Combining Shrinkage and Sparsity in Conjugate Vector Autoregressive Models

Niko Hauzenberger, Florian Huber and Luca Onorante

Seminar Paper

Philipp Page

Seminar Econometrics
Faculty of Management, Economics and Social Sciences
University of Cologne

February, 2022

Contents

1 Introduction

This seminar paper deals with the estimation of large-scale Bayesian vector autoregressive (BVAR) models. It presents key insights of the paper "Combining shrinkage and sparsity in conjugate vector autoregressive models" by Hauzenberger et al. (2021), focusing on achieving sparsity in a shrunk BVAR model and validating the approach in a practical forecasting application.

While the estimation of VARs has traditionally been conducted using ordinary-least-squares (OLS) the Bayesian literature addresses the common overfitting issue of OLS with informative priors. Bayesian inference relies on combining prior knowledge about model quantities with evidence found in the data to obtain full posterior distributions. In large dimensional models such as VAR models, shrinkage priors are a helpful tool to control for overfitting to the training set. In this work, the common Minnesota shrinkage prior invented by Litterman (1986) is used to shrink the VAR model coefficients in order to avoid overfitting. An important property of the Minnesota prior is that it can lead to a natural conjugate model so that important distributions, such as the marginal posterior distributions and the predictive density, are available in closed form and do not need to be approximated. This approach stands in contrast to non-conjugate models like the BVAR with stochastic-search-variable-selection (SSVS) prior (George et al., 2008). The main advantage of conjugate models in comparison to the non-conjugate competitor is computational speed because the posterior distributions in non-conjugate models need to be simulated using methods like Markov-Chain-Monte-Carlo (MCMC). At the same time, the non-conjugate SSVS prior model allows for more flexibility in regards to the prior restrictions the researcher can impose on the model parameters (Bańbura et al., 2010; George et al., 2008; G. Koop & Korobilis, Dimitris, 2009; G. M. Koop, 2013).

Hauzenberger et al. (2021) argue that shrinkage alone is not sufficient in conjugate models as the probability of setting parameters to exactly zero in the Minnesota prior setup is zero. To remove this upper bound of accuracy, they propose the post-processing of point-estimates to obtain a sparse model.

The remainder of this seminar paper is structured as follows: Section 2 introduces the topic of shrinkage and sparsity illustrated by the instance of the Bayesian lasso. In addition, a general definition of VAR models is given. Subsequently, Section 3 presents the key approach of Hauzenberger et al. (2021) towards combining shrinkage and sparsity in BVAR models. The theoretical approach presented in Section 3 is then validated in Section 4 by the example of a macroeconomic forecasting application using the McCracken and Ng (2015) dataset. Finally, the paper concludes in Section 5.

2 Theoretical Background

2.1 Shrinkage and Sparsity

The topic of both shrinkage and sparsity is related to parameter regularization where the researcher sacrifices a little bias to reduce the model variance (Tibshirani, 1996). This is also commonly referred to as the bias-variance trade-off which describes the dilemma of finding a balanced trade-off between model complexity and model fit. Typically, a complex model with many parameters has a smaller bias and a higher variance in comparison to a less complex model with fewer parameters. In the case of a linear regression model, the predictive error depends on both the bias and the variance. In forecasting scenarios, one is typically interested in a model which generalizes well to new data. This means that overly complex models lead to poor predictive performance due to an inflated variance while too simple models lead to poor predictive performance due to a larger bias (Hastie et al., 2009, pp. 223 f.).

A common way to address the bias-variance trade-off is through parameter regularization in the form of shrinkage and/or sparsity. Shrinkage refers to shrinking certain parameter estimates towards zero by penalizing large values. Sparsity refers to setting certain parameter estimates to exactly zero which leads to excluding them from the model at all. Both techniques aim at the same goal – reducing the predictive variance while sacrificing bias (Figueiredo, 2003; Tibshirani, 1996).

A popular example for both shrinkage and sparsity is the lasso regression model first introduced by Tibshirani in 1996. It extends the standard ordinary least squares (OLS) optimization problem by adding a penalization term for the absolute parameter values called the L_1-norm. The L_1-constrained least squares optimization problem for a regression model $y = X\beta + \epsilon$ with p parameters can be written as

$$\arg\min_{\beta} \underbrace{(y - X\beta)'(y - X\beta)}_{\text{Sum of squared residuals}} + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}_{L_1\text{-penalization}} \quad , \tag{1}$$

where $\lambda \geq 0$ controls the strength of shrinkage (Park & Casella, 2008; Tibshirani, 1996; Yuan & Lin, 2005).

In Bayesian inference, shrinkage and sparsity can be implemented through the choice of adequate prior distributions. For the lasso example, one can show that a Bayesian regression model with a Laplace prior on the coefficients centered at zero yields the frequentist lasso estimate at the posterior mode (proof in Appendix A). The general intuition behind centering the prior at

zero is to assume a priori that the coefficients are likely to be close to zero. This also pulls the coefficients' posterior distributions to zero if the evidence in the data does not show otherwise.

