Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis Econometrics & Management Science

Program: Econometrics

Prior sensitivity in time-varying parameter vector autoregressions

---

Author:
C.A. Vriends
440435

Supervisor:
dr. Annika Schnucker
Second assessor:
dr. Martina Zaharieva

Date:
January 31, 2021

Abstract

This paper investigates the (local) sensitivity of three commonly used priors (SVSS, Lasso & Horseshoe prior) in large time-varying parameter vector autoregression models (TVP-VARs). One of the hindrances in conducting a sensitivity analysis in a TVP-VAR is the proliferation of parameters that makes a MCMC approach often infeasible. Therefore, we opt for a Variational Inference (VI) approach that softens the burden of computation and makes a numerical based sensitivity approach practical. In this paper we derive the variational posteriors for two priors and make a TVP-VAR feasible by means of a Cholesky decomposition of the covariance matrix. First, we show for three different priors that a VI-based TVP-VAR is competitive with and often surpasses a common MCMC-based benchmark. Second, the local sensitivity of the priors is investigated in an empirical exercise. Where we find, in agreement with earlier literature, that there is a difference in sensitivity to shrinkage between short- and long-horizon forecasts. Long-horizon forecasts tend to be less susceptible to shrinkage than the short-horizon forecasts. This might be due to the fact that longer horizon forecasts converge to the unconditional mean of the system. Furthermore, given the parameter space for which we investigate the sensitivity, the Horseshoe prior seems to be the most sensitive to changes in the hyperparameters whereas the Lasso is the least sensitive. Although, the VI approach makes it practically feasible to use a numerical based approach to (local) sensitivity analysis, it still has its limitations due to the complexity of the function that we have to approximate the derivatives for. Nonetheless, this effort and these results are useful to a practitioner or academic to see if investigating local sensitivity is practical and worthwhile if one opts for similar models.

Acknowledgments

I would first like to thank my supervisor Dr. Annika Schnucker for her invaluable discussions on the methodology and help in formulating a research question. Her insightful feedback has pushed the contents of this thesis to a higher level. Furthermore, I would like to thank the co-reader Dr. Martina Zaharieva.

I would also like to thank the reviewers, Rick Lamers, Bart Overes, Willemijn Brus and Marc Stam, for their time and dedication to read through it and to force me to be more parsimonious in the formulation of ideas, the wording of certain paragraphs as well as the mathematical notation.

In addition, I would like to thank my parents for their continued support and being careful listeners whenever I needed. Finally, I could not have completed this thesis successfully without my friends, Rick Lamers, Kevin Visser, Jeroen Reunis and Tom Heideveld, whom provided me with numerous distractions and thoughtful discussions. Moreover, I would like to thank Martijn Hofstra and Marc Stam for the countless hours spent together in Polak and elsewhere, contemplating about the thesis.

Contents

1 Introduction

In the field of macroeconomics, the time-varying parameter vector autoregression (TVP-VAR) is a natural extension of the vector autoregression (VAR). The TVP-VAR is one of the models that has become more prevalent in the literature since the seminal work of Cogley and Sargent (2001, 2005) and Primiceri (2005). This is, in part, due to the fact that a TVP-VAR allows the macroeconomic relationship to change over time. From an inference perspective, it is important to allow this simultaneous time-varying element, as it otherwise might be uncertain whether the coefficients truly drift over time or that it is due to an incorrect homoskedastic specification of the volatility (Cogley & Sargent, 2005). Furthermore, several papers, [such as Clark (2011), D'Agostino et al. (2013), Koop and Korobilis (2013), Clark and Ravazzolo (2015)] show that allowing for time-variation in the autoregressive coefficients as well as the volatility is crucial to the forecasting performance of a model.

However, modeling time variation, in a macroeconomic setting, increases the number of unknown parameters substantially. The large number of parameters have to be estimated with relatively few observations as macroeconomic variables are often reported at a monthly or quarterly interval. Various approaches exist to deal with overparameterization, such as restricting parameters based on economic theory or using Bayesian shrinkage and selection based priors [(Litterman, 1986), (George et al., 2008)].

Although the Bayesian approach has seen empirical and theoretical success in TVP-VARs one of the common critiques is the element of subjectivity that is introduced by the use of a prior. A prior distribution encapsulates an *a priori* belief on the distribution of the parameters of interest. An academic or practitioner might be more or less resolute in its belief. There are several answers to this critique, the one that is opted for in this paper is a sensitivity analysis on the choice of prior (*global*) and the choice of its parameters (*local*). Other considerations might be that one opts for a non-informative prior (this does not induce shrinkage (Koop & Korobilis, 2010), thus it is not viable to resolve overparameterization), one might resort to hierarchical models or one might choose a more data-informed prior.

The feasibility of a sensitivity analysis is intimately linked to the method of estimation of a model. A common approach to sensitivity analysis is based on numerical finite-differences methods [Chan et al. (2019), Gelman et al. (2013)]. In this approach to sensitivity analysis, one estimates the derivative of an important hyperparameter with respect to the forecast performance (it might be an other measure) of a model. As it is a numerical method, one has to re-estimate the model for each small perturbation in one of the hyperparameters to approximate the derivative. In the case of estimating a TVP-BVAR[1], one often has to resort to simulation based techniques, such as Markov Chain Monte Carlo (MCMC), as the posterior is intractable. Although MCMC has the attractive characteristic that, given infinite computing resources, it converges to the true posterior (Bishop, 2006), it remains a computationally intensive method that makes it impractical for sensitivity

[1] A TVP-VAR that is estimated using Bayesian methods is referred to as a TVP-BVAR

analysis based on numerical finite-differences approaches. In addition, while there is ample empirical evidence that more variables in a VAR (and TVP-VAR) allows for better forecasting performance [Koop and Korobilis (2013), Carriero et al. (2019)], extending a TVP-VAR beyond 3 to 5 variables makes it practically infeasible to use MCMC. This is where Variational Inference (VI), or if it is used in a Bayesian context Variational Bayes (VB), comes in. This estimation technique has had several successes in the field of Machine Learning (ML). It approximates the intractable posterior distribution by a tractable family of distributions. Unlike MCMC, it does not guarantee that it converges to the true posterior distribution, it remains an approximation (Blei et al., 2017). Nonetheless, it is a deterministic algorithm that is (very) fast compared to MCMC (Gefang et al., 2019). This makes it feasible to opt for the more traditional numerical finite-differences approach to sensitivity analysis. Moreover, it has shown that it approximates the posterior well enough and that the difference in parameter estimates is negligible compared to MCMC [Hajargasht (2019), Koop and Korobilis (2020)], whilst, still being able to handle larger models that are infeasible with MCMC [Gefang et al. (2019), Koop and Korobilis (2020)].

The aim of this paper is twofold. First, we propose to use VB on an equation-by-equation specification of the TVP-VAR to address the challenge of estimation and investigate whether it might be an adequate substitute in the space of TVP-BVARs for MCMC-based models. Second, we investigate the local sensitivity of several priors.

On the first aspect, we conduct a simulation study, concerning several different data generating processes (DGPs), where, among other models, the TVP-BVAR (based on MCMC) is used as a benchmark. As mentioned before, opting for VB lowers the computational burden of estimating a TVP-BVAR compared to MCMC. On top of this, we reduce the computational burden even further by using a Cholesky decomposition approach, whilst not affecting the forecast performance in a meaningful manner. In the case of a TVP-VAR (or VAR), one can either estimate the system of equations as a whole or one can resort to a transformation of the system in a manner that estimating it as a whole is not required. This in turn lightens the computational burden. The limiting factor in breaking up the estimation of the system in completely separate estimations is the possible correlation among the errors of the different equations in the system. A transformation that incorporates these cross-correlations, while, at the same time reducing the estimation of the system as a whole to an equation-by-equation estimation problem, is a specific transformation that makes use of the Cholesky decomposition of the covariance matrix. In this manner, one still retains the realistic assumption of a complete covariance matrix, while materially reducing the computational burden at the same time. It is common in the literature to resort to such a transformation [(Lopes et al., 2018), (Carriero et al., 2019) (Huber et al., 2020)]. This transformation opens up the possibility to use a univariate TVP model. In their work, Koop and Korobilis (2020), suggest a specific VB-based TVP and show that it is a feasible alternative to MCMC and that the number of variables can be scaled to a number which is impractical using MCMC. Moreover, their suggestion for the factorization of the posterior is relatively general, as it allows for different priors on the shrinkage of the coefficients. This is a useful suggestion, as some problems favour shrinkage-based priors over

selection-based priors, or vice versa (Huber et al., 2020). This paper extends their model to a TVP-BVAR using the aforementioned Cholesky decomposition of the covariance matrix and implements, besides their own prior, two priors that have become more common in the literature of the past decades. These priors became more prevalent due to their favourable results in different forecasting exercises. The three different priors are: (*i*) the Stochastic Variable Search and Selection (SVSS) prior used in Koop and Korobilis (2020), (*ii)* the (Bayesian) Least Absolute Shrinkage and Selection Operator (Lasso) proposed by Belmonte et al. (2014) and (*iii*) the Horseshoe prior proposed by Huber et al. (2020). Each of these priors is implemented according to an hierarchical specification, the parameters, as of now referred to as hyperparameters, at the highest level of the hierarchy are used to conduct the sensitivity analysis. The remaining parts of the hierarchical structure are left untouched for each of the priors. The variational posterior derivations for the priors that are required for the use of VB are derived in this paper for the last two priors. These derivations rely on the results in the papers that they were first presented [Belmonte et al. (2014), Huber et al. (2020)]. The derivations for the first prior are already presented by Koop and Korobilis (2020).

On the second aspect, we conduct a sensitivity analysis in a TVP-BVAR with respect to the point forecast performance. The sensitivity of three different priors is evaluated. Moreover, for each prior the sensitivity of the hyperparameters is investigated by means of approximating the derivative using a numerical finite-differences approach. The empirical data for the sensitivity analysis is the US quarterly macroeconomic data from 1959Q4 through 2019Q4 for 248 variables from the Federal Reserve Bank of St. Louis'. The reason that this particular sensitivity analysis might be of interest to practitioners or academics is again twofold, (*i*) it shows that numerical based sensitivity analysis is feasible in more complex models if one is willing to resort to VB, (*ii*) it provides an indication of the overall sensitivity of a TVP-BVAR. In the case of VB, these might be important considerations before one would use a VB-based instead of MCMC-based TVP-BVAR in economic forecasting.

The remainder of the paper is structured in the following manner. Section 2 discusses the relevant literature in the area of TVP-VARs (e.g. larger systems and priors) and sensitivity analysis, section 3 explains the methodology of TVP-VARs in a Bayesian context, the exact hierarchical structure of the priors and in depth the use of VB. Section 4 describes the different DGPs for the simulation study and the exact macroeconomic series used in the sensitivity analysis. The results for the simulation study as well as the empirical exercise are presented in section 5. Section 6 is the conclusion and 7 concludes the paper with a discussion. The appendix contains the derivations for the two additional priors, pseudo-code of the complete estimation procedure, an addendum to the simulation exercise and a short discussion on reproducibility.

2 Literature

In this section different approaches to alleviate the issues in larger [2] TVP-VARs are discussed. Moreover, a short outline of the literature on sensitivity analysis and shrinkage priors is provided.

2.1 Large TVP-VARs

One of the first attempts to estimate a large system is from Koop and Korobilis (2013), it relies on forgetting factors. A TVP-VAR can be formulated in a state-space specification. In their approach, instead of estimating the error covariance matrix of the coefficients in the state equation at each time interval, they make it proportional to the previous error covariance matrix, this proportionality constant is the forgetting factor. As a state-space model is estimated in a recursive fashion, this is a considerable reduction in the number of parameters. Furthermore, this approach does not rely on a simulation scheme, such as MCMC, as it is an approximation to the posterior. It is at least able to handle up to 25 variables.

Koop et al. (2019) use the previously mentioned approach (Koop & Korobilis, 2013) with a form of dimension reduction on the coefficients as well as on the elements of the error covariance matrices. The dimension reduction technique that they opt for is random compression in combination with Bayesian model averaging (BMA). According to Koop et al. (2019) BMA sets a higher weight on the (random) compression that contributes the most to the explanation of the dependent variable. This dimension reduction technique is preferred over PCA as it is a supervised instead of an unsupervised reduction technique. Their approach is able to handle up to at least 129 variables in a TVP-BVAR. As it is based on the previous work of forgetting factors, it inherits the same estimation properties, it is a simulation-free approximation.

Huber et al. (2020), transform the error covariance matrix in the measurement equation to a diagonal matrix by means of a Cholesky decomposition approach, this makes it possible to estimate the TVP-VAR in an equation-by-equation manner. Furthermore, this allows for the use of MCMC and its inherent properties of convergence to an exact posterior. This approach is at least able to handle up to 30 variables.

Chan et al. (2020) note that in the empirical literature the error covariance matrix of the state equation is often near-singular. This trait is used to reduce the number of states in the state-space specification while at the same time retaining the same number of time-varying parameters. In essence, there is a lower number of states that drive the change in the time-varying parameters. In addition to the dimension reduction, some computational efficiencies are materialized by the use of a different (non-centered) specification and the use of a particle filter instead of the Kalman filter. Furthermore, this approach also relies on MCMC and is at least able to handle up to 15 variables.

In the mentioned literature different assumptions are made with regards to the structure of the TVP-VAR (e.g. diagonal or non-diagonal error covariance matrices) or some form of dimension

[2] Large differs between papers, as it varies from 25 (Koop & Korobilis, 2013) to 129 (Koop et al., 2019)

reduction is used and a trade-off is made between forecasting accuracy (Koop & Korobilis, 2013) and the possibility to conduct sensible inference (Chan et al., 2020).

Another avenue of research is the use of VB to estimate larger VAR models (also TVP-VARs). Gefang et al. (2019) resort to VB to estimate a large VAR that incorporates hierarchical shrinkage. They arrive at the conclusion that VB is a satisfactory alternative to MCMC as estimation is quite accurate compared to MCMC and the estimation, even in large systems, is relatively quick. Koop and Korobilis (2018) use VB in a TVP setting, due to their choice of factorization of the approximation distribution, standard techniques, such as the Kalman filter, still fit into this framework. They find a similar conclusion as Gefang et al. (2019), VB is a reasonable approach, as it is as accurate as MCMC in lower-dimensional settings and scales well compared to MCMC in higher-dimensional settings.

2.2 Sensitivity analysis

Sensitivity analysis is an integral component of a complete Bayesian modeling effort. As mentioned in the introduction, Bayesian statistics is inherently more subjective, compared to classical statistics, as it forces the academic or practitioner to make an informed choice about the prior. Although, this is the widely held view, some arguments can be made that even if one opts for a classical approach, the choices that one has to make in the modeling process, inevitably subjectivity slips in (e.g. choice of functional form). Lopes and Tobias (2011) make the argument, that it is incorrect to claim that the difference between the two schools of statistics can solely be based on the grounds of subjectivity, in part, it has to be based on the fact that the Bayesian approach is clearly different as prior information can be incorporated in a clear and formal manner in a way that is in line with the basic rules of probability. Furthermore, they support this line of reasoning with the fact that estimators of both schools are often asymptotically equivalent, suggesting that in neither school subjectivity is absent. Nonetheless, it is almost inevitable that the choice of a prior has an influence on the posterior and in turn on the inferences that are based on the model. Thus, the necessity for a sensitivity analysis is sensible.

Up until now we have discussed the sensitivity of the forecast ability (e.g. in a sense, the posterior) to the prior. It is referred to as posterior-to-prior sensitivity and is part of the domain of robust analysis. In the domain of robust analysis, one might also be interested in posterior-to-loss and to some lesser extent posterior-to-model sensitivity. In this paper, sensitivity analysis is synonymous with posterior-to-prior sensitivity analysis. According to Berger et al. (2000) posterior-to-prior sensitivity analysis can be deconstructed into three different approaches: (*i*) *a (naive) informal approach*, where different priors are considered and one investigates whether a simple estimation for each prior has a substantial effect on the measure that one is interested in (e.g. posterior mean). (*ii*) A *global approach*, where one considers, as many as practically reasonable, priors that are compatible with the type of prior information and one investigates the differences between the chosen priors. (*iii*) A *local approach* where one is interested in the rate of change in the measure of interest with respect to changes in the prior, in this approach one resorts to differential

techniques.

According to Gustafson (2000), the landscape of frameworks to assess local sensitivity is quite fractured. However, a crude stratification is based on three characteristics: (*i*) the type of output one is interested in (a summary statistic of the posterior or the distribution of the posterior), (*ii*) the type of perturbation in the prior (linear, non-linear, parametric or geometric) and (*iii*) whether the worst-case sensitivity is measured in absolute or relative terms (if investigated). The degree of sophistication also varies. The more sophisticated approaches adopt ideas from classical robustness theory and redefine contamination in such a manner that it is appropriate to use the idea of an influence function (Gustafson, 2000). However, as these sophisticated approaches are quite involved, recent literature, especially with respect to BVARs, favours simpler approaches. Chan et al. (2019) investigate the sensitivity of point forecasts and the predictive quantiles with respect to several hyperparameters in a BVAR. They expand on the proposed (general) automatic differentiation (AD) approach of Jacobi et al. (2018). They conclude that the forecasts are quite sensitive to the strength of the shrinkage for the coefficients, but that the prior mean and prior covariance matrix have little to no meaningful effect on the predictive measures. Their work is the first to delve into local sensitivity in a BVAR. In other literature on BVARs (or TVP-BVARs) the (naive) informal approach is still the standard approach to investigate the sensitivity. Nevertheless, the conclusion of Chan et al. (2019), that the strength of the shrinkage has the most profound effect, is in line with the informal sensitivity results that are presented as an addendum in other literature [e.g. Koop and Korobilis (2013), Koop and Korobilis (2020)].

2.3 Priors

As mentioned in earlier paragraphs, the choice of prior is of fundamental importance. It receives some additional attention in a VAR (and by extension a TVP-VAR) as the informativeness of the prior is the main reason why one resorts to the Bayesian approach to solve the problem of parameter proliferation (Koop & Korobilis, 2010). In the earliest work on VARs, the choice of priors was quite limited due to the inability, or infeasibility, to simulate from the posterior, as MCMC was not widely used until the beginning of the millennia (Berger et al., 2000). Thus, this restricted the choice of prior to the class of priors that have analytical results. However, one of the most notable earlier developments that laid the groundwork for further development is the Minnesota (or Litterman) prior [Doan et al. (1984), Litterman (1986)]. It is based on an approximation of the prior for the covariance matrix, all the while, maintaining analytical results and having a higher degree of flexibility compared to a simple conjugate specification. Even though the Minnesota prior and its variants are still used to this day, it is not a full Bayesian approach, as it does not allow for uncertainty in the covariance matrix (Koop & Korobilis, 2010). A more recent development in the literature is the use of hierarchical global-local priors [e.g. Belmonte et al. (2014), Bitto and Frühwirth-Schnatter (2019), Gefang et al. (2019) Huber et al. (2020)]. The prior is characterized by, as the name implies, a shrinkage component that is similar for all parameters (i.e. global) and a shrinkage component that is specific for each parameter (i.e. local). These type of priors can

commonly be represented as scale mixtures of Gaussians where the scale has a global as well as a local component (Huber et al., 2020). The type of local shrinkage that is unique to each prior is referred to as a local shrinkage rule. The diagonalization of the covariance matrix allows us to focus on the priors that are used in a TVP context.

One of the earlier proposed priors is the Stochastic Variable Search and Selection (SVSS) by George and McCulloch (1993). It is an hierarchical model that models the coefficients as a mixture of Gaussians with different variances, the mixture component is a latent (binary) variable. The fundamental idea that this prior elicits is that in an hierarchical specification each possible subset of restrictions is treated as a distinct submodel, where priors are used to describe uncertainty across all the submodels. It is computationally difficult to calculate all the posterior probabilities. However, many of the high probability submodels can be found by stochastic search using MCMC (George et al., 2008). Research shows that it is able to retrieve the correct data-generating model quite often in several simulated examples compared to an unrestricted Bayesian VAR (George et al., 2008).

Another established idea of shrinkage and selection in a regression context is the Least Absolute Shrinkage and Selection Operator or Lasso proposed by Tibshirani (1996). It is an extension of the objective function that is minimized in ordinary least squares. The term, referred to as the "penalty", that is included, is the sum of the absolute values of the coefficients multiplied by a constant. Therefore, Lasso is said to estimate the coefficients through an L_1-constrained least squares (George et al., 2008). This objective function results in a smooth, continuous, shrinkage of the coefficients, similar to Ridge regression (or L_2-constrained least squares). However, in contrast to Ridge regression, it forces certain coefficients to be exactly zero (thus it induces selection). As Tibshirani (1996) noted, Lasso can also be viewed from a Bayesian perspective. The Lasso estimate is retrieved as the posterior mode under independent double-exponential (or Laplace) priors for the coefficients. Based on this insight several Lasso-like estimators were derived [George et al. (2008), De Mol et al. (2008)] that exhibit similar behavior as the traditional Lasso (e.g. similar shrinkage paths). However, as is traditional in the case of regression, one considers a quadratic-loss function, the optimal estimator in this case is the posterior mean (Greenberg, 2012). This results in the (inconvenient) fact that these estimators do not exhibit the well-known selection property of the traditional Lasso.

Another prior that can be expressed as a scale mixture of Gaussians is the Horseshoe prior. It is first proposed by Carvalho et al. (2010) and has demonstrated to be quite flexible, as it performs well across a multitude of situations. This is due to several useful properties that are not commonly present in other local shrinkage rules, such as Lasso or a Jeffrey's based prior. In a sparse setting, the sampling density converges at a super-efficient rate to the true values. While, in a non-sparse setting it is able to filter the noise and leave the signal unshrunk (Carvalho et al., 2010).

The main difference between each of the priors is the manner in which each prior defines its local shrinkage rule. Each prior has its theoretical advantages or disadvantages, although all have been used successfully in a simulation or empirical exercise [Gefang et al. (2019), Huber et al. (2020)].

In essence, there is a dichotomy between the priors that are used in this thesis. Even if, all of them are, strictly speaking, shrinkage priors, the SVSS shows a more selection like behaviour as it is able to set the local scale parameter for a coefficient to near zero, which results in the fact that the mean posterior estimate of this coefficient is essentially zero.

