International Journal of Forecasting 41 (2025) 1589–1619

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecasting macroeconomic data with Bayesian VARs: Sparse or dense? It depends! ☆

Luis Gruber *, Gregor Kastner

Department of Statistics, University of Klagenfurt, Austria

ARTICLE INFO

Article history:
Dataset link: https://github.com/luisgruber/GK-IJF-BVAR

Keywords:
Density forecasting
Hierarchical priors
Illusion of sparsity
(Semi-)global–local shrinkage
Stochastic volatility

ABSTRACT

Vector autoregressions (VARs) are widely applied when it comes to modeling and forecasting macroeconomic variables. In high dimensions, however, they are prone to overfitting. Bayesian methods—more concretely, shrinkage priors—have been shown to be successful at improving prediction performance. In the present paper, we introduce the semi-global framework, in which we replace the traditional global shrinkage parameter with group-specific shrinkage parameters. We show how this framework can be applied to various shrinkage priors, such as global–local priors and stochastic search variable selection priors. We demonstrate the virtues of the proposed framework in an extensive simulation study and in an empirical application forecasting data on the US economy. Further, we shed more light on the ongoing 'illusion of sparsity' debate, finding that forecasting performances under sparse/dense priors vary across evaluated economic variables and across time frames. Dynamic model averaging, however, can combine the merits of both worlds.

© 2025 The Author(s). Published by Elsevier B.V. on behalf of International Institute of Forecasters. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Recent literature suggests that predictions of macroeconomic variables benefit from exploiting large datasets. Vector autoregressions (VARs), where the number of free parameters is relatively large compared to the limited length of macroeconomic time series, are especially prone to overfitting. Bayesian methods have been shown to be an effective regularization technique to reduce estimation uncertainty by imposing additional structure on the model (e.g., Chan, 2021b; Giannone et al., 2015; Koop, 2013; Litterman, 1986; Luo & Griffin; Sims & Zha, 1998).

Prior elicitation for VARs is a long-standing issue which goes back at least to the Minnesota prior in Litterman (1986). It stipulates that the lagged coefficients of the respective dependent variables are shrunk less than the lagged coefficients of other variables, and that coefficients are generally shrunk more strongly with an increasing number of lags. The Minnesota

☆ Variants of this paper were presented at the 12th European Seminar on Bayesian Econometrics (ESOBE 2022), the 5th Vienna Workshop on High-Dimensional Time Series in Macroeconomics and Finance 2022, the 2022 ISBA World Meeting, the 16th International Conference on Computational and Financial Econometrics 2022, the Bayesian Young Statisticians Meeting (BAYSM 2023), and the 44th International Symposium on Forecasting (ISF 2024). The authors thank all participants, in particular Sylvia Frühwirth-Schnatter, Neil Shephard, Massimiliano Marcellino, Florian Huber, Sylvia Kaufmann, Hedibert Lopes, Mike West, as well as anonymous referees and the associate editor for crucially valuable comments and suggestions.

This research was funded in part by the Austrian Science Fund (FWF) [10.55776/ZK35]. For open access purposes, the authors have applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

The numerical results presented in this manuscript were reproduced by the Editor-in-Chief on 31 January 2025.

* Correspondence to: Department of Statistics, University of Klagenfurt, Universitätsstraße 65-67, 9020 Klagenfurt am Wörthersee, Austria.
E-mail addresses: luis.gruber@aau.at (L. Gruber), gregor.kastner@aau.at (G. Kastner).

https://doi.org/10.1016/j.ijforecast.2025.02.001
0169-2070/© 2025 The Author(s). Published by Elsevier B.V. on behalf of International Institute of Forecasters. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

prior, therefore, is a classic example of a prior based on domain knowledge. On the other side of the spectrum, hierarchical priors such as stochastic search variable selection (SSVS) priors or global–local (GL) priors—where the latter have recently attracted increasing attention in combination with large VARs (e.g., Follett & Yu, 2019; Huber & Feldkircher, 2019; Kastner & Huber, 2020)—offer off-the-shelf shrinkage, since they are neither problem-specific nor designed for VARs per se.

The first contribution of this paper is the introduction of the class of semi-global priors, which aims at combining the merits of domain knowledge with off-the-shelf shrinkage. In short, this framework is characterized by shrinkage parameters acting on pre-specified subgroups of the parameter space, i.e., on the semi-global level. It can easily be applied to already existing priors. For example, it can be applied to any GL prior or to all kinds of SSVS priors. The framework is very flexible and nests other extensions to GL priors, like the extensions to the normal-gamma prior of Huber and Feldkircher (2019) and Chan (2021b), as special cases. We propose a grouping of the coefficients which mimics some features of the Minnesota prior: namely, the discrimination between lags in general; and within lags, the discrimination between own-lag and cross-lag coefficients.

Second, we provide a concise comparison of shrinkage priors that are commonly used in the VAR literature, but are scattered across different works and consequently lack a systematic comparison. Here, our attempt is to shed more light on the 'prior zoo' by an in-depth comparison focusing on the two most important features of shrinkage priors: (a) the concentration at zero for shrinking noise, and (b) the tail robustness for detecting the relevant signals. This in turn gives guidance for practitioners in meaningfully selecting hyperparameters. The priors under scrutiny are: the Minnesota prior, a semi-hierarchical version of the Minnesota prior (SHM), the stochastic search variable selection (SSVS) prior as in George et al. (2008), the horseshoe (HS) prior (Carvalho et al., 2010), the normal-gamma (NG) prior (Brown & Griffin, 2010), the Dirichlet–Laplace (DL) prior (Bhattacharya et al., 2015), and the R^2-induced Dirichlet decomposition (R2D2) prior (Zhang et al., 2022). Concerning the latter and to the best of our knowledge, we are the first who investigate the R2D2 prior in the context of VARs.

Third, we contribute to the ongoing 'illusion of sparsity' (IoS) debate (Cross et al., 2020; Fava & Lopes, 2021; Giannone et al., 2021). Loosely speaking, within the Bayesian framework, there are two contrasting approaches to reducing parameter uncertainty in high dimensions. On the one hand, there are sparse modeling techniques that aim at selecting a small set of important predictors. On the other hand, dense modeling techniques support the assumption that all possible explanatory variables might be important; their individual impact for prediction, however, is expected to be small. Giannone et al. (2021) design one specific prior—which is basically a discrete mixture prior with point mass at zero—in order to detect whether specific datasets are best summarized by means of many equally important covariates, or rather by a small subset of covariates. Analyzing the posterior results, they conclude that macroeconomic data are rather dense. Fava and Lopes (2021) show that with little changes to the prior used by Giannone et al. (2021), i.e., using fatter tailed distributions, the posterior results appear to be sparser.[1]

In light of these results, it becomes clear that the level of posterior sparsity can be very sensitive to prior assumptions. Thus, we augment the narrative by quantifying the sparseness of our considered priors *before* looking at the data. By applying the Hoyer sparseness measure (Hoyer, 2004) to simulated data from the prior distributions, we demonstrate that they can be ordered from sparse to dense. Sparsity is best expressed by GL priors; denseness is best expressed by Minnesota(-type) priors. In addition, we find that prior sparseness translates to posterior sparseness. In other words, the prior and the posterior sparsity ranking are similar.

Other than Fava and Lopes (2021) and Giannone et al. (2021), who investigate a linear model with a single response variable, the flexible framework of VARs allows us to analyze results for different response variables with only one dataset. In an extensive simulation study based on various data generating processes (DGPs), we test the validity of our claims. In sparse DGPs, the variants of GL priors have the highest concentration around the true model parameters. In purely dense DGPs, the Minnesota priors seem to be superior to all competing models.

The gold standard of model evaluation in economic and financial applications is comparing forecasting performance: A model is considered good if it performs well at predicting the future. In our empirical application, we adopt a variant of the quarterly dataset of the US economy proposed by Stock and Watson (2012) and provided by McCracken and Ng (2021). Overall, the semi-global versions of GL priors, with specific shrinkage for own-lag and cross-lag coefficients, achieve the highest support from the data in terms of one-step-ahead predictive likelihoods. Nevertheless, we find heterogeneity in model performance over time. The heterogeneous results demonstrate that there is no prior, and hence no sparse or dense modeling approach, that performs best during the whole evaluation period and for all variables. To combine the merits of different priors, we also discuss the possibility of dynamic model averaging, as proposed by Raftery et al. (2010), and the loss discounting framework proposed by Bernaciak and Griffin (2024).

Overall, this paper provides a broad overview of specification choices with respect to Bayesian VARs in their reduced forms featuring stochastic volatility (SV). Although all of our considered priors were originally proposed or put forward to the VAR framework using the reduced form, lately, many authors opt for using the structural form (e.g., Chan, 2021b; Cross et al., 2020). Placing priors on the structural form coefficients is alluring because then one can employ a faster Markov chain Monte Carlo (MCMC) algorithm. Nevertheless, placing the prior on the structural coefficients constitutes a different model where posteriors are (potentially substantially) different as well. Beyond that, there is early empirical

[1] We note at this point that sparsity in the VAR coefficients might imply unrealistic constraints from a purely economic point of view. However, as detailed by Sims (1980), these can nevertheless be useful for forecasting purposes, as restrictions can reduce forecast errors even if they are false.

evidence that modeling reduced-form coefficients leads to better out-of-sample forecasting performance (Bernardi et al., 2024). Modeling reduced-form coefficients, however, comes at the cost of a higher computational burden. To render MCMC computation feasible, we apply the corrected triangular algorithm, as in Carriero et al. (2022).[2]

Cross et al. (2020) investigate and compare a similar set of prior distributions for VARs. We differ in three important aspects. First, we demonstrate how the forecasting performance of GL priors can be improved with the semi-global shrinkage approach. Second, the aforementioned paper places the priors on the structural-form VAR coefficients. Third, Cross et al. (2020) evaluate forecasting performance only on a three-variate subset of considered variables, whereas we also consider the full-system predictive performance in addition to further robustness checks.

The remainder of the paper is organized as follows. Section 2 describes the econometric framework and lays out the various prior specifications, while Section 3 compares the shrinkage behavior and sparsity of the priors. Section 4 presents the results of an extensive simulation study considering different time series lengths within sparse and dense data-generating scenarios. In Section 5, we apply the different model setups to data on the US economy. After inspecting the posterior distributions, we perform a forecasting exercise to assess the predictive performance of the considered models. Finally, Section 6 concludes the article.

2. Econometric framework

For $t = 1, \dots, T$, let $\boldsymbol{y}_t$ denote an M-dimensional column vector containing observations on M time series variables. In a VAR model of order p, VAR(p), $\boldsymbol{y}_t$ is determined by[3]

$$\boldsymbol{y}_t' = \sum_{j=1}^{p} \boldsymbol{y}_{t-j}' \boldsymbol{A}_j + \boldsymbol{\varepsilon}_t', \tag{1}$$

where $\boldsymbol{A}_j$ is an $M \times M$ matrix of coefficients, the lag p a positive integer, and $\boldsymbol{\varepsilon}_t$ an M-dimensional vector of errors.

We assume the distribution of the errors to be multivariate Gaussian with a time-varying variance–covariance matrix, i.e., $\boldsymbol{\varepsilon}_t \sim \boldsymbol{N}(0, \Sigma_t)$, and follow Cogley and Sargent (2005) in applying the decomposition in the form of

$$\Sigma_t = \boldsymbol{U}'^{-1} \boldsymbol{D}_t \boldsymbol{U}^{-1}, \tag{2}$$

where $\boldsymbol{U}^{-1}$ is an upper triangular matrix with ones on the diagonal, and $\boldsymbol{D}_t$ is a diagonal matrix. We denote the free off-diagonal elements in $\boldsymbol{U}$ as $\boldsymbol{u} = (u_{12}, u_{13}, \dots, u_{(M-1)M})'$. The transformed errors $\boldsymbol{\xi}_t = \boldsymbol{U}'\boldsymbol{\varepsilon}_t$ then have a diagonal variance–covariance matrix $\boldsymbol{D}_t$. Borrowing from Jacquier et al. (1994) and Kim et al. (1998), the orthogonalized errors are assumed to follow univariate SV models: For $i = 1, \dots, M$,

$$\xi_{it} = \exp\left(\frac{h_{it}}{2}\right) \epsilon_{it}, \tag{3}$$

$$h_{it} = \mu_i + \rho_i (h_{i(t-1)} - \mu_i) + \sigma_i \eta_{it}, \tag{4}$$

where $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$ are assumed to be i.i.d. $N(\boldsymbol{0}, \boldsymbol{I}_M)$. Hence, the ii-th element of $\boldsymbol{D}_t$ is $d_{ii,t} = \exp(h_{it})$. The log-variance process $\boldsymbol{h}_i = (h_{i1}, \dots, h_{iT})'$ is initialized by $h_{i0} \sim N\left(\mu_i, \sigma_i^2/(1-\rho_i^2)\right)$. Here, μ_i is the level of log variance, $\rho_i \in (-1, 1)$ the persistence of log variance, and σ_i the volatility of log variance.

To facilitate prior implementation, it proves to be convenient to rewrite the model in matrix form. Define a $K \times 1$ vector of predictors $\boldsymbol{x}_t = (\boldsymbol{y}_{t-1}', \dots, \boldsymbol{y}_{t-p}')'$ and a $K \times M$ matrix of coefficients $\boldsymbol{\Phi} = (\boldsymbol{A}_1, \dots, \boldsymbol{A}_p)'$, where $K = mp$ is the number of vector-autoregressive coefficients per equation. Then, the VAR can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Phi} + \boldsymbol{E}, \tag{5}$$

where $\boldsymbol{Y} = (\boldsymbol{y}, \dots, \boldsymbol{y}_T)'$, $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_T)'$, and $\boldsymbol{E} = (\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_T)'$. $\boldsymbol{Y}$ and $\boldsymbol{E}$ are $T \times M$ matrices, and $\boldsymbol{X}$ is a $T \times K$ matrix. Further, let $\boldsymbol{\phi} = \text{vec}(\boldsymbol{\Phi}) = (\phi_1, \dots, \phi_n)'$, where $n = KM$ denotes the number of autoregressive coefficients.

As our approach to inference is Bayesian, we have to specify prior distributions. The generic prior for the coefficient vector is multivariate normal: $\boldsymbol{\phi} \sim \boldsymbol{N}(\boldsymbol{0}, \underline{\boldsymbol{V}})$. For growth rates and/or approximately stationary transformed data, it is common to center the prior at zero (e.g., Cross et al., 2020; George et al., 2008; Kastner & Huber, 2020; Koop & Korobilis, 2010), whereas for data in levels, often the prior mean of own-lag coefficients in the first lag is set to one (Litterman, 1986; Sims & Zha, 1998). Moreover, $\underline{\boldsymbol{V}}$ is a diagonal $n \times n$ matrix with diagonal elements $v_1, \dots, v_n$. The shrinkage priors under scrutiny, detailed in Sections 2.1 to 2.3, distinguish themselves in their treatment of $\underline{\boldsymbol{V}}$.

In view of the next sections, we want to highlight that the prior for the VAR coefficients does not depend on Σ_t and hence neither on $\boldsymbol{U}$. This implies that all considerations in Sections 2.1 to 2.3 and Sections 3.1 to 3.2 also apply to other specifications of the error covariance matrix: e.g., the order-invariant factor stochastic volatility specification without restrictions on the factor loadings proposed by Kastner and Huber (2020), the order-invariant stochastic volatility

[2] Carriero et al. (2022) present a correction to the broadly applied algorithm based on equation-per-equation estimation put forward by Carriero et al. (2019).

[3] For simplicity of exposition, we omit the intercept in the following (which we nonetheless include in the empirical application).

specification by Chan et al. (2024), or the order-invariant specification based on the eigendecomposition of the error covariance matrix proposed by Wu and Koop (2022).

Instead of modeling the reduced-form coefficients, one could also model the coefficients $\boldsymbol{B} = \boldsymbol{\Phi U}$ of the VAR in structural form: $\boldsymbol{y}_t'\boldsymbol{U} = \boldsymbol{x}_t'\boldsymbol{B} + \boldsymbol{\xi}_t'$, where $\boldsymbol{B} = \boldsymbol{\Phi U}$ and $\boldsymbol{\xi}_t = \boldsymbol{\epsilon}_t\boldsymbol{U} \sim N(\boldsymbol{0}, \boldsymbol{D}_t)$. Often, this is done to speed up computations, as it allows for embarrassingly parallel equation-per-equation estimation in the spirit of Carriero et al. (2019). Discussing whether it is better to shrink the reduced-form or the structural-form coefficients is beyond the scope of this paper. However, note that prior distributions are not invariant under reparametrization per se, and any comparison amongst the different settings must be treated with great care (cf. Appendix D for a demonstration). Performance of state-of-the art shrinkage priors in the structural setting are discussed by Cross et al. (2020) and Chan (2021b).

2.1. Shrinkage based on domain knowledge: Minnesota priors

Traditional Minnesota prior. The prior proposed by Litterman (1986)—referred to as the traditional Minnesota prior (MP_LIT) in the following—is mainly characterized by two assumptions based on domain knowledge: First, own-lags are assumed to account for most of the variation of a given variable. Second, recent lags are assumed to be more important for predicting current values than distant lags. Hence, $\underline{\boldsymbol{V}}$ is structured in a way such that the diagonal elements of the blocks in $\boldsymbol{\Phi}$ corresponding to $\boldsymbol{A}_j$ for $j = 1, \ldots, p$ (the entries of the own-lag coefficients) are shrunk less than the off-diagonal elements. And coefficients associated with more recent lags are shrunk less than the ones associated with more distant lags. We follow Koop and Korobilis (2010) in setting the notation: Denote by $\underline{\mathbf{V}}_i$ the block of $\underline{\mathbf{V}}$ that corresponds to the K coefficients in the ith equation, and let $\underline{\mathbf{V}}_{i,jj}$ be its diagonal elements. The diagonal elements are set to

$$\underline{\mathbf{V}}_{i,jj} = \begin{cases} \frac{\lambda_1}{r^2} & \text{for coefficients on own-lag } r \text{ for } r = 1, \ldots, p, \\ \frac{\lambda_2 \hat{\sigma}_i^2}{r^2 \hat{\sigma}_j^2} & \text{for coefficients on lag } r \text{ of variable } j \neq i, \end{cases} \tag{6}$$

where $\hat{\sigma}_i^2$ is the OLS variance of a univariate AR(6) model of the ith variable. We set $\lambda_1 > \lambda_2$ to regularize cross-lag coefficients more heavily. The term r^2 in the denominator automatically imposes more shrinkage on the coefficients towards their prior mean as lag length increases. The term $\hat{\sigma}_i^2/\hat{\sigma}_j^2$ not only adjusts for different scales in the data; it is also intended to account for different scales of the responses of one economic variable to another (Litterman, 1986).