To illustrate how the choice of a prior shrinks posterior estimates an empirical example has been conducted. The task was to predict the horsepower of cars in R's default *mtcars*[1] dataset based on information such as fuel consumption or weight. Figure 1 shows the coefficients' posterior distributions for each regressor used to predict car horsepower. One can observe that smaller

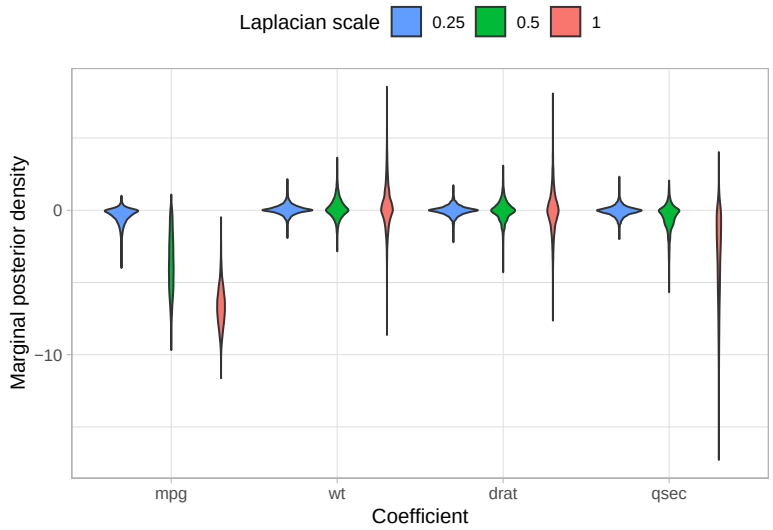

Figure 1: Posterior distributions for parameter estimates grouped by different Laplacian scale parameters.

Laplacian scale parameters lead to stronger shrinkage. This reflects the fact that smaller scale parameters lead to a sharper Laplace prior distribution with more probability mass around zero which pulls the posterior estimates to zero as well.

This specific example leads to both a shrunk and a sparse model. The posteriors of *wt* (weight) and *drat* (rear-axle ratio) show strong evidence that both regressors have little explanatory power because the distributions have most probability mass around zero. A point-estimate like the posterior mode becomes close to or exactly zero which effectively excludes the regressors from the model. With increased shrinkage, the same behavior can be observed for *qsec* (quarter-mile seconds). Finally, *mpg* (miles per gallon) seems to be the most important regressor to predict horsepower because its posterior distributions are strongly different from zero even for increased shrinkage.

[1]https://www.rdocumentation.org/packages/datasets/versions/3.6.2/topics/mtcars

2.2 Vector Autoregressive (VAR) Models

Vector autoregressive (VAR) models can be used to model multivariate time series. They gained popularity in the macroeconomic field after the influential paper by Sims (1980). This section defines VAR models following the notation of Hauzenberger et al. (2021).

A VAR(p) model with p lags has the following form:

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + C + \epsilon_t \ , \tag{2}$$

where $y_t = (y_{1t}, \ldots, y_{mt})'$ represents an m-dimensional vector of observations at time $t = 1, \ldots, T$. A_j for $j = 1, \ldots, p$ is the $(m \times m)$ matrix of coefficients at the j'th lag of y_t. The model's intercept is denoted by the $(m \times 1)$ vector C. Finally, $\epsilon_t \sim \mathcal{N}(0, \Sigma)$ is the Gaussian shock vector with mean zero and Σ as its $(m \times m)$-dimensional variance-covariance matrix.

Furthermore, define $\alpha = vec\{(A_1, \ldots, A_p, C)'\}$ as the vector of vectorized coefficients which has dimension $k = m(mp + 1)$ and $x_t = (y'_{t-1}, \ldots, y'_{t-p}, 1)'$ as the $(n = pm + 1)$-dimensional vector of explanatory variables. With these definitions at hand it is possible to rewrite Equation (2) as a regression model

$$y_t = (\mathcal{I}_m \otimes x'_t)\alpha + \epsilon_t \ . \tag{3}$$

Another form of notation commonly found in literature is the full matrix notation of VAR models (G. Koop & Korobilis, Dimitris, 2009). Define $A = (A_1 \ldots, A_p, C)'$ and E as $(T \times m)$-dimensional matrix of stacked shocks with tth row ϵ'_t. With tth row y'_t in Y and tth row x'_t in X the matrix notation reads

$$Y = XA + E \ . \tag{4}$$

A bivariate ($m = 2$) VAR model with one ($p = 1$) lag can be written as follows:

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \underbrace{\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}}_{m \times m} \underbrace{\begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix}}_{m \times 1} + \underbrace{\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}}_{m \times 1} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix} \ . \tag{5}$$

This instance of the simplest possible VAR model already has 6 coefficients and 4 variance-covariance values to be estimated which underlines the fact that VAR models suffer from a fast-growing number of parameters. This issue, which is also referred to as the curse of dimensionality, motivates the upcoming contents of this seminar paper, which address this problem by describing shrinkage and sparsification techniques for VAR models.

3 Combining Shrinkage and Sparsity in Bayesian VAR Models

VAR models are traditionally estimated by generalized least squares (GLS) as described in Lütkepohl (2005, pp. 69 ff.). With the rise of Bayesian statistics – thanks to computational advances – Bayesian VAR estimation of larger models became feasible and gained popularity in macroeconomic research (G. Koop & Korobilis, Dimitris, 2009, p. 269).