3 Methodology

3.1 TVP-VAR

A TVP-VAR can be formulated in the following state-space specification (Koop & Korobilis, 2010):

$$\begin{aligned} \boldsymbol{y}_t &= (I_M \otimes \boldsymbol{X}_t')\boldsymbol{\beta}_t + \varepsilon_t; \quad \varepsilon_t \sim N_M(\boldsymbol{0}, \boldsymbol{\Sigma}_t) \ , \\ \boldsymbol{\beta}_{t+1} &= \boldsymbol{\beta}_t + \eta_t; \qquad\qquad\quad \eta_t \sim N_k(\boldsymbol{0}, \boldsymbol{Q}_t) \ . \end{aligned} \tag{1}$$

This dynamic system consists of two relationships: the equation concerning $\boldsymbol{y}_t$ is the observation equation and the equation concerning $\boldsymbol{\beta}_{t+1}$ is the state equation. The state equation is unobserved and behaves according to a first-order Markov process. $\boldsymbol{y}_t$ contains the observations of M dependent variables at time t, hence it is $M \times 1$. $\boldsymbol{X}_t$ contains all the lagged observations of the M variables and it includes a constant at time t, hence it is $(M \times p + 1) \times 1$, where p is the number of lags. $\boldsymbol{\beta}_t$ is the corresponding vector of coefficients, and is of dimension $k \times 1$, where $k = M(M \times p + 1)$. ε_t and η_s are independent from one another for all t and s (thus past, future and present). This is a heteroskedastic specification as $\boldsymbol{\Sigma}_t$ and $\boldsymbol{Q}_t$ are time-varying and need not be diagonal. The total number of coefficients is $k \times T$, this quickly exceeds the total number of observations $M \times T$ if the number of variables or lags is increased.

Furthermore, this is a Gaussian linear state-space model (SSM). Standard techniques, such as the Kalman filter and smoother, are relevant in estimation. Well-known results exists to estimate this model in a Bayesian context.

There is ample empirical evidence that allowing the $\boldsymbol{\Sigma}$ to vary over time is beneficial for the predictive capability of a model [Koop and Korobilis (2013), Giannone et al. (2015), Chan and Eisenstat (2018)]. However, there are certain specifications, such as stochastic volatility (Durbin & Koopman, 2012), that might result in a non-linear Gaussian state-space model. In this case $\boldsymbol{y}_t$ is not linear in the states ($\boldsymbol{\beta}_t$) and additional parameters ($\boldsymbol{\Sigma}_t$). Estimating a non-linear model is in itself a difficult endeavour, as one has to opt for a linear approximation (Koop & Korobilis, 2018), an alternative to the Kalman filter (e.g. the particle filter (Koop & Korobilis, 2010)) or one could resort to other alternative specifications that maintain linearity for $\boldsymbol{\Sigma}_t$, such as forgetting factors (Koop & Korobilis, 2013).

3.1.1 Cholesky decomposition of $\boldsymbol{\Sigma}_t$ for equation-by-equation estimation

As mentioned in the literature review, although the TVP-VAR is in favour among academics and practitioners it comes at the cost of increased computational complexity. The number of

parameters to be estimated in the model is a function of the number of series and lags included (quadratic in the number of series and linear in the number of lags) in a TVP-VAR. This in turn (greatly) increases computation as the most computationally intensive operation, that is common in (Bayesian) estimation of a state-space model, the matrix inversion in the Kalman filter. It has to be computed for each timestep t and has a complexity of $\mathcal{O}(k^3)$[3]. Thus, the number of operations required to invert the matrix grows at a cubic rate.

One unrealistic assumption that one could make is that $\boldsymbol{\Sigma}_t$ is diagonal then it is possible to estimate the TVP-VAR on an equation-by-equation basis. Although, this reduces the computational complexity as $\boldsymbol{\beta}_t$ can be separated into M independent blocks (as $\boldsymbol{Q}_t$ is assumed to be diagonal). The assumption defeats the purpose of using a TVP-VAR. One could resort to M separate TVP-ARs, with the lags of the other variables as exogenous regressors, in this instance.

It is however possible to estimate a TVP-VAR on an equation-by-equation basis while still allowing for an unconstrained covariance matrix $\boldsymbol{\Sigma}_t$. This is achieved by augmenting the ith equation with (i - 1) contemporaneous values of $\boldsymbol{y}_t$ (Huber et al., 2020) (or a similar approach is augmenting the ith equation with the residuals [Koop et al. (2019), Carriero et al. (2019)]). Estimating on an equation-by-equation basis is based on the Cholesky decomposition of $\boldsymbol{\Sigma}_t = \boldsymbol{A}_t \boldsymbol{D}_t \boldsymbol{A}_t'$[4]. Where $\boldsymbol{D}_t$ is a diagonal matrix and $\boldsymbol{A}_t$ is a lower unit triangular matrix. $\boldsymbol{A}_t \boldsymbol{D}_t^{1/2}$ is the lower triangular Cholesky decomposition of $\boldsymbol{\Sigma}_t$. The Cholesky decomposition is used in several different manners in the context of VARs, in this instance, its sole purpose is to ease the process of estimation. This allows for draws from the variational posterior for the states. The use of the Cholesky decomposition in this instance should not be confused with the use for the identification of shocks, as the uniqueness of the decomposition depends on the ordering of the variables in $\boldsymbol{y}_t$, more on this in the next subsection. Following the idea of (Huber et al., 2020), transforming the (reduced form) observation equation (1) by the lower unit-triangular $\boldsymbol{A}_t^{-1}$, allows for the estimation of M independent equations that are augmented with the contemporaneous values of $\boldsymbol{y}_t$. This transformation has true separability as opposed to the transformation in Carriero et al. (2019) the estimation of the ith equation still relies on the residuals of the previous i-1 equations. To be more specific about Huber et al. (2020), let $\alpha_{(i,j),t}$ be the elements of $\boldsymbol{A}_t^{-1}$ for $i = 2, \ldots, M$ and $j = 1, \ldots, i-1$.

$$\begin{aligned} \boldsymbol{A}_t^{-1}\boldsymbol{y}_t &= \boldsymbol{A}_t^{-1}(I_M \otimes \boldsymbol{X}_t')\boldsymbol{\beta}_t + \boldsymbol{A}_t^{-1}\varepsilon_t; & \varepsilon_t &\sim N(0, \boldsymbol{\Sigma_t}) \\ &= \boldsymbol{Z}_t\boldsymbol{\beta}_t + \upsilon_t; & \upsilon_t &\sim N(0, \boldsymbol{D}_t)\ . \end{aligned} \tag{2}$$

[3]This is dependent on the algorithm that is used to compute the inverse, in this case it is based on Gauss-Jordan elimination

[4]There are several possible decompositions, the decomposition used here is similar to the decomposition of Primiceri (2005) and allows for more flexibility, $\boldsymbol{A}_t$ is time-varying, while Cogley and Sargent (2005) keep $\boldsymbol{A}$ time-invariant.

This can be reformulated in terms of $\boldsymbol{y}_t$ and if we separate it out in independent equations, each equation would have the following structure:

$$\begin{aligned}
&y_{1,t} = z_{1,t}\beta_{1,t} + \upsilon_{1,t} \\
&y_{2,t} + \alpha_{(2,1),t}y_{1,t} = z_{2,t}\beta_{2,t} + \upsilon_{2,t} \\
&y_{3,t} + \alpha_{(3,1),t}y_{1,t} + \alpha_{(3,2),t}y_{2,t} = z_{3,t}\beta_{3,t} + \upsilon_{3,t} \\
&\quad\vdots \\
&y_{M,t} + \alpha_{(M,1),t}y_{1,t} \cdots + \alpha_{(M,M-1),t}y_{M-1,t} = z_{M,t}\beta_{M,t} + \upsilon_{M,t} \ ,
\end{aligned} \tag{3}$$

where $\upsilon_{i,t} \sim N(0, \boldsymbol{D}_{(i,i),t})$ for $i = 1 \cdots M$. Bringing all the $\alpha_{i,j}$ terms to the left, one can see that a regular TVP with lags and exogenous variables shows up. It can be formulated in a more general way for $i > 1$:

$$y_{i,t} = z_{i,t}\beta_{i,t} + \sum_{j=1}^{M-1} -\alpha_{(i,j),t} \times y_{j,t} + \upsilon_{i,t}; \quad \upsilon_{i,t} \sim N(0, \boldsymbol{D}_{(i,i),t}) \ . \tag{4}$$

To recover the reduced form coefficients and the full covariance matrix of equation (1). One has to reconstruct the matrix $\boldsymbol{A}_t^{-1}$ after the process of estimation and multiply the estimated (structural) coefficients and estimated variances (i.e. $\boldsymbol{A}_t\boldsymbol{\beta}_t$ and $\boldsymbol{A}_t\boldsymbol{D}_t\boldsymbol{A}_t'$). Keeping in mind that the sign still has to be altered, as one estimates $-\alpha_{(i,j),t}$. These reduced form coefficients can subsequently be used for forecasting. Thus, to incorporate it into a complete state-space specification, for the ith variable (for $i > 1$) in the set of M variables, the model to be estimated is:

$$\begin{aligned}
y_{i,t} &= \boldsymbol{Z}_t\boldsymbol{\beta}_{i,t} + \sum_{j=1}^{i-1} -\alpha_{(i,j),t} \times y_{j,t} + \upsilon_{i,t} \\
&= \tilde{\boldsymbol{Z}}_{i,t}\tilde{\boldsymbol{\beta}}_{i,t} + \upsilon_{i,t}; && \upsilon_{i,t} \sim N(0, \boldsymbol{D}_{(i,i),t}) \ , \\
\tilde{\boldsymbol{\beta}}_{i,t} &= \tilde{\boldsymbol{\beta}}_{i,t} + \omega_{i,t}; && \omega_{i,t} \sim N(0, \tilde{\boldsymbol{Q}}_{i,t}) \ ,
\end{aligned} \tag{5}$$

where, $\alpha_{(i,j),t}$ are the elements A_t^{-1} for $i = 2, \ldots, M$ and $j = 1, \ldots, i-1$, $\tilde{\boldsymbol{Z}}_{i,t} = [\boldsymbol{Z}_t', (y_{1,t}, \ldots, y_{i-1,t})]'$, $\tilde{\boldsymbol{\beta}}_{i,t} = [\boldsymbol{\beta}_{i,t}', (-\alpha_{(i,1)}, \ldots, -\alpha_{(i,i-1)})]'$ and $\omega_{i,t} = [\eta_{i,t}^{'}, (\varkappa_{1,t}, \ldots, \varkappa_{i-1,t})]$, where $\varkappa_{i,t} \sim N(0, h_{i,t})$ for $i = 2, \ldots, i-1$ are the error terms of the additional states. Note that, $\text{diag}(\boldsymbol{D}_t) = (\sigma_{1,t}^2, \ldots, \sigma_{M,t})'$. Thus $\boldsymbol{D}_{(i,i),t}$ and $\sigma_{i,t}^2$ will be used interchangeably. The non-zero parameters of $\boldsymbol{A}_t^{-1}$ are incorporated in the same manner as the coefficients of the lags (or exogenous components). The benefits are that these parameters do not need additional consideration regarding the choice of prior or estimation procedure. These can be treated as regular coefficients. Further, note that the diagonal elements in $\boldsymbol{D}_t$ remain unconstrained. Additional assumptions or restrictions with respect to these elements are note necessary. Hence, any model for the observation variance can still be considered (e.g. log-volatility).

For the state equation, the sole difference is the number of states that have to be modeled. As

the initial model is augmented, these augmented components are also included in the state vector and thus follow the similar Markov process behavior. Therefore, the state covariance matrix also needs a slight alteration. The number of states increase, naturally, the number of variances that need to be estimated also increases. Thus, $\tilde{\boldsymbol{Q}}_{i,t} = diag([q'_{i,t}, (h_{1,t}, \ldots, h_{i-1,t})]')$, it is the expanded diagonal covariance matrix, as is clear from the definition of $\omega_{i,t}$, it includes the original variances and the variances of the additional states. The assumption of diagonality remains. In the end, the number of states, that have to be estimated at once, is reduced from $M(M \times p+1)$ for the complete system to $(M \times p + 1) + (i - 1)$ for the ith equation.

3.1.1.1 Ordering of the variables

One drawback of the Cholesky decomposition approach (or similar decompositions of the observation covariance matrix) is that the order of the variables matters, this is often problematic in circumstances where inference is of importance (e.g. impulse response analysis). For example, if one opts for an Inverse-Wishart prior for $\boldsymbol{\Sigma}_t$ in the estimation of a complete system, this prior is invariant to the ordering of the variables. However, if one opts for a diagonalization of $\boldsymbol{\Sigma}_t$, the implied prior for $\boldsymbol{\Sigma}_t$ is a combination of the priors on the elements in $\boldsymbol{A}_t \boldsymbol{D}_t \boldsymbol{A}_t'$ (elements of $\boldsymbol{A}_t^{-1}$ are treated as "regular" coefficients and the diagonal elements of $\boldsymbol{D_t}$ are treated as "regular" variances), although the decomposition need not be unique (as $\boldsymbol{\Sigma}_t$ might be positive-semidefinite), it is unlikely that the variances in the original $\boldsymbol{\Sigma}_t$ have the same linear combination in two different decompositions, thus it is not invariant to the ordering of the variables. This problem of different implied priors for $\boldsymbol{\Sigma}_t$ is referred to as the "prior ordering problem" (Carriero et al., 2019). As the focus of this thesis is on predictive results, not inference, in both simulation and empirical exercises it has shown to not have a detrimental effect on the predictive densities (Carriero et al., 2019).

3.1.2 Bayesian inference in a state-space model

A short introduction to a state-space model in a Bayesian context is necessary. Due to the equation-by-equation estimation discussed in the previous section, this section is restricted to the univariate case of a state-space model. The factorization of the full posterior for the parameters, unobserved coefficients and observed data of the model in equation (5) is the following[5]:

$$\begin{aligned} p(\tilde{\boldsymbol{\beta}}_{i,0:T}, \sigma^2_{i,1:T}, \tilde{\boldsymbol{Q}}_{i,1:T} | \boldsymbol{y}_{i,1:T}) &= p(\boldsymbol{y}_{i,1:T}, \tilde{\boldsymbol{\beta}}_{i,0:T}, \sigma^2_{i,t}, \tilde{\boldsymbol{Q}}_{i,1:T}) / p(\boldsymbol{y}_{i,1:T}) \\ &\propto p(\tilde{\boldsymbol{\beta}}_{i,0} | \underline{\mu}_0, \underline{\tilde{Q}}_0) \prod_{t=1}^{T} p(\boldsymbol{\beta}_{i,t} | \tilde{\boldsymbol{\beta}}_{i,t-1}, \tilde{\boldsymbol{Q}}_{i,t}) p(\boldsymbol{y}_{i,t} | \tilde{\boldsymbol{X}}'_{i,t} \tilde{\boldsymbol{\beta}}_{i,t}, \sigma^2_{i,t}) \\ &\times p(\tilde{\boldsymbol{Q}}_{i,t}) p(\sigma^2_{i,t}) \ , \end{aligned} \tag{6}$$

where $p(\boldsymbol{\beta}_{i,0}) \sim N(\underline{\mu}_0, \underline{Q}_0)$. Estimation of this model traditionally requires a simulation approach such as importance sampling or MCMC. One approach that is often used is based on Gibbs sampling. The non-trivial part in a Gibbs-based scheme is the conditional posterior for the states

[5]The complete derivation of the full joint probability to which the full posterior is proportional is in Appendix A.

$p(\tilde{\boldsymbol{\beta}}_{i,1:T}|\boldsymbol{y}_{i,t}, \sigma^2_{i,1:T}, \tilde{\boldsymbol{Q}}_{i,1:T})$, as it requires sampling from a high-dimensional Gaussian. To reduce this computational burden several algorithms exist, such as a Kalman Filter and Smoother (KFS) based algorithm, which de Jong and Shephard (1995) show is sufficient to draw efficiently from the conditional posterior (Koop, 2003). Once the states $\tilde{\boldsymbol{\beta}}_{i,1:T}$ are known, sampling for the other conditional posteriors, $p(\sigma^2_{i,1:T}|\boldsymbol{y}_{i,t}, \tilde{\boldsymbol{\beta}}_{i,1:T}, \tilde{\boldsymbol{Q}}_{i,1:T})$ and $p(\tilde{\boldsymbol{Q}}_{i,1:T}|\boldsymbol{y}_{i,t}, \tilde{\boldsymbol{\beta}}_{i,1:T}, \sigma^2_{i,1:T})$, is relatively simple as the model reduces to a linear regression, well-known results exist in this case (given that the priors for the variances are proper and conjugate).

3.2 Priors

In this section, the specification of each of the priors is formulated. The notation with each prior is kept as similar as possible to the original literature on which the specifications in this paper are based [Koop and Korobilis (2020), Belmonte et al. (2014), Huber et al. (2020)]. Thus, it is possible to refer back to the original work, this is especially useful if one is interested in the relevant posterior derivations. Furthermore, up until this point, there was an explicit distinction between the notation for the complete TVP-BVAR and the equation-by-equation basis TVP-BVAR. This was necessary to clearly state the differences between the two. However, this comes at a cost of convoluted notation. As of this point all the results are with respect to a TVP, the subscripts are left out of the results and derivations for the remaining part of section 3. This is possible, as each variable in the equation-by-equation basis will be estimated in an homogeneous manner. There are no differences between the variables for the priors, hyperparameters and variational posteriors.

3.2.1 Stochastic Variable Search and Selection (SVSS)

Koop and Korobilis (2020) extend the static SVSS of George and McCulloch (1993) to a dynamic SVSS as it is able to vary the selection over time. This dynamic SVSS specification is specified for a TVP. In a TVP the coefficients are (often, not necessarily) with respect to exogenous variables, in the TVP-VAR specification of equation (1) the coefficients are with respect to the lags of the variables itself. The dynamic SVSS is the following specification:

$$\begin{aligned}
\tilde{\beta}_{j,t}|\gamma_{j,t}, \tau^2_{j,t} &\sim (1-\gamma_{j,t})N(0, \underline{c} \times \tau^2_{j,t}) + \gamma_{j,t}N(0, \tau^2_{j,t}) \ , \\
\gamma_{j,t}|\pi_t &\sim Bernoulli(\pi_{0,t}) \ , \\
\frac{1}{\tau^2_{j,t}} &\sim Gamma(\underline{g}_0^{svss}, \underline{h}_0^{svss}) \ \forall_{j,t} \ , \\
\pi_{0,t} &\sim Beta(1,1) \ ,
\end{aligned} \tag{7}$$

where $\underline{c}$ is one of the hyperparameters that has to be set. However, if one would like to adhere to selection as much as possible, it should be set to near zero, such that the mass of the posterior for the specific coefficient is centered around zero. In this instance, it is set to 10^{-4}. The other parameters each have their own role. $\gamma_{j,t}$ is the posterior selection of a coefficient, $\pi_{0,t}$ is the posterior inclusion probability, $\tau^2_{j,t}$ is the level of shrinkage on a selected coefficient and this shrinkage is set by $\underline{g}_0^{svss}$

and $\underline{h}_0^{svss}$.

3.2.2 Least Absolute Shrinkage and Selection Operator (Lasso)

The (Bayesian) Lasso inspired prior, as proposed by Belmonte et al. (2014), that emulates Lasso shrinkage. This is an hierarchical mixture of normal priors with an exponential mixing density. Which is the following specification:

$$\begin{aligned} \tilde{\beta}_{j,t}|\iota_{j,t}^2 &\sim N(0, \iota_{j,t}^2) \ , \\ \iota_{j,t}^2|\psi_t^2 &\sim Exp\Big(\frac{\psi_t^2}{2}\Big) \ , \\ \psi_t^2 &\sim Gamma(\underline{g}_0^{lasso}, \underline{h}_0^{lasso}) \ \forall_t \ , \end{aligned} \tag{8}$$

where the parameter $\iota_{j,t}^2$ is the local shrinkage on the coefficient. This shrinkage is in turn affected by the global parameter ψ_t^2 that determines the rate of the exponential distribution. ψ_t^2 is influenced by the $\underline{g}_0^{lasso}$ (shape) and $\underline{h}_0^{lasso}$ (scale) parameters of the Gamma distribution.

3.2.3 Horseshoe

The original specification is quite difficult to implement in an efficient manner in a TVP context. Therefore, the following prior specification, based on the auxilliary variable specification as suggested by Huber et al. (2020), is used:

$$\begin{aligned} \tilde{\beta}_{j,t}|\lambda_t, \phi_{j,t} &\sim N(0, \lambda_t \phi_{j,t}) \ , \\ \lambda_t|\varphi_t &\sim IG(1/2, 1/\varphi_t) \ , \\ \phi_{j,t}|\nu_{j,t} &\sim IG(1/2, 1/\nu_{j,t}) \ , \\ \nu_{j,t}, \varphi_t &\sim IG(\underline{g}_0^{horseshoe}, \underline{h}_0^{horseshoe}) \ \forall_{j,t} \ , \end{aligned} \tag{9}$$

where λ_t is the global shrinkage parameter and ϕ_t the local shrinkage component. These are in turn affected by the global (φ_t) and local parameters ($\nu_{j,t}$) that control their distributions. These hyperparameters are all given the exact same distribution and the $\underline{g}_0^{horseshoe}$ (shape) and $\underline{h}_0^{horseshoe}$ (scale). This prior is often advocated as a more objective prior as the hyperparameters at the highest level $\underline{g}_0^{horseshoe}$ and $\underline{h}_0^{horseshoe}$ are often set to 1/2 and 1 respectively (a non-informative choice) or are chosen based on a Jeffrey's like approach. However, as the focus of this thesis is on the sensitivity of the choices of those specific hyperparameters, it is not sensible to set these to a data-informed or predetermined value.