Semi-hierarchical Minnesota prior. The semi-hierarchical Minnesota (SHM) prior is a generalization of the traditional Minnesota prior. Instead of fixing the shrinkage parameters λ_1 and λ_2, SHM treats them as unknown quantities and learns them in a data-based fashion. That is, the strong assumption that $\lambda_1 > \lambda_2$ is now replaced with the weaker assumption that own-lags and cross-lags might account for different amounts in the variation of a given variable. We follow Huber and Feldkircher (2019) in imposing gamma priors on λ_1 and λ_2. That is, $\lambda_i \sim G(c_i, d_i)$ for $i = 1, 2$, where $G(\alpha, \beta)$ denotes the probability density function of the gamma distribution with shape α and rate β.

As default choice, we set $c_i = d_i = 0.01$ for $i = 1, 2$. The resulting gamma distribution has expectation 1 and variance 100. Moreover, the density diverges to infinity as $\lambda_j \to 0$. As most mass of this distribution is at zero, the prior can heavily shrink the parameters if necessary, though as soon as there are some moderate to large nonzero coefficients (in absolute values), the shrinkage will be only moderate to weak. Since the prior variances of cross-lag coefficients still have to be scaled with $\hat{\sigma}_i^2/\hat{\sigma}_j^2$, it is not a fully hierarchical prior, and we refer to it as the semi-hierarchical Minnesota prior.

2.2. Off-the-shelf shrinkage: Global–local priors

Other than the Minnesota priors, global–local (GL) shrinkage priors are not built upon domain knowledge and are specifically designed neither for VARs nor for forecasting macroeconomic data. They are very popular in sparse and high-dimensional applications, and therefore also well suited for large VARs (Follett & Yu, 2019; Huber & Feldkircher, 2019; Kastner & Huber, 2020). Following Polson and Scott (2011), GL priors can be written as continuous scale mixture distributions. Let $K(\delta)$ denote a symmetric unimodal density with variance δ. Then, a typical GL prior is of the following form:

$$\phi_i | \vartheta_i \zeta \sim K(\vartheta_i \zeta), \quad \vartheta_i \sim f, \quad \zeta \sim g \quad \text{for } i = 1, \ldots, n, \tag{7}$$

where ζ represents the global shrinkage and ϑ_i the local shrinkage. While the global parameter determines the overall shrinkage, the local parameters act to detect the relevant signals. Hence, g is usually a density with substantial mass at zero. In case of ζ being very small, f must be a heavy-tailed density such that ϑ_i can override the effect of ζ for a nonzero coefficient. GL priors distinguish themselves in their choices for f and g, as discussed in the next section.

2.3. Structured semi-global(-local) shrinkage

In this section, we aim at building a framework that allows us to provide more structure on off-the-shelf shrinkage priors, but with as few restrictions as possible.

Even though many applications have shown that a simple prior like the traditional Minnesota prior—where all hyperparameters are treated as known and therefore are not estimated from the data in a hierarchical fashion—can be very competitive, it lacks flexibility. First, due to the fact that own-lag shrinkage and cross-lag shrinkage are fixed a priori, the traditional Minnesota prior is not (automatically) adaptable. Especially in higher dimensions, well-established default values are difficult to determine, which in turn makes selecting them hard to justify. This issue is partly dealt with by considering semi-hierarchical versions of the prior (cf. Section 2.1). Second, it relies on tuning the prior variances with OLS estimates in order to adapt for the different scales of the endogenous variables. Third, shrinkage of higher-order lags is deterministic. That is, prior variances are additionally scaled by r^x, where r denotes the lag and usually $x = 2$. Theoretically, GL priors could remedy those shortcomings with their local scales, which individually adjust the regularization for each coefficient. A problem arises, however, when the noise level and the absolute level of the coefficients are different in different parts of the parameter space. Then, the adjustment is probably not strong enough. In what follows, we introduce a simple framework that tackles this problem. It only needs little input from the researcher and does not require any pre-tuning of hyperparameters.

Let $\mathcal{A}_j$ denote the generic index set that labels the coefficients of the jth group in $\boldsymbol{\phi}$ (e.g., the first group could be the own-lag coefficients associated with the first lag, the second group could be the cross-lag coefficients associated with the first lag, etc.), and let n_j denote the number of elements in $\mathcal{A}_j$. Then, as a semi-global local prior with k groups, we define the following hierarchical representation:

$$\phi_i \sim K(\vartheta_i \zeta_j), \quad \vartheta_i \sim f, \quad \zeta_j \sim g, \quad i \in \mathcal{A}_j, \quad j = 1, \dots, k. \tag{8}$$

The only additional input required is the partitioning of $\boldsymbol{\Phi}$ into k subgroups. We propose a partitioning that mimics two features of the Minnesota prior, namely the distinction between own-lag and cross-lag coefficients and the distinction between lags in general: In each lag, the diagonal elements (the own-lags) and the off-diagonal elements (the cross-lags) constitute separate groups, which makes $2p$ groups in total. The vanilla global–local prior, i.e., no partitioning of $\boldsymbol{\Phi}$, is nested as the special case where $k = 1$.

The semi-global framework nests other modifications to global–local priors within the VAR framework as special cases. For example, Huber and Feldkircher (2019) consider both equation-specific and covariate-specific shrinkage. The first means that the covariates of each equation form M separate groups (individual degree of shrinkage for each equation) and the latter that the specific covariates across all equations form K separate groups (individual degree of shrinkage for each covariate across all equations). The multiplicative lag-wise specification considered in the same paper, however, is not nested in the semi-global framework, since we assume independence across the k groups. The adaptive Minnesota prior in Chan (2021b) could also be cast in the semi-global framework. It is the unification of the Minnesota prior with a global–local prior, namely the normal-gamma prior (Brown & Griffin, 2010). That is, there are two semi-global shrinkage parameters: one for own-lags, and one for cross-lags. Additionally, each coefficient is assigned a local scale. The prior variances are rounded off with the scaling term $\hat{\sigma}_i^2/(r^2\hat{\sigma}_j^2)$ of the Minnesota prior in analogy to Eq. (6). Besides using quantities estimated from the data to tune the prior variances, the adaptive Minnesota prior differs from our approach, as it is a prior for the structural coefficients. Another class of priors which could be extended to semi-global priors are the asymmetric conjugate priors for homoskedastic VARs proposed by Chan (2022). Similar to with the adaptive Minnesota prior, the asymmetric conjugate priors exploit the structural form of the VAR.

In the remainder, we briefly describe several state-of-the art off-the-shelf shrinkage priors and how they can (in one case even have to) be modified, such that the pooling strategy introduced by the semi-global framework materializes.

DL prior. The Dirichlet–Laplace (DL) prior, introduced by Bhattacharya et al. (2015) and put forward to the VAR framework by Kastner and Huber (2020), takes the following hierarchical form:

$$\phi_i | \varrho_i, \omega_j \sim DE(\varrho_i \omega_j), \quad \boldsymbol{\varrho} \sim Dir(a^\gamma, \dots, a^\gamma), \quad \omega_j \sim G(n_j a^\gamma, 1/2), \tag{9}$$

where $DE(\delta)$ denotes the double exponential (sometimes called Laplace) distribution with variance $2\delta^2$, and $Dir(\alpha_1, \dots, \alpha_K)$ denotes the Dirichlet distribution on the open $(K-1)$-simplex with concentration parameters $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$. Letting $\sqrt{\vartheta_i/2} = \varrho_i \omega_j$ for $i \in \mathcal{A}_j$ and $j = 1, \dots, k$, it can be shown that $\sqrt{\vartheta_i/2} \sim G(a^\gamma, 1/2)$ independently for $i \in \mathcal{A}_j$, $j = 1, \dots, k$ (cf. Lemma IV.3 in Zhou & Carin, 2012). Using this result, it becomes clear that the proposed semi-global pooling strategy is not compatible with the vanilla DL prior, since there is no information sharing on the (semi-)global level. Therefore, to make things work, we replace the global shape parameter a^γ with the semi-global shape parameter $a_j^\gamma, j = 1, \dots, k$. Applying the lemma, the semi-global–local DL prior has the following form:

$$\phi_i | \vartheta_i, \zeta_j \sim DE(\sqrt{\vartheta_i \zeta_j/2}), \quad \sqrt{\vartheta_i/2} \sim G(a_j^\gamma, 1/2), \quad \zeta_j = 1, \tag{10}$$

where $Var(\phi_i | \vartheta_i, \zeta_j) = \vartheta_i \zeta_j$. Posterior results can be very sensitive to the choice of a_j^γ, as it is the parameter that controls the strength of regularization (cf. Section 3). Hence, we place i.i.d. discrete hyperpriors $Pr(a_j^\gamma = \tilde{a}_r^\gamma) = p_r^\gamma$ on a_j^γ for

$j = 1, \dots, k$, where $\tilde{\boldsymbol{a}}^{\gamma} = (\tilde{a}_1^{\gamma}, \dots, \tilde{a}_K^{\gamma})'$ is the vector of strictly positive support points. Updates of a_j^{γ} are straightforward, as the full conditional posterior is again discrete on the same support points.

Last but not least, the DL prior can be expressed as a Gaussian scale mixture by introducing the auxiliary scaling parameter $\psi_i \sim Exp(1/2)$, s.t. $\phi_i|\vartheta_i, \zeta_j \sim DE(\sqrt{\vartheta_i\zeta_j/2}) \iff \phi_i|\psi_i, \vartheta_i, \zeta_j \sim N(0, \psi_i\vartheta_i\zeta_j/2)$. Consequently, for DL, $v_i = \psi_i\vartheta_i\zeta_j/2$.

NG prior. The normal-gamma prior proposed by Brown and Griffin (2010) is nowadays well established in different kinds of multivariate time series modeling setups (e.g., Bitto & Frühwirth-Schnatter, 2019; Huber & Feldkircher, 2019; Kastner, 2019). It takes the following form:

$$\phi_i|\vartheta_i\zeta_j \sim N(0, \vartheta_i\zeta_j), \quad \vartheta_i \sim G(a_j^{\delta}, a_j^{\delta}/2), \quad \zeta_j^{-1} \sim G(b^{\delta}, c). \tag{11}$$

That is, for NG, $v_i = \vartheta_i\zeta_j$. As with DL, we place on a_j^{δ}, $j = 1, \dots, k$, i.i.d. discrete hyperpriors $Pr(a_j^{\delta} = \tilde{a}_r^{\delta}) = p_r^{\delta}$, where $\tilde{\boldsymbol{a}}^{\delta} = (\tilde{a}_1^{\delta}, \dots, \tilde{a}_R^{\delta})'$ is the vector of strictly positive support points.

R2D2 prior. Only recently, Zhang et al. (2022) proposed the R^2-induced Dirichlet decomposition (R2D2) prior. It is based on a beta prior on $R^2 := \frac{W}{W+1}$, where W denotes the sum of all prior variances. The induced prior then takes the following form:

$$\phi_i|\varrho_i, \omega_j \sim DE\left(\sqrt{\varrho_i\omega_j/2}\right), \quad \boldsymbol{\varrho}_j = (\varrho_i)_{i\in\mathcal{A}_j} \sim Dir(a^{\pi}, \dots, a^{\pi}), \quad \omega_j|\xi_j \sim G(n_j a^{\pi}, \xi_j), \quad \xi_j \sim G(b^{\pi}, 1). \tag{12}$$

Apart from a higher level of hierarchy, it has a similar structure to the DL prior. A major difference is that, for R2D2, the prior variance follows a gamma distribution, whereas for DL, the prior scale follows a gamma distribution. Using the same result as before, conditional on ξ_j, $\varpi_i = \varrho_i\omega_j \sim G(a_{\pi}, \xi_j)$ independently. Exploiting the scaling property of the gamma distribution, the R2D2 prior can be cast in the form of Eq. (8):

$$\phi_i|\vartheta_i, \zeta_j \sim DE\left(\sqrt{\vartheta_i\zeta_j/2}\right), \quad \vartheta_i \sim G(a^{\pi}, a^{\pi}/2), \quad \zeta_j^{-1} \sim G(b^{\pi}, a^{\pi}/2), \tag{13}$$

where $\zeta_j^{-1} = \frac{2\xi_j}{a_{\pi}}$ and $\vartheta_i = \frac{\varpi_i}{\zeta_j}$. By introducing auxiliary scaling parameters $\psi_i \sim Exp(1/2)$, the R2D2 prior can be written as a Gaussian scale mixture, s.t. $\phi_i|\vartheta_i, \zeta_j \sim DE\left(\sqrt{\vartheta_i\zeta_j/2}\right) \iff \phi_i|\psi_i, \vartheta_i, \zeta_j \sim N\left(0, \psi_i\vartheta_i\zeta_j/2\right)$; i.e., for R2D2, $v_i = \psi_i\vartheta_i\zeta_j/2$.

The R2D2 prior and the NG prior are closely related. For R2D2, the variance of a double exponential distribution follows a gamma distribution, whereas for NG, the variance of a normal distribution follows a gamma distribution. Hence, the special case of NG with $c = \frac{a_j^{\delta}}{2}$ could be seen as an R^2-induced Dirichlet decomposition prior with a normal kernel.

As with DL and NG, we place on $a_j^{\pi}, j = 1, \dots, k$ i.i.d. discrete hyperpriors $Pr(a_j^{\pi} = \tilde{a}_r^{\pi}) = p_r^{\pi}$, where $\tilde{\boldsymbol{a}}^{\pi} = (\tilde{a}_1^{\pi}, \dots, \tilde{a}_R^{\pi})'$ is the vector of strictly positive support points.

Horseshoe. The horseshoe (HS) prior proposed by Carvalho et al. (2010) is given by

$$\phi_i|\vartheta_i\zeta_j \sim N(0, \vartheta_i\zeta_j), \quad \sqrt{\vartheta_i} \sim \mathcal{C}^{+}(0, 1), \quad \sqrt{\zeta_j} \sim \mathcal{C}^{+}(0, 1), \tag{14}$$

where $\mathcal{C}^{+}(\cdot, \cdot)$ denotes the half-Cauchy distribution. Unlike the NG and R2D2, the HS places the hyperprior distributions on the standard deviation rather than on the variance. Compared to all other priors discussed so far, it has the advantage that no tuning parameter has to be specified by the user.

Stochastic search variable selection. Another popular prior for Bayesian VARs is the stochastic search variable selection prior (SSVS, George et al., 2008), which is a discrete mixture of two normal distributions. Although this prior is neither a global–local shrinkage prior nor a continuous scale mixture prior, it can be modified in order to comply with the semi-global–local framework. In hierarchical form and adapted to the semi-global–local framework, the prior is given through

$$\phi_i|\gamma_i \sim N\left(0, (1-\gamma_i)\tau_{0i}^2 + \gamma_i\tau_{1i}^2\right), \quad \gamma_i|\underline{p}_j \sim Bernoulli(\underline{p}_j), \tag{15}$$

where γ_i is an auxiliary dummy variable, τ_{0i} should be small (close to zero), and it must hold that $\tau_{0i} \ll \tau_{1i}$. Intuitively speaking, if ϕ_i is assigned to the spike variance τ_{0i}^2, it is virtually constrained to zero. If ϕ_i is assigned to the slab variance τ_{1i}^2, it is shrunken depending on how τ_{1i}^2 is chosen. Further, we assume that the prior inclusion probability $\underline{p}_j$ might differ between the k groups. In order to learn the parameter in a data-based fashion, we impose i.i.d. beta priors on each $\underline{p}_j$, $j = 1, \dots, k$: $\underline{p}_j \sim Beta(s_1, s_2)$. Note, that for prediction purposes and for inference on the joint posterior $p(\phi_i, \gamma_i|\bullet)$, learning $\underline{p}_j$ in a data-based fashion only has an effect if there is information sharing between coefficients. In the case where each ϕ_i constitutes a separate group, i.e., each ϕ_i is assigned an independent $\underline{p}_i$ (as in, e.g., Cross et al., 2020), it is straightforward to show that the full conditional posterior $p(\gamma_i|\bullet)$ only depends on the prior expectation of $\underline{p}_i$.

2.4. Priors for the variance–covariance matrix

To complete the models, we have to specify the priors on the decomposed variance–covariance matrix $\boldsymbol{\Sigma}_t = \boldsymbol{U}'^{-1}\boldsymbol{D}_t\boldsymbol{U}^{-1}$. Note that the precision matrix—the inverse of the variance–covariance matrix—can be written as $\boldsymbol{\Sigma}_t^{-1} =$

$\boldsymbol{U}\boldsymbol{D}_t^{-1}\boldsymbol{U}'$. It is well known that the precision matrix should be a sparse matrix, as zero entries on the off-diagonals imply conditional independence among the respective equations (West, 2020). Hence, for $\boldsymbol{u}$, we choose the HS prior: $u_{ij} \sim N(0, \vartheta_{ij}^u \zeta^u)$, $\sqrt{\vartheta_{ij}^u} \sim \mathcal{C}^+(0, 1)$, $\sqrt{\zeta^u} \sim \mathcal{C}^+(0, 1)$, $i = 1, \ldots, M-1$ and $j = 2, \ldots M$.[4] Following Kim et al. (1998), for $i = 1, \ldots, M$, we choose a normal prior for the level of the log variance, $\mu_i \sim N(0, 100^2)$, and for the persistence parameter, ρ_i, a beta distribution on $\frac{\rho_i+1}{2} \sim Beta(20, 1.5)$. Finally, we follow Kastner and Frühwirth-Schnatter (2014) by imposing a gamma prior on the variance of the log variance, $\sigma_i^2 \sim G(1/2, 1/2)$.

Regarding our prior choices for the SV parameters and related to the finding by Rossini et al. (2024) that forecasting performance of homoskedastic VARs in conjunction with a Wishart prior on $\boldsymbol{\Sigma}^{-1}$ can be negatively influenced by choosing an unfavorable default value for the degrees-of-freedom hyperparameter, we acknowledge that choosing different priors for the vector of SV parameters $(\mu_i, \rho_i, \sigma_i^2)$ can have a non-negligible effect on the posterior distribution of those parameters. However, the impact on the estimated log variances h_{it}, which are most relevant for forecasting, is typically small.

2.5. Posterior estimation

Efficient computer implementations of all posterior samplers discussed in this paper are conveniently bundled into the R (Team, 2023) package `bayesianVARs`, interfacing to C/C++ via `Rcpp` (Eddelbuettel & François, 2011) and `RcppArmadillo` (Eddelbuettel & Sanderson, 2014) for increased computational efficiency. The package is available under the General Public License (GPL $\geq$ 3) from the Comprehensive R Archive Network (CRAN) at https://cran.r-project.org/package=bayesianVARs. We thus refrain from describing the various posterior samplers in detail. Instead, we provide their main building blocks, the necessary conditionals used for Gibbs sampling, in Appendix B.

