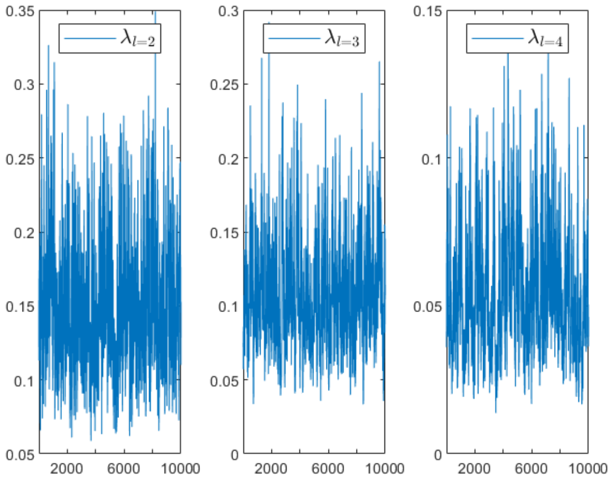

Figure 4: Posterior Draws $\lambda_{\{l\}}$ after burn-in

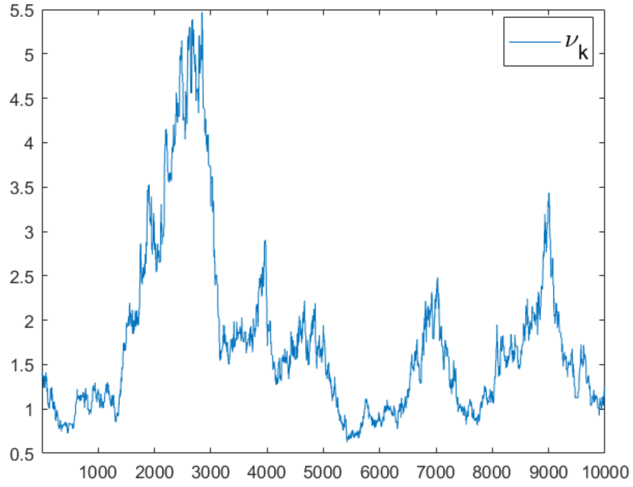

Figure 5: Posterior Draws ν_k after burn-in

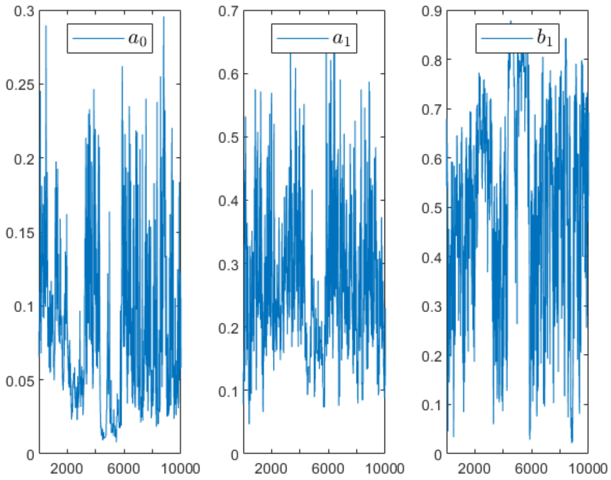

Figure 6: Posterior Draws of GARCH parameters (GDP) after burn-in

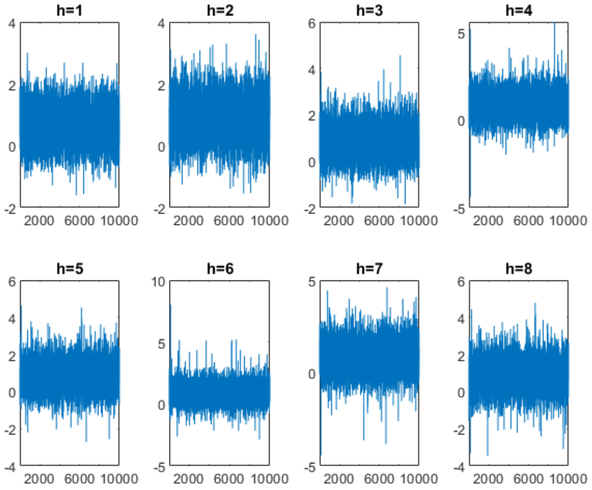

Figure 7: Posterior Draws of Forecast (GDP) after burn-in

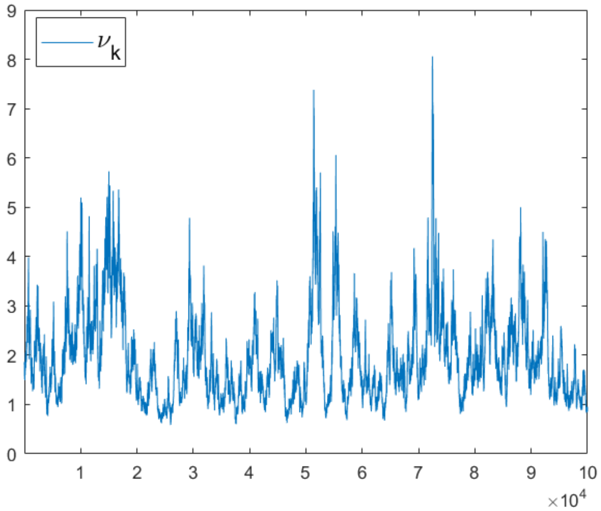

Figure 8: Posterior Draws of ν_k after burn-in long chain

Affirmation

I hereby declare that I have composed my Seminar paper "Forecasting US GDP with Bayesian VARs" independently using only those resources mentioned, and that I have as such identified all passages which I have taken from publications verbatim or in substance. I agree that the work will be reviewed using plagiarism testing software. Neither this paper, nor any extract of it, has been previously submitted to an examining authority, in this or a similar form.

Date: 05.01.2025
Signature: ____________