3.3 Variational Inference (VI)

MCMC is the traditional estimation technique used in Bayesian inference. Even though it is difficult to gauge convergence, there are several tests and best practices, to get an idea whether convergence is attained using MCMC. This lack of certainty with respect to true convergence might give the

researcher a false sense of convergence, or pseudo-convergence (Geyer, 2011). Another drawback of MCMC is the fact that in more complex models it becomes computationally infeasible. Therefore, instead of an exact (intractable) posterior, it is necessary to resort to an approximation of the (intractable) posterior. Variational Inference (VI) is one of these deterministic approximation techniques, that assumes that the posterior can be factorized in a certain manner (e.g. mean-field assumption) and has an analytical, thus tractable, approximation to the intractable posterior (Bishop, 2006). It has similarities to Expectation-Maximization (EM), as EM can be seen as a special case of VI, and Gibbs sampling (Blei et al., 2017). In the case of EM, it has the E-step and the M-step, alternating between estimating a parameter ϕ and averaging over other parameters γ. In a VI context, the parameters are partitioned in two parts, ϕ and γ. The approximating distribution for ϕ is required to be a point mass. Thus, $q(\phi)$ is equal to a point estimate of ϕ and the approximating distribution for γ is unconstrained. In other words $q(\gamma) = q(\gamma|\phi, y)$ is conditional on the most recent update of ϕ (Gelman et al., 2013).[6]

The following explanation is based on the textbook explanation of Bishop (2006). $\boldsymbol{Y}$ is the data, $\boldsymbol{Z} = (\tilde{\boldsymbol{\beta}}, \sigma^2, \tilde{\boldsymbol{Q}}, \theta)$ are the latent parameters, where θ are the remaining parameters in model (5), such as $\boldsymbol{\gamma}$ in the SVSS prior. The log marginal probability, $\ln p(\boldsymbol{Y})$, can be decomposed into the sum of two functionals, the lower bound $\mathcal{L}(q)$, or sometimes referred to as the evidence lower bound (ELBO)[7], and the forward Kullback-Leibler divergence $KL(q||p)$[8]. $q(\boldsymbol{Z})$ is the tractable approximation to the intractable posterior. $p(\boldsymbol{Z}|\boldsymbol{Y})$ is the intractable posterior. $p(\boldsymbol{Y}, \boldsymbol{Z})$ is the joint probability of the data and the latent variables.

$$\ln p(\boldsymbol{Y}) = \mathcal{L}(q) + KL(q||p) \ , \tag{10}$$

where

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Y}, \boldsymbol{Z})}{q(\boldsymbol{Z})} \right\} d\boldsymbol{Z} \ , \tag{11}$$

$$KL(q||p) = -\int q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{Y})}{q(\boldsymbol{Z})} \right\} d\boldsymbol{Z} \ . \tag{12}$$

As the posterior is intractable, it is not possible to minimize the KL divergence in a direct manner with respect to $q(\boldsymbol{Z})$. However, it is possible to maximize the lower bound $\mathcal{L}(q)$ with respect to $q(\boldsymbol{Z})$. This is equivalent to minimizing the KL divergence[9]. There are no assumptions with respect to the shape or form of the approximation distribution $q(\boldsymbol{Z})$. To be certain that it is a feasible approximation distribution, it is necessary to restrict the family of possible distributions,

[6] These symbols are just for explanation, they have no relation to the other equations that are presented in the thesis

[7] The marginal probability $p(\boldsymbol{Y})$ is in the machine learning literature referred to as the "evidence", as $KL(q||p) \geq 0$, the $\mathcal{L}(q)$ sets a lower bound on $p(\boldsymbol{Y})$ and $KL(q||p) = 0$ if and only if $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{Y})$

[8] It is a divergence as it is not symmetric $KL(q||p) \neq KL(p||q)$

[9] As $p(\boldsymbol{Y})$ does not depend on $q(\boldsymbol{Z})$, it is just a constant, it does affect the level but not the optimal solution of maximizing $\mathcal{L}(q)$ w.r.t. $q(\boldsymbol{Z})$.

such that these are tractable. At the same time, the family should not be too restrictive as the approximation of the true posterior will suffer (Bishop, 2006).

One often used assumption that restricts the family of distributions of $q(\boldsymbol{Z})$ is the assumption that the approximation distribution can be factorized into M independent, distinct, groups.

$$q(\boldsymbol{Z}) = \prod_{i=1}^{M} q_i(\boldsymbol{Z}_i) \ . \tag{13}$$

This is referred to as the mean-field assumption (MF). In this setting, VI with the MF assumption (MFVI), maximizing the $\mathcal{L}(q)$ with respect to all the distributions that are of the family specified by (13) is a variational optimization of $\mathcal{L}(q)$ with respect to all factorized distributions $q_i(\boldsymbol{Z}_i)$. In essence, it is a deterministic iterative optimization procedure, where the optimal solution for each factorized distribution $q_i(\boldsymbol{Z}_i)$[10] is the following:

$$\ln q_j^{\star}(\boldsymbol{Z}_j) = \mathop{\mathbb{E}}_{i \neq j} \big[\ln p(\boldsymbol{Y}, \boldsymbol{Z}) \big] + constant \ . \tag{14}$$

It states that the optimal solution to the natural logarithm of $q_j(\boldsymbol{Z})$ is the expectation of the natural logarithm of the joint distribution with respect to all the other factors $\{q_i\}$, except where $i \neq j$. As it is dependent on the expectations with respect to the other factors, which are themselves not yet optimal, an iterative procedure is required. In the iterative procedure each subsequent optimal expectation is updated and slotted into the next iteration. These optima are all analytical expressions. Furthermore, the VB algorithm monotonically decreases the KL divergence with each iteration until a predetermined threshold of convergence (decrease between subsequent iterations in KL divergence is smaller than a set threshold) is reached (Beal, 2003).

Equation (14) is also where the similarities with Gibbs sampling arise. As an example, each parameter set in $\boldsymbol{Z}$ is often a block in a Gibbs sampler scheme (θ is often further separated in blocks). As mentioned in section 3.1.2, to sample for each block the conditional posterior for each parameter set has to be analytically derived. In VI, for each parameter set the derivation is similar, as the joint posterior reduces to a conditional posterior, the essential difference is that one has to take the expectation with respect to the other parameters in $\boldsymbol{Z}$ to completely define the posterior density. Therefore, it can be seen as an average density instead of a conditional density for Gibbs sampling in the typical sense (Gelman et al., 2013).

Although (14) is not the exact density of the optimal solution, it is not always necessary to derive it completely. It will be visible later on that the density, and in turn the normalizing constant, can be derived by inspection, as is common in Bayesian statistics.

[10] The exact derivation is also in the book of Bishop (2006) and requires variational calculus to determine the *functional derivative* (Feynman et al., 1984) that is required to maximize the lower bound $\mathcal{L}(q)$

3.3.1 Variational Inference (VI) in a TVP

To clarify the complete process of estimating the model, which involves the use of several techniques that are intertwined, the complete pseudo-code for the algorithm is in appendix B.

The novel addition of Koop and Korobilis (2020), is that their factorization of the posterior approximation distribution makes it possible to rely on well-known results. They model the dynamic SVSS prior as a prior for latent data, where $p(\tilde{\beta}_{j,t}|\gamma_{j,t})$ can be written as $p(z_{j,t}|\tilde{\beta}_{j,t}, \upsilon_{j,t}) \equiv N(\tilde{\beta}_{j,t}, \upsilon_{j,t})$ for the latent observations $z_{j,t} = 0, \forall j, t$ and where $\upsilon_{j,t} = (1-\gamma_{j,t})^2\underline{c} \times \tau_{j,t}^2 + \gamma_{j,t}^2\tau_{j,t}^2$. V_t is a $p \times p$ diagonal matrix with the elements of $\upsilon_{j,t}$ following Wang et al. (2016). This way of modeling a prior generalizes to the other global-local priors mentioned in section 3.2, the form of $\upsilon_{j,t}$ changes depending on the prior.

The posterior for this model where the volatility is allowed to vary with time and the latent data approach for modeling a dynamic prior is the following:

$$\begin{aligned} p(\tilde{\boldsymbol{\beta}}_{1:T}, \tilde{\boldsymbol{Q}}_{1:T}, V_{1:T}, \sigma_{1:T}^2|y_{1:T}, z_{1:T}) \propto \prod_{t=1}^{T} p(\tilde{\boldsymbol{\beta}}_{\boldsymbol{t}}|\tilde{\boldsymbol{\beta}}_{t-1}, \tilde{\boldsymbol{Q}}_t)p(y_t|\tilde{\boldsymbol{\beta}}_t, \sigma_t^2)p(z_t|\tilde{\boldsymbol{\beta}}_t, V_t) \\ \times p(\gamma_t)p(\tau_t^2)p(\tilde{\boldsymbol{Q}}_t)p(\sigma_t^2) \ . \end{aligned} \tag{15}$$

This is the posterior that one would like to approximate with $q(\tilde{\boldsymbol{\beta}}_{1:T}, \tilde{\boldsymbol{Q}}_{1:T}, V_{1:T}, \sigma_{1:T}^2)$. A naive mean-field approximation is to assume independence between each set of parameters at the same moment in time as well as over time. This naive approximation is too unrealistic, as it is essential that elements are dependent over time, since it is a model that specifically models time. Koop and Korobilis (2020) propose a factorization that takes into account the dependence over time, which is the following factorization:

$$p(\tilde{\boldsymbol{\beta}}_{1:T}, \tilde{\boldsymbol{Q}}_{1:T}, V_{1:T}, \sigma_{1:T}^2) = q(\tilde{\boldsymbol{\beta}}_{1:T}) \prod_{t=1}^{T} q(\sigma_t^2) \prod_{j=1}^{p} q(\upsilon_{j,t})q(q_{j,t}) \ , \tag{16}$$

where, $q_{j,t}$ are elements of $\tilde{\boldsymbol{Q}}_t$. Not to be confused with the approximation distribution $q(\cdot)$. To estimate a TVP with one of the aforementioned priors in a VB manner they rely on the insight that the two conditional priors (the state equation has an implicit prior due to the Markov characteristic) and a prior such as SVSS, can be combined by rewriting the state equation of (5) in the following manner[11]:

$$\tilde{\boldsymbol{\beta}}_t = \tilde{\boldsymbol{F}}_t\tilde{\boldsymbol{\beta}}_{t-1} + \tilde{\eta}_t; \quad \tilde{\eta}_t \sim N(\boldsymbol{0}, \tilde{\boldsymbol{W}}_t) \ , \tag{17}$$

where $\tilde{\boldsymbol{F}}_t = \tilde{\boldsymbol{W}}_t \times \mathbb{E}(\boldsymbol{W}_t)^{-1}$, with $\tilde{\boldsymbol{W}}_t = (\mathbb{E}(\boldsymbol{W}_t)^{-1} + \mathbb{E}(\boldsymbol{V}_t)^{-1})^{-1}$, where $\boldsymbol{W}_t = diag(q_{1,t}, \ldots, q_{k,t})$ and $\boldsymbol{V}_t = diag(\upsilon_{1,t}, \ldots, \upsilon_{k,t})$. The expectation operators are with respect to $q(\tilde{\boldsymbol{\beta}}_t|y_{1:t})$. The MFVI approach described here is the most simple variant of VI. There are some limitations, such as that the likelihood has to be part of the exponential family and the priors have to be conditionally

[11] Details on the exact derivations for the transformed state equation are in the Technical Appendix of Koop and Korobilis (2018)

conjugate, this guarantees analytical solutions for each approximation distribution. The earlier mentioned priors in section 3.2 meet this conditionally conjugate requirement.

3.3.1.1 Allowing for time-varying volatility

The common approach to modeling time-varying volatility is to specify a model for the volatility itself. One specific approach is referred to as stochastic volatility. One of the issues that stochastic volatility induces, as mentioned earlier, is that the model is not linear in the parameters and conjugacy with respect to the likelihood is not preserved. This, in turn, introduces complexities in the derivation of the variational posteriors.

One approach that one can opt for is the use of forgetting factors or as it is sometimes referred to variance discounting, this does not require the specification for a separate model for the time-varying volatility. It is an approach that was already mentioned in section 2 to reduce the computational burden of estimating the state covariance matrix as the number of variables increases. Variance discounting can also be used to simplify the estimation of the volatility dynamics (West & Harrison, 1997). In this thesis the same approach is utilised as in Koop and Korobilis (2020). Define $\chi_t = \frac{1}{\sigma_t^2}$ to be the inverse variance (precision) and assume that the time $t-1$ posterior and the time t prior of χ have the following forms:

$$\begin{aligned} \chi_{t-1}|y_{1:t-1} &\sim Gamma(a_{t-1}, b_{t-1}) \ , \\ \chi_t|y_{1:t-1} &\sim Gamma(\delta a_{t-1}, \delta b_{t-1}) \ , \end{aligned} \tag{18}$$

where δ is the variance discounting factor and is constrained to be $0 < \delta < 1$. Furthermore, the hyperparameters in these priors are recursive and are subject to a choice of $\underline{a}_0$ and $\underline{b}_0$. The discounting factor can be seen as an exponential discounting factor and thus is a hyperparameter that indicates whether one gives more weight to recent versus older observations. The variational posterior that is necessary to update χ_t will be shown in the next section. Moreover, in line with West and Harrison (1997), Koop and Korobilis (2020) utilise a smoothing approach on the filtered estimates to obtain a precise estimate of σ_t^2. In the VI-based TVP-BVARs that we estimate this prior is only set for the elements in the diagonal matrix $\boldsymbol{D}_t$, the other variances in the reduced form are treated as regular coefficients.

As Koop and Korobilis (2020) allude to, in the case that one is interested in forecasting and not in causality, it is beneficial for the forecasting ability of a model to allow for variation in volatility over time. Furthermore, it does not seem to differ severely in what manner one allows for this time-varying volatility. Either specifying a model or using a variance discounting approach seems to result in the same forecasting ability. Therefore, in this thesis, the time-varying volatility is modeled using the above described approach of variance discounting.

3.3.2 Variational posteriors

The VI algorithm requires the update of the approximation densities at each iteration of the algorithm. As mentioned earlier, a given approximation density of any parameter $\boldsymbol{Z}_j$ relies on the other parameters $\boldsymbol{Z}_{i\neq j}$. These approximation densities, such as $q(\tilde{\boldsymbol{\beta}}_t|y_{1:t})$, are referred to as variational posteriors. The expectation of these variational posteriors is used in the iterative scheme to update the approximation densities until a set level of convergence is attained. The variational posteriors are for the measurement equation (5) and altered state equation (17).

3.3.2.1 Posterior of $\tilde{\boldsymbol{\beta}}$

This variational posterior $q(\tilde{\boldsymbol{\beta}}_t|y_{1:t})$ is non-trivial to derive, as opposed to the other variational posteriors that are simpler to derive when $\tilde{\boldsymbol{\beta}}_t$ is known. In Beal (2003) the complete treatment and derivation of a linear dynamical system, which is analogous to a state-space model, and its states is presented. Although the derivations are too involved to present here, the results are essential. $q(\tilde{\boldsymbol{\beta}}_t|y_{1:t})$ is normally distributed with mean $\boldsymbol{m}_{t|T}$ and covariance $\boldsymbol{P}_{t|T}$. These moments are given by the forward and backward passes of the Kalman filter and Rauch-Tung-Striebel smoother [Beal (2003), Koop and Korobilis (2020)]. Hence, given all the other parameters, gathered in θ, the expectation of the variational posterior for $q(\tilde{\boldsymbol{\beta}}_t|y_{1:t})$ is $\hat{\boldsymbol{\beta}}_t = \mathbb{E}_{q(\theta|y_{1:t})}(\tilde{\boldsymbol{\beta}}_t|y_{1:t}) = \boldsymbol{m}_{t|T}$.

3.3.2.2 Posterior of $\tilde{\boldsymbol{Q}}$

Following Koop and Korobilis (2020) the variational posterior of $q(q_{j,t}|y_{1:t})$ is Inverse-Gamma distributed. The expectation of this distribution is the following:

$$\hat{q}_{j,t} = \mathbb{E}_{q(\theta|y_{1:t})}(q_{j,t}|y_{1:t}) = \frac{d_{j,t}}{c_{j,t}} \ , \tag{19}$$

where, $c_{j,t} = \underline{c}_0 + 1/2$ and $d_{j,t} = \underline{d}_0 + \boldsymbol{L}_{(j,j),t}$. $\boldsymbol{L}_{(j,j),t}$ are the diagonal elements of $\boldsymbol{L}_t$. Furthermore, $\boldsymbol{L}_t = \hat{\boldsymbol{P}}_{t:t} + \hat{\boldsymbol{\beta}}_{t:t}\hat{\boldsymbol{\beta}}'_{t:t} + \big(\hat{\boldsymbol{P}}_{t-1|t-1} + \hat{\boldsymbol{\beta}}_{t-1:t-1}\hat{\boldsymbol{\beta}}'_{t-1:t-1}\big)\big(I_k - 2\tilde{\boldsymbol{F}}\big)$. $\hat{\boldsymbol{P}}_{t:t}$ is the time t filtered posterior estimate of the variance of $\tilde{\boldsymbol{\beta}}_t$ and $\hat{\boldsymbol{\beta}}_t$ is the time t filtered posterior estimate of the posterior mean of $\tilde{\boldsymbol{\beta}}_t$. These are obtained in the update step of the variational posterior of $\tilde{\boldsymbol{\beta}}_t$.

3.3.2.3 Posterior of $\boldsymbol{\sigma}^2$

The variational posterior of $q(\chi_t|y_{1:t})$ is (also) Gamma distributed and the expectation of the filtered inverse variance is:

$$\hat{\chi}_t = \mathbb{E}_{q(\theta|y_{1:t})}(\chi_t|y_{1:t}) = \frac{1/2 + \delta a_{t-1}}{\frac{1}{2}\big[(y_t - \boldsymbol{X}'_t\hat{\boldsymbol{\beta}}_t)^2 + \boldsymbol{X}'_t\boldsymbol{P}_{t:T}\boldsymbol{X}_t + \delta b_{t-1}\big]} \ . \tag{20}$$

In a similar vein as West and Harrison (1997), the filtered inverse variance is smoothed to obtain a more precise estimate. The recursive filter is $\tilde{\chi}_t = (1-\delta)\hat{\chi}_t + \delta\tilde{\chi}_{t+1}$ for $t = T-1,\dots,1$, where $\tilde{\chi}_t = \mathbb{E}_{q(\theta|y_{1:t})}(\chi_t|y_{t+1})$ and $\tilde{\chi}_T = \hat{\chi}_T$.

3.3.2.4 Posteriors of the SVSS prior

The updating steps for the SVSS prior are described here. Furthermore, the expectation $\mathbb{E}[\cdot]$ is taken with respect to the other VB posteriors, that are gathered in θ, on the right-hand side of that particular equation. As opposed to the derivations of the subsequent two priors, one has to refer to Koop and Korobilis (2020) for the details. The variational posterior for each parameter is the following:

$$\hat{\tau}_{j,t}^2 = \mathbb{E}_{q(\theta|y_{1:t})}(\tau_{j,t}^2|y_{1:t}) = (\underline{h}_0^{svss} + \hat{\beta}_{j,t}^2)/(\underline{g}_0^{svss} + 1/2) \ , \tag{21}$$

$$\hat{\gamma}_{j,t} = \mathbb{E}_{q(\theta|y_{1:t})}(\gamma_{j,t}|y_{1:t}) = \frac{N(\hat{\beta}_{j,t}|0, \hat{\tau}_{j,t}^2)\hat{\pi}_{0,t}}{N(\hat{\beta}_{j,t}|0, \hat{\tau}_{j,t}^2)\hat{\pi}_{0,t} + N(\hat{\beta}_{j,t}|0, \underline{c} \times \hat{\tau}_{j,t}^2)(1-\hat{\pi}_{0,t})} \ , \tag{22}$$

$$\hat{\upsilon}_{j,t}^{svss} = \mathbb{E}_{q(\theta|y_{1:t})}(\upsilon_{j,t}^{svss}|y_{1:t}) = (1-\hat{\gamma}_{j,t})^2\underline{c} \times \hat{\tau}_{j,t}^2 + \hat{\gamma}_{j,t}\hat{\tau}_{j,t}^2 \ , \tag{23}$$

$$\hat{\pi}_{0,t} = \mathbb{E}_{q(\theta|y_{1:t})}(\pi_{0,t}|y_{1:t}) = \left(1 + \sum_{j=1}^{k} \hat{\gamma}_{j,t}\right)/(2+k) \ , \tag{24}$$

for $t = 1, \ldots, T$ and $j = 1, \ldots, k$.

3.3.2.5 Posteriors of the Lasso prior

Similar to SVSS, the updating steps for the Horseshoe prior are presented in this section. Where, again, the expectation $\mathbb{E}[\cdot]$ is taken with respect to the other VB posteriors, that are gathered in θ, on the right-hand side of that particular equation. The derivations for these variatonal posteriors are in Appendix A.3. The variational posterior for each parameter is the following:

$$\hat{\iota}_{j,t}^2 = \mathbb{E}_{q(\theta|y_{1:t})}(\iota_{j,t}^2|y_{1:t}) = \left(\sqrt{\frac{\hat{\psi}_t^2}{\hat{\beta}_{j,t}^2 + \hat{P}_{jj,t}}}\right)^{-1} \ , \tag{25}$$

$$\hat{\psi}_t^2 = \mathbb{E}_{q(\theta|y_{1:t})}(\psi_t^2|y_{1:t}) = \frac{k + \underline{g}_0^{lasso}}{\left(\sum_{j=1}^{k} \hat{\iota}_{t,j}^2\right)/2 + \underline{h}_0^{lasso}} \ , \tag{26}$$

$$\hat{\upsilon}_{j,t}^{lasso} = \mathbb{E}_{q(\theta|y_{1:t})}(\upsilon_{j,t}^{lasso}|y_{1:t}) = \hat{\iota}_{j,t}^2 \ , \tag{27}$$

for $t = 1, \ldots, T$ and $j = 1, \ldots, k$.