This seminar paper focuses on conjugate BVAR models which have the property that the prior distributions are of the same family of distributions as the corresponding posterior distributions (Gelman et al., 2013, pp. 35 f.). The key advantage of conjugacy compared to non-conjugate models is that posterior distributions are available in closed form and do not need to be approximated via sampling methods like MCMC.

The remainder of this section outlines the conjugate BVAR model setup proposed by Hauzenberger et al. (2021) and describes their sparsification algorithms. If not stated otherwise, the reference used is the aforementioned paper.

3.1 Conjugate Shrinkage Prior Setup

The VAR model specification outlined in Section 2.2 is a general one and not related to estimation. In order to estimate the defined VAR model in a Bayesian manner, it is necessary to specify prior distributions for all unknown parameters for which the researcher wishes to calculate a posterior distribution. The prior setup described below makes use of informative priors to realize shrinkage on the parameters.

The two unknown components to be estimated in this VAR model setup are the vectorized coefficients α and the variance-covariance matrix Σ. The prior for α is conditional on Σ in the conjugate setup and can be written as

$$\alpha|\Sigma \sim \mathcal{N}(\alpha_0, \Sigma \otimes V_0(\delta)) \,, \tag{6}$$

where a common choice for α_0 is zero for stationary time series. The prior for Σ is an inverted-Wishart distribution with degrees of freedom s_0 and scaling matrix S_0 as hyperparameters. This allows the researcher to control the amount of shrinkage imposed on the variance-covariance matrix:

$$\Sigma \sim \mathcal{W}^{-1}(s_0, S_0) \,. \tag{7}$$

Shrinkage on α in this prior setup is realized through the $V_0(\delta)$ term in Equation (6). Hauzenberger et al. (2021) use a variant of the Minnesota prior

for V_0 implemented using a set of dummy observations as outlined in Bańbura et al. (2010). The intuition of the Minnesota prior is to shrink all equations of the VAR model towards a random walk. This implies that the coefficients of the first own lags of any y_t in Y are shrunk towards one whereas the others are shrunk towards zero (Bańbura et al., 2010; Kadiyala & Karlsson, 1997; G. M. Koop, 2013). This captures the notion that the first own lags of a variable are a priori considered to be important. Details about the exact usage of the Minnesota prior and how one can control the degree of shrinkage using the δ vector of hyperparameters can be found in Hauzenberger et al. (2021, p. 306).

The priors detailed in Equations (6) and (7) are natural conjugate priors because they reflect the likelihood $p(y|\alpha, \Sigma)$ in a natural way: It can be shown that the likelihood is the product of a normal distribution conditional on the variance-covariance matrix and an inverted-Wishart distribution for Σ. Hence, multiplying this likelihood with priors of the same distributions yields posteriors of the same distributional form (Kadiyala & Karlsson, 1997; G. Koop & Korobilis, Dimitris, 2009).

While this conjugate prior setup comes with the advantage of computationally cheaper calculations in comparison to non-conjugate prior setups it also comes with a cost. The Kronecker structure in Equation (6) implies that prior variances for the VAR coefficients across equations must be proportional to each other. The prior variance is given by $\sigma_{jj}^2 V_0(\delta)$ where σ_{jj}^2 is the (j,j)th element in Σ. This means that the researcher cannot differentiate between coefficients on own and other lags which is a restrictive property.

3.2 Ex-Post Sparsification

The prior setup described in Section 3.1 leads to posterior distributions available in closed form. The full posterior distributions with their moments can be found in the original paper (Hauzenberger et al., 2021, p. 307).

In particular, the one-step ahead posterior predictive density $p(y_{T+1}|Y, X)$ is also available in closed from. It is a multivariate t-distribution with variance denoted by

$$Var(y_{T+1}|Y,X) = \frac{1}{s_1 - 2} \underbrace{\left(1 + \sum_{i=1}^{n}\sum_{j=1}^{n}(x_{iT+1}x_{jT+1}\nu_{ij})\right)}_{\text{Parameter uncertainty}} S_1 \ , \tag{8}$$

where s_1 and S_1 are the mean and variance of the inverted-Wishart posterior for Σ and ν_{ij} is the (i,j)th element in the Minnesota-dummy augmented posterior variance of α.

The predictive variance denoted in Equation (8) reveals two problems that amplify each other. Firstly, the highlighted parameter uncertainty adds up with the number of parameters. As explained in Section 2.2, VAR models suffer from the curse of dimensionality such that even the shrunk parameters add up to a serious amount. Secondly, the parameter uncertainty is further amplified by the predictive variance of the reduced-form shocks in ϵ_{T+1} denoted by the scale-matrix S_1 of the inverted-Wishart posterior for Σ. This problem becomes particularly striking in macroeconomic research because typical macroeconomic datasets have many (correlated) variables which leads to an inflated error-variance in the estimates (S_1).

This issue serves as motivation for the upcoming subsections describing how Hauzenberger et al. (2021) approach the sparsification of Σ and α. In a sparse model, ν_{ij} becomes exactly zero for excluded covariates such that predictive parameter uncertainty is reduced.