Concerning the scalability of our approach, we present some rough estimates of computation time. The estimates are based on an Intel® Core™ i7-10610U processor using one core. For the dimensions of our empirical application, i.e., $M = 21$, $T = 198$, and $p = 1, \ldots, 5$, generating 100 draws from the posterior distribution takes approximately 1.9, 5.7, 11.4, 19.1, and 28.6 s, respectively. Since the complexity of the underlying correct triangular algorithm (Carriero et al., 2022) is $O(M^4)$, the estimation of larger datasets up to $M \approx 50$ is surely possible. For even higher dimensions, approximate estimation techniques or a different structure of the error variance–covariance matrix could be considered. Instead of using the correct triangular algorithm, one could use the equation-per-equation algorithm described by Carriero et al. (2019) as an approximation. In that paper, a VAR with $p = 13$ lags is successfully estimated using a time series with $M = 113$ and $T = 648$, which is similar in dimensionality to the application in Bańbura et al. (2010). The approach of the latter, however, builds on a substantially different model, namely a homoskedastic and conjugate VAR. Variational inference as proposed in Bernardi et al. (2024), another approximate estimation technique, can also reduce the estimation time of VARs which are similar in spirit to those presented in this paper.

Concerning the error structure, instead of using the Cholesky stochastic volatility specification in (2) and (3), one could use the factor stochastic volatility (FSV) specification proposed by Kastner and Huber (2020). Their variant is also compatible with the priors for the VAR coefficients described in Sections 2.1 to 2.3 and Sections 3.1 to 3.2. By conditioning on the factors and their loadings, this approach makes conditional equation-per-equation estimation straightforward. In situations where $K \gg T$, equation-per-equation estimation can be further combined with the algorithm for high-dimensional regressions proposed by Bhattacharya et al. (2016). This then has complexity $O(pM^2T^2)$, and Kastner and Huber (2020), for example, demonstrate that MCMC inference for a VAR with $p = 5$ lags using a time series with $M = 215$ and $T = 228$ becomes feasible.

3. Shrinkage behavior and sparsity of various priors

In Section 3.1, we analyze both the tail behavior and the concentration at zero of the various shrinkage priors. In Section 3.2, we characterize the sparsity of the priors.

3.1. Univariate analysis

The limiting behavior of univariate global–local prior densities is typically analyzed conditionally on the global scale (e.g., Brown & Griffin, 2010; Carvalho et al., 2010), i.e. marginalized over the local scales only. Since the global parameters often introduce dependence among all ϕ_i, $i \in \mathcal{A}_j$, $j = 1, \ldots, k$, we follow this line of thinking and conduct the analysis in Section 3.1 in terms of univariate conditionally independent prior densities $p(\phi_i|\cdot)$, where $\cdot$ stands for all global hyperparameters.

[4] In Section 5.5, we provide some robustness checks regarding different priors for $\boldsymbol{u}$.

Theorem 3.1 (*Limiting behavior of shrinkage priors*)**.**

(a) ***Tail behavior***

For $|\phi_i| \to \infty$, the marginal densities satisfy

$$p^{\infty}_{DL}(\phi_i|a^{\gamma}_j) = O\left(|\phi_i|^{a^{\gamma}_j/2-3/4}\exp\left\{-\sqrt{2|\phi_i|}\right\}\right), \quad \text{for any } a^{\gamma}_j > 0, \tag{16}$$

$$p^{\infty}_{NG}(\phi_i|\zeta_j, a^{\delta}_j) = O\left(|\phi_i|^{a^{\delta}_j-1}\exp\left\{-\sqrt{\frac{a^{\delta}_j}{\zeta_j}}|\phi_i|\right\}\right), \quad \text{for any } a^{\delta}_j > 0, \tag{17}$$

$$p^{\infty}_{R2D2}(\phi_i|\zeta_j, a^{\pi}_j) = O\left(|\phi_i|^{\frac{2a-2}{3}}\exp\left\{-3\left(\frac{\phi_i^2 a^{\pi}_j}{4\zeta_j}\right)^{1/3}\right\}\right), \quad \text{for any } a^{\pi}_j > 0, \tag{18}$$

$$p^{\infty}_{HS}(\phi_i|\zeta_j) = O\left(\frac{1}{\phi_i^2}\right), \tag{19}$$

$$p^{\infty}_{SHM}(\phi_i|\underline{\mathbf{V}}_{i,jj}) = O\left(\exp\left\{\frac{-\phi_i^2}{2\underline{\mathbf{V}}_{i,jj}}\right\}\right), \quad (\underline{\mathbf{V}}_{i,jj}, \textit{see}\,(6)) \tag{20}$$

$$p^{\infty}_{SSVS}(\phi_i|\underline{p}_j, \tau_{0i}, \tau_{1i}) = O\left(\exp\left\{-\frac{\phi_i^2}{2(\underline{p}_j\tau_{1i}^2 + (1-\underline{p}_j)\tau_{0i}^2)}\right\}\right). \tag{21}$$

(b) ***Concentration at zero***

For $|\phi_i| \to 0$, the marginal densities satisfy

$$p^{0}_{DL}(\phi_i|a^{\gamma}_j) = \begin{cases} O\left(\frac{1}{|\phi_i|^{1-a^{\gamma}_j}}\right) & \text{if } 0 < a^{\gamma}_j < 1, \\ O\left(\log\left(\frac{1}{|\phi_i|}\right)\right) & \text{if } a^{\gamma}_j = 1, \\ \textit{no singularity} & \text{if } a^{\gamma}_j > 1, \end{cases} \tag{22}$$

$$p^{0}_{NG}(\phi_i|\zeta_j, a^{\delta}_j) = \begin{cases} O\left(\frac{1}{|\phi_i|^{1-2a^{\delta}_j}}\right) & \text{if } 0 < a^{\delta}_j < \frac{1}{2}, \\ O\left(\log\left(\frac{1}{|\phi_i|}\right)\right) & \text{if } a^{\delta}_j = \frac{1}{2}, \\ \textit{no singularity} & \text{if } a^{\delta}_j > \frac{1}{2}, \end{cases} \tag{23}$$

$$p^{0}_{R2D2}(\phi_i|\zeta_j, a^{\pi}_j) = O\left(\frac{1}{|\phi_i|^{1-2a^{\pi}_j}}\right), \quad \text{for } 0 < a^{\pi}_j < \frac{1}{2}, \tag{24}$$

$$p^{0}_{HS}(\phi_i|\zeta_j) = O\left(\log\left(\frac{1}{|\phi_i|}\right)\right). \tag{25}$$

The marginal densities of SHM/MP_LIT and SSVS have no singularity at zero.

For the proofs of Theorem 3.1(*a*)(16) and the first part of 3.1(*b*)(22), we refer to Zhang et al. (2022), and for 3.1(*a*)(19) and 3.1(*b*)(25), we refer to Carvalho et al. (2010). The remaining proofs can be found in Appendix A.

Some noteworthy remarks follow. All global–local priors share a pole at $\phi_i = 0$, at least for certain hyperparameter settings, which is important for handling vectors with many zero elements. The Minnesota priors and SSVS do not diverge to infinity at zero, which could be seen as a disadvantage in sparse settings. For $a^{\gamma}_j < 1$, $a^{\delta}_j < \frac{1}{2}$, and $a^{\pi}_j < \frac{1}{2}$, DL, NG, and R2D2, respectively, diverge to infinity with polynomial speed; this is faster than HS. Concerning the tails: HS has the heaviest (Cauchy-like) tails. It is followed by DL and R2D2, respectively, which are still heavy-tailed in the sense that they have heavier tails than the exponential distribution. NG is lighter-tailed than DL and R2D2, but still heavier-tailed than the Minnesota priors and SSVS, which have Gaussian-like tails.[5]

Fig. 1 depicts the univariate densities conditional on the global scales and the prior densities for the global scales separately. For reasons of comparability, we show the densities for $\sqrt{\zeta_j}$ for all priors. It can easily be verified that for

[5] A marginal note regarding R2D2: Zhang et al. (2022) claim that R2D2 is polynomial in both regions at zero and in the tails, which seems contradictory to Theorem 3.1(*a*)(18). For the avoidance of doubt, we want to clarify that the univariate density of R2D2 is polynomial in the tails only if the global scale ζ_j is integrated out. This is somewhat misleading, since for all other priors under scrutiny in both papers, i.e. here and in Zhang et al. (2022), the global scale is fixed.

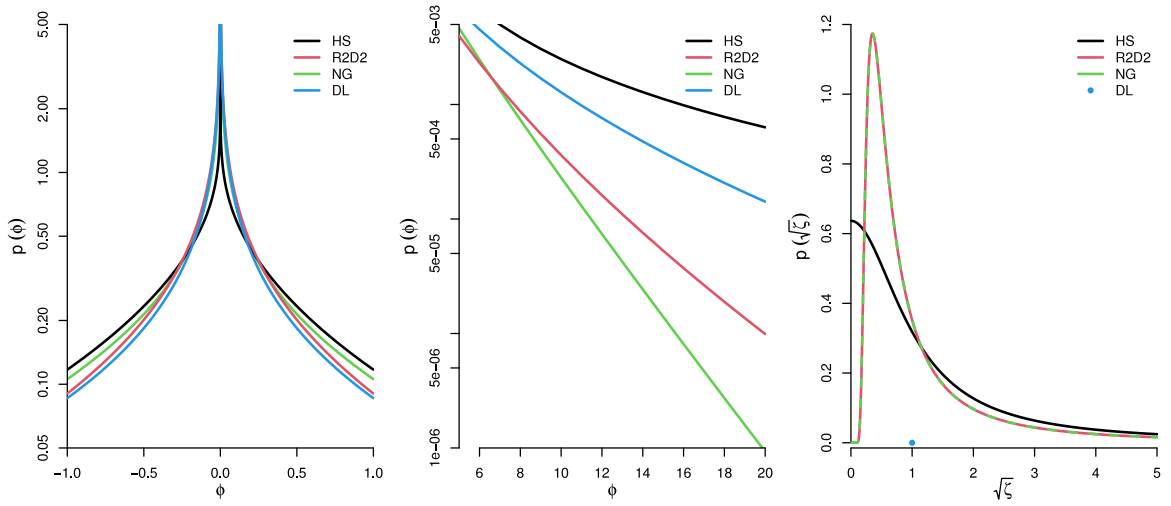

Fig. 1. Left and middle panel: Visualization of univariate prior densities conditional on the global scales. Right panel: Visualization of prior densities for the global scales. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

NG and R2D2, the gamma hyperpriors on ζ_j imply generalized gamma hyperpriors on $\sqrt{\zeta}_j$. Clearly, the division of tasks among global and local scales varies among different GL priors. Whereas the global scale of HS has considerable mass at zero, the global scale of NG and R2D2 is bounded away from zero. The global scale of DL is a Dirac mass at one. That is, for DL, NG, and R2D2, the task of the local scales is twofold: in addition to detecting signals, they must be able to overwhelm the global scale in shrinking the zero coefficients to zero.

3.2. Multivariate analysis

In order to describe the implied level of sparsity of the different priors, we now analyze their multivariate distributions. We begin by characterizing the sparsity of the priors qualitatively and continue by quantifying it via a sparseness measure.

Fig. 2 depicts three-dimensional scatter plots of prior simulations. The star-like shapes of GL priors (here, the HS prior) have most of their mass where all coefficients are almost zero or where at most one is substantial different from zero, indicating sparseness. By contrast, the ball-like shape of SHM, which does not allow for outliers, indicates denseness. SSVS is somewhere in the middle of the contrasting approaches. The figure also demonstrates the effect of structured shrinkage. Assume $\phi_1, \phi_2 \in \mathcal{A}_1$ and $\phi_3 \in \mathcal{A}_2$. Then, for all priors there is relatively more mass where $\phi_1 = \phi_2 = 0 \;\;\&\;\; \phi_3 \neq 0$ or else $\phi_3 = 0 \;\;\&\;\; \phi_1 \neq 0 \;\;\&\;\; \phi_2 \neq 0$; i.e., the structuring introduces discrimination between the groups. For GL priors, this implies discrimination between groups while being sparse within groups. On the other hand, SHM only discriminates between own-lag and cross-lag coefficients. Within own-lag and within cross-lag coefficients, it remains dense, which explains the spinning-top shape.

We quantify the sparseness of each prior by means of the Hoyer sparseness measure (Hoyer, 2004). The Hoyer measure for the vector $\boldsymbol{\phi}_j = (\phi_i)_{i \in \mathcal{A}_j}$,

$$H(\boldsymbol{\phi}_j) = \frac{\sqrt{n_j} - \left(\sum_{i \in \mathcal{A}_j} |\phi_i|\right) / \sqrt{\sum_{i \in \mathcal{A}_j} \phi_i^2}}{\sqrt{n_j} - 1}, \quad j = 1, \ldots, k, \tag{26}$$

is a normalized measure where $H = 1$ indicates maximum sparseness, i.e., only one single nonzero value, and $H = 0$ indicates maximum denseness, i.e., the absolute values of all components are equal. More importantly, it can express sparseness when all elements of $\boldsymbol{\phi}_j$ are nonzero, but when almost all are very close to zero and only very few are large (in absolute values). This is important, as we analyze continuous prior distributions which never strictly exclude single coefficients.

We estimate the Hoyer measure through Monte Carlo simulation. Table 1 reports the arithmetic means stemming from a Monte Carlo simulation with 10,000 iterations and simulated vectors of length $n = 1000$. Interestingly, the simulation suggests that for MP_LIT/SHM, i.e., all priors of the form $\boldsymbol{\phi} \sim N(\mathbf{0}, \boldsymbol{I}c)$, $c \sim f$, regardless the specific choice of f, the expectation of the Hoyer measure converges to 0.21 for $n = 1000$. This behavior of the measure is convenient, since sparsity should not depend on the overall noise level. For the remaining priors, results can be sensitive to the choice of hyperparameters. Thus, for each prior, we consider two different scenarios. In scenario A, the hyperparameters for DL, R2D2, and NG are selected to ensure that the concentration at zero is comparable to the one of HS. In scenario B, we select the hyperparameters such that the concentration at zero is higher compared to scenario A: for DL we set $a^\gamma = 0.001$, for

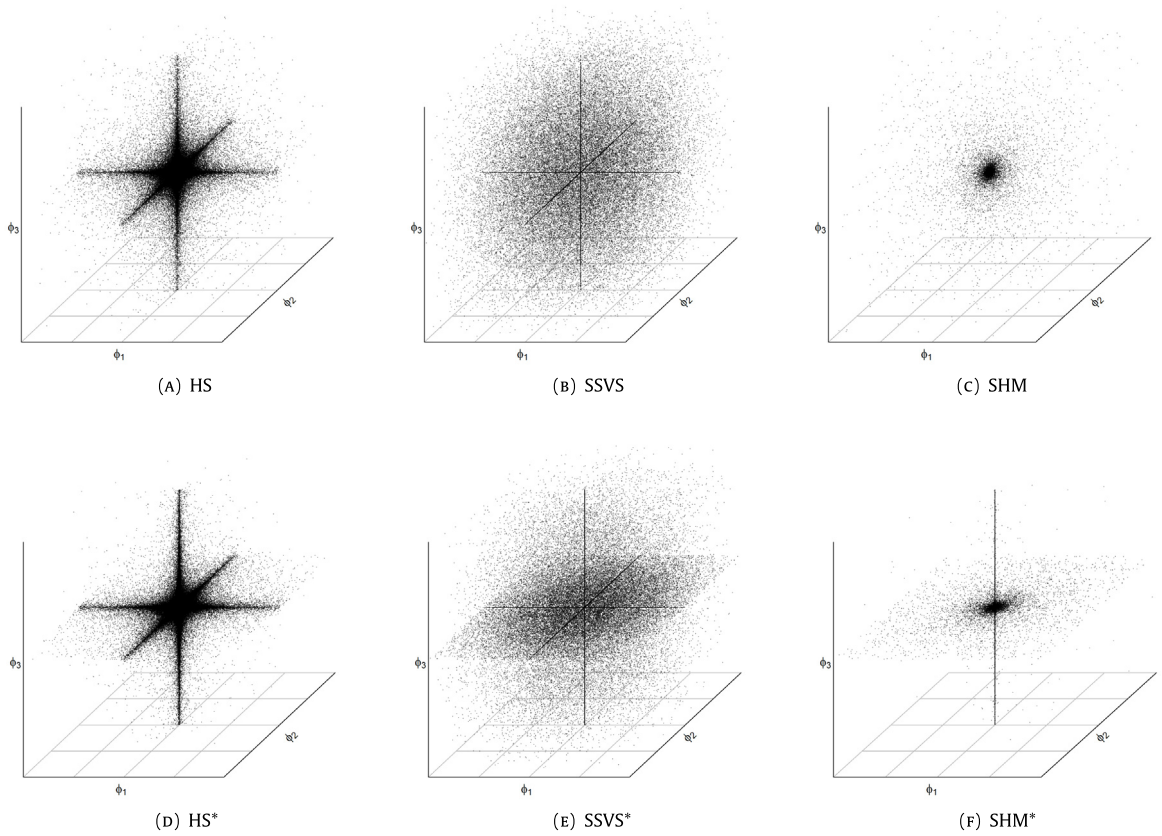

Fig. 2. Exemplary visualization of structured shrinkage. 100,000 samples from the HS prior representing the global–local priors, SSVS prior, and SHM prior. In HS, SSVS, and SHM, all coefficients belong to the same group, whereas in HS*, SSVS*, and SHM*, ϕ_1 and ϕ_2 are from the same, whereas ϕ_3 is from a different group.

Table 1
Hoyer sparseness measure for SHM/MP_LIT, HS, DL, R2D2, NG, and SSVS. The reported values are the means stemming from 10,000 simulations of vectors of length 1000. In scenario A, for DL $a^\gamma = 1$, for R2D2 $a^\pi = a^\gamma/2$, for NG $a^\delta = a^\gamma/2$, for SSVS $\tau_0 = 0.01$, $\tau_1 = 100$ and $\underline{p} = 0.5$. In scenario B, for DL $a^\gamma = 0.001$, for R2D2 $a^\pi = a^\gamma/2$, for NG $a^\delta = a^\gamma/2$, for SSVS $\tau_0 = 0.01$, $\tau_1 = 100$ and $\underline{p} = 0.01$. For SHM/MP_LIT and HS, results do not depend on hyperparameter choices.

	MP_LIT/SHM	HS	DL	R2D2	NG	SSVS
A	0.21	0.89	0.51	0.45	0.37	0.45
B	–	–	0.99	0.98	0.98	0.95

R2D2 we set $a^\pi = 0.0005$, and for NG we set $a^\delta = 0.0005$, which results in the same behavior around the origin for those three priors. Our numerical experiments suggest that the sparseness of DL, R2D2, and NG does not depend upon choices for the global hyperparameters. For SSVS, we set $\tau_0 = 0.01$ and $\tau_1 = 100$ in both scenarios. We set $\underline{p} = 0.5$ in scenario A and $\underline{p} = 0.01$ in scenario B.