3.3.2.6 Posteriors of the Horseshoe prior

Similar to SVSS, the updating steps for the Horseshoe prior are presented in this section. Where, again, the expectation $\mathbb{E}[\cdot]$ is taken with respect to the other VB posteriors, that are gathered in θ,

on the right-hand side of that particular equation. The derivations for these variatonal posteriors are in Appendix A.2. The variational posterior for each parameter is the following:

$$\hat{\lambda}_t = \mathbb{E}_{q(\theta|y_{1:t})}(\lambda_t|y_{1:t}) = \frac{\hat{\varphi}_t^{-1} + 1/2 \sum_j^k \hat{\phi}_{t,j}^{-1}(\hat{\beta}_{j,t}^2 + \hat{P}_{jj,t})}{k + 1/2} \ , \tag{28}$$

$$\hat{\phi}_{j,t} = \mathbb{E}_{q(\theta|y_{1:t})}(\phi_{j,t}|y_{1:t}) = \frac{\hat{\nu}_{j,t} + 1/2(\hat{\beta}_{j,t}^2 + \hat{P}_{jj,t})\hat{\lambda}_t^{-1}}{1 + 1/2} \ , \tag{29}$$

$$\hat{\nu}_{j,t} = \mathbb{E}_{q(\theta|y_{1:t})}(\nu_{j,t}|y_{1:t}) = \frac{\underline{h}_0^{horseshoe} + \hat{\phi}_{j,t}^{-1}}{\underline{g}_0^{horseshoe}} \ , \tag{30}$$

$$\hat{\varphi}_t = \mathbb{E}_{q(\theta|y_{1:t})}(\varphi_t|y_{1:t}) = \frac{\underline{h}_0^{horseshoe} + \hat{\lambda}_t^{-1}}{\underline{g}_0^{horseshoe}} \ , \tag{31}$$

$$\hat{\upsilon}_{j,t}^{horseshoe} = \mathbb{E}_{q(\theta|y_{1:t})}(\upsilon_{j,t}^{horseshoe}|y_{1:t}) = \hat{\lambda}_t\hat{\phi}_{j,t} \ , \tag{32}$$

for $t = 1, \dots, T$ and $j = 1, \dots, k$.

3.3.3 Convergence

In a VI estimation setting, the natural metric that one might opt for to gauge convergence is the (unscaled) KL divergence. However, as this is not straightforward to derive in a TVP-BVAR, we use the convergence of the estimated parameters as a proxy. That means that when the absolute difference between parameters in subsequent iterations is below a certain threshold, we state that the estimates of the parameters have converged. This is also a commonly used criteria in VI (Gelman et al., 2013). The convergence criteria for a VI-based TVP-BVAR is as follows, it is either $|\boldsymbol{\beta}_{1:T}^i - \boldsymbol{\beta}_{1:T}^{i-1}| < 10^{-6}$, where i refers to the current iteration of the VI algorithm or if the limit of 200 iterations is reached. In the case of the simulation study, convergence is achieved in 30 to 50 iterations, in the empirical exercise as few as 10 to 20 iterations are required. Moreover, besides the convergence criteria, a sanity check on in-sample fit (compared to TVP-BVAR) is also part of the repertoire to get an idea whether the model is working in a correct manner.

3.4 Derivative approximation

Local sensitivity analysis is the change in a summary statistic of the posterior (e.g. predictive mean) of a model to an infinitesimal perturbation in the hyperparameter of a prior. This is equivalent to the idea of calculating (or deriving) derivatives of the model with respect to a set of hyperparameters. There are a number of approaches, the most well known are (*i*) finite differencing (FD), (*ii*) automatic differentiation (AD) and (*iii*) symbolic differentiation (SD) (Nocedal & Wright, 2006).

One of the limitations of MCMC is that the more traditional approach to local sensitivity analysis is often based on FD. This requires the re-estimation of the model with slight perturbations

to the hyperparameter, as MCMC might be a computationally intensive method, this is quite impractical. Although, there are some recent developments, such as Chan et al. (2019) that are based on AD in conjunction with MCMC. This is also a promising avenue for VI, more in this in the discussion in section 7.

This is where the benefits of MFVI become apparent as it is a deterministic algorithm that iterates sequentially over the different (analytical) expressions for the optimal factorized distributions it is (very) fast compared to MCMC. This makes it feasible to opt for the more traditional FD approach to sensitivity analysis.

As a large TVP-BVAR has quite a number of variables, it is sensible to look at the sensitivity of the hyperparameters with respect to a measure of the forecast performance of a model. In this case it will be the mean of the squared forecast error (MSFE). The derivative at point p^* for $f(\cdot)$ is calculated in the following manner:

$$\Delta f(p^*) = \frac{[f(p^* + error) - f(p^* - error)]}{2 \times error} \ , \tag{33}$$

where the most common choice for the error is $\sqrt{1.1 \times 10^{-16}}$ (Nocedal & Wright, 2006). In this thesis, the $f(\cdot)$ is the complete estimation of the model and out-of-sample prediction which results in the MSFE. The p^* is one of the hyperparameters in each prior. Thus, the derivative that is approximated is the MSFE with respect to the hyperparameter of a prior. The degree to which the true derivative is approximated on an interval, such as $[a, b]$, is dependent on the number of points that are evaluated between a and b. The higher the number of points that are evaluated, the better the approximation, this is sometimes referred to as the "resolution", a higher resolution is analogous to more evaluations on the interval $[a, b]$. However, as one can see, this has a drawback in increased computation. As for each derivative approximation, two function evaluations have to be calculated. If $f(\cdot)$ is time consuming to calculate, which is often the case in MCMC, this approximation technique is too cumbersome or infeasible for some set of models. To determine the minimal resolution that is required, we iterate over different resolutions and based on visual inspection determine whether it is sufficient. Thus, if the resolution is increased and the derivative does not seem to differ in a meaningful manner compared to a lower resolution, we assert that the lower resolution for the approximation is optimal.

4 Data

4.1 Simulation exercise

To verify if VI is a viable alternative to the more common models that are utilised in a macroeconomic setting, a simulation exercise is set up. The mean squared deviation (MSD) is used to measure how well the model is able to retrieve the DGP values. The MSD is defined in the following

manner:

$$\text{MSD} = \frac{\sum_t^T \sum_j^P (\boldsymbol{\beta}_{j,t} - \hat{\boldsymbol{\beta}}_{j,t})^2}{P \times T} \ , \tag{34}$$

where P is the number of $\boldsymbol{\beta}$ coefficients associated with the lags in the DGP, not the contemporaneous, or augmented, values of $\boldsymbol{y}$. Furthermore, besides a comparison on MSD an additional comparison is made between different models based on predictive performance, the *h-step* MSFE and average log-predictive likelihood (ALPL). The *h-step* MSFE is defined in the following manner:

$$h\text{-}step \text{ MSFE} = \frac{\sum_t^{T^h} \sum_j^M (\boldsymbol{y}_{j,t} - \hat{\boldsymbol{y}}_{j,t}^h)^2}{M \times T^h} \ , \tag{35}$$

where $h = 1, \ldots, 8$ and T^h is the length for the available forecasts. A measure based on the log-predictive likelihood, such as ALPL, is often favoured over the MSFE as it is a more complete measure of overall predictive accuracy. This favouritism is in part due to its connection to the Kullbeck-Leibler divergence. As Gelman et al. (2013) state: "in the limit of large sample sizes, the model with the lowest Kullbeck-Leibler information — and thus, the highest expected log predictive probability — will have the highest posterior probability"[12]. As the model in (5) is a normal linear model, it allows for quite a simple calculation of the ALPL (as is also in line with Koop and Korobilis (2013)). The ALPL is defined in the following manner:

$$\text{ALPL} = \frac{1}{8} \sum_h^8 \frac{1}{T^h} \sum_t^{T^h} \ln \left[p_M(\boldsymbol{y}_t | \hat{\boldsymbol{y}}_t^h, \hat{\boldsymbol{\Sigma}}_h) \right] \ , \tag{36}$$

where $p_M(\cdot)$ is the *pdf* of the multivariate normal distribution, $\boldsymbol{y}_t^h$ is the *h-step ahead* forecast for all M variables at time t and $\hat{\boldsymbol{\Sigma}}_h$ is the sample covariance of the variables that are in the *h-step ahead* forecast sample.

The DGP that is used in the simulation exercise is of the following specification:

$$\begin{aligned}
&\boldsymbol{y}_t = (I_M \otimes \boldsymbol{X}_t')\boldsymbol{\beta}_t + \varepsilon_t; && \varepsilon_t \sim N_M(\boldsymbol{0}, \boldsymbol{\Sigma_t}) \\
&\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \eta_t; && \eta_t \sim N_k(\boldsymbol{0}, \boldsymbol{Q_t}) \\
&\text{vec}^{-1}(\boldsymbol{\beta}_0)_{(M,M)_{(i,j)}} = \zeta_{(i,j)} \times \underline{c}_1, \text{ where } i \neq j && \zeta_{(i,j)} \sim Ber(1 - \underline{p}_0) \\
&\text{vec}^{-1}(\boldsymbol{\beta}_0)_{(M,M)_{(i,j)}} = \underline{c}_2, \text{ where } i = j && \\
&\boldsymbol{\Sigma}_{(i,j),0} = \underline{c}_3, \text{ where } i = j && \\
&\boldsymbol{\Sigma}_{(i,j),0} = \underline{c}_4, \text{ where } i < j && \\
&\boldsymbol{\Sigma}_{(i,j),t} = |\boldsymbol{\Sigma}_{(i,j),t-1} + \iota_{(i,j),t}|, \text{ where } i = j \text{ and } t = 1, \ldots, T && \iota_{(i,j),t} \sim N(0, \underline{\varrho}_1) \\
&\boldsymbol{\Sigma}_{(i,j),t} = \boldsymbol{\Sigma}_{(i,j),t-1} + \iota_{(i,j),t}, \text{ where } i < j \text{ and } t = 1, \ldots, T && \iota_{(i,j),t} \sim N(0, \underline{\varrho}_1) \\
&diag(\boldsymbol{Q}_t) = |diag(\boldsymbol{Q}_{t-1}) + \xi_t| && \xi_t \sim N_k(\boldsymbol{0}, \underline{\varrho}_2 \times \boldsymbol{S}_k) \\
&\text{vec}^{-1}(\boldsymbol{S}_k)_{(k,k)_{(i,j)}} = \zeta_{(i,j)} &&
\end{aligned} \tag{37}$$

[12] One of the drawbacks of using an uncorrected *lpl*, is that model size is not taken into account, more on this specific aspect in the discussion

The following parameters are set $M \in \{3, 7\}$, $T \in \{100, 200\}$, $\underline{c}_1 = 0.5$, $\underline{c}_2 = 0.25$, $\underline{c}_3 = 1$, $\underline{c}_4 = 0.5$, $\underline{p}_0 \in \{60\%, 80\%\}$, $\underline{\varrho}_1 = 1e-2$, $\underline{\varrho}_2 = 1e-9$. The lag p is set to 1, as it is a DGP for a TVP-BVAR, it is difficult to extend it to more lags while still maintaining a non-explosive (or non-implosive) system. Some parameters, especially $\underline{\varrho}_2$, which controls the evolution of $\boldsymbol{\beta}_t$, and $\underline{p}_0$, which controls the sparsity of $\boldsymbol{\beta}_t$ (a higher percentage refers to a higher degree of sparsity), are quite sensitive and even a slight increase or decrease lead to either an implosive or explosive process. With these parameter settings, the eigenvalues of $\text{vec}^{-1}(\boldsymbol{\beta_0})_{(M,M)}$ are larger than 1, which guarantees stationarity. Furthermore, the lower triangular values of $\boldsymbol{\Sigma}_t$ are mirrored across the diagonal. The sample for forecasting (i.e. test sample) is respectively from 75:100 for $T = 100$ and 175:200 for $T = 200$. The sparsity parameter $\underline{p}_0$ determines the sparsity at the initialization of the series, this sparsity is time-invariant. To make sure that it is not just sparsity at the moment of initialization the sparsity is maintained by setting the corresponding variances of the sparse states to zero in $\boldsymbol{Q}_{1:T}$.

In total, there are eight different scenarios ($|\{3, 7\}| \times |\{100, 200\}| \times |\{60\%, 80\%\}| = 2^3$). The reasoning for the eight different scenarios is as follows. In the case of the different number of variables in the system, as mentioned in the literature section, the forecast for specific variables (e.g. GDP) often becomes better if the system is extended with more macroeconomic variables (Koop & Korobilis, 2013). The differences in sparsity might be warranted by the fact that in different compositions of systems of variables there might be no interrelationship present. This is reflected by differences in sparsity, if the coefficients matrix is dense, it means that the interrelationships are common among most of the variables, if it is sparse, it is the other way around. This can also be the case in macroeconomic modelling, certain macroeconomic variables might have no meaningful effect, while others have. The difference in length of time is due to the practicality that the history of macroeconomic data might not be that long. For example, the Case-Shiller 20-City Composite Home Price Index has quarterly data as of the year 2000.

The total number of iterations for the simulation is constrained at 200. This is in part motivated by time considerations and in part by the level of the standard deviation of the MSFE (it is considered to be reasonably small at this number of iterations, an extra 100 iterations would have marginally reduced the variation).

4.1.1 Models used for comparison and hyperparameter settings

Several models are used in the simulation study to compare the ability of the VI-based TVP-BVAR to retrieve the true coefficients and to forecast at several timesteps into the future. The three different models, all implemented in `R`[13], are the following:

- **VAR** — a standard VAR with a lag of 1. It is based on the specification described in Pfaff (2008a) and implemented using the `vars` library (Pfaff, 2008b). Besides the lag, there are no other parameters to be specified.

[13] A (statistical) programming language developed by the R Core Team (2020). Furthermore, as mentioned earlier, all the code is available at the GitLab repository, all results should be reproducible. More on this in appendix D

- **BVAR** — a Bayesian VAR with a Minnesota prior, estimated using MCMC (in this case Hamiltonian MC with default library parameters). It is based on an hierarchical approach proposed by Giannone et al. (2015) and implemented in the `BVAR` library (Kuschnig & Vashold, 2020). The hyperparameters in this specification λ (controls the tightness of the prior), α (controls the decay of the variance with increasing lag order) and ψ (controls the prior's standard deviation on lags of variables other than the dependent variables), are set in an automatic fashion in line with Giannone et al. (2015) (for exact details of the automatic procedure see the documentation). Furthermore, the prior variance is $\boldsymbol{\Sigma}_0 = 1e+7 \times \boldsymbol{I}$ and prior mean is $\boldsymbol{\beta}_0 = \boldsymbol{1}$. Moreover, the number of lags is set to 1, the number of draws for the MCMC procedure is set to 10,000 for $M = 3$ and 5,000 for $M = 7$, the number of burn-in draws is half of the total number of draws. Thinning is set to 1.

- **TVP-BVAR** — a Bayesian TVP with a gamma-gamma prior (or double gamma prior) that is modeled in the same manner as the VI-based TVP-BVAR with augmented values of $\boldsymbol{y}_t$, it is also estimated using MCMC (in this case based on a variation of a data augmentation scheme (DA), referred to as the ancillarity–sufficiency interweaving strategy (ASIS)). It is the exact specification of Bitto and Frühwirth-Schnatter (2019), implemented in the `shrinkTVP` library (Knaus et al., 2019). This specification has several hyperparameters, the following are automatically inferred from the data a^{ξ} (controls the local variance of the states), κ^2 (controls the global variance of the states), a^{τ} (controls the local location of the initial states) and λ^2 (controls the global level of shrinkage for the location of the initial states), whereas the values for the remaining parameters (e.g. c_0, g_0, b^{ξ}, b^{τ}) are the default values (these are the values that often work well in practice according to Bitto and Frühwirth-Schnatter (2019), once again see the documentation for details). The observation variance is allowed to vary, the model assumed for the variance is of a stochastic volatility specification, which follows an AR(1) process. The number of MCMC draws is set to 2,000 for $M = 3$ and 1,000 for $M = 7$, similar to the BVAR, the number of burn-in draws is half of the total number of draws. Thinning is set to 1.

- **RW** — A random walk model, which is one of the simplest benchmarks there is. The previous value will be the forecast for the subsequent period. In univariate and multivariate cases, whereas the model is simple in nature, it has shown to be quite a competitive benchmark. It is specified in the following manner: $\boldsymbol{y}_{T+h} = \boldsymbol{y}_T \ \forall_h$.

A point of concern that one might raise is the relatively low number of MCMC draws for the BVAR and TVP-BVAR, it might be the case that an acceptable level of convergence is not yet achieved. The reason that these draws are quite low is quite simple, as it reduces the computation time quite considerably. However, we could gauge convergence using the Gelman-Rubin statistic, which uses within-chain variation and between-chain variation to provide an indication of convergence. For the BVAR in both cases, for $M = 3$ and $M = 7$, the statistic indicates convergences. In the case of the TVP-BVAR, except in the case of the longer time horizon scenarios for $M = 7$, in all

the other scenarios the statistic is below the threshold of 1.2 that is an indication for convergence. These statistics are shown in table 5 in Appendix C.

The hyperparameters for each of the priors in the VI-based TVP-BVAR are based on earlier literature. The important hyperparameters for the SVSS prior, $\underline{g}_0$, $\underline{h}_0$, $\underline{\pi}_0$ and $\underline{c}$ are set to the values utilised in Koop and Korobilis (2020). Thus, $\underline{g}_0^{svss} = 1$, $\underline{h}_0^{svss} = 12$ and $\underline{c} = 10^{-4}$. This is an uninformative specification for $\tau_{j,t}^2$ and $\gamma_{j,t}$. The hyperparameters for the Lasso prior $\underline{h}_0^{lasso}$ and $\underline{b}_g^{lasso}$ are set according to Belmonte et al. (2014), where this specification is promising in their empirical exercise. Thus, $\underline{h}_0^{lasso} = 0.001$ and $\underline{g}_0^{lasso} = 0.001$, is also an uninformative prior. The hyperparameters for the Horseshoe prior are set to $\underline{h}_0^{horseshoe} = 2$ and $\underline{g}_0^{horseshoe} = 1$. This is a more informative prior than suggested in Belmonte et al. (2014), this slightly stronger shrinkage is required to attain a practical rate of convergence in the simulation exercise.

The other hyperparameters that are of importance, but not strictly tied to any of the specific priors mentioned earlier are: $\underline{c}_{j,0}$, $\underline{d}_{j,0}$ (the hyperparameters on the prior variance for $\tilde{\boldsymbol{\beta}}_t$), $\underline{a}_0$, $\underline{b}_0$ (the hyperparameters on the prior for the time-varying volatility), δ, $\boldsymbol{\beta}_{j,0}$ and $\boldsymbol{P}_{j,0}$. These are also set to the prior specification that is recommended in Koop and Korobilis (2020). Thus, $\underline{c}_{j,0} = 100$, $\underline{d}_{j,0} = 1$, $\underline{a}_0 = 0.01$, $\underline{b}_0 = 0.01$, $\delta = 0.8$, $\tilde{\boldsymbol{\beta}}_{j,0} = 0$ and $\boldsymbol{P}_{j,0} = 4$. The specification for $\underline{c}_{j,0}$, $\underline{d}_{j,0}$ is more conservative and limits the ability to model erratic jumps in $\tilde{\boldsymbol{\beta}}_{j,t}$ (Koop & Korobilis, 2020).

4.2 Empirical exercise

The data is from the Federal Reserve Bank of St. Louis' FRED-QD dataset[14], this is quarterly US data from 1959Q4 through 2019Q4 for 248 variables. The FRED-QD dataset is a large macroeconomic dataset that is in part designed to mimic a common macroeconomic dataset used in the academic literature, in particular the one in Stock and Watson (2012) [Huber et al. (2020), Koop and Korobilis (2020)]. The monthly series are averaged to a quarterly frequency if necessary. Furthermore, following recommendations in Carriero et al. (2015) the series are transformed to attain stationarity (some variables might remain in levels, others might be transformed into growth rates). If x_t is the original variable and z_t is the transformed variable, the following transformations are possible to achieve stationarity for the FRED-QD data: (1) $z_t = x_t$, (2) $z_t = \Delta x_t$, (3) $z_t = \Delta^2 x_t$, (4) $z_t = \ln x_t$, (5) $z_t = \Delta \ln x_t$, (6) $z_t = \Delta^2 \ln x_t$ and (7) $\Delta x_t / x_{t-1} - 1$. Although, the computation burden is drastically reduced in a VI-based TVP-BVAR, the number of model evaluations, for a meaningful approximation of the derivative, is still substantial. Therefore, the sensitivity analysis is carried out on a smaller model of $M = 3$. The series included in this model are shown in table 1. This limitation is further explored in the discussion, section 7.

Table 1: *Series included in the sensitivity analysis*

	Series*	Transformation	Description
1	GDPC1	5	Real Gross Domestic Product, 3 Decimal (Billions of Chained 2012 Dollars)
2	CPIAUCSL	6	Consumer Price Index for All Urban Consumers: All Items (Index 1982-84=100)
3	FEDFUNDS	2	Effective Federal Funds Rate (Percent)

*This is equivalent to the mnemonics in the FRED-QD dataset

5 Results

In this section the results of the simulation study, where we discuss if the proposed VI-based TVP-BVAR, as of now referred to as a TVP-VI, is a competent alternative to the more common models, and the results of the empirical exercise will be reviewed.

5.1 Simulation study

The TVP-VI models as discussed here are a viable alternative to the MCMC-based TVP-BVAR and other common benchmarks. There is enough evidence to support the idea that it seems to be at least competitive, based on ALPL, and in some cases even strictly better, based on MSFE, than the TVP-BVAR in the different aspects (i.e. sparsity, time horizon and different number of series) that are scrutinised, as can be seen in tables 2 & 3. Moreover, the ability to retrieve the true coefficients is quite similar to the MCMC-based TVP-BVAR as the MSD between the TVP-VI models and the

[14]Publicly available on https://research.stlouisfed.org/econ/mccracken/fred-databases/

MCMC-based TVP-BVAR is reasonably close. The relative performance of time-invariant models is in the same vicinity as the time-varying models. However, the naive random walk is not able to cope with the complexity of the DGPs.