3.2.1 Sparsification of α

The straightforward way to obtain a sparse representation of α is by randomly setting values to zero and comparing the restricted model with its unrestricted version. However, the solution space of 2^k possible solutions is unfeasible to explore even on modern computers. Therefore, Hauzenberger et al. (2021) utilize an approach first introduced by Hahn and Carvalho (2015) and later modified by Ray and Bhattacharya (2018). The goal of this technique is to solve the following optimization problem to obtain a sparse estimator $\hat{\alpha}^*$ for an estimator $\hat{\alpha}$ obtained with the Minnesota prior presented in Section 3.1:

$$\hat{\alpha}^* = \arg\min_{\alpha} \left\{ \frac{1}{2} \|(Z\hat{\alpha} - Z\alpha)\|_2^2 + \sum_{j=1}^{\kappa} \kappa_j |\alpha_j| \right\} , \tag{9}$$

with $Z = (\mathcal{I}_m \otimes X)$, α being the sparse counterpart of the point-estimate $\hat{\alpha}$ and $\|m\|_2^2$ the notation for the Euclidean norm of a vector m. Intuitively, the first part of this optimization problem is the distance between an unrestricted (non-sparse) and a restricted (sparse) model while the second part is a penalty for non-zero values of α.

Similar to the lasso example in Section 2.1, it is required to carefully choose the penalization strength denoted here by κ_j. While Hahn and Carvalho (2015) propose to employ a cross-validation approach – which is computationally not feasible for large κ – Ray and Bhattacharya (2018) adopt the signal adaptive variable selection (SAVS) estimator. This estimator utilizes the coordinate descent algorithm (Friedman et al., 2007) and yields a closed form solution when dividing the optimization problem of Equation (9) into multiple separate

problems for each column Z_j of Z. The solution to this problem is given by

$$\hat{\alpha}_j^* = \text{sign}(\hat{\alpha}_j)\|Z_j\|^{-2}(|\hat{\alpha}_j|\|Z_j\|^2 - \kappa_j)_+ \,, \tag{10}$$

for $j = 1, \ldots, k$, sign(c) returning the sign of a real number c and $(c)_+ = \max\{c, 0\}$.

An advantage of this ex-post sparsification is that κ_j is a variable-specific penalization that relaxes the restriction of the conjugate Minnesota prior setup where it is only possible to impose a proportional strength of shrinkage on coefficients across equations.

Setting κ_j can be realized by making it dependent on the non-sparse estimate $\hat{\alpha_j}$:

$$\kappa_j = \frac{\lambda}{|\hat{\alpha}_j|^\zeta} \,, \tag{11}$$

with $\lambda > 0$ and $\zeta \geq 1$. Intuitively, the larger ζ the more weight is placed on small coefficients $\hat{\alpha}_j$ to be zeroed out in post-processing. Empirically, Ray and Bhattacharya (2018) found $\zeta = 2$ to be a good choice. Choosing λ is more involved. Hauzenberger et al. (2021) modify the SAVS estimator to choose λ in a way to discriminate between the number of lags and whether it is an own lag or the lag of another variable for each VAR equation. They set λ as

$$\lambda_{l,ij} = \begin{cases} \lambda(l-1)^2 & \text{if} \quad i = j \\ \lambda l^2 & \text{if} \quad i \neq j \end{cases} \,, \tag{12}$$

for $l = 1, \ldots, p; i = 1, \ldots, m; j = 1, \ldots, m$. The first case refers to own lags of variables and sets the penalty to zero for the first own lag. The penalty increases with the lag number capturing the notion that importance dies out over time (as expected in AR processes). The second case increases the penalty quadratically in lag number for other variables, respectively. In this setup, λ is now considered a time-invariant scaling parameter where higher values impose a stronger shrinkage and lower values a weaker shrinkage.

3.2.2 Sparsification of Σ

As depicted in Equation (8), the second driver of estimation uncertainty is the variance-covariance matrix Σ and its posterior variance S_1. Sparsifying Σ can further counteract predictive uncertainty. Hauzenberger et al. (2021) use a standard approach to ex-post sparsify the precision matrix Σ^{-1} called the graphical lasso (Bashir et al., 2019; Friedman et al., 2008). The loss

function to find the best sparse estimator $\hat{\Omega}^*$ for the precision matrix Σ^{-1} with elements given by ω_{ij} is defined as follows:

$$\hat{\Omega^*} = \arg\min_{\Omega} \left\{ \text{tr}(\Omega \hat{S}) - \log(\det(\Omega)) + \sum_{i \neq j} \rho_{ij} |\omega_{ij}| \right\} , \tag{13}$$

with $\hat{S}$ representing a variance-covariance matrix estimate, ρ_{ij} a parameter-specific penalty, $\log(\det(\bullet))$ the log-determinant of a matrix and $\text{tr}(\bullet)$ the trace of a square matrix. Generally, one can observe that this problem is similar to the one in Equation (9). The first part of this optimization problem measures the negative expected model fit while the second part imposes a parameter specific penalty on non-zero precision values. Again, the problem can be solved as a set of independent optimization problems using the coordinate descent algorithm applied on each off-diagonal element.

The penalty term ρ_{ij} is chosen as

$$\rho_{ij} = \frac{w}{|\hat{s}^*_{ij}|^{\frac{\kappa}{2}}} , \tag{14}$$

where $|\hat{s}^*_{ij}|$ denotes the absolute value of the (i,j)th element in $\hat{S}^{-1}$ and w is a lag-invariant scalar penalty parameter. $\kappa \geq 1$ controls the penalization strength on small precision parameters.