The exercise corroborates our preceding considerations. From sparse to dense, the priors can be ranked in the following order: DL, R2D2, NG, SSVS, HS, and at some distance SHM/MP_LIT. Moreover, the sparseness of DL, R2D2, NG, and SSVS is very sensitive to the choices of the parameters that control the concentration at zero, namely a^γ, a^π, a^δ, and $\underline{p}$, respectively.

4. Synthetic data exercises

This section aims at comparing the performances of the various priors on simulated data. Before that, the merits of the semi-global framework are illustrated on synthetic data.

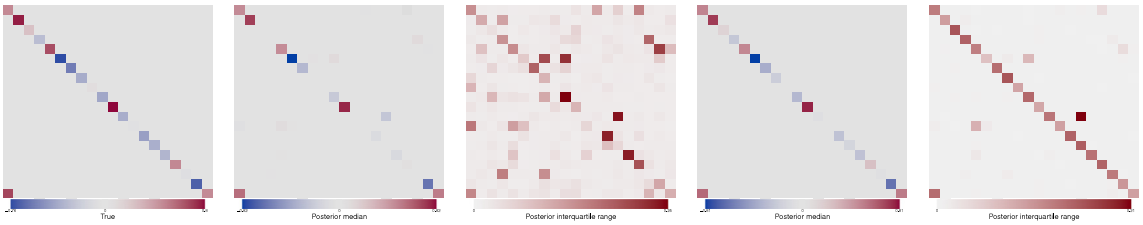

Fig. 3. Exemplary visualization of true and estimated VAR coefficients. From the left to the right: true values, posterior median HS, interquartile-range HS, posterior median HS semi-global–local, posterior interquartile-range HS semi-global–local. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.1. Structured/informed shrinkage: An illustration

Consider the following example of a VAR(1): Assume that all diagonal elements of $\boldsymbol{\Phi}$ are nonzero, whereas there is only one nonzero off-diagonal element. That is, taken as a whole, $\boldsymbol{\Phi}$ is not sparse. Fig. 3 demonstrates that a vanilla global–local prior—here, the HS prior—overshrinks the diagonal elements and does not shrink the off-diagonal elements strongly enough. The structured semi-global local version, however, adapts the degree of shrinkage in the two different subspaces of $\boldsymbol{\Phi}$.

4.2. Simulation study

In this section, we compare the performances of the different priors on simulated data. The structure of the simulation study is borrowed from Kastner and Huber (2020). We consider dense and sparse data generating processes (DGPs), as well as different dimensions in $T \in \{50, 100, 200\}$. Dimensionality is $M = 20$ throughout. In each setting, we sample 20 replicates of the DGP. In all scenarios, we distinguish between own- and cross-lags. The nonzero coefficients of the DGPs are drawn from Gaussians with mean μ_i and standard deviation σ_i for $i \in \{ol, cl\}$, where ol and cl denote own-lag and cross-lag, respectively. In both scenarios, the probability of own-lag coefficients being nonzero is 0.8; i.e., own-lags are always considered to be dense. The probability of cross-lag coefficients being nonzero is 0.01 in the sparse scenario, 0.1 in the medium scenario, and 0.8 in the dense scenario. Further, the scenarios differ regarding the signal-to-noise ratio. In the sparse scenario, we set $\mu_{ol} = \sigma_{ol} = \mu_{cl} = \sigma_{cl} = 0.3$. In the medium scenario, we set $\mu_{ol} = \sigma_{ol} = 0.15$ and $\mu_{cl} = \sigma_{cl} = 0.1$. In the dense scenario, we set $\mu_{ol} = \sigma_{ol} = 0.15$ and $\mu_{cl} = \sigma_{cl} = 0.01$. To simulate stable VAR processes, we control that none of the eigenvalues of the companion form have a modulus greater than one (Lütkepohl, 2005).

Elements in $\boldsymbol{u}$ are nonzero with probability 0.1. Furthermore, we set $\mu_u = \sigma_u = 0.001$. The AR(1) processes driving the log variances of the orthogonalized errors have mean $\mu_{\sigma i} = -10$, persistencies $\rho_{\sigma i}$ in the range $[0.85, 0.98]$, and standard deviations $\sigma_{\sigma i}$ in the range $[0.1, 0.3]$.

The root mean squared error (RMSE) between the posterior mean of the coefficients and the true parameter values is a common measure to evaluate the performance:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{R}\sum_{r=1}^{R}\phi_i^{(r)} - \phi_i\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\phi}_i - \phi_i)^2}, \tag{27}$$

where $\phi_i^{(r)}$ denotes the rth posterior sample, and $\hat{\phi}_i = \frac{1}{R}\sum_{r=1}^{R}\phi_i^{(r)}$ is the posterior mean of the ith VAR coefficient.

The following prior specifications are considered: DL, NG, and R2D2 denote the vanilla global–local implementations with the concentration parameters being fixed. In DL, $a^\gamma = \frac{1}{K}$, the value proposed by Kastner and Huber (2020). In NG and R2D2, $a^\delta = a^\pi = \frac{1}{2K}$; i.e. DL, NG, and R2D2 are specified to have the same concentration at zero. Further, in R2D2, $b^\pi = 0.5$, the value proposed by Zhang et al. (2022). In NG, $b^\delta = 0.5$ and $c = \frac{a^\delta}{2}$, such that NG and R2D2 differ only in their respective kernels: whereas NG defines hyperpriors for the variance of a normal distribution, R2D2 defines hyperpriors for the variance of a double-exponential distribution. In DL_a, NG_a, and R2D2_a, the subscript 'a' indicates that the following discrete prior is placed on the corresponding concentration parameter: The support points of the discrete priors $\tilde{\boldsymbol{a}}^\gamma = \tilde{\boldsymbol{a}}^\delta = \tilde{\boldsymbol{a}}^\pi = \tilde{\boldsymbol{a}} = (\tilde{a}_1 = \frac{1}{1000}, \ldots, \tilde{a}_{1000} = 1)'$ are equally spaced, and the probabilities are proportional to the density of an exponential distribution with rate 1. In MP_LIT, $\lambda_1 = 0.16$ and $\lambda_2 = 0.004$. In SHM, $c_i = d_i = 0.01$ for $i = 1, 2$. For all SSVS priors, we closely follow the semi-automatic approach of George et al. (2008) by setting $\tau_{0i} = \frac{1}{100}\sqrt{\widehat{var}(\phi_i)}$ and $\tau_{1i} = 100\sqrt{\widehat{var}(\phi_i)}$. Here, $\widehat{var}(\phi_i)$ denotes the variance of the posterior distribution resulting from a conjugate flat normal-Wishart prior, which is available in closed form. George et al. (2008) use the variance of the ordinary least squares (OLS) estimator. However, in situations where the number of covariates per equation exceeds the number of observations, the OLS estimator might not exist. For SSVS, the prior inclusion probability is fixed at $\underline{p} = 0.5$, whereas for SSVS_p, the subscript 'p' indicates that a Beta(1, 1) is placed on $\underline{p}$. A superscript '*' following a prior indicates a semi-global

Table 2
100×RMSE for simulated data. The reported values are the medians stemming from 20 simulations per setting. The smallest value in each column is indicated in bold font.

Model	Sparse			Medium			Dense		
	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$
MP_LIT	5.63	4.68	4.03	**5.09**	4.49	4.08	3.18	2.88	2.60
SHM	5.51	4.70	4.04	5.35	4.56	4.07	3.50	**2.50**	**2.01**
DL	5.38	3.27	2.22	6.02	4.63	3.57	4.50	3.77	2.82
DL_a	5.83	3.49	2.37	6.17	4.69	3.56	5.07	4.02	2.94
DL_a^*	5.29	3.22	2.19	6.07	4.58	3.40	4.79	3.85	2.79
HS	5.22	3.31	2.25	5.75	4.63	3.55	4.05	3.52	2.62
HS*	**3.89**	**2.52**	**1.77**	5.26	**4.18**	**3.30**	**3.14**	2.63	2.10
NG	5.25	3.24	2.19	5.73	4.62	3.55	4.16	3.67	2.78
NG_a	5.52	3.45	2.36	5.74	4.62	3.54	4.44	3.82	2.85
NG*	4.83	3.04	2.11	5.56	4.63	3.49	3.99	3.51	2.72
NG_a^*	4.65	2.94	2.15	5.47	4.36	**3.30**	4.08	3.44	2.57
R2D2	5.31	3.19	2.19	5.72	4.67	3.57	4.09	3.59	2.73
$R2D2_a$	5.49	3.39	2.34	5.73	4.60	3.54	4.35	3.80	2.83
R2D2*	4.82	2.98	2.09	5.57	4.65	3.49	3.96	3.48	2.60
$R2D2_a^*$	4.56	2.91	2.10	5.47	4.36	3.31	4.05	3.36	2.50
SSVS	5.48	3.67	2.68	5.96	5.17	4.29	4.11	3.65	2.90
$SSVS_p$	5.93	4.38	2.60	6.11	5.51	4.85	4.18	3.91	3.42
$SSVS_p^*$	5.57	3.56	2.25	6.15	5.45	4.69	4.18	3.81	3.08

modification, e.g., the semi-global–local HS prior is denoted by HS*. Sub- and superscripts can also be combined: For instance, NG_a^* stands for the semi-global–local NG prior where the concentration parameters $a_j^\delta, j = 1, \ldots, 2p$, are treated hierarchically, whereas NG* stands for the semi-global–local NG prior where the concentration parameters $a_j^\delta = 1/(2K)$, $j = 1, \ldots, 2p$, are fixed.

Table 2 reveals the results of the simulation study. HS* has the lowest RMSE in six out of nine scenarios: it is best in all sparse scenarios, in the medium scenarios with at least 100 observations, and in the dense scenarios with only 50 observations. In the dense scenarios with at least 100 observations, SHM has the lowest scores. Concerning GL priors, the semi-global versions perform better than the vanilla versions in all scenarios. In the sparse and the medium scenarios, GL priors in general, i.e., the vanilla and the semi-global versions, tend to perform better than the Minnesota priors (with one notable exception). In the dense scenarios, SHM is overall best, closely followed by HS*. SHM is considerably better than MP_LIT in the dense scenarios with at least 100 observations, whereas in the remaining scenarios their performance in terms of the RMSE is comparable. Turning to the various specifications of SSVS, we tend to observe competitive performance only in the sparse scenarios.

5. Empirical application to data of the US economy

In Section 5.1, we shortly summarize the dataset, and present the model specifications and the forecasting design. In Section 5.2, we inspect the posterior distributions of VAR coefficients arising from different prior distributions. In Section 5.3, we assess out-of-sample forecasting accuracy of the different models. In Section 5.4, we demonstrate the merits of combining forecasts of different models in a dynamic fashion via dynamic model averaging (DMA).

5.1. Data overview, model specification and forecasting design

The aim of the empirical application is to forecast economically relevant US time series. We use the quarterly dataset provided by McCracken and Ng (2021), which is based on the well-known Stock and Watson (2012) dataset. The sample period ranges from 1959:Q4 to 2020:Q1. All in all, we include $M = 21$ quarterly time series with the intention to cover the most important segments of the US economy. The complete list of all variables used in the following illustration is provided in Table 6 (Appendix C).

Similar to Huber and Feldkircher (2019), Cross et al. (2020), and Chan (2021b), most variables enter the models in log differences to be interpreted as growth rates, except interest rates, which are already defined in rates and hence taken in levels. In our experience, this choice results in more stable predictive densities compared to data taken in (log) levels. Another stream of literature deals explicitly with the problem of potentially non-stationary data by eliciting priors accordingly (cf., e.g., Giannone et al., 2019; Sims & Zha, 1998; Villani, 2009). It has to be acknowledged that results about the performances of the priors might be different when considering different transformations. A preliminary analysis (available upon request), however, suggested that many of the results were qualitatively similar when the data were taken in log levels.

A very important and often underemphasized choice when modeling macroeconomic time series with VARs is the number of lags of endogenous variables included as predictors. In the Bayesian framework, fixing p at a certain number implies prior information that higher lags do not carry any important information. Hence, Litterman (1986) suggests

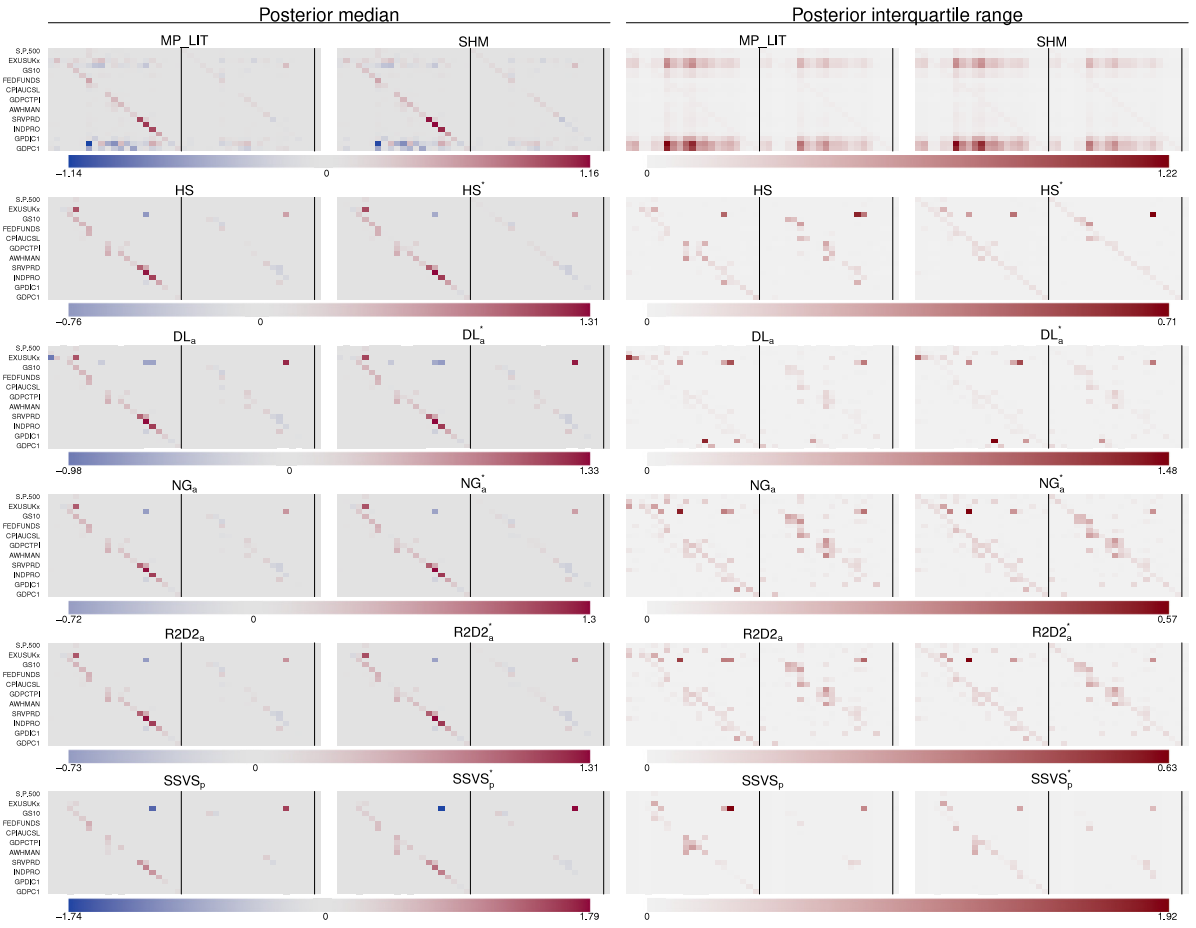

Fig. 4. Posterior summaries for VAR(2) coefficients arising from different prior distributions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

estimating as many lags as is computationally feasible, in which the prior has to be designed such that irrelevant coefficients get shrunk to zero. However, the more lags we include, the severer the problem of overfitting becomes, and the harder it will be to regularize the parameter space—not to mention the increased computational burden. Typical choices for quarterly data in similar dimensions are either four or five lags (Chan, 2021b; Cross et al., 2020; Giannone et al., 2015; Huber & Feldkircher, 2019). We estimate VAR(p)s for $p \in \{1, 2, 3, 4, 5\}$ to investigate to what degree longer lags improve/worsen forecasting performance.

The forecasting performance is evaluated by means of a recursive pseudo-out-of-sample forecasting exercise. Based on the initial estimation window ranging from 1959:Q4 to 1980:Q1, one-step-ahead and four-step-ahead predictive densities are evaluated. After that, the estimation window is expanded by one quarter and the model is re-estimated. For each estimation step, we run 10 independent MCMC chains. Per chain, we keep 15,000 draws after a burn-in of 5000 iterations. This procedure is consequently repeated until the end of the sample 2020:Q1 is reached.

The prior specifications we consider in the empirical application are the same as in Section 4.2. In what follows, we also include an intercept term. The prior for the intercept is independent normal centered at zero with variance 1000 for all models.

5.2. Inspecting the posterior distributions

Before presenting the results of the forecasting exercise, it is worth analyzing the posterior distributions arising from the different priors. Fig. 4 depicts posterior medians and posterior interquartile ranges of different VAR(2) models.[6] All models, even the vanilla SSVS and GL priors, detect many signals associated with the first own-lag and few signals in the remaining part of the coefficient space. Concerning the cross-lag coefficients, the posteriors arising from the Minnesota

[6] Although we find signals with higher lag orders, the out-of-sample results indicate that two lags are usually enough (cf. Section 5.3).