An additional note of caution is that these results are limited to the specific implementation of VI in a TVP-VAR. There are several other ways to implement VI (e.g. different decomposition of the approximation distribution, allowing for time-varying volatility in a different manner), and the several VAR based DGPs. Nonetheless, these results are in line with earlier literature that compared VI (or VB) to MCMC based approaches [Gefang et al. (2019), Koop and Korobilis (2020)]. With regards to the aspect of computation, as mentioned earlier, an increase in dimensionality results in a material increase in computation time. In Appendix B, in 4 the median runtime for all the different scenarios are stated. As one can conclude, is that in the case of $M = 7$, the benefits of VI are apparent. The fastest converging prior, Lasso, is more than ten times as fast as the TVP-BVAR (with a relatively low number of draws).

The **bold** cells in tables 2 & 3 are the best *h-step* MSFE results of all the models. To ease the process of comparing, the TVP-BVAR is chosen as the baseline for the MSFE and ALPL measures. The main results are discussed in the following paragraph, the extensive details of these findings is in the subsequent sections. Moreover, a table of the comparison of the average differences in relative performance between each of the important aspects is in appendix C. To repeat the scenarios that are scrutinised, scenarios 1 — 4 is for $M = 3$ and scenarios 5 — 8 is for $M = 7$. Subsequently, scenarios 1 & 2 and 5 & 6 are for the shorter time horizon ($T = 100$), whereas scenarios 3 & 4 and 7 & 8 are for the longer time horizon ($T = 200$). The differences in sparsity are within each time horizon, where the first scenario is the less sparse scenario. For example, scenario 1 has 60% sparsity, whereas scenario 2 has 80% sparsity.

5.1.1 Outcome of different levels of sparsity and ability to retrieve true coefficients

A noticeable difference is the asymmetry in MSD between the TVP-VI priors, The SVSS prior has a relatively weak result compared to the other priors, even in the most sparse case. Where one would think the SVSS has a slight advantage. It could be the consequence of how sparsity is introduced in the system, as the SVSS prior has the ability to model differences in sparsity over time, whereas the sparsity in the DGP is time-invariant. The MSD for all scenarios is lowest for the BVAR, it might be the case that this is a result of, again, the manner in which sparsity is introduced in the system. As it does not vary with time, this model (and the VAR) might be better able to set the sparse coefficients to zero and it offsets the error that is induced on the time-varying coefficients that are modeled in a time-invariant manner.

Furthermore, the MSD for all models stays relatively similar in all scenarios for $M = 3$ and varies slightly in $M = 7$, the TVP-BVAR is better able, about twice as well, to model the DGP than the TVP-VI alternatives, although it is still an order of magnitude off from the time-invariant models as can be seen in tables 2 & 3.

In the case of sparsity, for the lower dimensional case, $M = 3$, the differences in sparsity do not

seem to have a material effect on the relative performance of the *h-step* MSFE in the majority of the scenarios, although slightly on the absolute performance, of the models. In scenario 1 versus 2 the average difference for the TVP-VI models is -0.01 versus a difference of 0.004 for the BVAR and VAR. In scenario 3 versus 4 the average difference in relative performance for the TVP-VI models is -0.07 versus a difference of 0.008 for the BVAR and VAR. This might be in part explained by the fact that the difference in absolute sparsity is also not that numerous between the percentages of sparsity (the number of coefficients set to zero in the case of 60% sparsity is (on average) 3 compared to (on average) 5 in the case of 80%). In the higher dimensional case, $M = 7$, the difference in sparsity has a more pronounced effect on the time-invariant models. In scenario 5 versus 6 the average difference for the TVP-VI models is -0.015 versus a difference of -0.009 for the BVAR and VAR. In scenario 7 versus 8 the average difference for the TVP-VI models is -0.02 versus a difference of 0.194 for the BVAR and VAR. Overall, the TVP-VI models seem to handle the differences in sparsity quite well, as opposed to the time-invariant models that suffer if the time horizon is extended. However, the relative performance of scenario 7 is an outlier for the time-invariant models, this reduction in relative performance is visible in each of the aspects that we compare.

If we delve further into the difference between the TVP-VI priors, the differences are not that pronounced. However, there seems to be a dichotomy there. The results, for MSFE and ALPL as well as MSD, from the Lasso and Horseshoe prior are quite similar, whereas the results for SVSS differ in a material manner. More specifically, the *1-step* MSFE, ALPL and MSD across all scenarios differ substantially between SVSS and Lasso/Horseshoe. The SVSS prior is lacking in predictive performance in the *1-step ahead* forecast, while bettering the other two priors in all the other *h-step ahead* forecasts in almost all scenarios (except 3 & 7). It is unclear to what extent this difference is caused by either the philosophy behind the shrinkage in this prior (it shrinks differently compared to the other priors) or the specific choice of hyperparameters in this instance. In the next subsection this degree of sensitivity is investigated for each prior. Another reason why the difference between the TVP-VI priors is not that pronounced is that the prior might be a relatively minor part in the model. The performance might be, in part or even wholly, attributed to the underlying manner in which VI in a TVP is achieved. It could be the specific decomposition of the approximation (e.g. mean-field approximation) or the manner in which time-varying volatility is modeled. It might be the case that these model considerations have a far greater influence on the predictive performance than the choice of prior in this configuration.

5.1.2 Consequence of different time horizons

Although the prevailing notion is that an increase in length of time, would likely reflect in a lower absolute MSFE. In none of the scenarios this is the case, the increase is less severe for the lower dimensional case than for the higher dimensional case. However, this is likely to be a result of less predictions that have to be made at each moment in time (3 versus 7), thus there are less occasions to make an error. Nonetheless, for the time-varying models this could be a consequence of the

increase in the number of parameters, as it scales with time. Although, there is extra information present, it might be difficult to model this in such an efficient manner that it negates an increase in absolute MSFE. In this instance, it could also be a consequence of the specific DGP. This increase in absolute MSFE for the time-invariant models might be a result of the error that is inherently present due to the fact that the coefficients vary over time. It is not unlikely that this error increases with time as the coefficients have more opportunity to vary. If one looks at the ALPL, in the lower dimensional case for the TVP-VI models there is a substantial difference between the different time horizons (scenario 1 & 2 versus 3 & 4), while this is not present in the higher dimensional case. Suggesting that the TVP-BVAR has, at least based on the ALPL, a better overall predictive fit in the longer time horizon versus the shorter time horizon, this makes the argument for the fact that a MCMC based TVP-BVAR approach is still a competitive model in the lower dimensional case. In the case of the BVAR and VAR, the reduction in ALPL from the shorter to the longer time horizon is present in the lower- as well as the higher dimensional case. If we look at the relative performance based on the *h-step* MSFE of the models between the different scenarios, there is no material difference for the majority of the scenarios, except for the time-invariant models in the higher dimensional case.

5.1.3 Differences in dimensionality

An increase in M from 3 to 7 is a substantial difference in the number of coefficients that have to be estimated of 4400 $[3(3+1) \times 100 - 7(7+1) \times 100]$ for the shorter time horizon (scenarios 1 & 2 versus 5 & 6) and 8800 for the longer time horizon (scenarios 3 & 4 versus 7 & 8). The differences that exist between the same scenarios are likely ascribable to the increase in the number of variables. In relative performance there does seem to be a material difference between scenario 3 and 7, as the relative performance for all the TVP-VI models is better in scenario 7 than 3. The BVAR and VAR are also not able to cope with the increase in complexity. However, comparing scenario 7 to scenario 8, it is difficult to ascertain that this result is strictly due to the difference in dimensionality, as it also seems to be a consequence of the difference in sparsity. Another interesting observation is the difference in ALPL between all the scenarios in the lower dimensional and higher dimensional setup for the TVP-VI models. A higher ALPL is preferable, as it is indicative of overall better predictive fit. Therefore, it is quite interesting to see that, in the lower dimensional case, the point measure and density measure of predictive accuracy are inversely related to one another for the TVP-VI models. While, for these models, in the higher dimensional case the two measures are pointing in the same direction. As mentioned earlier, this is perhaps indicative of the fact that, as opposed to the lower dimensional setting, in a higher dimensional setting a MCMC based TVP-BVAR approach is lacking the ability to cope with the increase in the number of parameters and this is reflected mainly in the ALPL.

Table 2: *Simulation results for* $M = 3$

Scenario 1: $T = 100$, $p_0 = 60\%$											
Time-varying	*MSFE*	*1-step*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	MSD	ALPL
TVP-VI											
- *SVSS*	1.335	0.98	**0.95**	**0.91**	**0.87**	**0.83**	**0.78**	**0.75**	**0.70**	4.6×10^{-2}	-4.64
- *Lasso*	1.331	**0.70**	0.95	0.95	0.91	0.87	0.82	0.79	0.74	2.4×10^{-2}	-4.38
- *Horseshoe*	**1.311**	0.71	0.94	0.93	0.89	0.85	0.81	0.77	0.73	2.6×10^{-2}	-4.61
TVP-BVAR*	1.581		-	-	-	-	-	-	-	1.2×10^{-2}	0.00
Time-invariant											
BVAR	1.670	0.97	1.05	1.07	1.07	1.07	1.07	1.08	1.08	$\mathbf{1.3 \times 10^{-3}}$	8.36
VAR	1.668	0.96	1.05	1.07	1.07	1.07	1.07	1.08	1.08	1.5×10^{-3}	5.06
RW	2.949	1.37	1.75	1.91	1.96	1.98	1.96	1.98	1.99	-	-
Scenario 2: $T = 100$, $p_0 = 80\%$											
Time-varying	*MSFE*	*1-step*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	MSD	ALPL
TVP-VI											
- *SVSS*	1.321	0.98	0.95	**0.91**	**0.87**	**0.83**	**0.79**	**0.75**	**0.71**	4.5×10^{-2}	-4.97
- *Lasso*	1.341	**0.72**	0.97	0.97	0.92	0.88	0.84	0.80	0.76	2.5×10^{-2}	-4.63
- *Horseshoe*	**1.318**	0.73	**0.95**	0.95	0.90	0.86	0.82	0.78	0.74	2.4×10^{-2}	-4.86
TVP-BVAR*	1.564	1.00	-	-	-	-	-	-	-	1.9×10^{-2}	0.00
Time-invariant											
BVAR	1.644	0.97	1.05	1.06	1.06	1.07	1.07	1.07	1.07	$\mathbf{1.3 \times 10^{-3}}$	7.26
VAR	1.647	0.98	1.05	1.06	1.06	1.06	1.07	1.07	1.07	1.5×10^{-3}	4.62
RW	2.918	1.38	1.74	1.91	1.95	1.98	1.97	1.98	2.00	-	-
Scenario 3: $T = 200$, $p_0 = 60\%$											
Time-varying	*MSFE*	*1-step*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	MSD	ALPL
TVP-VI											
- *SVSS*	1.845	0.91	0.94	**0.91**	**0.88**	**0.84**	**0.79**	**0.74**	**0.70**	4.5×10^{-2}	-6.67
- *Lasso*	1.828	**0.68**	**0.91**	0.92	0.90	0.87	0.83	0.78	0.73	2.5×10^{-2}	-6.98
- *Horseshoe*	**1.814**	0.69	0.90	0.92	0.89	0.86	0.82	0.77	0.72	2.3×10^{-2}	-7.30
TVP-BVAR*	2.206	1.00	-	-	-	-	-	-	-	2.3×10^{-2}	0.00
Time-invariant											
BVAR	2.346	0.91	1.04	1.07	1.09	1.10	1.10	1.10	1.10	$\mathbf{1.1 \times 10^{-3}}$	2.34
VAR	2.348	0.91	1.04	1.07	1.09	1.09	1.10	1.10	1.10	1.2×10^{-3}	0.85
RW	4.246	1.35	1.73	1.91	2.02	2.07	2.08	2.10	2.12	-	-
Scenario 4: $T = 200$, $p_0 = 80\%$											
Time-varying	*MSFE*	*1-step*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	MSD	ALPL
TVP-VI											
- *SVSS*	**1.762**	0.86	0.93	**0.89**	**0.86**	**0.82**	**0.78**	**0.73**	**0.69**	6.3×10^{-2}	-7.85
- *Lasso*	1.759	**0.67**	0.91	0.91	0.89	0.86	0.82	0.77	0.73	2.3×10^{-2}	-8.12
- *Horseshoe*	1.772	0.68	**0.91**	0.92	0.89	0.86	0.83	0.78	0.75	2.3×10^{-2}	-8.41
TVP-BVAR*	2.141	1.00	-	-	-	-	-	-	-	1.7×10^{-2}	0.00
Time-invariant											
BVAR	2.261	0.92	1.04	1.06	1.08	1.08	1.09	1.09	1.09	$\mathbf{1 \times 10^{-3}}$	2.16
VAR	2.260	0.92	1.04	1.06	1.08	1.08	1.08	1.09	1.09	1.1×10^{-3}	0.44
RW	4.109	1.36	1.75	1.91	2.02	2.06	2.07	2.07	2.11	-	-

*It is the baseline for the comparison of the *h-step* MSFE & ALPL

Table 3: *Simulation results for* $M = 7$

Scenario 5: $T = 100$, $p_0 = 60\%$											
Time-varying	*MSFE*	*1-step*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	MSD	ALPL
TVP-VI											
- *SVSS*	2.317	1.04	0.96	**0.90**	**0.85**	**0.81**	**0.77**	**0.73**	**0.69**	3.0×10^{-2}	4.34
- *Lasso*	2.225	**0.74**	0.91	0.90	0.87	0.82	0.79	0.74	0.70	1.7×10^{-2}	5.17
- *Horseshoe*	**2.212**	0.74	0.90	0.89	0.86	0.82	0.78	0.74	0.69	1.7×10^{-2}	4.91
TVP-BVAR*	2.756	1.00	-	-	-	-	-	-	-	2.0×10^{-2}	0.00
Time-invariant											
BVAR	2.759	0.90	0.99	1.01	1.02	1.02	1.02	1.03	1.02	$\mathbf{2.0 \times 10^{-3}}$	3.42
VAR	2.754	0.91	0.99	1.01	1.01	1.01	1.02	1.02	1.02	2.7×10^{-3}	1.87
RW	4.761	1.27	1.61	1.74	1.80	1.83	1.85	1.85	1.83	-	-
Scenario 6: $T = 100$, $p_0 = 80\%$											
Time-varying	*MSFE*	*1-step*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	MSD	ALPL
TVP-VI											
- *SVSS*	2.159	1.02	0.96	**0.91**	**0.87**	**0.82**	**0.78**	**0.74**	**0.69**	2.8×10^{-2}	4.86
- *Lasso*	2.115	**0.76**	0.95	0.93	0.89	0.84	0.80	0.76	0.71	1.8×10^{-2}	5.34
- *Horseshoe*	**2.101**	0.76	**0.94**	0.92	0.88	0.83	0.80	0.75	0.70	1.8×10^{-2}	5.12
TVP-BVAR*	2.553	1.00	-	-	-	-	-	-	-	1.6×10^{-2}	0.00
Time-invariant											
BVAR	2.578	0.94	1.02	1.03	1.02	1.02	1.02	1.02	1.01	$\mathbf{2.5 \times 10^{-3}}$	4.23
VAR	2.574	0.96	1.02	1.02	1.01	1.01	1.01	1.01	1.02	3.5×10^{-3}	2.34
RW	4.512	1.32	1.67	1.81	1.86	1.88	1.88	1.88	1.85	-	-
Scenario 7: $T = 200$, $p_0 = 60\%$											
Time-varying	*MSFE*	*1-step*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	MSD	ALPL
TVP-VI											
- *SVSS*	3.974	0.96	0.94	0.90	0.85	0.81	**0.76**	**0.72**	**0.67**	2.8×10^{-2}	5.90
- *Lasso*	3.726	**0.68**	**0.83**	**0.86**	0.84	**0.80**	0.76	0.72	0.68	1.4×10^{-2}	5.34
- *Horseshoe*	**3.722**	0.70	0.83	0.86	**0.83**	0.80	0.76	0.72	0.67	1.4×10^{-2}	5.04
TVP-BVAR*	4.836	1.00	-	-	-	-	-	-	-	2.5×10^{-2}	0.00
Time-invariant											
BVAR	5.779	0.78	0.95	1.03	1.10	1.20	1.31	1.46	1.66	$\mathbf{1.7 \times 10^{-3}}$	1.28
VAR	5.794	0.78	0.95	1.03	1.10	1.20	1.31	1.47	1.67	1.9×10^{-3}	0.25
RW	9.437	1.21	1.59	1.81	1.96	2.09	2.19	2.28	2.35	-	-
Scenario 8: $T = 200$, $p_0 = 80\%$											
Time-varying	*MSFE*	*1-step*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	MSD	ALPL
TVP-VI											
- *SVSS*	3.373	0.94	0.92	**0.89**	**0.85**	**0.82**	**0.77**	**0.74**	**0.69**	2.4×10^{-3}	5.55
- *Lasso*	3.296	**0.71**	0.89	0.90	0.87	0.83	0.79	0.76	0.71	1.5×10^{-3}	4.63
- *Horseshoe*	**3.274**	0.72	**0.88**	0.89	0.86	0.83	0.79	0.75	0.70	1.5×10^{-3}	4.40
TVP-BVAR*	4.087	1.00	-	-	-	-	-	-	-	1.7×10^{-3}	0.00
Time-invariant											
BVAR	4.063	0.84	0.97	1.00	1.01	1.02	1.03	1.03	1.04	$\mathbf{1.7 \times 10^{-3}}$	1.00
VAR	4.063	0.85	0.97	1.00	1.013	1.02	1.03	1.03	1.03	1.9×10^{-3}	-0.33
RW	7.179	1.27	1.60	1.75	1.83	1.88	1.89	1.91	1.91	-	-

*It is the baseline for the comparison of the *h-step* MSFE & ALPL

5.2 Empirical application

In the empirical application of the TVP-VI, a sensitivity analysis is carried out on real world data. As described in section 4 the derivatives with respect to the hyperparameters for each prior are estimated with a numerical approximation technique, in this case a finite difference (FD) approach. According to Koop and Korobilis (2020), there are several important parameters to their implementation of VI in a TVP setting. A few of these, the $\underline{c}_{j,0}$ and $\underline{d}_{j,0}$, the hyperparameters for the prior of $\tilde{\boldsymbol{Q}}_t$, have a considerable effect on the MSFE and the eventual shrinkage that is necessary on the coefficients. However, in this empirical exercise the sole focus is on the shrinkage priors for the coefficients. All the hyperparameters that are not scrutinized for their sensitivity are kept the same as in the simulation study, for the exact values, see section 4.1.1. As one might observe, some of the derivatives with respect to the hyperparameters show slightly erratic behaviour. Although, it is still possible to infer the sensitivity of any hyperparameter, it is likely to be a consequence of the complexity of the function for which we try to approximate the derivative or it is due to inherent limitations in numerical precision. The series are, besides being transformed, standardized for the SVSS prior. The scale between the transformed series differs in several orders of magnitude (GDP versus Federal Funds Rate), this hinders fast convergence, as point further explored in section 7.

The derivatives show the sensitivity of the different priors. In the case of the overall MSFE, the Horseshoe prior seems to be more sensitive than the Lasso prior, this can be seen in figures 3 and 2. It is difficult to compare the SVSS prior as the variables are standardized in this instance. Nonetheless, when we look at the derivatives on a *h-step* basis, they all exhibit similar behaviour as longer horizon forecast seem to be less susceptible to shrinkage as the lower horizon forecasts. In hindsight, one of the drawbacks that is apparent in the use of the finite differences approach to approximate a derivative is the fact that it might suffer from numerical precision or that the resolution needs to be increased to such a high degree that it becomes infeasible due to the high number of model re-estimations, even in the case of VI. This is especially apparent in the case of the SVSS prior, as the derivatives for $\underline{h}_0^{svss}$ are quite erratic for the overall MSFE and nonsensical for the *h-step* MSFE. However, in the case of the other priors, it provides an adequate picture of the local sensitivity of the model to the hyperparameters in each of the priors. These findings are also explained in detail in the subsequent subsections.

5.2.1 Sensitivity analysis

In the case of the Lasso prior, in figure 2, the simplest specification of any of the priors. The derivative of the shape $\underline{g}_0^{lasso}$ and (inverse) scale $\underline{h}_0^{lasso}$ parameters of the hierarchical distribution for ψ_t^2 are quite similar in shape and the derivative is relatively smooth. The propagation of $\underline{g}_0^{lasso}$, keeping $\underline{h}_0^{lasso}$ fixed, into the hierarchical structure can be interpreted as an increase in this parameter makes for a weaker prior on ψ_t^2 and this in turn makes for a weaker prior on $\iota_{j,t}^2$, resulting in less shrinkage on the coefficients. The propagation of $\underline{h}_0^{lasso}$, keeping $\underline{g}_0^{lasso}$ fixed, is the exact opposite of $\underline{g}_0^{lasso}$, an increase results in stronger shrinkage on the coefficients. Although, it is more

nuanced as an increase in either of the parameters results in a different distribution (an increase of 0.1 in $\underline{g}_0^{lasso}$ results in a different shape as an increase of 0.01 in $\underline{h}_0^{lasso}$). The effect for the shape parameter is less pronounced than the scale parameter. However, both of the derivatives are negative (or in part negative), this suggests that a better selection of the hyperparameters would likely have a better (out-of-sample) MSFE. The series that are used are not standardized, to provide a frame of reference for the interpretation of the scale of the derivatives, the actual MSFE is 1.5×10^{-2}. Based on the absolute scale between the derivatives of the two hyperparameters, 10^{-7} versus 10^{-5}, we can state that the prior is more sensitive to the scale than the shape parameter of the hierarchical distribution. This is not an uncommon result, as when one looks at the actual pdf of the Gamma distribution, one can observe that this parameter has a more drastic effect on the matter of how the density is dispersed (e.g. more informative versus uninformative) than the shape parameter in the space for these parameters. Another interesting result here is that as one increases the size of the scale parameter, the prior becomes more informative and also improves the MSFE. However, this is up to a certain point, such that more shrinkage is not by definition beneficial to the out-of-sample MSFE.