3.2.3 Approximation of Sparse Posteriors

The ex-post sparsification problems described in the preceding two sections yield exactly one sparsified point estimate obtained by solving the optimization problems defined in Equations (9) and (13). However, using point-estimates only misses the whole point of applying Bayesian estimation. The key advantage of Bayesian estimation is that the researcher can make statements about the predictive uncertainty by looking at the variance of parameter-specific posterior distributions. This is not possible with point-estimates only.

To counteract this issue, Hauzenberger et al. (2021) propose a way to approximate a sparsified posterior distribution similar in nature to the way how one would approximate a posterior distribution using MCMC if the analytical form is unknown. Instead of taking one point-estimate like the posterior mode and sparsifying it, the authors propose to draw repeatedly from the (non-sparse but shrunk) posterior and *sparsify each draw*. Drawing from the posterior is easy. First, one can sample R values $\Sigma^{(r)}$ from the marginal posterior of $\Sigma|Y,X \sim \mathcal{W}^{-1}$ and subsequently sample R values $\alpha^{(r)}$ from the conditional posterior of $\alpha|\Sigma,Y,X \sim \mathcal{N}$ plugging in each sample $\Sigma^{(r)}$ consecutively.

Sparsifying each draw from the posterior and taking a sample point-estimate from the population of sparsified draws is approximately similar to simply taking one point-estimate from the analytical posterior and sparsifying it. However, the major advantage of sampling and sparsifying repeatedly is that one can make statements about the predictive uncertainty of the obtained sparse model. Intuitively, one approximates a sparse joint posterior distribution $p(\hat{\alpha}^*, \hat{\Omega}^* | Y, X)$.

This result is a core result in this paper by Hauzenberger et al. (2021) and a great advantage over comparable approaches (Hahn & Carvalho, 2015) because it allows the quantification of predictive uncertainty in post-processed sparse BVAR models.

4 Macroeconomic Forecasting Application

To validate the model behavior in a real-world scenario Hauzenberger et al. (2021) conduct an empirical forecasting application on a popular macroeconomic dataset, namely the McCracken and Ng (2015) dataset.

This dataset spans $m = 165$ macroeconomic and financial variables in quarterly resolution from 1959:Q1 up until 2018:Q4. The target variables of the forecasting exercise are the traditional ones: Output (GDPC1), consumer price inflation (CPIAUCSL), and the Federal Funds Rate (FEDFUNDS).

4.1 Forecasting Exercise Setup

To evaluate the performance of sparse BVAR models, the authors compare different sparse models with different hyperparameters to other competitive models specified below.

They rely on a recursive forecasting design and compute $h \in \{1, 4, 8\}$-step-ahead forecasts. Recursive means that the training set, which spans the initial observations from 1959:Q1 to 1989:Q4 (first 30 years), is expanded one step at a time to recursively compute out-of-sample forecasts for the next h time steps.

The metrics used to evaluate the model performance on the hold-out set are the root-mean-squared-forecast-error (RMSE) for point-estimates and log predictive likelihoods (LPLs) for predictive densities (see Appendix B.1 and B.2).

In order to give meaning to these metrics Hauzenberger et al. (2021) compare different models with different hyperparameter setups against each other. Ideally, sparsification should select the most important variables automatically which is why the **L-VAR** model features all $m = 165$ variables of the dataset. This goes in line with existing literature highlighting that it is important to exploit larger information sets (Bańbura et al., 2010; Giannone et al., 2015; G. M. Koop, 2013). This model size is then compared to other model sizes:

S-VAR This model is considered the baseline model and all metrics are computed relative to the metrics of this model. It is a simple BVAR model featuring only the three target variables.

M-VAR The total number of variables included in this model are $m = 21$ variables. This extends the **S-VAR** by 18 financial market variables.

FA-VAR This factor-augmented (FA) model attempts to exploit the full information set by extending the **S-VAR** with three principal components extracted from all other variables except the three target variables.

All models feature $p = 5$ lags and are estimated with the Minnesota prior setup presented in Section 3.1. In addition, a non-conjugate BVAR model with the SSVS prior is considered (George et al., 2008). However, this model is not computed for the **L-VAR** model size due to the additional computational burden of the non-conjugate prior (meaning that the posterior needs to be simulated).

The choice of hyperparameters differs for the **L-VAR** in comparison to the other model sizes. For all models except for the **L-VAR** the θ_1 hyperparameter of the Minnesota prior (which is part of the δ hyperparameter tuple specified in Section 3.1) is determined by maximizing the marginal likelihood over a grid of search values. This procedure is unfeasible for the **L-VAR** which is why results are reported for $\theta_1 \in \{0.025, 0.05, 0.075\}$. Large θ_1 values put less weight on shrinkage while small values put more weight on shrinkage. As a consequence, the SAVS estimator used to ex-post sparsify estimates leads to a more sparse model if θ_1 is small. The values of λ (see Section 3.2.1) and w (see Section 3.2.2) are also reported over a grid for $\lambda \in \{0.01, 0.1, 0.5, 1\}$ and $w = \frac{\lambda}{10}$.