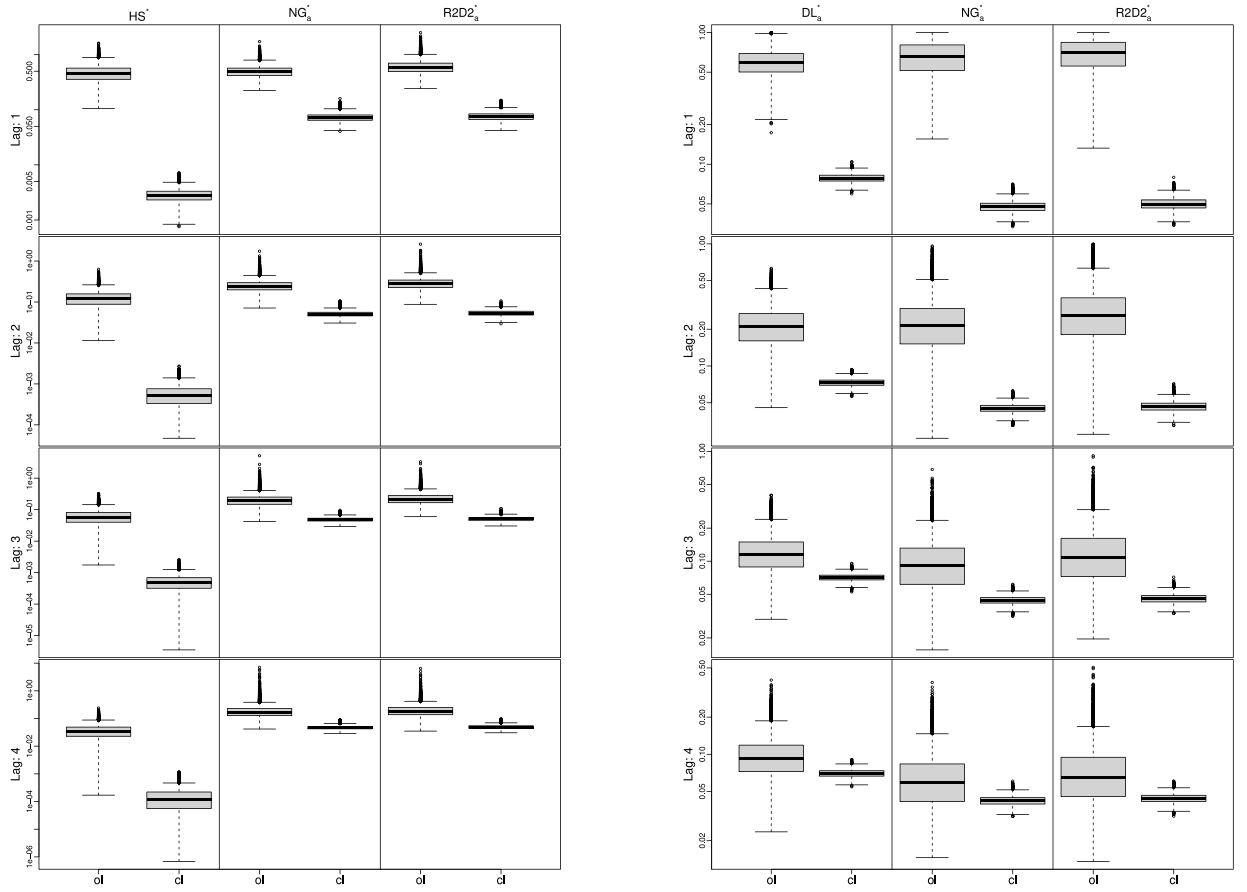

Fig. 5. Box plots of the posterior distributions of the semi-global hyperparameters for a VAR(4). Left panel: Posteriors of $\sqrt{\zeta_j}$ for HS* (left), NG_a^* (middle), R2D2_a^* (right), $j \in \{\text{lag: } 1, \ldots, 4\} \times \{\text{own-lag, cross-lag}\}$. Right panel: Posteriors of a_j^γ for DL_a^* (left), a_j^δ for NG_a^* (middle), a_j^π for R2D2_a^* (right), $j \in \{\text{lag: } 1, \ldots, 4\} \times \{\text{own-lag, cross-lag}\}$.

priors show relatively high uncertainty, whereas the remaining priors only find sparse signals. The differences between the vanilla and the semi-global versions of the SSVS and GL priors are hardly detectable by visual exploration only.

Table 3 displays the posterior mean of the Hoyer measure.[7] In the first lag, the posteriors under all priors are dense, whereas for the Minnesota priors, own-lags in general are dense. Moreover, the Minnesota priors are always denser than the remaining priors for all considered lag lengths with respect to both own-lag and cross-lag coefficients. Last but not least, especially on the off-diagonals, the posteriors under all priors except SHM and MP_LIT can get extremely sparse. The heterogeneous results regarding sparsity show that posterior inference is strongly affected by the prior assumptions, which stresses the importance of careful prior elicitation. This corroborates the findings of Jarociński and Maćkowiak (2017): In order to answer the question whether one variable is relevant for another variable, the authors derive a closed-form expression for the posterior probability of Granger non-causality and a generalization thereof in a homoskedastic VAR with a conjugate prior. They find that the posterior probability for zero restrictions increases as the prior gets looser. Since prior choices matter, we think it can be delusive to draw general conclusions about sparseness mainly by analyzing posterior quantities. Fava and Lopes (2021) elaborate on how such posterior results often just spread illusions.

Comparing the sparseness between GL priors, we observe that the own-lag coefficients of the semi-global modifications are denser than their vanilla GL counterparts. HS* discriminates the strongest between own-lags and cross-lags in terms of sparsity, especially in more distant lags. In this context, it is worth analyzing the posterior distributions of the (semi-)global hyperparameters. Fig. 5 depicts, exemplary for a VAR(4), box plots of the posterior distribution of $\sqrt{\zeta_j}$ for HS*, NG_a^*, and R2D2_a^*, and the posterior distribution of a_j^γ for DL_a^*, a_j^δ for NG_a^*, and a_j^π for R2D2_a^*, $j \in \{\text{lag} : 1, \ldots, 4\} \times \{\text{own-lag, cross-lag}\}$. For all models there is a clear discrimination in all lags between the own-lag and cross-lag. That is, shrinkage is always

[7] We compute the Hoyer measure for every single coefficient posterior draw, which then provides us with the posterior distribution of the measure.

Table 3
Posterior means of the Hoyer measure, separately for own-lag (ol) and cross-lag (cl) coefficients for VARs of order $p \in \{1, \ldots, 5\}$.

(a) DL_a

p	ol/cl	$l = 1$	$l = 2$	$l = 3$	$l = 4$	$l = 5$
1	ol	0.30	–	–	–	–
	cl	0.78	–	–	–	–
2	ol	0.37	0.48	–	–	–
	cl	0.82	0.84	–	–	–
3	ol	0.37	0.53	0.68	–	–
	cl	0.82	0.84	0.85	–	–
4	ol	0.37	0.56	0.71	0.75	–
	cl	0.82	0.85	0.85	0.88	–
5	ol	0.38	0.56	0.74	0.76	0.78
	cl	0.82	0.85	0.85	0.86	0.84

(b) DL_a^*

p	ol/cl	$l = 1$	$l = 2$	$l = 3$	$l = 4$	$l = 5$
1	ol	0.26	–	–	–	–
	cl	0.78	–	–	–	–
2	ol	0.33	0.40	–	–	–
	cl	0.82	0.84	–	–	–
3	ol	0.33	0.44	0.64	–	–
	cl	0.83	0.84	0.86	–	–
4	ol	0.32	0.45	0.65	0.72	–
	cl	0.83	0.85	0.85	0.89	–
5	ol	0.33	0.45	0.70	0.74	0.76
	cl	0.82	0.85	0.85	0.87	0.85

(c) SHM

p	ol/cl	$l = 1$	$l = 2$	$l = 3$	$l = 4$	$l = 5$
1	ol	0.23	–	–	–	–
	cl	0.60	–	–	–	–
2	ol	0.30	0.19	–	–	–
	cl	0.61	0.61	–	–	–
3	ol	0.31	0.22	0.22	–	–
	cl	0.61	0.61	0.61	–	–
4	ol	0.31	0.21	0.22	0.25	–
	cl	0.61	0.61	0.61	0.61	–
5	ol	0.31	0.22	0.22	0.25	0.26
	cl	0.61	0.61	0.61	0.61	0.61

(d) HS

p	ol/cl	$l = 1$	$l = 2$	$l = 3$	$l = 4$	$l = 5$
1	ol	0.31	–	–	–	–
	cl	0.75	–	–	–	–
2	ol	0.38	0.54	–	–	–
	cl	0.83	0.82	–	–	–
3	ol	0.38	0.65	0.76	–	–
	cl	0.85	0.81	0.86	–	–
4	ol	0.38	0.77	0.81	0.77	–
	cl	0.86	0.83	0.84	0.85	–
5	ol	0.40	0.77	0.81	0.76	0.85
	cl	0.87	0.84	0.84	0.86	0.81

(e) HS^*

p	ol/cl	$l = 1$	$l = 2$	$l = 3$	$l = 4$	$l = 5$
1	ol	0.25	–	–	–	–
	cl	0.80	–	–	–	–
2	ol	0.32	0.27	–	–	–
	cl	0.84	0.89	–	–	–

(*continued on next page*)

Table 3 (*continued*).

3	ol	0.32	0.30	0.37	–	–
	cl	0.85	0.91	0.93	–	–
4	ol	0.32	0.29	0.40	0.46	–
	cl	0.85	0.91	0.92	0.88	–
5	ol	0.32	0.29	0.41	0.48	0.58
	cl	0.85	0.89	0.91	0.87	0.85

(f) MP_LIT

p	ol/cl	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$
1	ol	0.24	–	–	–	–
	cl	0.60	–	–	–	–
2	ol	0.29	0.21	–	–	–
	cl	0.60	0.61	–	–	–
3	ol	0.29	0.22	0.22	–	–
	cl	0.60	0.61	0.61	–	–
4	ol	0.29	0.22	0.23	0.26	–
	cl	0.60	0.61	0.61	0.61	–
5	ol	0.29	0.23	0.23	0.26	0.25
	cl	0.60	0.61	0.61	0.61	0.61

(g) NG_a

p	ol/cl	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$
1	ol	0.29	–	–	–	–
	cl	0.75	–	–	–	–
2	ol	0.37	0.47	–	–	–
	cl	0.79	0.78	–	–	–
3	ol	0.37	0.55	0.66	–	–
	cl	0.80	0.77	0.81	–	–
4	ol	0.37	0.63	0.68	0.71	–
	cl	0.80	0.76	0.79	0.79	–
5	ol	0.38	0.63	0.71	0.73	0.75
	cl	0.80	0.76	0.78	0.78	0.78

(h) NG_a^*

p	ol/cl	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$
1	ol	0.25	–	–	–	–
	cl	0.77	–	–	–	–
2	ol	0.32	0.31	–	–	–
	cl	0.80	0.79	–	–	–
3	ol	0.31	0.36	0.57	–	–
	cl	0.80	0.78	0.82	–	–
4	ol	0.31	0.37	0.57	0.68	–
	cl	0.80	0.78	0.81	0.81	–
5	ol	0.32	0.37	0.63	0.69	0.73
	cl	0.80	0.77	0.80	0.80	0.80

(i) $R2D2_a$

p	ol/cl	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$
1	ol	0.29	–	–	–	–
	cl	0.76	–	–	–	–
2	ol	0.37	0.48	–	–	–
	cl	0.80	0.79	–	–	–
3	ol	0.37	0.56	0.66	–	–
	cl	0.81	0.78	0.82	–	–
4	ol	0.37	0.63	0.69	0.72	–
	cl	0.82	0.77	0.80	0.80	–
5	ol	0.38	0.64	0.73	0.74	0.76
	cl	0.82	0.77	0.79	0.79	0.79

(*continued on next page*)

Table 3 (*continued*).

(j) $R2D2_a^*$

p	ol/cl	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$
1	ol	0.25	–	–	–	–
	cl	0.77	–	–	–	–
2	ol	0.32	0.31	–	–	–
	cl	0.81	0.80	–	–	–
3	ol	0.32	0.35	0.55	–	–
	cl	0.81	0.78	0.83	–	–
4	ol	0.32	0.36	0.55	0.66	–
	cl	0.81	0.79	0.82	0.82	–
5	ol	0.32	0.37	0.62	0.68	0.72
	cl	0.81	0.78	0.82	0.81	0.81

(k) SSVS

p	ol/cl	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$
1	ol	0.33	–	–	–	–
	cl	0.82	–	–	–	–
2	ol	0.40	0.64	–	–	–
	cl	0.84	0.83	–	–	–
3	ol	0.38	0.73	0.60	–	–
	cl	0.85	0.81	0.82	–	–
4	ol	0.36	0.75	0.42	0.57	–
	cl	0.84	0.74	0.64	0.87	–
5	ol	0.38	0.73	0.47	0.45	0.67
	cl	0.84	0.70	0.64	0.81	0.58

(l) $SSVS_p^*$

p	ol/cl	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$
1	ol	0.32	–	–	–	–
	cl	0.81	–	–	–	–
2	ol	0.36	0.62	–	–	–
	cl	0.87	0.89	–	–	–
3	ol	0.34	0.72	0.55	–	–
	cl	0.84	0.58	0.73	–	–
4	ol	0.34	0.70	0.52	0.41	–
	cl	0.85	0.60	0.70	0.41	–
5	ol	0.36	0.70	0.41	0.37	0.43
	cl	0.85	0.68	0.68	0.46	0.31

stronger for cross-lag coefficients than for own-lag coefficients in all lags. The strength of discrimination between the own-lag and cross-lag decreases with higher lags, and is the strongest for HS. Last but not least, the posterior distributions of the semi-global hyperparameters concentrate around smaller values for more distant lags than for closer lags; i.e., shrinkage increases with increasing lags.

5.3. Model evidence: Log predictive likelihoods

Following Geweke and Amisano (2010) we evaluate forecasting performance in terms of log predictive likelihoods, treating the period ranging from 1959:Q4 to 1980:Q1 as the training sample (initial estimation window). Evaluation of the predictive densities begins in 1980:Q2 or 1981:Q1 for one-step-ahead forecasts and four-step-ahead forecasts, respectively. The estimation window is recursively expanded by one observation until the end of the evaluation period 2020:Q1 is reached.

Table 4 depicts sums of one- and four-step-ahead log predictive likelihoods relative to the traditional Minnesota prior with four lags (MP_LIT(4)). We start by inspecting the one-step-ahead scores. The most important findings are the following: The overall best performance is achieved by the semi-global modification of HS with four lags (HS*(4)). In almost all cases, the semi-global modifications clearly improve forecasting performance compared to their vanilla GL counterparts. Moreover, the semi-global modifications usually perform better than the Minnesota priors, whereas the vanilla GL priors do not. SHM priors perform well in general. Comparing the scores of SHM and MP_LIT reveals that learning the shrinkage parameters of the Minnesota prior from the data always improves forecasting performance, which corroborates findings of Giannone et al. (2015) and Cross et al. (2020). Among all considered priors, SSVS priors achieve the lowest scores.

Concerning the performance of GL priors compared to Minnesota priors: The vanilla GL priors do not produce more accurate forecasts than SHM. This contradicts results of Huber and Feldkircher (2019) at first sight. The weaker

Table 4
Sum of one- and four-step-ahead log predictive likelihoods relative to MP_LIT(4), which itself has scores of 12602 (one step) and 11522 (four steps). The highest value in each column is indicated in bold font.

Model	One-step-ahead					Four-step-ahead				
	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$
MP_LIT	−67	−17	−8	0	−13	−64	−10	−1	0	−1
SHM	−55	0	10	7	4	−61	−10	−4	2	5
HS	−58	−27	−23	−36	−43	−39	17	14	4	10
DL	−44	10	11	−6	−23	−25	45	31	29	22
NG	−46	11	9	−2	−14	−30	40	41	30	27
R2D2	−40	8	9	3	−13	−30	44	35	30	25
DL_a	−41	16	6	−12	−53	−21	39	36	12	−13
NG_a	−44	9	−2	−28	−53	−27	42	40	16	8
$R2D2_a$	−45	8	−6	−20	−43	−25	42	37	16	−1
NG*	−47	17	16	4	−14	−30	40	43	33	33
R2D2*	−39	13	18	6	−11	−27	37	39	41	22
HS*	−33	24	**41**	**50**	**38**	−39	21	43	**45**	**53**
DL_a^*	−30	32	32	13	−29	−20	**52**	43	21	1
NG_a^*	−28	**33**	34	19	−16	−27	**52**	**56**	36	20
$R2D2_a^*$	−28	**33**	40	17	−10	−23	50	53	36	21
SSVS	−63	−29	−29	−49	−90	−45	−3	−11	−32	−42
$SSVS_p$	−87	−61	−62	−88	−106	−53	−23	−14	−36	−38
$SSVS_p^*$	−84	−39	−36	−63	−84	−69	−35	−15	−39	−43

performance of SHM in Huber and Feldkircher (2019) can be explained (at least to some extent) by the vague prior imposed on $\boldsymbol{u}$, a normal prior with zero mean and variance 10. The fact that the prior choice for $\boldsymbol{u}$ indeed plays a crucial role in terms of forecasting performance is demonstrated in Section 5.5. Interestingly, among the vanilla GL priors, HS has the lowest scores. For all orders, it has even lower scores than MP_LIT(4) concerning one-step-ahead forecasts. The story changes radically, though, when looking at the semi-global modifications. For $p \in \{3, 4, 5\}$, HS* performs best. DL_a^*, $R2D2_a^*$, and NG_a^* with $p \in \{2, 3, 4\}$ also perform considerably better than the best Minnesota prior, SHM(3).

Regarding four-step-ahead predictions, we again observe that the semi-global modifications usually perform better than the vanilla implementations. Here, the overall best performance is achieved by $NG_a^*(3)$. What stands out compared to one-step-ahead forecasts is that even the vanilla GL priors are doing better than the Minnesota priors. In other words, at longer forecasting horizons there is stronger evidence in favor of sparsity inducing priors.

Turning the focus to lag length, we find that for the dataset at hand, competitive forecasts require at least two lags of the endogenous variables as predictors. Most priors perform best with $p = 2$ or $p = 3$ lags. HS* and MP_LIT perform best with $p = 4$ lags. Interestingly, concerning one-step-ahead forecasts, the performances of HS, DL, DL_a, DL_a^*, NG, NG_a, NG_a^*, R2D2, $R2D2_a$, and $R2D2_a^*$ deteriorate with $p > 3$. In this respect, HS* and SHM—the only priors without a clear drop in scores for $p = 5$—are the most robust priors.

The forecasting exercise also reveals both the potential and limitations of the hierarchical treatment of hyperparameters. The Minnesota prior clearly benefits from treating both shrinkage parameters as random variables. For GL priors, the story is more complex. On the one hand, DL, NG, and R2D2 in their vanilla implementations do not improve by placing priors on the concentration parameters a^γ, a^δ, and a^π, respectively. On the other hand, the semi-global modifications of both NG and R2D2 unfold their full potential only if both a_j^δ and a_j^π, $j = 1, \ldots, 2p$, are equipped with hyperpriors. This finding confirms our conjecture that GL priors have limitations when the noise level and the sparsity differ in different regions of the parameter space.

Fig. 6 depicts how the forecasting scores evolve over time for an illustrative selection of different priors. It reveals that the time between the early-2000s recession and the financial crisis of 2007/2008 marks a turning point in the evolution of the scores. Especially during the time of the zero lower bound (ZLB), which we define as the period after the financial crisis when the FFR is below 20 basis points (cf. Mavroeidis, 2021), the sparse models—the semi-global modifications of the GL priors—accumulate much higher scores than the denser SHM. The sparsity-inducing priors seem to capture the ZLB considerably well, by virtually zeroing almost all cross-lag coefficients in the equations of the interest rates. Before the crisis, however, SHM is very competitive. The only sparse prior on par with SHM before the crisis is HS*. Overall, there is no single model that performs best over the whole evaluation period. The findings suggest that there may be no general answer to the question of whether sparse or dense modeling techniques perform better at forecasting economic data.