The Horseshoe prior, in figure 3 is a more hierarchical prior than the Lasso, this allows for a higher degree of uncertainty in the other hyperparameters of the hierarchical structure. The distinct difference between the Horseshoe and the Lasso prior, is the fact that the last prior in the hierarchy, for which the hyperparameters are set, is an Inverse-Gamma distribution as opposed to a Gamma distribution. The propagation of $\underline{g}_0^{horseshoe}$, keeping $\underline{h}_0^{horseshoe}$ fixed, into the hierarchical structure can be interpreted as an increase in this parameter makes for a stronger prior on the parameters $\nu_{j,t}$ and φ_t, this in turn affects the priors on $\phi_{j,t}$ and λ_t and causes these priors to be less informative. This results in less shrinkage on the coefficients. The propagation of $\underline{h}_0^{horseshoe}$, keeping $\underline{g}_0^{horseshoe}$ fixed, is the exact opposite of $\underline{g}_0^{horseshoe}$. An increase in $\underline{h}_0^{horseshoe}$ results in a less informative prior on $\phi_{j,t}$ and λ_t and this propagates further into a more informative prior and stronger shrinkage on the coefficients. Thus, interpreting the results of 3, the shape $\underline{g}_0^{horseshoe}$ and scale $\underline{h}_0^{horseshoe}$ parameters have a similar effect in terms of absolute change in the MSFE. The sign of the derivatives with respect to both parameters, are different for each parameter. This makes sense, given the propagation of the parameters, they both hint at the fact that an increase in shrinkage would be beneficial to the out-of-sample MSFE. To provide a frame of reference for the MSFE, the absolute MSFE in this instance is quite close to the Lasso prior, namely 1.5×10^{-2}. As a direct comparison is possible in terms of the level of change, the Horseshoe prior is more sensitive than the Lasso prior, although, the absolute effect on the MSFE is still marginal.

In the case of the SVSS prior, in figure 1, the behaviour of the derivatives is more erratic. This might be a consequence of the increased complexity due to the hierarchical structure, or to the earlier mentioned inherent drawbacks of the use of a numerical approximation technique. The derivative of the shape $\underline{g}_0^{svss}$ and (inverse) scale $\underline{h}_0^{svss}$ parameters of the hierarchical distribution for $\frac{1}{\tau_{j,t}^2}{}_t$. The propagation of $\underline{g}_0^{svss}$, keeping $\underline{h}_0^{svss}$ fixed, is such that an increase in $\underline{g}_0^{svss}$ makes the prior on $\frac{1}{\tau_{j,t}^2}{}_t$ less informative, resulting in less shrinkage on the coefficients. The propagation of $\underline{h}_0^{svss}$,

keeping $\underline{g}_0^{svss}$ fixed, is such that an increase in $\underline{h}_0^{svss}$ results in a stronger prior on $\frac{1}{\tau_{j,t}^2 t}$ and in turn results in more shrinkage on the coefficients. In absolute differences between the derivatives of the hyperparameters, the scale parameter has a more material effect than the shape parameter, 10^{-3} versus 10^{-5}. What is interesting, is that the derivatives are positive for the shape parameter and steadily decreasing, this suggests that a decrease in shrinkage would likely be a slight detriment to the out-of-sample MSFE. This is corroborated by the results for $\underline{h}_0^{svss}$, as an increase in shrinkage would result in a better out-of-sample MSFE. Due to the fact that the series are standardized, the reference for the interpretation of the derivatives, the actual MSFE is 3×10^{-1}.

Another interesting result is the different derivatives with respect to each *h-step* forecast window, to gauge whether the sensitivity also varies across different forecast steps into the future. The scale parameter for the SVSS prior ($\underline{h}_0^{svss}$) is omitted as the derivative was too erratic, even in the case of a smaller parameter space and increased resolution of the derivative approximation. The results for the derivatives on the basis of the *h-step* are presented in figure 4, 5 & 6. One of the first observations that one can make is the difference between the shorter forecast windows and the longer forecast windows. This is present in each of the priors. Aside from the $\underline{g}_0^{horseshoe}$, in the other priors the shorter forecast windows, especially 1-step and 2-step, seem to be more sensitive to shrinkage, whereas the longer forecast windows do not exhibit the same sensitivity. For example, in the case of $\underline{g}_0^{lasso}$, all the derivatives of the forecast windows above the 2-step window are relatively small in size, while this is not the case for the 1-step and 2-step forecast windows. This is also present in $\underline{h}_0^{lasso}$, although the derivatives of the forecast windows are still positive, the size of the derivative for the 1-step until 4-step forecast windows is materially different. Furthermore, in the case of the SVSS prior, for the $\underline{g}_0^{svss}$ parameter, it is the case that the 1-step until the 4-step forecast windows are relatively susceptible to change in the level of shrinkage, while this is not the case for the 5-step until 8-step forecast windows. The insensitivity increases the longer the forecasts horizon is extended, this is especially visible in the Lasso prior. This monotonic decrease in sensitivity is also corroborated by the results in Chan et al. (2020). One of the possible explanations according to Chan et al. (2019) is the fact that the longer horizon forecasts converge on the unconditional mean of the system, this can be more precisely estimated than the individual coefficients. Thus, the general structure of the relationships among the variables is more important, in the case of longer horizon forecasts, than a precise estimate of the coefficients. Although the initial parameters cause the derivatives to hint at different optima between the priors (the Lasso prior would benefit from weaker shrinkage, while the Horseshoe would benefit from stronger shrinkage) they all exhibit the behaviour that there is a distinction between the shorter- and longer horizon forecasts.

Figure 1: Derivatives for the SVSS prior

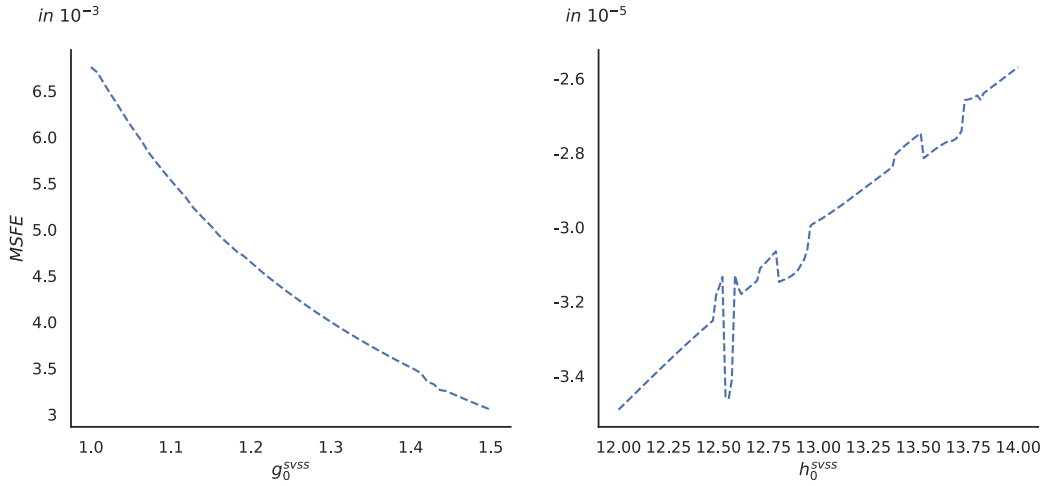


Figure 2: Derivatives for the Lasso prior

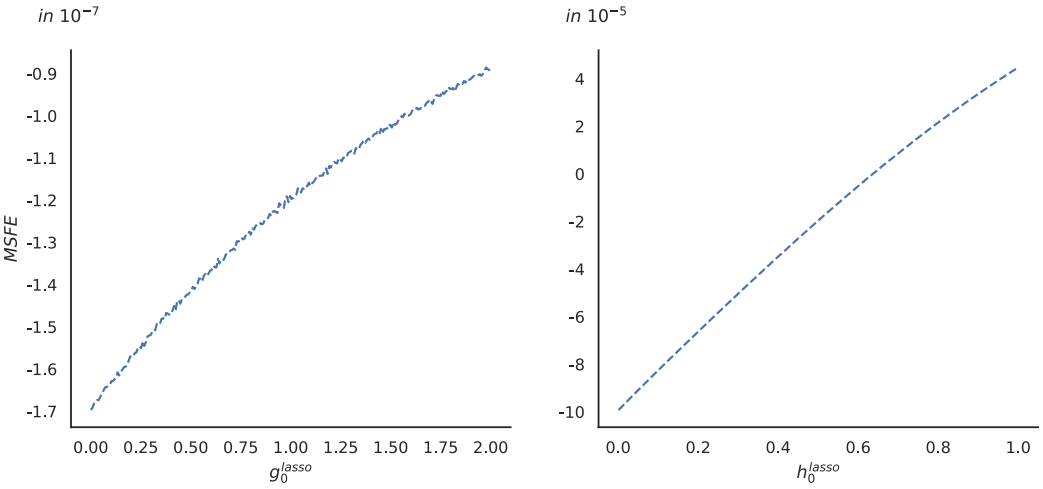


Figure 3: Derivatives for the Horseshoe prior

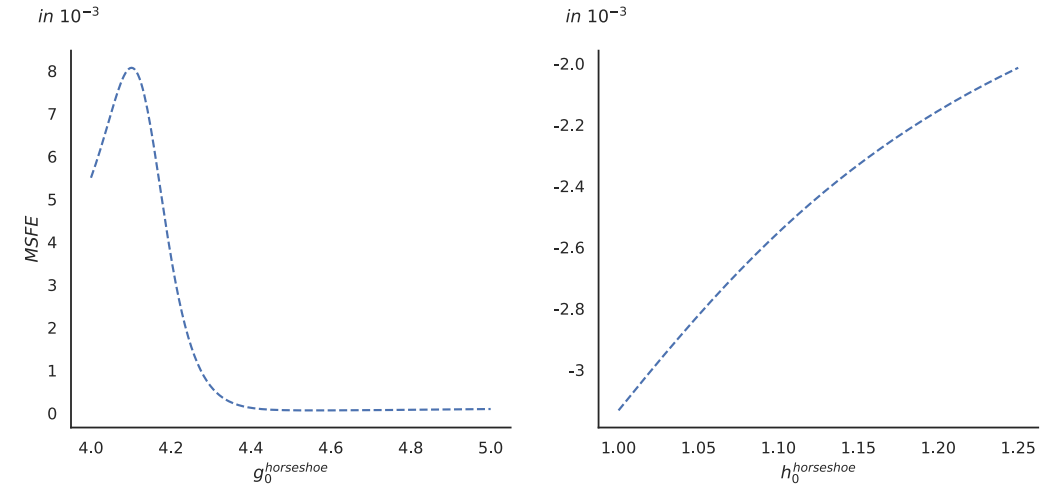


Figure 4: Derivatives on *h-step* basis for the SVSS prior

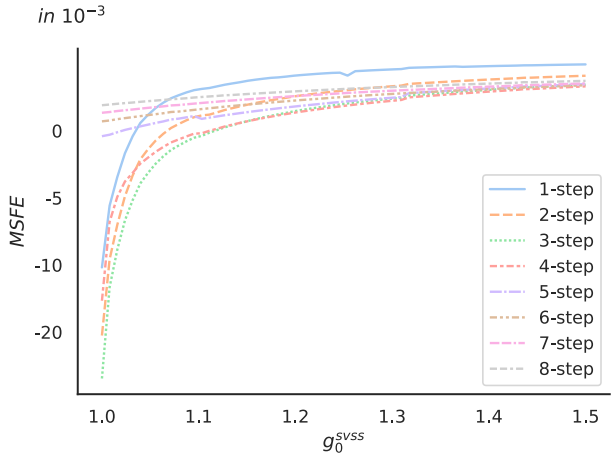


Figure 5: Derivatives on *h-step* basis for the Lasso prior

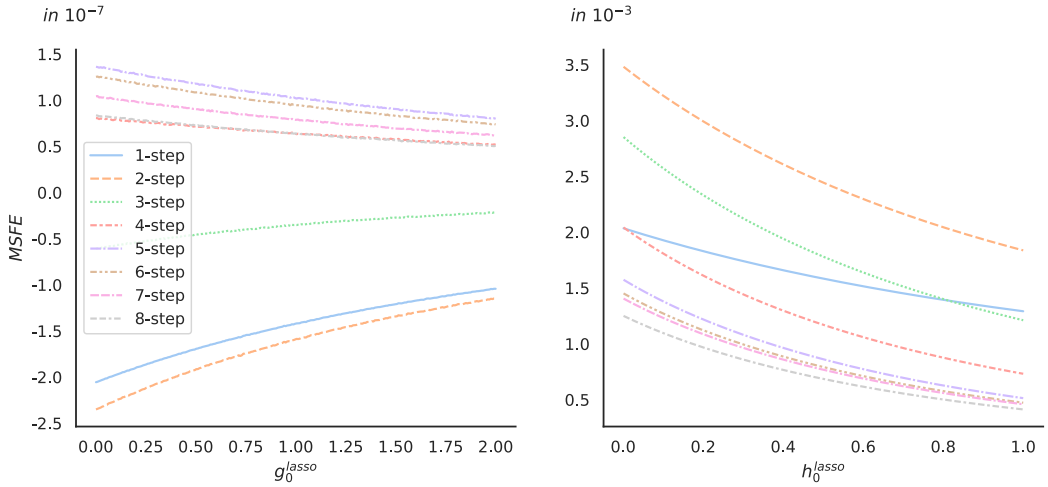
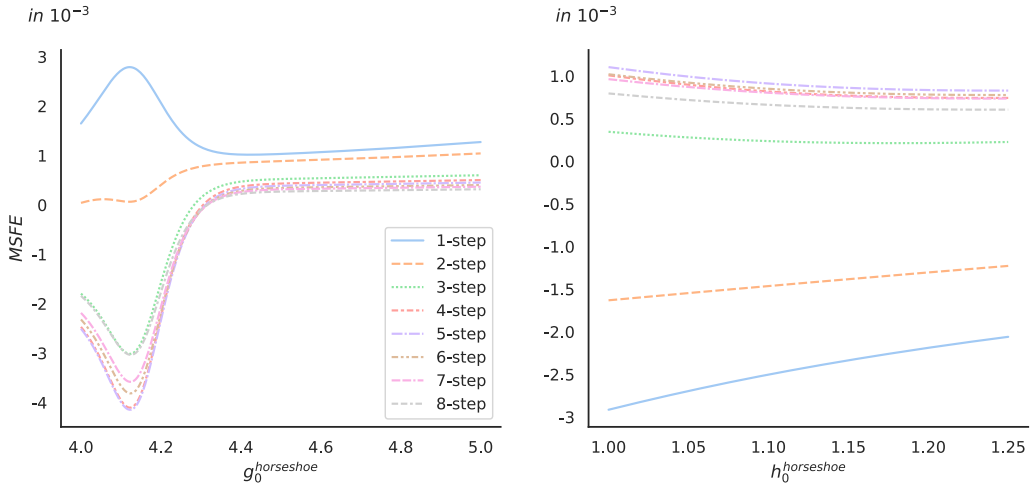


Figure 6: Derivatives on *h-step* basis for the Horseshoe prior



6 Conclusion

The aim of this paper is twofold. First, to determine if VI is a viable alternative in the space of TVP-BVARs for MCMC-based models. Second, to conduct a sensitivity analysis on global-local shrinkage priors that are common in the literature, in this case the SVSS, Lasso and Horseshoe prior. The contribution of this paper is that it shows that a VI-based TVP-BVAR, using a Cholesky decomposition approach to estimate the system in an equation-by-equation manner, is an adequate substitute. Moreover, derivations for the variational posteriors of the Lasso and Horseshoe prior are also a useful addition. In addition to the previous contributions, another contribution is that it also shows that a numerical based sensitivity analysis is feasible in a more complex set of models.

In the simulation exercise, a VI-based TVP-VAR surpasses the point forecast accuracy (MSFE) of the traditional benchmark of a TVP-BVAR (estimated using MCMC), all the while being remarkably faster in the estimation of a model in larger and even smaller systems. This corroborates findings in previous literature [Gefang et al. (2019), Koop and Korobilis (2020)]. Based on the predictive density (ALPL) a TVP-BVAR is still the better model in a smaller system, while this advantage vanishes in a larger system. However, note that this result is limited to the different DGPs that were scrutinized and the specific implementation of VI in a TVP-VAR as the implementation can differ in a multitude of ways. For example, a different decomposition of the approximation distribution, a different use of the Cholesky decomposition of the covariance matrix or allowing for time-varying volatility in a different manner. Nonetheless, it is without doubt that the performance of the three priors in the TVP-VI in the simulation exercise surpasses the TVP-BVAR and is quite similar between the priors, the SVSS prior is slightly better overall. This indifference between the priors might in part be attributed to the fact that the prior is a minor part in the complete TVP-VI specification.

In the empirical exercise, we looked at the local sensitivity of the hyperparameters with respect to the (overall and *h-step* basis) point forecast accuracy. It shows that for the specific parameter spaces that we used, the Horseshoe prior is more sensitive to a change in the parameters than is the case for the Lasso prior. It is difficult to compare the SVSS prior as the data is standardized due to a more favourable speed of convergence. With respect to the *h-step* basis point forecast accuracy, the priors exhibited very similar behaviour to one another as in almost each prior, although the Horseshoe prior deviates slightly, the longer forecast horizons are less susceptible to a change in shrinkage than the shorter forecast horizons. This is in line with earlier literature [Chan et al. (2019)]. One plausible explanation, according to Chan et al. (2019), is that the longer forecast horizons converge on the unconditional mean of the system, this can be more precisely estimated than the individual coefficients. This suggests that the general structure of the relationship is more important for longer horizon forecasts than a precise estimate of the coefficients. Although, VI facilitates the inspection of the local sensitivity of the priors, the numerical approach to derivative approximation — finite differences — has its drawbacks and numerical precision limitations. In this instance, the function that we approximate the derivative of is quite complex. The results of the

approximation procedure were sometimes nonsensical or it would require a too high resolution of the parameter space (the number of points at which one evaluates the derivative) that it becomes an infeasible effort. Nonetheless, we reviewed the local sensitivity of several priors and arrive at similar conclusion with respect to forecast horizons presented in earlier literature Chan et al. (2019).

7 Discussion

Although VI has several advantages, there are some drawbacks that hinder the ease of experimentation with this technique. For each instance that one would like to use VI, the variational posteriors for each set of parameters has to be derived analytically. Sometimes, in simpler models this can be rather straightforward, and it is often very similar to the derivations for a MCMC scheme. Nonetheless, this is an additional step that one has to consider when one makes a choice for a specific model. The reason that it is not identical to the case where one would like to resort to MCMC, is the fact that there are programming languages, specifically probabilistic programming languages (PPLs), such as Stan and libraries, such as PyMC4, that ease the use of MCMC. In these languages, the analytical derivations for sampling are not necessary or at least very limited, as one can specify the model in the syntax itself. It has to be noted that, sometimes the MCMC algorithms that are supported in those languages are not sufficient for complex models, where it can be difficult to attain convergence with the default algorithms (Hamiltonian MCMC or NUTS). An interesting development that might open the avenue to a better experience in experimentation with, and subsequent adoption of VI is Automatic Differentiation Variational Inference (ADVI) (Kucukelbir et al., 2016). As one might infer from the name, it frees the researcher or practitioner from deriving the analytical posteriors. In hindsight, this might have been a better path than the derivation by hand, and it is perhaps less prone to errors in derivation as well as in technical implementation. Although, the latest implementation in Stan is used by researchers and practitioners alike, it is still in the experimental stage.

Although VI opens up the possibility to model high-dimensional problems in Bayesian approaches, in this instance, it still had its limitations due to the variational posterior of $\boldsymbol{\beta}_t$. If one estimates a TVP-VAR as a complete system, the Kalman filter is the limiting factor, as is well known, the Kalman filter does not scale that well with an increase in the number of states, due to the inversion of the covariance matrix of the states. There is some work in this area that might be of interest for a further exploration of the idea of VI in a TVP-VAR context. Such as, the Ensemble Kalman Filter (EnKF) (Evensen, 2003), this is an approximation of the result that one would obtain with the (original) Kalman filter. Although, resorting to an approximation (EnKF) upon an approximation (VI) is perhaps too far from the true posterior distribution.

As mentioned in the section of the simulation results, in hindsight, a better metric could have been the Mean Percentage Error (MPE). As it is scale-invariant, this would have made the absolute comparisons between the results with the different number of variables more appropriate than the use of absolute MSFE. Furthermore, relying on any measure that uses a log-predictive likelihood

requires to some extent a correction for the size of a model (ceteris paribus, a larger model will likely have a better predictive fit). Alternatives might be the Watanabe–Akaike information criterion (WAIC), this comes with its own limitations as this requires data partitioning in a certain manner and that is not that trivial to achieve in a time-series context.

As mentioned earlier, the derivative approximation approach of finite differences, has some drawbacks and limitations that are inherent to a numerical approach. For example, in the case of the empirical exercise, it was still interesting to inspect the sensitivity in a larger system. However, after extensive experimentation with different components (resolution, the error and parameter space) it was not possible to retrieve sensible results in a larger system. This could be due to the limits of numerical precision as the changes between subsequent evaluations in the parameter space were too small. An interesting avenue of research with a similar setup TVP-VI with different priors might be feasible for larger systems if one opts for automatic differentiation as is done in Chan et al. (2019) for VARs.

References

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference* (Doctoral dissertation). UCL (University College London).

Belmonte, M. A., Koop, G., & Korobilis, D. (2014). Hierarchical Shrinkage in Time-Varying Parameter Models. *Journal of Forecasting*, *33*(1), 80–94. https://doi.org/10.1002/for.2276

Berger, J. O., Insua, D. R., & Ruggeri, F. (2000). Bayesian Robustness. In D. R. Insua & F. Ruggeri (Eds.), *Robust Bayesian Analysis* (pp. 1–32). Springer New York. https://doi.org/10.1007/978-1-4612-1306-2_1

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag. https://www.microsoft.com/en-us/research/people/cmbishop/prml-book/

Bitto, A., & Frühwirth-Schnatter, S. (2019). Achieving shrinkage in a time-varying parameter model framework. *Journal of Econometrics*, *210*(1), 75–97. https://doi.org/10.1016/j.jeconom.2018.11.006

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, *7*(4), 434–455.