4.2 Average Point- and Density Forecast Performance

The average point- and density forecast performances as evaluated by the RMSE and the LPLs in comparison to the baseline **S-VAR** only reveal mixed evidence in favor of larger sparse models. The average forecast performance is the out-of-sample model performance averaged over the complete hold-out set (1990:Q1 to 2018:Q4). It is reported separately for each forecasting horizon $h \in \{1, 4, 8\}$ and the grid of hyperparameters specified in Section 4.1. In the following, the results are summarized and the detailed results can be found in Hauzenberger et al. (2021, Tables 2, 3, 4, and 5).

Generally, one can observe that multi-step-ahead forecasts for $h \in \{4, 8\}$ benefit from larger information sets. **L-VAR** and **M-VAR** models perform better on average for these forecasts. Regarding sparse versus non-sparse models, one can observe that using ex-post sparsification and the Minnesota prior works well, especially for longer run forecasts of output and interest rate. Sparse models using the non-conjugate SSVS prior never dominate for point-estimates and only in rare cases for density forecasts. The main driver of bad performance of both sparse and large models is revealed by the marginal LPL of inflation (Hauzenberger et al., 2021, p. 319, Table 4). The **S-VAR**

and **FA-VAR** perform best while the density forecasting performance of **L-VAR** models and especially sparse **L-VAR** models drops sharply. This is a well-known result in central bank practice (Giannone et al., 2015) and investigated further in Section 4.3. For output and interest rate, this behavior is reversed meaning that larger and more sparse models perform better overall.

Finally, it is important to note that sparsification only very rarely hurts forecasting performance in a significant manner which the authors validate by computing Model Confidence Sets (MCS) on a 25% significance level. For point-forecasting performance, the MCS contain between 26 and 33 out of 33 models in total with large shares of models using SAVS across all target variables. In terms of density forecasting performance, the MCS are smaller for output and interest rate with larger shares of models using SAVS. For inflation, the MCS is larger again and favors smaller models as discussed previously.

4.3 Density Forecast Performance over time

To investigate further how large and sparse models using SAVS perform, it is interesting to consider the forecasting performance over time. Discriminating between different time periods allows identifying reasons for performance drops.

Figure 2 depicts the cumulative sum of log predictive likelihoods for one-year-ahead ($h = 4$) forecasts over time for the complete hold-out period. The left-hand side is the period before the financial crisis while the right-hand side illustrates the period during and after the financial crisis. One can observe that, before the financial crisis (left plot), large models using SAVS and the **M-VAR** using SAVS perform very well. Especially the large models with moderate shrinkage and SAVS (solid blue and magenta lines) are superior to all other models. This means that there is strong evidence in favor of large and sparse models before the financial crisis.

On the other hand side, this pattern is reversed during the financial crisis as depicted by the gray shaded background in the right plot. All **L-VARs** with SAVS exhibit by far the strongest drop in density forecasting performance. This reflects the previously observed fact in Section 4.2 that inflation forecasts are the main driver of bad performance. The heavy misestimation of inflation seems to impact the joint forecasting performance of the sparse **L-VAR** models dramatically. During the time of the financial crisis, models without SAVS and the **M-VAR** and **FA-VAR** with SAVS are more stable. Nonetheless, after the financial crisis, the larger models, such as all **L-VARs** and the **M-VAR** with SAVS, return to the good performance like before the crisis which can be seen by the steep increase of the curve after the end of the crisis.

(b) *One-year-ahead*

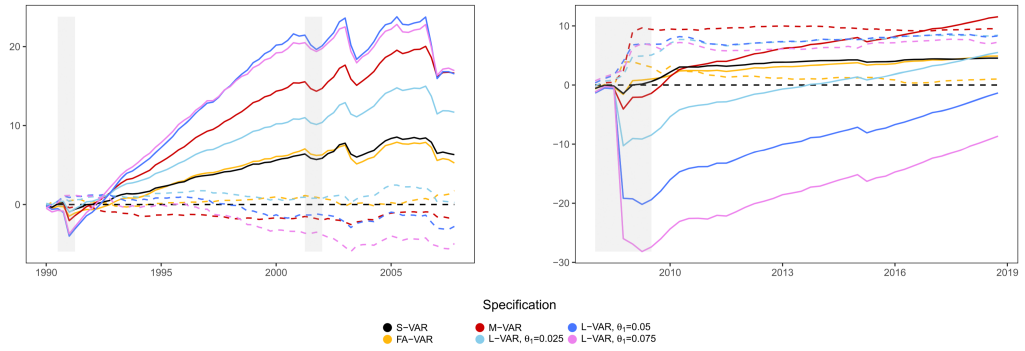

Figure 2: Cumulative sum of LPLs for one-year-ahead forecasts before (left), during, and after (right) the financial crisis (Hauzenberger et al., 2021, Figure 2). Solid lines indicate models with SAVS while dashed lines represent the non-sparse counterpart without SAVS. Large values for θ_1 put less weight on shrinkage leading to a more dense model after SAVS. Higher values are better.

Similar results can be observed for one-quarter-ahead forecasts (Hauzenberger et al., 2021, Figure 2).