In real applications, there might be interest in forecasts of only a small set of variables (cf., e.g., Bańbura et al., 2010; Feldkircher et al., 2024). Fig. 7 depicts cumulative log predictive likelihoods where instead of the full 21-dimensional predictive density, the joint predictive density of three focal variables, namely GDP, CPI and FFR, is evaluated. Similar to before, we observe that SHM dominates until the mid-2000s, and the semi-global–local shrinkage priors since the mid-2000s.

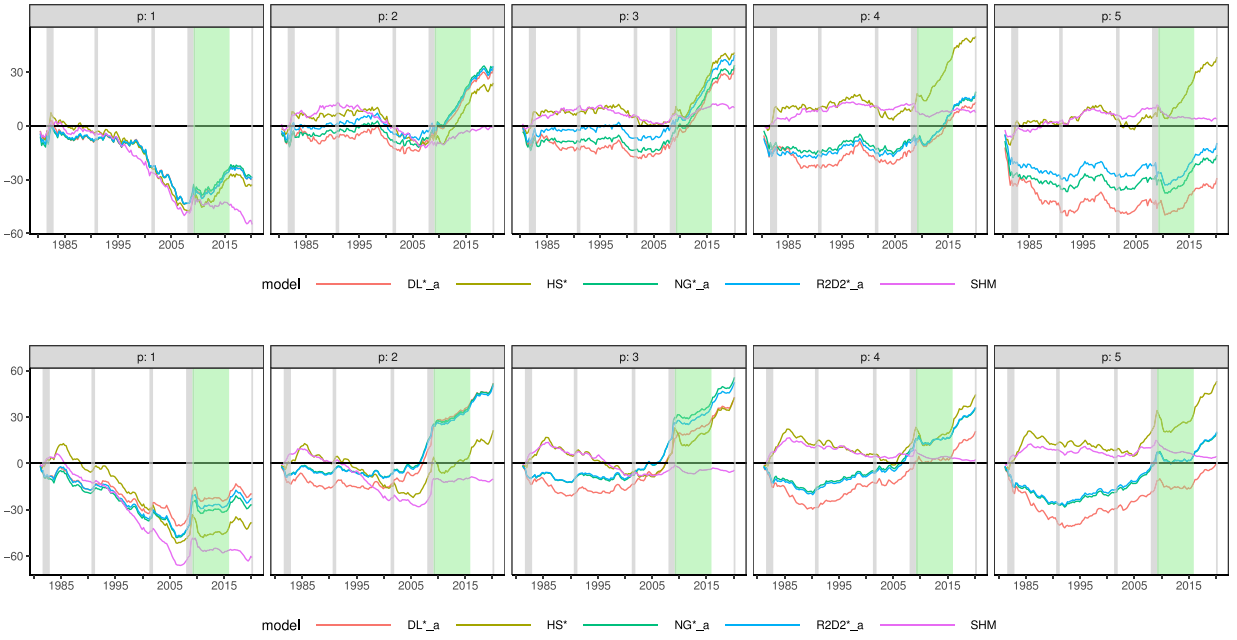

Fig. 6. One-step-ahead (top panels) and four-step-ahead (bottom panels) cumulative log predictive likelihoods relative to MP_LIT with $p = 4$ considering the full 21-dimensional predictive density. Gray shaded areas are recessions dated by the National Bureau of Economic Research. The green shaded area denotes the period of the zero lower bound. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

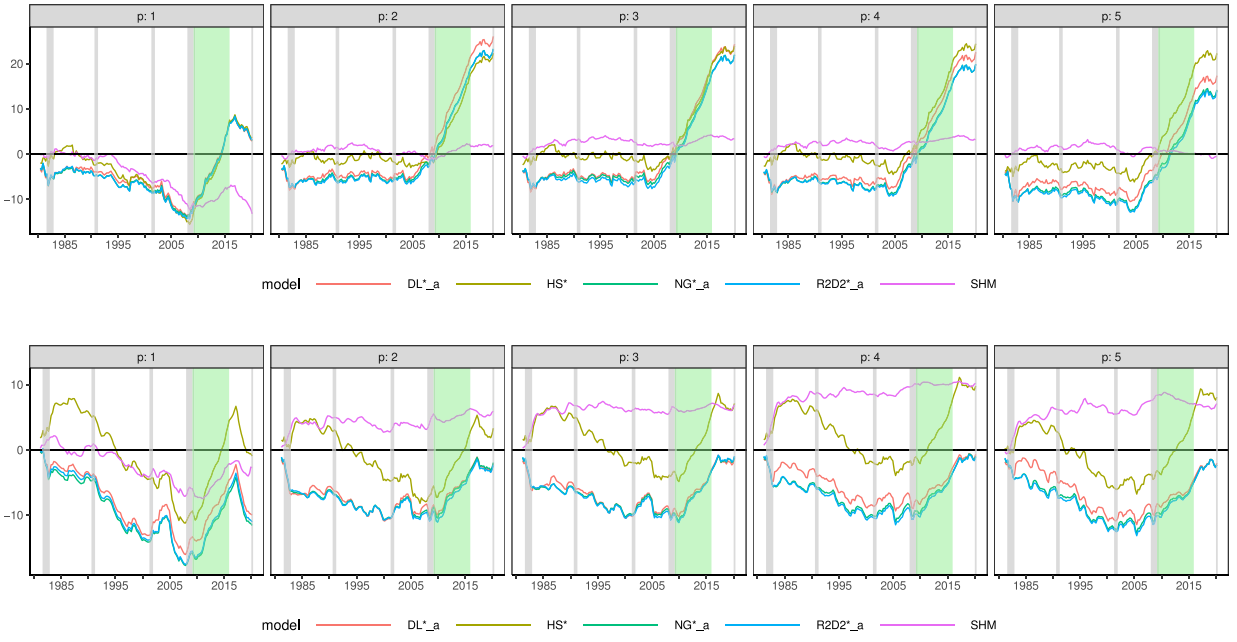

Fig. 7. One-step-ahead (top panels) and four-step-ahead (bottom panels) cumulative log predictive likelihoods relative to MP_LIT with $p = 4$ considering the joint predictive density of GDP, CPI, and FFR. Gray shaded areas are recessions dated by the National Bureau of Economic Research. The green shaded area denotes the period of the zero lower bound. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.4. Dynamic model averaging

The previous sections highlighted the heterogeneity of model performance over time. The question arises whether combining weighted forecasts can improve forecasting performance.

Dynamic model averaging (DMA, Koop & Korobilis, 2012; Onorante & Raftery, 2016; Raftery et al., 2010) is a straightforwardly implementable averaging scheme, where the weights are sequentially updated as new observations become available, depending on the forecasting performance up to this point in time. A model with strong support from the data for a given period will receive a higher weight for the following period. By contrast, a model performing relatively worse will receive a lower weight.

In a nutshell, DMA works as follows. We denote by $PL_{t|t-1,i}$ the one-step-ahead predictive likelihood in t for model i within model space M. The predicted weight $\omega_{t|t-1,i}$ associated with model i is computed as follows:

$$\omega_{t|t-1,i} = \frac{\omega^{\alpha}_{t-1|t-1,i}}{\sum_{i\in\mathcal{M}} \omega^{\alpha}_{t-1|t-1,i}}, \tag{28}$$

where $0 \leq \alpha \leq 1$ denotes a discount factor. Then, the model-updating equation is

$$\omega_{t|t,i} = \frac{\omega_{t|t-1,i} PL_{t|t-1,i}}{\sum_{i\in\mathcal{M}} \omega_{t|t-1,i} PL_{t|t-1,i}}. \tag{29}$$

To shed more light on the discount factor: standard (static) Bayesian model averaging is achieved by setting $\alpha = 1$, as $\omega_{t|t-1,i}$ would be proportional to the (training-sample) marginal likelihood of model i using data through time $t-1$. We follow Raftery et al. (2010) in specifying $\alpha = 0.99$, indicating that $PL_{t-1|t-2}$ receives 99% as much weight as $PL_{t|t-1}$.

The sparse models that we consider for DMA are DL^*_a, HS*, NG^*_a, and R2D2^*_a. MP_LIT and SHM, by contrast, represent the dense models. Concerning lag length, we average over VARs of order $p \in \{1, 2, 3, 4, 5\}$. Last but not least, we initialize the weights to be equally distributed over all considered models for the first prediction.

Turning towards the results, we first consider the one-step-ahead predictive density of all 21 variables. The top panel of Fig. 8 shows the evolution of model weights and cumulative log predictive likelihoods of the considered models relative to DMA over the whole evaluation period. Until the end of the 2000s, SHM with $p \in \{2, 3, 4, 5\}$ and HS* with $p = 4$ are dominating alternately. From 2009 until the end of the evaluation period, HS* with $p = 4$ receives the majority of the weights, the remaining semi-global–local priors have some weights, and SHM has quasi-zero weights. Analyzing the cumulative log predictive likelihoods, it turns out that DMA is practically never worse than any single model specification.

DMA can also improve forecasts if only a subset of variables is considered. The bottom panel of Fig. 8 depicts DMA model weights and cumulative log predictive likelihoods relative to DMA when only the predictive density of GDP, CPI, and FFR is taken into account. By and large, the evolution of the predictive scores is similar to when all variables are considered. However, the analysis of model weights reveals that the distribution of weights over the models is now more dispersed. $\text{DL}^*_a(2)$, which is practically not apparent when evaluating the full 21-dimensional predictive density, now receives the largest weights after 2010.

The result of DMA, of course, depends on the choice of α. To address this issue, Bernaciak and Griffin (2024) introduce the loss discounting framework (LDF), which extends DMA to general loss functions and more general discounting dynamics. We consider the following variant of LDF: Let $S_\alpha = \{\alpha_1, \ldots, \alpha_s\}$ denote a set of different discount factors. In a first layer, we apply DMA with discount factor α_i for $i = 1, \ldots, s$ to our model pool. This yields s dynamically averaged meta-models. In the second layer, we apply DMA with discount factor α_i for $i = 1, \ldots, s$ to the s meta-models. This process can be repeated arbitrarily. In the final meta-model layer, a single discount factor needs to be chosen. Adding more layers has a diminishing effect on the sequence of predictive distributions and therefore on the final result. It can be shown that the weights converge as the number of layers approaches infinity. In this respect, the final result is least affected by the choice of the discount factor in the last layer when a large number of layers L is chosen. Concretely, we set $S_\alpha = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99, 1\}$ and $l = 4$, which appears to be sufficient in our case in the sense that the exact choice of the discount factor in the final layer shows negligible impact. According to Fig. 8, LDF outperforms all single model specifications and DMA throughout the whole evaluation period with respect to the full 21-dimensional predictive density. Focusing on the joint predictive distribution of the three focal variables, LDF is on par with DMA and not worse than any single model in the long run.

To sum up, model performance is heterogeneous not only over time but also across the evaluated variables. Model averaging is a very effective tool for combining the merits of various modeling techniques, in contrast to the apparently unrewarding search for the single model that can do them all.

5.5. Sensitivity of results with respect to the prior for $\boldsymbol{\Sigma}_t$

In this section, we investigate the sensitivity of results with respect to different prior choices for $\boldsymbol{\Sigma}_t$. The main focus is on different prior choices for $\boldsymbol{u}$. Additionally, we briefly discuss a homoskedastic specification of the variance–covariance matrix.

The elements in $\boldsymbol{u}$ are often interpreted as contemporaneous coefficients. Hence, an obvious choice for $\boldsymbol{u}$ would be to use a prior of the same family as the prior on $\boldsymbol{\phi}$; e.g., DL/DL^*_a on $\boldsymbol{\phi}$ is combined with DL on $\boldsymbol{u}$. For DL on $\boldsymbol{u}$, for simplicity, we

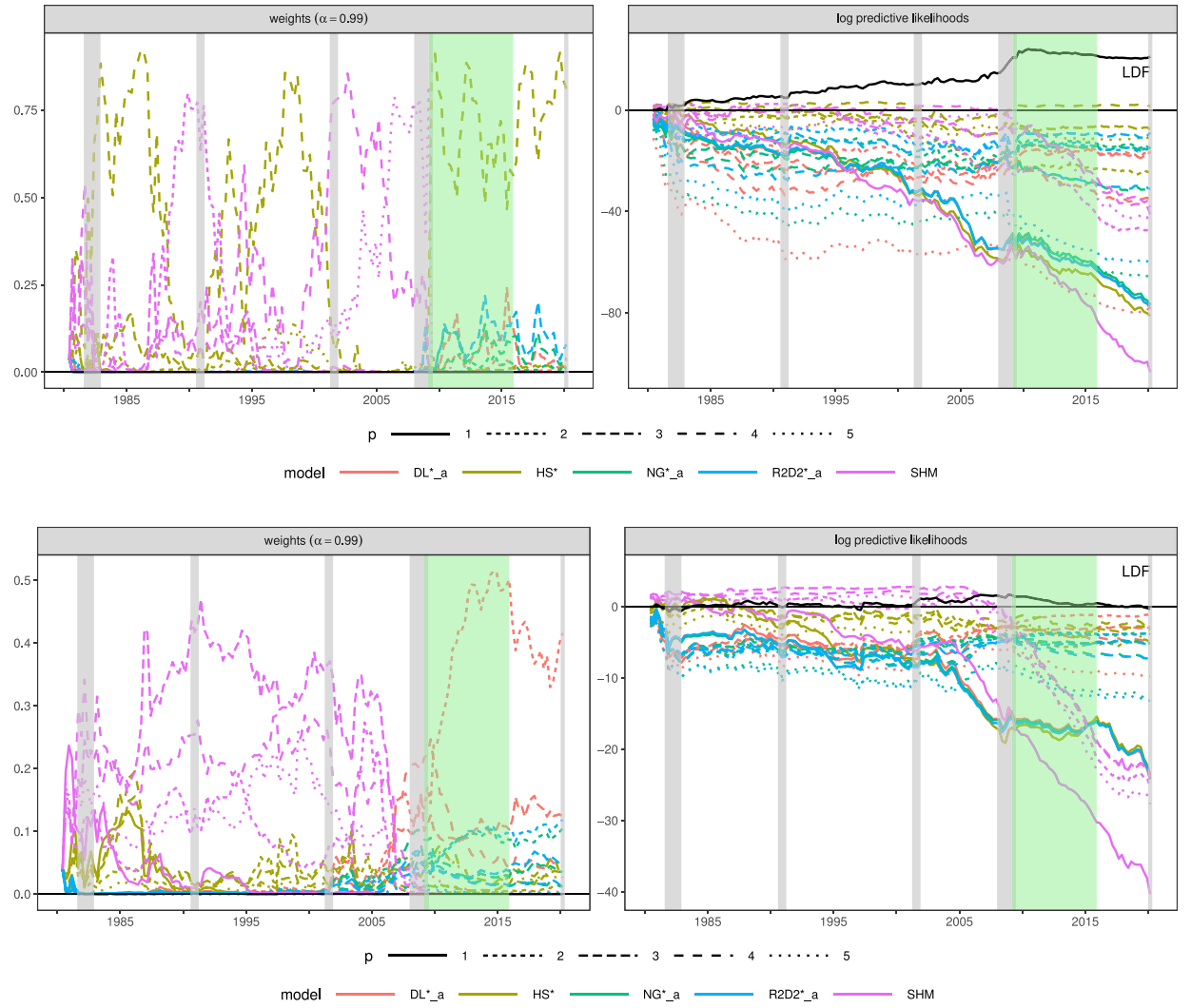

Fig. 8. Dynamic model averaging (DMA) for one-step-ahead predictions: model weights (left) and cumulative log predictive likelihoods relative to DMA (right) with discount factor $\alpha = 0.99$. Top panels: Weights and cumulative log predictive likelihoods considering the full 21-dimensional predictive density. Bottom panels: Weights and cumulative log predictive likelihoods considering the joint predictive density of GDP, CPI, and FFR. The black solid line stands for the cumulative log predictive likelihood of the loss discounting framework (LDF) relative to DMA. The gray shaded areas are recessions dated by the National Bureau of Economic Research. The green shaded area denotes the period of the zero lower bound. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

specify $a^{\gamma} = 1/n_u$, although this parameter could be learned from the data as well. Here, n_u is the number of elements in $\boldsymbol{u}$. For NG and R2D2 on $\boldsymbol{u}$, we set $a^{\delta} = a^{\pi} = 1/(2n_u)$, such that for DL, NG, and R2D2, the concentration at zero is matched. For SSVS on $\boldsymbol{u}$, we specify $\tau_{0i} = 1/1000$ and $\tau_{1i} = 1$. As a second alternative, we investigate a relatively uninformative normal prior with zero mean and variance 10 for each element in $\boldsymbol{u}$, independently, which we refer to as the FLAT prior. As a third option, we investigate a global(-but-not-local) shrinkage (GS) prior. More specifically, the prior is conditionally normal with zero mean and a global gamma hyperprior on the variance parameter: $u_{ij}|\lambda_g \sim N(0, \lambda_g)$, $\lambda_g \sim G(0.01, 0.01)$. GS is similar in spirit to SHM. In combination with SHM on $\boldsymbol{\phi}$, then the contemporaneous coefficient $\boldsymbol{u}$ represents the third distinct group besides own-lag and cross-lag coefficients.

Table 5 shows sums of one-step-ahead log predictive likelihoods relative to a VAR with $p = 4$ lags with MP_LIT on $\boldsymbol{\phi}$ and HS on $\boldsymbol{u}$ for an illustrative selection of VARs. GS or FLAT on $\boldsymbol{u}$ clearly deteriorates forecasting performance of all models compared to the HS used in the main analysis. Interestingly, GS is even worse than FLAT, which indicates that there are few strong signals that GS simply overshrinks. GL priors and SSVS on $\boldsymbol{u}$ are serious contenders to the HS. Overall, the results highlight the importance of a sparsity-inducing prior for $\boldsymbol{u}$. The sensitivity of forecasting performance with respect to the prior choice for $\boldsymbol{u}$ strengthen our view to use the same prior on $\boldsymbol{u}$ for all models in the main analysis, and could also

Table 5
Sum of relative one-step-ahead log predictive likelihoods for several VARs equipped with different priors on $\boldsymbol{u}$. The benchmark model is MP_LIT on $\boldsymbol{\phi}$ with $p = 4$ lags in combination with HS on $\boldsymbol{u}$. The flag */* indicates that the prior on $\boldsymbol{u}$ is of the same family as the prior on $\boldsymbol{\phi}$. The highest value in each row of the columns corresponding to $p \in \{2, 3, 4\}$ is indicated in bold font.