Carriero, A., Clark, T. E., & Marcellino, M. (2015). Bayesian VARs: Specification Choices and Forecast Accuracy. *Journal of Applied Econometrics*, *30*(1), 46–73. https://doi.org/10.1002/jae.2315

Carriero, A., Clark, T. E., & Marcellino, M. (2019). Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors. *Journal of Econometrics*, *212*(1), 137–154. https://doi.org/10.1016/j.jeconom.2019.04.024

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480. http://www.jstor.org/stable/25734098

Chan, J. C. C., & Eisenstat, E. (2018). Bayesian model comparison for time-varying parameter VARs with stochastic volatility. *Journal of Applied Econometrics*, *33*(4), 509–532.

Chan, J. C. C., Eisenstat, E., & Strachan, R. W. (2020). Reducing the state space dimension in a large TVP-VAR. *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2019.11.006

Chan, J. C. C., Jacobi, L., & Zhu, D. (2019). How Sensitive Are VAR Forecasts to Prior Hyperparameters? An Automated Sensitivity Analysis. *Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling: Part A (Advances in Econometrics)* (pp. 229–248). https://doi.org/10.1108/S0731-90532019000040A010

Clark, T. E. (2011). Real-Time Density Forecasts From Bayesian Vector Autoregressions With Stochastic Volatility. *Journal of Business & Economic Statistics*, *29*(3), 327–341. https://doi.org/10.1198/jbes.2010.09248

Clark, T. E., & Ravazzolo, F. (2015). Macroeconomic Forecasting Performance under Alternative Specifications of Time-Varying Volatility. *Journal of Applied Econometrics*, *30*(4), 551–575. https://doi.org/10.1002/jae.2379

Cogley, T., & Sargent, T. J. (2001). Evolving Post-World War II U.S. Inflation Dynamics. *NBER Macroeconomics Annual*, *16*, 331–373. https://doi.org/10.1086/654451

Cogley, T., & Sargent, T. J. (2005). Drifts and volatilities: monetary policies and outcomes in the post WWII US. *Review of Economic Dynamics*, *8*(2), 262–302. https://doi.org/10.1016/j.red.2004.10.009

D'Agostino, A., Gambetti, L., & Giannone, D. (2013). Macroeconomic forecasting and structural change. *Journal of Applied Econometrics*, *28*(1), 82–101. https://doi.org/10.1002/jae.1257

De Mol, C., Giannone, D., & Reichlin, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? [Honoring the research contributions of Charles R. Nelson]. *Journal of Econometrics*, *146*(2), 318–328. https://doi.org/10.1016/j.jeconom.2008.08.011

de Jong, P., & Shephard, N. (1995). The Simulation Smoother for Time Series Models. *Biometrika*, *82*(2), 339–350. https://doi.org/10.2307/2337412

Doan, T., Litterman, R., & Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, *3*, 1–100. https://doi.org/10.1080/07474938408800053

Durbin, J., & Koopman, S. J. (2012). *Time Series Analysis by State Space Methods* (2nd ed.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199641178.001.0001

Evensen, G. The ensemble kalman filter: Theoretical formulation and practical implementation. In: *Seminar on recent developments in data assimilation for atmosphere and ocean, 8-12 september 2003*. ECMWF. Shinfield Park, Reading: ECMWF, 2003, 221–264. https://www.ecmwf.int/node/9321

Feynman, R. P., Leighton, R. B., & Sands, M. (1984). *The Feynman Lectures of Physics* (Vol. 2). Addison-Wesley. https://www.feynmanlectures.caltech.edu/II_19.html

Gefang, D., Koop, G., & Poon, A. (2019). *Variational Bayesian Inference in Large Vector Autoregressions with Hierarchical Shrinkage* (working paper). University of Leicester and University of Strathclyde. https://bit.ly/2ySyH5W

Gelman, A., Carlin, J. B., Stern, H. S., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Taylor & Francis. http://www.stat.columbia.edu/~gelman/book/

George, E. I., & McCulloch, R. E. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, *88*(423), 881–889. https://doi.org/10.1080/01621459.1993.10476353

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

Geyer, C. (2011). Introduction to MCMC. In S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 3–48). Chapman & Hall/CRC. https://doi.org/10.1201/b10905

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

Greenberg, E. (2012). *Introduction to Bayesian Econometrics* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139058414

Gustafson, P. (2000). Local robustness in bayesian analysis. In D. R. Insua & F. Ruggeri (Eds.), *Robust bayesian analysis* (pp. 71–88). Springer New York. https://doi.org/10.1007/978-1-4612-1306-2_4

Hajargasht, R. (2019). *Approximation Properties of Variational Bayes for Vector Autoregressions* (working paper). Swinburne Business School. https://arxiv.org/abs/1903.00617

Huber, F., Koop, G., & Onorante, L. (2020). Inducing Sparsity and Shrinkage in Time-Varying Parameter Models. *Journal of Business & Economic Statistics*, *0*(0), 1–15. https://doi.org/10.1080/07350015.2020.1713796

Jacobi, L., Joshi, M., & Zhu, D. (2018). Automated sensitivity analysis for bayesian inference via markov chain monte carlo: Applications to gibbs sampling. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2984054

Knaus, P., Bitto-Nemling, A., Cadonna, A., & Frühwirth-Schnatter, S. (2019). *shrinkTVP: Efficient Bayesian Inference for Time-Varying Parameter Models with Shrinkage* [R package version 1.1.1]. https://CRAN.R-project.org/package=shrinkTVP

Koop, G. (2003). *Bayesian Econometrics* (1st ed.). John Wiley & Sons.

Koop, G., & Korobilis, D. (2010). 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., & Korobilis, D. (2013). time-varying parameter VARs. *Journal of Econometrics*, *177*(2), 185–198. https://doi.org/10.1016/j.jeconom.2013.04.007

Koop, G., & Korobilis, D. (2018). *Variational Bayes inference in high-dimensional time-varying parameter models* (working paper). University of Strathclyde and University of Essex. https://bit.ly/34qANWo

Koop, G., & Korobilis, D. (2020). *Bayesian Dynamic Variable Selection in High Dimensions* (Revise & Resubmit). Journal of Econometrics. https://doi.org/10.2139/ssrn.3246472

Koop, G., Korobilis, D., & Pettenuzzo, D. (2019). Bayesian compressed vector autoregressions. *Journal of Econometrics*, *210*(1), 135–154. https://doi.org/10.1016/j.jeconom.2018.11.009

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2016). Automatic differentiation variational inference.

Kuschnig, N., & Vashold, L. (2020). *BVAR: Hierarchical Bayesian vector autoregression* [R package version 1.0.0]. https://CRAN.R-project.org/package=BVAR

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

Lopes, H. F., McCulloch, R. E., & Tsay, R. S. (2018). *Parsimony inducing priors for large scale state-space models* (Revise & Resubmit). Bayesian Analysis. https://bit.ly/3iaiuKp

Lopes, H. F., & Tobias, J. L. (2011). Confronting prior convictions: On issues of prior sensitivity and likelihood robustness in bayesian analysis. *Annual Review of Economics*, *3*(1), 107–131. https://doi.org/10.1146/annurev-economics-111809-125134

Nocedal, J., & Wright, S. J. (2006). *Numerical Optimization*. Springer. https://link.springer.com/book/10.1007/978-0-387-40065-5

Pfaff, B. (2008a). Var, svar and svec models: Implementation within R package vars. *Journal of Statistical Software*, *27*(4). http://www.jstatsoft.org/v27/i04/

Pfaff, B. (2008b). Var, svar and svec models: Implementation within R package vars. *Journal of Statistical Software*, *27*(4). http://www.jstatsoft.org/v27/i04/

Primiceri, G. E. (2005). Time Varying Structural Vector Autoregressions and Monetary Policy. *The Review of Economic Studies*, *72*(3), 821–852. http://www.jstor.org/stable/3700675

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Stock, J. H., & Watson, M. W. (2012). *Disentangling the channels of the 2007-2009 recession* (Working Paper No. 18094). National Bureau of Economic Research. https://doi.org/10.3386/w18094

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. http://www.jstor.org/stable/2346178

Wang, H., Yu, H., Hoy, M., Dauwels, J., & Wang, H. (2016). Variational Bayesian dynamic compressive sensing. *2016 IEEE International Symposium on Information Theory (ISIT)*, 1421–1425. https://doi.org/10.1109/ISIT.2016.7541533

West, M., & Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models* (2nd ed.). Springer-Verlag. https://link.springer.com/book/10.1007/b98971

Appendices

Appendix A Derivations

A.1 Full joint pdf of an (homoskedastic) SSM

For reference, this derivation is also presented in multiple standard textbooks, such as Durbin and Koopman (2012). In the univariate case, $\mathbf{\Sigma}$ is often interchanged with $\boldsymbol{\sigma}^2$.

$$
\begin{aligned}
p(y_{1:T}, \beta_{0:T}, \Sigma, Q_{1:T}) &= p(y_{1:T}|\beta_{0:T}, \Sigma, Q_{1:T})p(\beta_{0:T}, \Sigma, Q_{1:T}) \\
&= p(\beta_{0:T}|\Sigma, Q_{1:T})p(\Sigma, Q_{1:T})p(y_1|y_{2:T}, \beta_{0:T}, \Sigma)p(y_{2:T}|\beta_{0:T}, \Sigma) \\
&= p(\beta_{0:T}|\Sigma, Q_{1:T})p(\Sigma, Q_{1:T})p(y_1|\beta_1, \Sigma)p(y_{2:T}|\beta_{0:T}, \Sigma) \\
&= p(\beta_{0:T}|\Sigma, Q_{1:T})p(\Sigma, Q_{1:T})\prod_{t=1}^{T} p(y_t|\beta_t, \Sigma) \\
&= p(\beta_0)p(\beta_{1:T}|Q_{1:T})p(\Sigma)p(Q_{1:T})\prod_{t=1}^{T} p(y_t|\beta_t, \Sigma) \\
&= p(\beta_0)p(\beta_1|\beta_{2:T}, Q_{1:T})p(\beta_{2:T}|Q_{1:T})p(\Sigma)p(Q_{1:T})\prod_{t=1}^{T} p(y_t|\beta_t, \Sigma) \\
&= p(\beta_0)p(\beta_1|\beta_2, Q_1)p(\beta_{2:T}|Q_{1:T})p(\Sigma)p(Q_{1:T})\prod_{t=1}^{T} p(y_t|\beta_t, \Sigma) \\
&= p(\beta_0)p(\Sigma)\prod_{t=1}^{T} p(\beta_t|\beta_{t-1}, Q_t)p(y_t|\beta_t, \Sigma)p(Q_t)
\end{aligned}
\tag{38}
$$

A.2 Variational posteriors: Horseshoe prior

For each variational posterior we have to derive the optimal distribution. This optimal distribution can be derived based on the result in equation (14). Furthermore, as we factorize along these parameters, we have the same starting point for each variational posterior. The expectation $\mathbb{E}_{q(\theta|y_{1:t})}[\cdot]$ is with respect to all the parameters that we are not deriving the variational posterior for in that instance. Thus, in the case of variational posterior $q(\lambda_t|y_{1:t})$, the parameters gathered in θ are $\{\tilde{\boldsymbol{\beta}}_{0:t}, \tilde{\boldsymbol{Q}}_{1:}, \phi_{1:t}, \nu_{1:t}, \varphi_{1:t}, \sigma^2_{1:t}\}$. Furthermore, the process of deriving these variational posteriors is very similar for all the parameters. Therefore, only λ_t will be annotated.

A.2.1 $q(\lambda_t|y_{1:t})$

First, we start of with the full posterior:

$$
\begin{aligned}
q(\lambda_t|y_{1:t}) &\propto \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(p(\tilde{\beta}_{0:t}, \tilde{Q}_{1:t}, \lambda_{1:t}, \varphi_{1:t}, \phi_{1:t}, \nu_{1:t}, \sigma^2_{1:t}|y_{1:t}, z_{1:t})]) \\
&= \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(p(\tilde{\beta}_0) \prod_{i=1}^{t} p(y_i|\tilde{\beta}_i, \sigma^2_i) p(\lambda_i|\varphi_i) p(\varphi_i) p(\sigma^2_i) \prod_{j=1}^{k} p(\tilde{\beta}_{j,i}|\tilde{\beta}_{i-1,j}, \tilde{Q}_{j,i}) \\
&\times p(z_{j,i}|\beta_{j,i}, \lambda_i\phi_{j,i}) p(\phi_{j,i}|\nu_{j,i}) p(\nu_{j,i}) p(\tilde{Q}_{j,i}))]) \ .
\end{aligned}
$$

Then we can absorb all the terms that are not relevant into the constant, these are all the distributions not related to time t or related to λ_t. Following the definition of $p(z_{j,t}|\beta_{j,y}, \lambda_t\phi_{j,y})$, we can re-specify it in terms of the actual distributions:

$$
\begin{aligned}
&\propto \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(\prod_{j=1}^{k} p(z_{j,t}|\tilde{\beta}_{j,y}, \lambda_t\phi_{j,y}) p(\lambda_t|\varphi_t))]) \\
&= \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(\prod_{j=1}^{k} N(\tilde{\beta}_{j,t}|0, \lambda_t\phi_{j,t}) IG(\lambda_t|1/2, 1/\varphi_t))]) \ .
\end{aligned}
$$

Writing out the distributions and again absorbing all the terms that are not directly related to λ_t into the constant. Furthermore, the essential part here is that λ_t can be taken out of the expectation $\mathbb{E}_{q(\theta|y_{1:t})}[\cdot]$ as this expectation is not taken with respect to λ_t. As a consequence the $\exp(\cdot)$ disappears due to the $ln(\cdot)$. We have the following:

$$
\begin{aligned}
&\propto \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(\lambda_t^{-k} \exp(-\lambda_t^{-1} \frac{1}{2} \sum_{j=1}^{k} (\phi_{j,t}^{-1} \tilde{\beta}^2_{j,t})) \times \lambda_t^{-1-1/2} \exp(-[\varphi_t\lambda_t]^{-1}))]) \\
&= \lambda_t^{-1-(k+1/2)} \exp(\mathbb{E}_{q(\theta|y_{1:t})}[-\lambda_t^{-1} \frac{1}{2} \sum_{j=1}^{k} (\phi_{j,t}^{-1} \tilde{\beta}^2_{j,t}) - [\varphi_t\lambda_t]^{-1}]) \ .
\end{aligned}
$$

As the expectation is linear (*expected value of the sum of random variables is the sum of the expected values*), we can separate out the expectation for each parameter. The expectation of $\tilde{\beta}^2_{j,t}$ can be derived using the elementary result $\mathbb{E}[X^2] = \mathbb{E}[X]^2 + Var(X)$. After inspection the familiar

form of the Inverse-Gamma distributions presents itself:

$$= \lambda_t^{-1-(k+1/2)} \exp(-\lambda_t^{-1}(\frac{1}{2}\sum_{j=1}^{k} \mathbb{E}_{q(\phi_{j,t}|y_{1:t})}[\phi_{j,y}^{-1}]\, \mathbb{E}_{q(\tilde{\beta}_{j,t}|y_{1:t})}[\tilde{\beta}_{j,t}^2] + \mathbb{E}_{q(\varphi_t|y_{1:t})}[\varphi_t^{-1}]))$$

$$= \lambda_t^{-1-(k+1/2)} \exp(-\lambda_t^{-1}(\frac{1}{2}\sum_{j=1}^{k} \hat{\phi}_{j,t}^{-1}(\hat{\beta}_{j,t}^2 + \hat{P}_{jj,t}) + \hat{\varphi}_t^{-1}])) \ ,$$

$$q(\lambda_t|y_{1:t}) \sim IG(k + 1/2, \hat{\varphi}_t^{-1} + \frac{1}{2}\sum_{j=1}^{k} \phi_{j,t}^{-1}(\hat{\beta}_{j,t}^2 + \hat{P}_{jj,t})) \ . \tag{39}$$

A.2.2 $q(\phi_{j,t}|y_{1:t})$

$$\begin{aligned}
q(\phi_{j,t}|y_{1:t}) &\propto \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(p(\tilde{\beta}_{0:t}, \tilde{Q}_{1:t}, \lambda_{1:t}, \varphi_{1:t}, \phi_{1:t}, \nu_{1:t}, \sigma_{1:t}^2|y_{1:t}, z_{1:t})])\\
&= \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(p(\tilde{\beta}_0)\prod_{i=1}^{t} p(y_i|\tilde{\beta}_i, \sigma_i^2)p(\lambda_i|\varphi_i)p(\varphi_i)p(\sigma_i^2)\prod_{j=1}^{k} p(\tilde{\beta}_{j,i}|\tilde{\beta}_{i-1,j}, \tilde{Q}_{j,i})\\
&\times p(z_{j,i}|\tilde{\beta}_{j,i}, \lambda_i\phi_{j,i})p(\phi_{j,i}|\nu_{j,i})p(\nu_{j,i})p(\tilde{Q}_{j,i}))])\\
&\propto \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(p(z_{j,t}|\tilde{\beta}_{j,t}, \lambda_t\phi_{j,t})p(\phi_{j,t}|\nu_{j,t}))])\\
&= \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(N(\tilde{\beta}_{j,t}|0, \lambda_t\phi_{j,t})IG(\phi_{j,t}|1/2, 1/\nu_{j,t})])))\\
&\propto \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(\phi_{j,t}^{-1}\exp(-\frac{1}{2}[\lambda_t\phi_{j,t}]^{-1}\tilde{\beta}_{j,t}^2) \times \phi_{j,t}^{-1-1/2}\exp(-[\phi_{j,t}\nu_{j,t}]^{-1}))])\\
&= \phi_{j,t}^{-1-(1+1/2)}\exp(-\frac{1}{2}\phi_{j,t}^{-1}(\mathbb{E}_{q(\lambda_{j,t}|y_{1:t})}[\lambda_t^{-1}]\, \mathbb{E}_{q(\tilde{\beta}_{j,t}|y_{1:t})}[\tilde{\beta}_{j,t}^2] + \mathbb{E}_{q(\nu_{j,t}|y_{1:t})}[\nu_{j,t}^{-1}]))\\
&= \phi_{j,t}^{-1-(1+1/2)}\exp(-\frac{1}{2}\phi_{j,t}^{-1}(\hat{\lambda}_t^{-1}(\hat{\beta}_{j,t}^2 + \hat{P}_{t,jj}) + \hat{\nu}_{j,t}^{-1}))
\end{aligned}$$

$$q(\phi_{j,t}|y_{1:t}) \sim IG(1 + 1/2, \hat{\nu}_{j,t}^{-1} + \frac{1}{2}\frac{\hat{\beta}_{j,t} + \hat{P}_{t,jj}}{\hat{\lambda}_t}) \tag{40}$$

A.2.3 $q(\nu_{j,t}|y_{1:t})$

$$\begin{aligned}
q(\nu_{j,t}|y_{1:t}) &\propto \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(p(\tilde{\beta}_{0:t}, \tilde{Q}_{1:t}, \lambda_{1:t}, \varphi_{1:t}, \phi_{1:t}, \nu_{1:t}, \sigma_{1:t}^2|y_{1:t}, z_{1:t})])\\
&= \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(p(\tilde{\beta}_0)\prod_{i=1}^{t} p(y_i|\tilde{\beta}_i, \sigma_i^2)p(\lambda_i|\varphi_i)p(\varphi_i)p(\sigma_i^2)\prod_{j=1}^{k} p(\tilde{\beta}_{j,i}|\tilde{\beta}_{i-1,j}, \tilde{Q}_{j,i})\\
&\times p(z_{j,i}|\tilde{\beta}_{j,i}, \lambda_i\phi_{j,i})p(\phi_{j,i}|\nu_{j,i})p(\nu_{j,i})p(\tilde{Q}_{j,i}))])\\
&\propto \exp(\mathbb{E}_{q(\phi_{j,t}|y_{1:t})}[\ln(p(\phi_{j,t}|\nu_{j,t})p(\nu_{j,t}))])\\
&= \exp(\mathbb{E}_{q(\phi_{j,t}|y_{1:t})}[\ln(IG(\phi_{t,j|1/2,1/\nu_{j,t}})IG(\nu_{j,t}|\underline{g}_0^{horseshoe}, \underline{h}_0^{horseshoe}))])\\
&\propto \exp(\mathbb{E}_{q(\phi_{j,t}|y_{1:t})}[\ln(\exp(-[\phi_{j,t}\nu_{j,t}]^{-1}) \times \nu_{j,t}^{-1-\underline{g}_0^{horseshoe}}\exp(-\nu_{j,t}^{-1}\underline{h}_0^{horseshoe}))])\\
&= \nu_{j,t}^{-1-\underline{g}_0^{horseshoe}}\exp(-\nu_{j,t}(\underline{h}_0^{horseshoe} + \mathbb{E}_{q(\phi_{j,t}|y_{1:t})}[\phi_{j,t}^{-1}]))\\
&= \nu_{j,t}^{-1-\underline{g}_0^{horseshoe}}\exp(-\nu_{j,t}(\underline{h}_0^{horseshoe} + \hat{\phi}_{j,t}^{-1}))
\end{aligned}$$

$$q(\nu_{j,t}|y_{1:t}) \sim IG(\underline{g}_0^{horseshoe}, \underline{h}_0^{horseshoe} + \hat{\phi}_{j,t}^{-1}) \tag{41}$$

A.2.4 $q(\varphi_t|y_{1:t})$

$$\begin{aligned}
q(\varphi_t|y_{1:t}) &\propto \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(p(\tilde{\beta}_{0:t}, \tilde{Q}_{1:t}, \lambda_{1:t}, \varphi_{1:t}, \phi_{1:t}, \nu_{1:t}, \sigma_{1:t}^2|y_{1:t}, z_{1:t})])\\
&= \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(p(\tilde{\beta}_0)\prod_{i=1}^{t} p(y_i|\tilde{\beta}_i, \sigma_i^2)p(\lambda_i|\varphi_i)p(\varphi_i)p(\sigma_i^2)\prod_{j=1}^{k} p(\tilde{\beta}_{j,i}|\tilde{\beta}_{i-1,j}, \tilde{Q}_{j,i})\\
&\times p(z_{j,i}|\tilde{\beta}_{j,i}, \lambda_i\phi_{j,i})p(\phi_{j,i}|\nu_{j,i})p(\nu_{j,i})p(\tilde{Q}_{j,i}))])\\
&\propto \exp(\mathbb{E}_{q(\lambda_t|y_{1:t})}[\ln(p(\lambda_t|\varphi_t)p(\varphi_t))])\\
&= \exp(\mathbb{E}_{q(\lambda_t|y_{1:t})}[\ln(IG(\lambda_t|1/2, 1/\varphi_t)IG(\varphi_t|\underline{g}_0^{horseshoe}, \underline{h}_0^{horseshoe}))])\\
&\propto \exp(\mathbb{E}_{q(\lambda_t|y_{1:t})}[\ln(\exp(-[\lambda_t\varphi_t]^{-1}) \times \varphi_t^{-1-\underline{h}_0^{horseshoe}} \exp(\varphi_t^{-1}\underline{g}_0^{horseshoe}))])\\
&= \varphi^{-1-\underline{g}_0^{horseshoe}} \exp(-\varphi_t^{-1}(\underline{h}_0^{horseshoe} + \mathbb{E}_{q(\lambda_t|y_{1:t})}[\lambda_t^{-1}]))\\
&= \varphi^{-1-\underline{g}_0^{horseshoe}} \exp(-\varphi_t^{-1}(\underline{h}_0^{horseshoe} + \hat{\lambda}_t^{-1}))
\end{aligned}$$

$$q(\varphi_t|y_{1:t}) \sim IG(\underline{g}_0^{horseshoe}, \underline{h}_0^{horseshoe} + \hat{\lambda}_t^{-1}) \tag{42}$$

A.3 Variational posteriors: Lasso

The derivations are done in almost an identical manner as the annotated derivation of λ_t in Appendix A.2.