A reason for the decline in forecasting performance during the financial crisis is most likely related to the bias-variance trade-off introduced in Section 2.1. A researcher typically applies shrinkage and sparsity techniques to a model to reduce the predictive variance while sacrificing bias. If the hyperparameters are chosen carefully this usually comes with the benefit of better generalization to new data. However, as one can see by the decline of performance during the financial crisis, a reduced predictive variance also means that the model is less likely to capture outliers in the data – like the financial crisis. This instance highlights the bias-variance trade-off very well: Sacrificing bias for a smaller predictive variance leads to great forecasting performance during "normal" times but comes with the drawback of a decline in performance during "unusual" times like the financial crisis.

5 Conclusion

This seminar paper presents the methods proposed by Hauzenberger et al. (2021) to combine shrinkage and sparsity in Bayesian VAR models while maintaining conjugacy using the well-known Minnesota prior setup. Maintaining conjugacy comes with the advantage of computational efficiency because key distributions such as the marginal posterior distributions and the predictive density are available in closed form and do not need to be approximated with MCMC methods. At the same time, the proposed ex-post sparsification of point-estimates relaxes the restrictive property of the Minnesota prior of not being able to have different predictors across VAR equations. The drawback of working with point-estimates only is solved by drawing and sparsifying repeatedly from the posterior as outlined in Section 3.2.3. This is computationally feasible because the utilized optimization algorithms converge very fast. In fact, the computational speed to estimate a conjugate model with the presented ex-post sparsification is by orders of magnitude faster than working with a non-conjugate SSVS prior BVAR model (Hauzenberger et al., 2021, Online Appendix).

The forecasting application in Section 4 shows that sparsification in large models does not yield noteworthy performance gains for point-estimates but leads to slightly better density estimates with the exception of inflation forecasts. As investigated further in Section 4.3, one cannot recommend the usage of large, sparse BVAR models to central bank researchers who are interested in inflation forecasts due to the huge drop in forecasting accuracy during the financial crisis. The bias-variance trade-off has been identified as a potential reason for this drop. A sparse model is less likely to capture outliers due to the reduced predictive variance and performs worse in times of crisis which can be interpreted as outliers.

To conclude, the sparsification approach presented in this work combines desirable properties of the conjugate and non-conjugate model setup. While keeping the computational burden moderate the ex-post sparsification relaxes restrictive properties of the traditional conjugate Minnesota prior. However, it is worth noting that the forecasting application cannot be reproduced on a standard desktop computer in a timely manner even for the proposed conjugate model setup. Further research might be needed to improve computational speed such as the work by Chan (2019) who adopt non-conjugate Minnesota-type adaptive hierarchical priors while keeping the computational burden very low using a custom posterior simulator.

A Bayesian Lasso Estimate

The following demonstrates that a Bayesian regression model with a Laplace prior on the coefficients centered at zero yields the frequentist lasso estimate at its posterior mode. Full specifications of the Bayesian lasso can be found for example in Figueiredo (2003), Park and Casella (2008), and Yuan and Lin (2005).

Consider the standard regression model $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$ with $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ and i.i.d. $\beta = \beta_1, \beta_2, \ldots, \beta_p$. The joint likelihood is normal and given by

$$p(y|X, \beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2\right) \tag{15}$$

with $\epsilon_i = y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}$. A Laplace prior centered at zero conditional on the scale parameter τ can be written as

$$p(\beta|\tau) = \prod_{j=1}^{p} \frac{1}{2\tau} \exp\left(-\frac{|\beta_j|}{\tau}\right) \propto \exp\left(-\frac{1}{\tau} \sum_{j=1}^{p} |\beta_j|\right). \tag{16}$$

Given the likelihood in Equation (15) and the prior in Equation (16), the posterior $p(\beta|y, X, \tau) \propto p(y|X, \beta)p(\beta|\tau)$ is

$$\begin{aligned} p(\beta|y, X, \tau) &\propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2\right) \exp\left(-\frac{1}{\tau} \sum_{j=1}^{p} |\beta_j|\right) \\ &= \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2 - \frac{1}{\tau} \sum_{j=1}^{p} |\beta_j|\right). \end{aligned} \tag{17}$$

Now, it will be shown that the optimization problem for finding the posterior mode of Equation (17) can be rewritten to the frequentist lasso optimization problem as specified in Section 2.1, Equation (1). Therefore, the posterior mode is calculated by maximizing the *log*-posterior:

$$\log p(\beta|y, X, \tau) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2 - \frac{1}{\tau} \sum_{j=1}^{p} |\beta_j| = -\left(\frac{1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2 + \frac{1}{\tau} \sum_{j=1}^{p} |\beta_j|\right).$$

Since the *log*-posterior is negative one can reformulate the maximization problem to a minimization problem by omitting the negative sign:

$$\arg\min_{\beta} \left[\frac{1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2 + \frac{1}{\tau} \sum_{j=1}^{p} |\beta_j|\right] = \arg\min_{\beta} \frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} \epsilon_i^2 + \frac{2\sigma^2}{\tau} \sum_{j=1}^{p} |\beta_j|\right].$$

Ignoring the normalizing constant $\frac{1}{2\sigma^2}$ and setting $\lambda = \frac{2\sigma^2}{\tau}$ yields the frequentist Lasso estimate:

$$\arg\min_{\beta} \left[\sum_{i=1}^{n} \epsilon_i^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right].$$

□

B Model Evaluation Metrics

B.1 Root-Mean-Squared-Forecast-Error (RMSE)

The root-mean-squared-forecast-error (RMSE) describes the square-root of the squared average deviation of forecasts from the true values. It can be written as follows (Bańbura et al., 2010):

$$\text{RMSE} = \sqrt{\frac{1}{T_1 - T_0 - H + 1} \sum_{T=T_0+H-h}^{T_1-h} (y_{i,T+h|T} - y_{i,T+h})^2},$$

where $y_{i,T+h|T}$ is a point estimate (e.g. posterior median) at time $T+h$ given the information set for the time span T and $y_{i,T+h}$ the true value at time $T+h$. T_0 and T_1 are the beginning and the end of the evaluation sample and H the longest forecast horizon.