Prior on $\boldsymbol{\phi}$	$p = 2$ Prior on $\boldsymbol{u}$				$p = 3$ Prior on $\boldsymbol{u}$				$p = 4$ Prior on $\boldsymbol{u}$			
	HS	GS	FLAT	*/*	HS	GS	FLAT	*/*	HS	GS	FLAT	*/*
MP_LIT	−17	−96	−64	–	−8	−87	−73	–	**0**	−89	−63	–
SHM	**0**	−83	−44	–	**10**	−75	−59	–	**7**	−73	−46	–
HS	−27	−106	−78	−27	−23	−103	−95	−23	−36	−107	−92	−36
DL	**10**	−59	−35	0	**11**	−71	−66	3	−6	−83	−74	−3
NG	**11**	−72	−51	9	9	−70	−64	**15**	−2	−86	−74	**1**
R2D2	**8**	−75	−46	7	9	−74	−64	**13**	3	−80	−71	**9**
SSVS	−29	−105	−86	−25	−29	−101	−101	−28	−49	−125	−108	−51
HS*	**24**	−63	−27	**24**	**41**	−49	−35	**41**	**50**	−46	−19	**50**
DL_a^*	**32**	−57	−26	29	**32**	−69	−61	22	**13**	−82	−95	5
NG_a^*	**33**	−44	−18	32	**34**	−59	−44	**34**	**19**	−77	−59	14
R2D2_a^*	33	−57	−21	**36**	**40**	−54	−36	32	**17**	−66	−70	16

explain to some extent the relatively weak performance of SHM in Huber and Feldkircher (2019), where it is combined with the FLAT prior on $\boldsymbol{u}$.

Concerning the role of SV in forecasting macroeconomic data, we also considered homoskedastic VARs in a preliminary exercise. We assume that the diagonal matrix $\boldsymbol{D}_t = \boldsymbol{D}$ is constant for $t = 1, \ldots, T$. In that case, we specify independent inverse-gamma priors with both shape and scale equaling 0.01 for the diagonal elements in $\boldsymbol{D}$. The prior on $\boldsymbol{u}$ is the HS, as described in Section 2.4. In terms of forecasting performance, the results (available upon request) were overwhelmingly in favor of stochastic volatility. Therefore, we did not continue down this route of research any further. More comprehensive treatments of this matter with similar findings can be found in Carriero et al. (2019), Clark and Ravazzolo (2015), Koop and Korobilis (2013), among others.

6. Conclusion

Careful prior elicitation in Bayesian vector autoregressions goes back at least to Litterman (1986) and has been refined ever since. Our main contribution here was a discussion of modern variants of prior choices, both established and novel, and a side-by-side comparison in terms of their structural (sparsity) behavior. In doing so, we borrowed from well-established (Minnesota-type) structural knowledge as well as from more automatic (shrinkage-type) approaches, aiming to combine the best of both worlds. As a result of this aim, we proposed what we term semi-global–local priors. We compare the various specifications in terms of their in- and out-of-sample performance on simulated data, as well as on a practically interesting (and notoriously hard) macroeconomic prediction problem. Overall, we found that the semi-global–local priors outperformed the vanilla global–local priors and were strong competitors to the semi-hierarchical Minnesota prior, which itself already constitutes an improvement over the traditional Minnesota prior. Finally, we investigated how dynamic model averaging can help shed light on the temporal patterns representing historical regime shifts. It became clear that no single model dominated the others. Rather, we found that some models were sometimes better than others, and that dynamically combining them was fruitful.

While trying to comprehensively and integratively shed light on the 'illusion of sparsity' debate, we also want to point out that many questions remain unanswered. Issues left for future research include an in-depth analysis where forecasting performance between modeling reduced-form and structural-form coefficients is compared side by side. Moreover, a systematic analysis is still missing of the various approaches to modeling reduced-form VARs with time-varying variance–covariance matrices (e.g. Cholesky-SV vs. Factor-SV). A first notable attempt in this direction is undertaken by Chan (2023). As mentioned by an anonymous referee, the semi-global approach could be useful for giving structure to the variance–covariance matrix as well, e.g., by imposing group-specific shrinkage according to industry and economic sectors (cf. Kastner, 2019). Furthermore, while our forecasting exercise took into account the entire predictive distribution, we did not specifically focus on the assessment of tail risk. Threshold- and quantile-weighted scoring rules, as discussed by Gneiting and Ranjan (2011) or the asymmetric continuous probability score (ACPS) proposed by Iacopini et al. (2023) could serve as starting points in this regard. Lastly, we believe that this work could be expanded to cater to VARs with time-varying regression parameters (e.g., Bitto & Frühwirth-Schnatter, 2019; Feldkircher et al., 2024; Frühwirth-Schnatter-Schnatter & Knaus, 2021; Huber et al., 2019) where the issue of choosing good shrinkage priors is exacerbated in the sense that in addition to shrinkage towards zero, shrinkage towards constancy needs to be dealt with.

CRediT authorship contribution statement

Luis Gruber: Writing – review & editing, Writing – original draft, Software, Methodology. **Gregor Kastner:** Writing – review & editing, Writing – original draft, Software, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Proof of Theorem 3.1

In order to prove Theorem 3.1, we first have to establish the densities of the priors marginalized over the local scale ϑ_i.

Proposition A.1. *Given the DL prior (10), the density of ϕ_i for any $i \in \mathcal{A}_j$, $j = 1, \dots, k$, marginalized over the local scale ϑ_i, is given by*

$$p_{DL}(\phi_i | a_j^\gamma) = \frac{|\phi_i|^{(a_j^\gamma - 1)/2} K_{1-a_j^\gamma}(\sqrt{2|\phi_i|})}{2^{(1+a_j^\gamma)/2}\Gamma(a_j^\gamma)}, \tag{A.1}$$

where $K_\nu(x)$ is the modified Bessel function of the second kind.[8]

Proof. Cf. Bhattacharya et al. (2015). □

Proposition A.2. *Given the NG prior (11), the density of ϕ_i for any $i \in \mathcal{A}_j$, $j = 1, \dots, k$, marginalized over the local scale ϑ_i, is given by*

$$p_{NG}(\phi_i | \zeta_j a_j^\delta) = \frac{2^{1/2-a_j^\delta}\left(\frac{a_j^\delta}{\zeta_j}\right)^{a_j^\delta/2+1/4}}{\sqrt{\pi}\Gamma(a_j^\delta)} K_{(a_j^\delta - 1/2)}\left(\sqrt{\frac{a_j^\delta}{\zeta_j}}|\phi_i|\right)|\phi_i|^{a_j^\delta - 1/2} \tag{A.2}$$

Proof.

$$\begin{aligned}
p_{NG}(\phi_i | \zeta_j a_j^\delta) &= \int_0^\infty \frac{1}{\sqrt{2\pi\vartheta_i\zeta_j}} \exp\left\{-\frac{\phi_i^2}{2\vartheta_i\zeta_j}\right\} \frac{\left(\frac{a_j^\delta}{2}\right)^{a_j^\delta}}{\Gamma(a_j^\delta)} \vartheta_i^{a_j^\delta - 1} \exp\left\{-\frac{a_j^\delta \vartheta_i}{2}\right\} d\vartheta_i \\
&= \frac{1}{\sqrt{2\pi}\zeta_j} \frac{\left(\frac{a_j^\delta}{2}\right)^{a_j^\delta}}{\Gamma(a_j^\delta)} \frac{2K_{(a_j^\delta - 1/2)}\left(\sqrt{\frac{a_j^\delta}{\zeta_j}}|\phi_i|\right)}{\left(\frac{a_j^\delta \zeta_j}{\phi_i^2}\right)^{\frac{a_j^\delta - 1/2}{2}}} \times \\
&\quad \overbrace{\int_0^\infty \frac{\left(\frac{a_j^\delta \zeta_j}{\phi_i^2}\right)^{\frac{a_j^\delta - 1/2}{2}}}{2K_{(a_j^\delta - 1/2)}\left(\sqrt{\frac{a_j^\delta}{\zeta_j}}|\phi_i|\right)} \vartheta_i^{a_j^\delta - 1/2 - 1} \exp\left\{-\frac{1}{2}\left(a_j^\delta \vartheta_i + \frac{\phi_i^2}{\vartheta_i \zeta_j}\right)\right\} d\vartheta_i}^{\text{density of GIG}\left(a_j^\delta - 1/2, a_j^\delta, \frac{\phi_i^2}{\vartheta_i \zeta_j}\right)} \\
&= \frac{2^{1/2-a_j^\delta}\left(\frac{a_j^\delta}{\zeta_j}\right)^{a_j^\delta/2+1/4}}{\sqrt{\pi}\Gamma(a_j^\delta)} K_{(a_j^\delta - 1/2)}\left(\sqrt{\frac{a_j^\delta}{\zeta_j}}|\phi_i|\right)|\phi_i|^{a_j^\delta - 1/2}. \quad \square
\end{aligned}$$

[8] Modified Bessel function of the second kind: $K_\nu(x) = \frac{\Gamma(\nu+1)(2x)^\nu}{\sqrt{2\pi}} \int_0^\infty \frac{\cos t}{(t^2+x^2)^{\nu+1/2}} dt$.

Proposition A.3. *Given the R2D2 prior (13), the density of ϕ_i for any $i \in \mathcal{A}_j$, $j = 1, \dots, k$, marginalized over the local scales ϑ_i, is given by*

$$p_{R2D2}(\phi_i|\zeta_j a_j^\pi) = \frac{\sqrt{a_j^\pi}}{2\sqrt{\pi\zeta_j}\Gamma(a_j^\pi)} G_{0,3}^{3,0}\left(\begin{matrix} - \\ \frac{1}{2}, 0, a_j^\pi - \frac{1}{2} \end{matrix} \middle| \frac{\phi_i^2 a_j^\pi}{4\zeta_j} \right), \tag{A.3}$$

where $G_{p,q}^{m,n}\left(\begin{matrix} b_1, \dots, b_q \\ a_1, \dots, a_p \end{matrix} \middle| z \right)$ is the Meijer-G function.[9]

Proof.

$$p(\phi_i|\zeta_j a_j^\pi) = \int_0^\infty \frac{1}{2\sqrt{\vartheta_i\zeta_j/2}} \exp\left\{-\frac{|\phi_i|}{\sqrt{\vartheta_i\zeta_j/2}}\right\} \frac{(a_j^\pi/2)^{a_j^\pi}}{\Gamma(a_j^\pi)} \vartheta_i^{(a_j^\pi - 1)} \exp\left\{-\frac{a_j^\pi}{2}\vartheta_i\right\} d\vartheta_i.$$

The Meijer-G function nests many functions as special cases, e.g., $G_{0,1}^{1,0}\left(\begin{matrix} - \\ 0 \end{matrix} \middle| x \right) = G_{1,0}^{0,1}\left(\begin{matrix} 1 \\ - \end{matrix} \middle| \frac{1}{x} \right) = \exp\{-x\}$. Hence,

$$\begin{aligned}
p(\phi_i|\zeta_j a_j^\pi) &= \frac{(a_j^\pi/2)^{a_j^\pi}}{\sqrt{2\zeta_j}\Gamma(a_j^\pi)} \int_0^\infty \vartheta_i^{(a_j^\pi - 3/2)} \exp\left\{-\frac{|\phi_i|}{\sqrt{\vartheta_i\zeta_j/2}}\right\} G_{1,0}^{0,1}\left(\begin{matrix} 1 \\ - \end{matrix} \middle| \frac{2}{a_j^\pi\vartheta_i} \right) d\vartheta_i \\
&= \frac{a_j^{\pi(3/2)}}{4\sqrt{\zeta_j}\Gamma(a_j^\pi)} \int_0^\infty \exp\left\{-\frac{|\phi_i|}{\sqrt{\vartheta_i\zeta_j/2}}\right\} \left(\frac{2}{a_j^\pi\vartheta_i}\right)^{(3/2 - a_j^\pi)} G_{1,0}^{0,1}\left(\begin{matrix} 1 \\ - \end{matrix} \middle| \frac{2}{a_j^\pi\vartheta_i} \right) d\vartheta_i \\
&= \frac{a_j^{\pi(3/2)}}{4\sqrt{\zeta_j}\Gamma(a_j^\pi)} \int_0^\infty \exp\left\{-\frac{|\phi_i|}{\sqrt{\vartheta_i\zeta_j/2}}\right\} G_{1,0}^{0,1}\left(\begin{matrix} \frac{5}{2} - a_j^\pi \\ - \end{matrix} \middle| \frac{2}{a_j^\pi\vartheta_i} \right) d\vartheta_i.
\end{aligned}$$

The last equality follows from 9.31.(5) in Zwillinger and Jeffrey (2007). Now, let $\theta_i = \frac{1}{\sqrt{\vartheta_j}}$. Then,

$$\begin{aligned}
p(\phi_i|\zeta_j a_j^\pi) &= \frac{a_j^{\pi(3/2)}}{4\sqrt{\zeta_j}\Gamma(a_j^\pi)} \int_0^\infty \exp\left\{-\frac{|\phi_i|\theta_i}{\sqrt{\zeta_j/2}}\right\} G_{1,0}^{0,1}\left(\begin{matrix} \frac{5}{2} - a_j^\pi \\ - \end{matrix} \middle| \frac{2\theta_i^2}{a_j^\pi} \right) \frac{2}{\theta^3} d\theta_i \\
&= \frac{\sqrt{2}}{\sqrt{\zeta_j}\Gamma(a_j^\pi)} \int_0^\infty \exp\left\{-\frac{|\phi_i|\theta_i}{\sqrt{\zeta_j/2}}\right\} G_{1,0}^{0,1}\left(\begin{matrix} \frac{5}{2} - a_j^\pi \\ - \end{matrix} \middle| \frac{2\theta_i^2}{a_j^\pi} \right) \left(\frac{2\theta_i^2}{a_j^\pi}\right)^{-3/2} d\theta_i \\
&= \frac{\sqrt{2}}{\sqrt{\zeta_j}\Gamma(a_j^\pi)} \int_0^\infty \exp\left\{-\frac{|\phi_i|\theta_i}{\sqrt{\zeta_j/2}}\right\} G_{1,0}^{0,1}\left(\begin{matrix} 1 - a_j^\pi \\ - \end{matrix} \middle| \frac{2\theta_i^2}{a_j^\pi} \right) d\theta_i \\
&= \frac{\sqrt{2}}{\sqrt{\zeta_j}\Gamma(a_j^\pi)} \frac{1}{\sqrt{\pi}\frac{|\phi_i|}{\sqrt{\zeta_j/2}}} G_{3,0}^{0,3}\left(\begin{matrix} 0, \frac{1}{2}, 1 - a_j^\pi \\ - \end{matrix} \middle| \frac{4\zeta_j}{a_j^\pi\phi_i^2} \right).
\end{aligned}$$

The last equality follows from 7.813.(2) in Zwillinger and Jeffrey (2007).

$$\begin{aligned}
p(\phi_i|\zeta_j a_j^\pi) &= \frac{a_j^{\pi\frac{1}{2}}}{2\sqrt{\zeta_j}\Gamma(a_j^\pi)} \frac{1}{\sqrt{\pi}} \left(\frac{4\zeta_j}{a_j^\pi\phi_i^2}\right)^{\frac{1}{2}} G_{3,0}^{0,3}\left(\begin{matrix} 0, \frac{1}{2}, 1 - a_j^\pi \\ - \end{matrix} \middle| \frac{4\zeta_j}{a_j^\pi\phi_i^2} \right) \\
&= \frac{a_j^{\pi\frac{1}{2}}}{2\sqrt{\zeta_j}\sqrt{\pi}\Gamma(a_j^\pi)} G_{3,0}^{0,3}\left(\begin{matrix} \frac{1}{2}, 1, \frac{3}{2} - a_j^\pi \\ - \end{matrix} \middle| \frac{4\zeta_j}{a_j^\pi\phi_i^2} \right) \\
&= \frac{\sqrt{a_j^\pi}}{2\sqrt{\zeta_j}\sqrt{\pi}\Gamma(a_j^\pi)} G_{0,3}^{3,0}\left(\begin{matrix} - \\ \frac{1}{2}, 0, a_j^\pi - \frac{1}{2} \end{matrix} \middle| \frac{\phi_i^2 a_j^\pi}{4\zeta_j} \right).
\end{aligned}$$

The last equality follows from 9.31.(2) in Zwillinger and Jeffrey (2007). □

Proposition A.4. *Given the HS prior (14), the density of ϕ_i for any $i \in \mathcal{A}_j$, $j = 1, \dots, k$, marginalized over the local scales ϑ_i, is given by*

[9] A detailed definition of the Meijer-G function can be found in Bateman (1953) and in the Appendix of Zhang et al. (2022).

$$p(\phi_i|\zeta_j) = \frac{1}{\sqrt{2}\pi^{3/2}} \exp\left\{\frac{\phi_i^2}{2\zeta_j}\right\} E_1\left(\frac{\phi_i^2}{2\zeta_j}\right), \tag{A.4}$$

where $E_1(\cdot)$ is the exponential integral function.

Proof. Cf. Carvalho et al. (2010). □

Proof of Theorem 3.1(a). For (16), cf. Zhang et al. (2022), and for (19), cf. Carvalho et al. (2010). (20) and (21) are standard and hence not derived.

(17): According to 10.25.3 in https://dlmf.nist.gov/10.25, when both ν and z are real, if $z \to \infty$, then $K_\nu(z) \approx \pi^{1/2}(2z)^{-1/2}\exp\{-z\}$. Then, as $|\phi| \to \infty$, using Proposition A.2,

$$p_{NG}^{\infty}(\phi_i|\zeta_j a_j^\delta) = \frac{2^{1/2-a_j^\delta}\left(\frac{a_j^\delta}{\zeta_j}\right)^{a_j^\delta/2+1/4}}{\sqrt{\pi}\Gamma(a_j^\delta)}\pi^{1/2}\left(2\sqrt{\frac{a_j^\delta}{\zeta_j}}|\phi_i|\right)^{-1/2}\exp\left\{-\sqrt{\frac{a_j^\delta}{\zeta_j}}|\phi_i|\right\}|\phi_i|^{a_j^\delta-1/2}$$

$$\propto |\phi_i|^{a_j^\delta-1}\exp\left\{-\sqrt{\frac{a_j^\delta}{\zeta_j}}|\phi_i|\right\}.$$

(18): Follows from Theorem 1.3.(i) in Mathai et al. (2009).

Proof of Theorem 3.1(b). For (25), cf. Carvalho et al. (2010).

(22): Case $0 < a_j^\gamma < 1$: Cf. Zhang et al. (2022).