A.3.1 $q(\iota_{j,t}^2|y_{1:t})$

$$\begin{aligned}
q(\iota_{j,t}^2|y_{1:t}) &\propto \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(p(\tilde{\beta}_{0:t}, \tilde{Q}_{1:t}, \iota_{1:t}^2, \psi_{1:t}^2, \sigma_t^2|y_{1:t}, z_{1:t})]))\\
&= \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(p(\tilde{\beta}_0)\prod_{i=1}^{t} p(y_i|\tilde{\beta}_i, \sigma_i^2)p(\sigma_i^2)\prod_{j=1}^{k} p(\tilde{\beta}_{j,i}|\tilde{\beta}_{j,i-1}, \tilde{Q}_{j,i})p(z_{j,i}|\\
tilde\beta_{j,i}, \iota_{j,i}^2)&\\
&\times p(\iota_{j,i}^2|\psi_i^2)p(\psi_i^2)p(\tilde{Q}_{j,i}))])\\
&\propto \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(\prod_{j=1}^{k} p(z_{j,t}|\tilde{\beta}_{j,t}, \iota_{j,t}^2)p(\iota_{j,t}^2|\psi_t^2))])\\
&\propto \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(\prod_{j=1}^{k} N(\tilde{\beta}_{j,t}|0, \iota_{j,t}^2)Exp(\iota_{j,t}^2|\frac{\psi_t^2}{2}))])
\end{aligned}$$

Based on results in Belmonte et al. (2014), the inverse of the above distribution is an *Inverse Gaussian*.

$$q\Big(\frac{1}{\iota_{j,t}^2}|y_{1:t}\Big) \sim \textit{Inverse-Gaussian}\Big(\sqrt{\frac{\hat{\psi}_t^2}{\hat{\boldsymbol{\beta}}_{j,t}^2 + \hat{P}_{jj,t}}}, \hat{\psi}_t^2\Big) \tag{43}$$

A.3.2 $q(\psi_t^2|y_{1:t})$

$$
\begin{aligned}
q(\psi_t^2|y_{1:t}) &\propto \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(p(\tilde{\beta}_{0:t}, \tilde{Q}_{1:t}, \iota_{1:t}^2, \psi_{1:t}^2, \sigma_t^2|y_{1:t}, z_{1:t})])) \\
&= \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(p(\tilde{\beta}_0)\prod_{i=1}^{t} p(y_i|\tilde{\beta}_i, \sigma_i^2)p(\sigma_i^2)\prod_{j=1}^{k} p(\tilde{\beta}_{j,i}|\tilde{\beta}_{j,i-1}, \tilde{Q}_{j,i})p(z_{j,i}|\tilde{\beta}_{j,i}, \iota_{j,i}^2) \\
&\times p(\iota_{j,i}^2|\psi_t^2)p(\psi_t^2)p(\tilde{Q}_{j,i}))]) \\
&\propto \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(p(\psi_t^t)\prod_{j=1}^{k} p(\iota_{j,t}^2|\psi_t^2))]) \\
&= \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(Gamma(\psi_t^t|\underline{g}_0^{lasso}, \underline{h}_0^{lasso})\prod_{j=1}^{k} Exp(\iota_{j,t}^2|\frac{\psi_t^2}{2}))]) \\
&= \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(\psi_t^{2[\underline{g}_0-1]}\exp(-\underline{h}_0^{lasso}\psi_t^2) \times \psi_t^{2+k}\exp(-\frac{\psi_t^2}{2}\sum_{j=1}^{k}\iota_{j,t}^2))]) \\
&= \exp(\mathbb{E}_{q(\theta|y_{1:t})}[\ln(\psi_t^{2[\underline{g}_0+k-1]}\exp(-\psi_t^2(\sum_{j=1}^{k}\iota_{j,t}^2/2 + \underline{h}_0^{lasso})))]) \\
&= \psi_t^{2[\underline{g}_0^{lasso}+k-1]}\exp(-\psi_t^2(\sum_{j=1}^{k}\hat{\iota}_{j,t}^2/2 + \underline{h}_0^{lasso})))
\end{aligned}
$$

$$
q(\psi_t^2|y_{1:t}) \sim Gamma\Big(\underline{g}_0^{lasso} + k, \Big(\sum_{j=1}^{k}\hat{\iota}_{j,t}^2\Big)/2 + \underline{h}_0^{lasso}\Big) \tag{44}
$$

Appendix B Algorithm pseudo-code

This pseudo-code extends the algorithm description in Koop and Korobilis (2020) to accommodate the estimation of a TVP-VAR and the several additional priors.

Algorithm 1 Variational Bayes algorithm for a TVP-VAR with the SVSS, Lasso and Horseshoe priors

for m = 1 **to** M **do**
  Choose values of $\boldsymbol{m}_0, \boldsymbol{P}_0, a_0, b_0, c_{j,0}, d_{j,0}, g_0, h_0$ (for the prior of choice) $\underline{c}$ and δ;
  i = 1
  while $|\boldsymbol{m}_{1:T|T}^{(i)} - \boldsymbol{m}_{1:T:|T}^{(i-1)}| > 10^{-6}$ **do**
    for t = 1 **to** T **do**
      $\tilde{\boldsymbol{W}}_t^{(i)} = diag\Big(\big(q_{1,t}^{-1\ (i-1)} + \upsilon_{k,t}^{-1\ (i-1)}\big)^{-1}, \ldots, \big(q_{k,t}^{-1\ (i-1)} + \upsilon_{k,t}^{-1\ (i-1)}\big)^{-1}\Big)$
      $\tilde{\boldsymbol{F}}_t^{(i)} = \tilde{\boldsymbol{W}}_t\Big(\boldsymbol{W}_t^{(i-1)}\Big)^{-1}$
      $\boldsymbol{m}_{t|t-1}^{i} = \tilde{\boldsymbol{F}}_t^{i}\boldsymbol{m}_{t-1|t-1}^{(i)}$ ▷ Predicted mean
      $\boldsymbol{P}_{t|t-1}^{(i)} = \tilde{\boldsymbol{F}}_t^{(i)}\boldsymbol{P}_{t-1|t-1}^{(i)}\tilde{\boldsymbol{F}}_t^{\prime(i)} + \tilde{\boldsymbol{W}}_t^{(i)}$ ▷ Predicted variance
      $\boldsymbol{K}_t^{(i)} = \boldsymbol{P}_{t|t-1}^{(i)}\boldsymbol{X}_t^{\prime}\Big(\boldsymbol{X}_t\boldsymbol{P}_{t|t-1}^{(i)}\boldsymbol{X}_t^{\prime} + \hat{\sigma}_t^{2\ (i-1)}\Big)^{-1}$ ▷ Kalman gain
      $\boldsymbol{m}_{t|t}^{(i)} = \boldsymbol{m}_{t|t-1}^{(i)} + \boldsymbol{K}_t^{(i)}\Big(y_t - \boldsymbol{X}_t\boldsymbol{m}_{t|t-1}^{(i)}\Big)$ ▷ Filtered mean of $\tilde{\boldsymbol{\beta}}_t$
      $\boldsymbol{P}_{t|t}^{(i)} = \Big(\boldsymbol{I}_k - \boldsymbol{K}_t^{(i)}\Big)\boldsymbol{P}_{t|t-1}^{(i)}$ ▷ Filtered variance of $\tilde{\boldsymbol{\beta}}_t$
    end for
    for T = T-1 **to** 1 **do**
      $\boldsymbol{C} = \boldsymbol{P}_{t|t}^{(i)}\tilde{\boldsymbol{F}}_t^{(i)}\Big(\boldsymbol{P}_{t+1|t}^{(i)}\Big)^{-1}$
      $\boldsymbol{m}_{t|T}^{(i)} = \boldsymbol{m}_{t|t}^{(i)} + \boldsymbol{C}\Big(\boldsymbol{m}_{t+1|T}^{(i)} - \boldsymbol{m}_{t+1|t}^{(i)}\Big)$ ▷ Smoothed mean of $\tilde{\beta}_t$
      $\boldsymbol{P}_{t|T}^{(i)} = \boldsymbol{P}_{t|t}^{(i)} + \boldsymbol{C}\Big(\boldsymbol{P}_{t+1|T}^{(i)} - \boldsymbol{P}_{t+1|t}^{(i)}\Big)\boldsymbol{C}^{\prime}$ ▷ Smoothed variance of $\tilde{\beta}_t$
    end for
    $\boldsymbol{L}_t = \boldsymbol{P}_{t|T}^{(i)} + \boldsymbol{m}_{t|T}^{(i)}\boldsymbol{m}_{t|T}^{(i)\prime} + \Big(\boldsymbol{P}_{t-1|T}^{(i)} + \boldsymbol{m}_{t-1|T}^{(i)}\boldsymbol{m}_{t-1|T}^{(i)\prime}\Big) \times$ ▷ Squared error in state eq.
    $\Big(\boldsymbol{I}_k - 2\tilde{\boldsymbol{F}}_t^{(i)}\Big)$
    $\boldsymbol{R}_t = \Big[\Big(y_t - \boldsymbol{X}_t\boldsymbol{m}_{t|T}^{(i)}\Big)^2 + \boldsymbol{X}_t\boldsymbol{P}_{t|T}^{(i)}\boldsymbol{X}_t^{\prime}\Big]$ ▷ Squared error in state eq.
    if prior = `svss` **then** ▷ Updating the variational posteriors of SVSS
      for t = 1 **to** T **do**
        for j=1 **to** k **do**
          $\tau_{j,t}^{2\ (i)} = (\underline{h}_0^{svss} + \boldsymbol{m}_{j,t|T}^{2\ (i)})/(\underline{g}_0^{svss} + 1/2)$ ▷ Posterior mean of $\tau_{j,t}^{2\ (i)}$
          $\hat{\gamma}_{j,t}^{(i)} = \frac{N(\boldsymbol{m}_{j,t|T}^{(i)}|0,\hat{\tau}_{j,t}^{2\ (i)})\hat{\pi}_{0,t}^{(i)}}{N(\boldsymbol{m}_{j,t|T}^{(i)}|0,\hat{\tau}_{j,t}^{2\ (i)})\hat{\pi}_{0,t}^{(i)} + N(\boldsymbol{m}_{j,t|T}^{(i)}|0,\underline{c}\times\hat{\tau}_{j,t}^{2\ (i)})(1-\hat{\pi}_{0,t}^{(i)})}$ ▷ Posterior mean of $\gamma_{j,t}$
          ▷ Code continues on the next page

$\hat{v}_{j,t}^{svss\ (i)} = (1-\hat{\gamma}_{j,t})^{2\ (i)}\underline{c} \times \hat{\tau}_{j,t}^{2\ (i)} + \hat{\gamma}_{j,t}^{(i)}\hat{\tau}_{j,t}^{2\ (i)}$ ▷ Posterior mean of $v_{j,t}^{svss}$

end for

$\hat{\pi}_{0,t}^{(i)} = \left(1 + \sum_{j=1}^{k}\hat{\gamma}_{j,t}^{(i)}\right)/(2+k)$ ▷ Posterior mean of $\pi_{0,t}$

end for

end if

if prior = `lasso` **then** ▷ Updating the variational posteriors of Lasso

for t = 1 **to** T **do**

for j=1 **to** k **do**

$\hat{\iota}_{j,t}^{2\ (i)} = \left(\sqrt{\frac{\hat{\psi}_t^{2\ (i)}}{\boldsymbol{m}_{j,t|T}^{2\ (i)} + \hat{P}_{jj,t}^{(i)}}}\right)^{-1}$ ▷ Posterior mean of $\iota_{j,t}^2$

$\hat{v}_{j,t}^{lasso\ (i)} = \hat{\iota}_{j,t}^{2\ (i)}$ ▷ Posterior mean of $v_{j,t}^{lasso}$

end for

$\hat{\psi}_t^{2\ (i)} = \frac{k+\underline{g}_0^{lasso}}{\left(\sum_{j=1}^{k}\hat{\iota}_{t,j}^{2\ (i)}\right)/2+\underline{h}_0^{lasso}}$ ▷ Posterior mean of ψ_t^2

end for

end if

if prior = `horseshoe` **then** ▷ Updating the variational posteriors of Horseshoe

for t = 1 **to** T **do**

for j=1 **to** k **do**

$\hat{\phi}_{j,t}^{(i)} = \frac{\hat{\nu}_{j,t}^{(i)} + 1/2(\boldsymbol{m}_{j,t|T}^{2\ (i)} + \hat{P}_{jj,t}^{(i)})\hat{\lambda_t}^{-1\ (i)}}{1+1/2}$ ▷ Posterior mean of $\phi_{j,t}$

$\hat{\nu}_{j,t}^{(i)} = \frac{\underline{h}_0^{horseshoe} + \hat{\phi}_{j,t}^{-1\ (i)}}{\underline{g}_0^{horseshoe}}$ ▷ Posterior mean of $\nu_{j,t}$

$\hat{v}_{j,t}^{horseshoe\ (i)} = \hat{\lambda}_t^{(i)}\hat{\phi}_{j,t}^{(i)}$ ▷ Posterior mean of $v_{j,t}^{horseshoe}$

end for

$\hat{\lambda}_t^{(i)} = \frac{\hat{\varphi_t}^{-1\ (i)} + 1/2\sum_j^k \hat{\phi}_{t,j}^{-1\ (i)}(\boldsymbol{m}_{j,t|T}^{2\ (i)} + \hat{P}_{jj,t}^{(i)})}{k+1/2}$ ▷ Posterior mean of λ_t

$\hat{\varphi}_t^{(i)} = \frac{\underline{h}_0^{horseshoe} + \hat{\lambda_t}^{-1}}{\underline{g}_0^{horseshoe}}$ ▷ Posterior mean of φ_t

end for

end if

for t = 1 **to** T **do** ▷ Updating the variational posteriors of the variances

for j=1 **to** k **do**

$\hat{q}_{j,t}^{(i)} = \frac{\underline{d}_0 + \boldsymbol{L}_{(j,j),t}^{(i)}}{\underline{c}_0 + 1/2}$ ▷ Posterior mean of $q_{j,t}$

end for

$\hat{\chi}_t^{(i)} = (\delta a_t + 1/2)/(\delta b_{t-1} + 1/2\boldsymbol{R}_t)$ ▷ Filtered mean of $\frac{1}{\sigma^2}$

end for

for T = T-1 **to** 1 **do**

$\tilde{\chi}_t^{(i)} = (1-\delta)\hat{\chi}_t^{(i)} + \delta\tilde{\chi}_{t+1}^{(i)}$ ▷ Smoothed mean of $\frac{1}{\sigma^2}$

end for

i = i + 1

end while

Set $\tilde{\beta}_{j,t}^{m} = \boldsymbol{m}_{j,t|T}^{(i)}$, $\boldsymbol{D}_{mm,t} = \tilde{\chi}_t^{-1\ (i)}$

end for ▷ Code continues on the next page

Note that $\tilde{\beta}_{M*p:(M*p+m-1),t} = -\boldsymbol{A}_t^{-1}[M, 1; M, m-1]$ for all $m > 1$ ▷ Submatrix notation
for t = 1 **to** T **do** ▷ Transform the estimates to retrieve the reduced form system
    Initialize $Q_t = \boldsymbol{I}_M$
    Initialize $\beta_t = \mathbf{0}_{(M,M*p+1)}$
    for m = 1 **to** M **do**
        if $m > 1$ **then**
            $Q_t[M, 1; M, m-1] = -\tilde{\beta}_{M*p:(M*p+m-1),t}$
        end if
        $\beta_t = Q_t^{-1}\tilde{\beta}_{(M,M*p+1),t}$ ▷ Retrieve the reduced form coefficients
        $\Sigma_t = Q_t^{-1}\boldsymbol{D}Q_t^{-1'}$ ▷ Retrieve the reduced form covariance matrix
    end for
end for

Appendix C Simulation addendum

In this section the additional information regarding the simulation exercise is presented. Table 4 shows the median runtime for one iteration of each specific model for each scenario in the study. As one can observe, the TVP-BVAR does not handle the increase in dimensionality as well as the TVP-VI or the time-invariant alternatives. Keep in mind that these runtimes are highly dependent on the hardware that is used.

Table 4: *Median runtime for one iteration (in seconds)*

	$T = 100$, $p_0 = 80\%$	$T = 100$, $p_0 = 60\%$	$T = 200$, $p_0 = 80\%$	$T = 200$, $p_0 = 60\%$
$M = 3$*				
Time-varying				
TVP-VI				
- *Svss*	197.50	184.50	**840.00**	**835.00**
- *Lasso*	34.00	34.50	77.00	87.50
- *Horseshoe*	52.50	52.00	114.00	116.0
TVP-BVAR	**228.75**	**212.23**	454.20	457.25
Time-invariant				
BVAR	165.28	190.37	174.89	180.15
VAR	0.97	0.88	0.81	0.89
$M = 7$*				
Time-varying				
TVP-VI				
- *Svss*	488.00	478.00	1272.50	1146.50
- *Lasso*	99.00	95.00	223.00	219.00
- *Horseshoe*	176.50	175.00	427.00	442.50
TVP-BVAR	**1226.07**	**1249.31**	**2379.22**	**2334.28**
Time-invariant				
BVAR	166.78	190.24	176.91	180.15
VAR	1.29	1.21	1.28	1.27

*It is the median runtime on an *Intel i9-9880H (4.80GHz)* with 16GB of RAM.

The multivariate Gelman-Rubin statistic (Brooks & Gelman, 1998), to be more precise the multivariate potential scale reduction factor (MSRPF), is calculated for 20 different simulation (randomly selected)[15] datasets, on all parameters that have to be estimated in the BVAR and TVP-BVAR. The final set of 20 statistics is averaged to have one statistic that provides an indication of convergence in the simulation exercise. The number of chains is set to 4 and the number of MCMC iterations (after burn-in) varies per model and dimensionality. The threshold that is used is 1.2, this is the one suggested by Brooks and Gelman (1998).

Table 5: *Gelman-Rubin statistic*

	$M = 3$*			
	T = 100, $p_0 = 60\%$	*T = 100,* $p_0 = 80\%$	*T = 200,* $p_0 = 60\%$	*T = 200,* $p_0 = 80\%$
Time-varying				
TVP-BVAR	1.071** (1000)	1.073** (1000)	1.116** (1000)	1.119** (1000)
Time-invariant				
BVAR	1.004** (5000)	1.010*** (5000)	1.004** (5000)	1.004** (5000)
	$M = 7$*			
Time-varying				
TVP-BVAR	1.144** (500)	1.137** (500)	1.271 (500)	1.270 (500)
Time-invariant				
BVAR	1.011** (2500)	1.013** (2500)	1.010** (2500)	1.009** (2500)

*It is the average runtime on an Intel i9-9880H (4.80GHz) with 16GB of RAM.
** It is below the threshold of 1.2

[15]The datasets are equal across the models and scenario's. At initialization of the procedure they are selected at random from the complete set of 200 simulation datasets

The results for the comparison between the different scenarios to isolate the effect of sparsity, different time horizons and the low- and high dimensional setup is presented here. We solely looked at the difference in relative performance based on the average *h-step* MSFE.

Table 6: Average differences in *h-step* MSFE

Sparsity		
Scenario	*time-varying*	*time-invariant*
1 vs. 2	-0.010	0.004
3 vs. 4	-0.070	0.008
5 vs. 6	-0.015	-0.009
7 vs. 8	-0.002	0.194
Time horizon		
Scenario	*time-varying*	*time-invariant*
1 vs. 3	0.090	-0.006
2 vs. 4	0.030	-0.003
5 vs. 7	0.030	-0.188
6 vs. 8	0.020	0.016
Dimensionality		
Scenario	time-varying	time-invariant
1 vs. 5	0.020	0.057
2 vs. 6	0.015	0.044
3 vs. 7	-0.037	-0.124
4 vs. 8	0.010	0.063

Appendix D Reproducibility and code

All results should be reproducible with the code that is publicly available on GitLab (link: https://gitlab.com/cavriends/tvp_vars_vb). However, there might be some slight differences due to the different RNGs that are used on different systems. Nonetheless, the results should be in the vicinity of the results presented in this paper. It should be noted, that, although this code is thoroughly tested and scrutinized, it is still possible that some errors are still in there. It should not hinder the reproducibility. If it hinders reproducibility, or if you have managed to find a serious error, please do not hesitate to reach out[16]. Furthermore, I would like to credit the work of Koop and Korobilis (2020) (and the working paper, [Koop and Korobilis (2018)]) for their ideas and MATLAB implementation, without their work to build upon, this would not have been possible.

[16] Via cavriends@gmail.com or www.linkedin.com/in/cornevriends/