B.2 Predictive Likelihood

The predictive likelihood is the posterior predictive density evaluated at the actual true outcome. The sum of log predictive likelihoods can be used to compare the forecasting performance of one model relative to other models. It is defined as follows (G. M. Koop, 2013):

$$\sum_{T=T_0}^{T_1-h} \log[p(y_{i,T+h|T} = y_{i,T+h}|\text{Data}_T)],$$

where $y_{i,T+h|T}$ is a point estimate (e.g. posterior median) at time $T+h$ given the information set for the time span T and $y_{i,T+h}$ the true value at time $T+h$. T_0 and T_1 are the beginning and the end of the evaluation sample and h the forecast horizon. Finally, Data_T is the information set available at time T.

References

Bańbura, M., Giannone, D., & Reichlin, L. (2010). Large bayesian vector auto regressions. *Journal of Applied Econometrics*, *25*(1), 71–92. https://doi.org/10.1002/jae.1137

Bashir, A., Carvalho, C. M., Hahn, P. R., & Jones, M. B. (2019). Post-processing posteriors over precision matrices to produce sparse graph estimates. *Bayesian Analysis*, *14*(4). https://doi.org/10.1214/18-BA1139

Chan, J. C. C. (2019). Minnesota-type adaptive hierarchical priors for large bayesian VARs. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3440949

Figueiredo, M. T. (2003). Adaptive sparseness for supervised learning [Place: Los Alamitos, CA, USA Publisher: IEEE Computer Society]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *25*(9), 1150–1159. https://doi.org/10.1109/TPAMI.2003.1227989

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441. https://doi.org/10.1093/biostatistics/kxm045

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization [Publisher: Institute of Mathematical Statistics]. *The Annals of Applied Statistics*, *1*(2), 302–332. https://doi.org/10.1214/07-AOAS131

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013, November 27). *Bayesian data analysis* (3rd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/b16018

George, E. I., Sun, D., & Ni, S. (2008). Bayesian stochastic search for VAR model restrictions. *Journal of Econometrics*, *142*(1), 553–580. https://doi.org/https://doi.org/10.1016/j.jeconom.2007.08.017

Giannone, D., Lenza, M., & Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics*, *97*(2), 436–451. https://doi.org/10.1162/REST_a_00483

Hahn, P. R., & Carvalho, C. M. (2015). Decoupling shrinkage and selection in bayesian linear models: A posterior summary perspective [Publisher: Taylor & Francis]. *Journal of the American Statistical Association*, *110*(509), 435–448. https://doi.org/10.1080/01621459.2014.993077

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Hauzenberger, N., Huber, F., & Onorante, L. (2021). Combining shrinkage and sparsity in conjugate vector autoregressive models [Publisher: John Wiley and Sons Ltd]. *Journal of Applied Econometrics*, *36*(3), 304–327. https://doi.org/10.1002/jae.2807

Kadiyala, K. R., & Karlsson, S. (1997). Numerical methods for estimation and inference in bayesian VAR-models [Publisher: Wiley Online Library]. *Journal of Applied Econometrics*, *12*(2), 99–132.

Koop, G., & Korobilis, Dimitris. (2009). Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends in Econometrics*, *3*(4), 267–358. https://doi.org/10.1561/0800000013

Koop, G. M. (2013). Forecasting with medium and large bayesian VARS. *Journal of Applied Econometrics*, *28*(2), 177–203. https://doi.org/https://doi.org/10.1002/jae.1270

Litterman, R. B. (1986). Forecasting with bayesian vector autoregressions: Five years of experience [Publisher: [American Statistical Association, Taylor & Francis, Ltd.]]. *Journal of Business & Economic Statistics*, *4*(1), 25–38. http://www.jstor.org/stable/1391384

Lütkepohl, H. (2005). *New introduction to multiple time series analysis* [OCLC: ocm61028971]. New York: Springer.

McCracken, M. W., & Ng, S. (2015). *FRED-MD: A monthly database for macroeconomic research*. https://doi.org/10.20955/wp.2015.012

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

Ray, P., & Bhattacharya, A. (2018). Signal adaptive variable selector for the horseshoe prior.

Sims, C. A. (1980). Macroeconomics and reality [Publisher: Wiley, Econometric Society]. *Econometrica*, *48*(1), 1–48. http://www.jstor.org/stable/1912017

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso [Publisher: Royal Statistical Society, Wiley]. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. http://www.jstor.org/stable/2346178

Yuan, M., & Lin, Y. (2005). Efficient empirical bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, *100*, 1215–1225. https://doi.org/10.1198/016214505000000367