Case $a_j^\gamma = 1$: According to 10.30.3 in https://dlmf.nist.gov/10.30, when $\nu = 0$, $z \to 0$, $K_\nu(z) \approx -\ln z$. Given $a_j^\gamma = 1$, $|\phi| \to 0$, using Proposition A.1,

$$p_{DL}^{0}(\phi_i|a_j^\gamma) \approx \frac{\ln\left((2|\phi_i|)^{-1/2}\right)}{2^{(1+a_j^\gamma)/2}\Gamma(a_j^\gamma)} \propto \ln\left(\frac{1}{|\phi_i|}\right).$$

(23): Case $0 < a_j^\delta < 0.5$: According to 10.27.3 in https://dlmf.nist.gov/10.27 $K_\nu(z) = K_{-\nu}(z)$. Moreover, according to 10.30.2 in https://dlmf.nist.gov/10.30, when $\nu > 0$, $z \to 0$ and z is real, $K_\nu(z) \approx \Gamma(\nu)(z/2)^{-\nu}/2$. Given $0 < a_j^\delta < 0.5$, $|\phi| \to 0$, using Proposition A.2,

$$p_{NG}^{0}(\phi_i|\zeta_j a_j^\delta) \approx \frac{2^{1/2-a_j^\delta}\left(\frac{a_j^\delta}{\zeta_j}\right)^{a_j^\delta/2+1/4}}{\sqrt{\pi}\Gamma(a_j^\delta)}\frac{1}{2}\Gamma(1/2-a_j^\delta)\left(\frac{\sqrt{\frac{a_j^\delta}{\zeta_j}}|\phi_i|}{2}\right)^{a_j^\delta-1/2}|\phi_i|^{a_j^\delta-1/2}$$

$$\propto |\phi_i|^{2a_j^\delta-1}.$$

Case $a_j^\delta = \frac{1}{2}$: According 10.30.3 in https://dlmf.nist.gov/10.30, when $\nu = 0$, $z \to 0$, $K_\nu(z) \approx -\ln z$. Given $a_j^\delta = \frac{1}{2}$, $|\phi| \to 0$, using Proposition A.2,

$$p_{NG}^{0}(\phi_i|\zeta_j a_j^\delta) \approx \frac{2^{1/2-a_j^\delta}\left(\frac{a_j^\delta}{\zeta_j}\right)^{a_j^\delta/2+1/4}}{\sqrt{\pi}\Gamma(a_j^\delta)}\ln\left(\left(\sqrt{\frac{a_j^\delta}{\zeta_j}}|\phi_i|\right)^{-1}\right) \propto \ln\left(\frac{1}{|\phi_i|}\right).$$

(24): Case $0 < a_j^\pi < 0.5$: Follows from Theorem 1.2.(i) in Mathai et al. (2009).

Appendix B. A note on posterior estimation

Efficient computer implementations of all posterior samplers discussed in this paper are conveniently bundled into the R package `bayesianVARs`, available from the Comprehensive R Archive Network (CRAN) at https://cran.r-project.org/package=bayesianVARs. In what follows, we provide the necessary conditionals used for Gibbs sampling.

The main challenge in terms of computational complexity is to sample from the conditional posterior distribution of $\boldsymbol{\phi}$, which is multivariate normal: $\boldsymbol{\phi}|\bullet \sim \boldsymbol{N}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{V}})$, where $\bar{\boldsymbol{V}} = \left(\underline{\boldsymbol{V}}^{-1} + \sum_{t=1}^{T}(\boldsymbol{\Sigma}_t^{-1} \otimes \boldsymbol{x}_t\boldsymbol{x}_t')\right)^{-1}$ and $\bar{\boldsymbol{\mu}} = \bar{\boldsymbol{V}}\left(\sum_{t=1}^{T}(\boldsymbol{I}_M \otimes \boldsymbol{x}_t)\boldsymbol{\Sigma}_t^{-1}\boldsymbol{y}_t\right)$. The complexity stems from the required manipulations of the $n \times n$ matrix variance–covariance matrix of the vector of coefficients. To render computation feasible, we use the corrected triangular algorithm, as in Carriero et al. (2022).

The factorization of the variance–covariance matrix in (3) makes sampling of the free off-diagonal elements in $\boldsymbol{U}$ straightforward: These can be sampled equation-per-equation using standard Bayesian linear regression results (Cogley & Sargent, 2005). Let $\boldsymbol{u}^{(j)} = (u_{1j}, \dots, u_{(j-1),j})'$ denote the $(j-1)$-dimensional column vector that collects the free off-diagonal elements of the jth column of $\boldsymbol{U}$ for $j = 2, \dots, M$. Assuming a normal prior $\boldsymbol{u}^{(j)} \sim N(\boldsymbol{0}, \underline{\boldsymbol{V}}_u^{(j)})$, the full conditional posterior is also normal: $\boldsymbol{u}^{(j)}|\bullet \sim N(\boldsymbol{\mu}_u^{(j)}, \bar{\boldsymbol{V}}_u^{(j)})$, where $\bar{\boldsymbol{V}}_u^{(j)} = \left(\tilde{\boldsymbol{X}}^{(j)'}\tilde{\boldsymbol{X}}^{(j)} + \underline{\boldsymbol{V}}_u^{(j)^{-1}}\right)^{-1}$ and $\boldsymbol{\mu}_u^{(j)} = \bar{\boldsymbol{V}}_u^{(j)}\tilde{\boldsymbol{X}}^{(j)'}\tilde{\boldsymbol{y}}^{(j)}$. Here, $\tilde{\boldsymbol{y}}^{(j)} = \left(\varepsilon_{j1}e^{-h_{j1}/2}, \dots, \varepsilon_{jT}e^{-h_{jT}/2}\right)'$ denotes the normalized observation vector, and

$$\tilde{\boldsymbol{X}}^{(j)} = \left(\left(\varepsilon_{11}e^{-h_{j1}/2}, \dots, \varepsilon_{1T}e^{-h_{jT}/2}\right)', \dots, \left(\varepsilon_{(j-1)1}e^{-h_{j1}/2}, \dots, \varepsilon_{(j-1)T}e^{-h_{jT}/2}\right)'\right)$$

is the $T \times (j-1)$ design matrix.

In the following paragraphs, we provide the conditionals for the latent variables introduced by the hierarchical shrinkage priors. We abstain from providing the conditionals if a hierarchical prior is placed on $\boldsymbol{u}$, because it is basically the same as for $\boldsymbol{\phi}$.

DL. Bhattacharya et al. (2015) propose an efficient algorithm for sampling the hyperparameters of the DL prior, shown in Eq. (9). An important feature of this algorithm is the joint update of $(\boldsymbol{\varrho}, \boldsymbol{\omega}, \boldsymbol{\psi})$, because one-at-a-time updates of Dirichlet vectors usually lead to slow mixing. However, we propose to sample the hyperparameters of the DL prior shown in Eq. (10). Let $\tilde{\vartheta}_i = \sqrt{\vartheta_i/2}$. Then, one important aspect of our algorithm is the joint update of $p(\tilde{\boldsymbol{\vartheta}}, \boldsymbol{\psi}|\boldsymbol{\phi}) = p(\tilde{\boldsymbol{\vartheta}}|\boldsymbol{\phi})p(\boldsymbol{\psi}|\boldsymbol{\phi}, \tilde{\boldsymbol{\vartheta}})$: A draw from the joint conditional posterior distribution is generated by first sampling $\tilde{\vartheta}_i|\phi_i \sim GIG\left(a_j^\gamma - 1, 1, 2|\phi_i|\right), i \in \mathcal{A}_j, j = 1, \dots, k$, where *GIG* denotes the generalized inverse Gaussian distribution[10] and then $\psi_i|\phi_i, \tilde{\vartheta}_i \sim GIG\left(\frac{1}{2}, 1, \left(\frac{\phi_i}{\tilde{\vartheta}_i}\right)^2\right), i \in \mathcal{A}_j, j = 1, \dots, k$. What is left is the full conditional posterior of a_j^γ, which is discrete with the same support points as the prior: $Pr(a_j^\gamma = \tilde{a}_r^\gamma|\bullet) = \frac{p_r^\gamma \prod_{i\in\mathcal{A}_j} f_G(\tilde{\vartheta}_i; \tilde{a}_r^\gamma, \frac{1}{2})}{\sum_{s=1}^R p_s^\gamma \prod_{i\in\mathcal{A}_j} f_G(\tilde{\vartheta}_i; \tilde{a}_s^\gamma, \frac{1}{2})}, i \in \mathcal{A}_j, j = 1, \dots, k$, where $f_G(x; \alpha, \beta)$ denotes the density of the gamma distribution with shape α and rate β evaluated at x.

Overall, our algorithm consists of one fewer step than the algorithm in Bhattacharya et al. (2015), because we directly update the product $\tilde{\vartheta}_i = \varrho_i\omega_j$. Note that if there is interest in the posterior of ω_j (Eq. (9)), draws of this distribution can be easily generated by transforming the posterior draws of the local scales: According to the prior in Eq. (9), it holds that $\omega_j = \sum_{i\in\mathcal{A}_j} \varrho_i\omega_j$, because a Dirichlet distributed vector always sums up to one. Above, we defined $\sqrt{\vartheta_i/2} = \varrho_i\omega_j$. Hence, we get $\omega_j = \sum_{i\in\mathcal{A}_j} \sqrt{\vartheta_i/2} = \sum_{i\in\mathcal{A}_j} \tilde{\vartheta}_i$.

NG. In order to draw from the posterior distribution under the NG prior shown in Eq. (11), we use the following equivalent representation:

$$\phi_i|\varpi_i \sim N(0, \varpi_i), \quad \varpi_i|\xi_j \sim G(a_j^\delta, \xi_j), \quad \xi_j \sim G\left(b^\delta, \frac{2c}{a_j^\delta}\right). \tag{B.1}$$

where $\varpi_i = \vartheta_i\zeta_j$ and $\xi_j = \frac{a_j^\delta}{2\zeta_j}$. The full conditional posterior distribution of ϖ_i is generalized inverse Gaussian: $\varpi_i|\bullet \sim GIG(a_j^\delta - \frac{1}{2}, 2\xi_j, \phi_i^2)$. The full conditional posterior of ξ_j follows a gamma distribution: $\xi_j|\bullet \sim G(n_j a_j^\delta + b^\delta, \frac{2c}{a_j^\delta} + \sum_{i\in\mathcal{A}_j} \varpi_i)$. The full conditional posterior of a_j^δ is discrete, with the same support points as the prior: $Pr(a_j^\delta = \tilde{a}_r^\delta|\bullet) = \frac{p_r^\delta \prod_{i\in\mathcal{A}_j} f_G(\varpi_i; \tilde{a}_r^\delta, \xi_j)}{\sum_{s=1}^R p_s^\delta \prod_{i\in\mathcal{A}_j} f_G(\varpi_i; \tilde{a}_s^\delta, \xi_j)}, i \in \mathcal{A}_j, j = 1, \dots, k$.

R2D2. Zhang et al. (2022) propose an algorithm to sample the hyperparameters of the R2D2 prior shown in Eq. (12), which relies on the joint update of the Dirichlet vector proposed by Bhattacharya et al. (2015). With the same arguments as above, we prefer to sample the hyperparameters of the R2D2 prior using an equivalent representation, which is similar in spirit to Eq. (13):

$$\phi_i|\psi_i, \varpi_i \sim N\left(0, \psi_i\varpi_i/2\right), \quad \psi_i \sim Exp(1/2), \quad \varpi_i|\xi_j \sim G(a_j^\pi, \xi_j), \quad \xi_j \sim G(b^\pi, 1). \tag{B.2}$$

The full conditional posteriors of ψ_i and ϖ_i are both generalized inverse Gaussian: $\psi_i| \bullet GIG(\frac{1}{2}, 1, \frac{2\phi_i^2}{\varpi_i})$ and $\varpi_i|\bullet \sim GIG(a_j^\pi - \frac{1}{2}, 2\xi_j, \frac{2\phi_i^2}{\psi_i})$, for $i \in \mathcal{A}_j, j = 1, \dots, k$. The conditional posterior of ξ_j follows a gamma distribution: $\xi_j|\bullet \sim G(n_j a_j^\pi + b^\pi, 1 + \sum_{i\in\mathcal{A}_j} \varpi_i), j = 1, \dots, k$. The full conditional posterior of a_j^π is discrete, with the same support points as the prior: $Pr(a_j^\pi = \tilde{a}_r^\pi|\bullet) = \frac{p_r^\pi \prod_{i\in\mathcal{A}_j} f_G(\varpi_i; \tilde{a}_r^\pi, \xi_j)}{\sum_{s=1}^R p_s^\pi \prod_{i\in\mathcal{A}_j} f_G(\varpi_i; \tilde{a}_s^\pi, \xi_j)}, i \in \mathcal{A}_j, j = 1, \dots, k$.

[10] The density of $GIG(\theta, \psi, \chi)$ is proportional to $x^{\theta-1}\exp\left\{-\frac{1}{2}\left(\psi x + \frac{\chi}{x}\right)\right\}$. An efficient algorithm to generate draws from GIG is implemented in the R package `GIGrvg` (Leydold & Hörmann, 2022).

Horseshoe. We follow Makalic and Schmidt (2016) in sampling the hyperparameters of the HS prior shown in Eq. (14). They propose the following scale-mixture representation of the prior:

$$\phi_i|\vartheta_i\zeta_j \sim N(0, \vartheta_i\zeta_j), \quad \vartheta_i|\nu_i \sim G^{-1}(1/2, 1/\nu_i), \quad \zeta_j|\tau_j \sim G^{-1}(1/2, 1/\tau_j),$$
$$\nu_i \sim G^{-1}(1/2, 1), \quad \tau_j \sim G^{-1}(1/2, 1), \tag{B.3}$$

where $G^{-1}(\alpha, \beta)$ denotes the inverse-gamma distribution with shape α and scale β. The scale-mixture representation has the advantage that the full conditional posterior distributions are of well-known forms. The full conditional posteriors of both the local parameters ϑ_i and the (semi-)global parameters ζ_j are inverse-gamma: $\vartheta_i|\bullet \sim G^{-1}(1, \frac{1}{\nu_i} + \frac{\phi_i^2}{2\tau_j})$, $\zeta_j|\bullet \sim G^{-1}(\frac{n_j+1}{2}, \frac{1}{\tau_j} + \frac{1}{2}\sum_{i=1}^{\mathcal{A}_j} \frac{\phi_i^2}{\vartheta_i})$. Also, the full conditional posteriors of both the local and (semi-)global auxiliary variables ν_i and τ_j are inverse-gamma: $\nu_i|\bullet \sim G^{-1}(1, 1 + \frac{1}{\vartheta_i})$, $\tau_j|\bullet \sim G^{-1}(1, 1 + \frac{1}{\zeta_j})$.

SHM. Huber and Feldkircher (2019) use a (log) random walk Metropolis step to sample the shrinkage parameters of SHM. The gamma distribution, however, arises as a special case of the generalized inverse Gaussian (GIG) distribution. It is well known that GIG is a conjugate prior for the variance parameter of a normal distribution. Hence, the shrinkage parameters λ_j for $j = 1, 2$ can be sampled from their full conditional posterior distributions. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the index sets that collect own-lag and cross-lag coefficients, respectively. Then, $\lambda_j|\bullet \sim GIG\left(\left(c_j - \frac{|\mathcal{M}_j|}{2}\right), 2d_j, \sum_{i\in\mathcal{M}_j} \frac{\phi_i^2}{\tilde{r}_i}\right)$, where $\tilde{r}_i$ is the constant term determined in (6). This has the benefit that no tuning of proposal distributions is required.

SSVS. The full conditional posterior distribution of γ_i follows a Bernoulli distribution: $\gamma_i|\bullet \sim Bernoulli(\bar{p}_i)$, $i = \mathcal{A}_j$, $j = 1, \ldots, k$, where

$$\bar{p}_i = Pr(\gamma_i = 1|\phi_i) = \frac{\frac{1}{\tau_{1i}} \exp\left(-\frac{\phi_i^2}{2\tau_{1i}^2}\right) \underline{p}_j}{\frac{1}{\tau_{1i}} \exp\left(-\frac{\phi_i^2}{2\tau_{1i}^2}\right) \underline{p}_j + \frac{1}{\tau_{0i}} \exp\left(-\frac{\phi_i^2}{2\tau_{0i}^2}\right)(1 - \underline{p}_j)}.$$

The full conditional posterior distribution of $\underline{p}_j$ is beta: $\underline{p}_j|\bullet \sim Beta(s_1 + \sum_{i\in\mathcal{A}_j} \gamma_i, s_2 + |\mathcal{A}_j| - \sum_{i\in\mathcal{A}_j} \gamma_i)$.

Latent volatilities. We sample all the parameters of (3) using the ancillarity-sufficiency interweaving strategy (Yu & Meng, 2011) as in Kastner and Frühwirth-Schnatter (2014) via the `R` package `stochvol` (Hosszejni & Kastner, 2021).

Appendix C. Data

See Table 6.

Appendix D. A note on priors for VAR coefficients

Many of the priors used in the Bayesian VAR literature—amongst them the Minnesota prior as in Litterman (1986), the stochastic search variable selection (SSVS) prior as in George et al. (2008), the Dirichlet–Laplace (DL) prior as in Kastner and Huber (2020), the adaptive normal-gamma prior as in Huber and Feldkircher (2019), or the horseshoe prior as in Follett and Yu (2019)—were proposed using the reduced-form VAR in (1). Nowadays, as it has become a trend using large datasets, many authors, however, opt for working with the structural VAR formulation. The independence of the structural VAR equations allows for more efficient computations (Chan, 2021b, 2022; Cross et al., 2020). The structural VAR is

$$\begin{aligned} \boldsymbol{y}_t'\boldsymbol{U} &= \boldsymbol{x}_t'\boldsymbol{B} + \boldsymbol{\xi}_t', \\ \boldsymbol{y}_t' &= \boldsymbol{x}_t'\boldsymbol{B} + \boldsymbol{y}_t'\tilde{\boldsymbol{U}} + \boldsymbol{\xi}_t', \end{aligned} \tag{D.1}$$

where $\boldsymbol{B} = \boldsymbol{\Phi}\boldsymbol{U}$ denotes the structural VAR coefficients, $\tilde{\boldsymbol{U}} = \boldsymbol{I} - \boldsymbol{U}$ the contemporaneous coefficients, and $\boldsymbol{\xi}_t = \boldsymbol{\epsilon}_t\boldsymbol{U}$ the orthogonalized errors. It is easy to see that, given $\boldsymbol{B}$ and $\boldsymbol{U}$, the reduced-form coefficients can be recovered by computing $\boldsymbol{\Phi} = \boldsymbol{B}\boldsymbol{U}^{-1}$.

However, we want to point out that priors in general are not invariant under reparametrization: Changing the form without adjusting the prior accordingly will result in two different models. Simply put, imposing a specific prior on the structural coefficients $\boldsymbol{B}$ is not the same as imposing it on the reduced-form coefficients $\boldsymbol{\Phi}$. We denote by b_{ij} the ij-th element of $\boldsymbol{B}$, and by u_{ij}^{-1} the ij-th element in $\boldsymbol{U}^{-1}$. Then as the 'recovered' matrix $\boldsymbol{\Phi} = \boldsymbol{B}\boldsymbol{U}^{-1}$, we have

$$\boldsymbol{\Phi} = \begin{pmatrix} b_{11} & b_{11}u_{12}^{-1} + b_{12} & \cdots & b_{11}u_{1M}^{-1} + \cdots + b_{1(M-1)}u_{(M-1)M}^{-1} + b_{1M} \\ b_{21} & b_{21}u_{12}^{-1} + b_{22} & \cdots & b_{21}u_{1M}^{-1} + \cdots + b_{2(M-1)}u_{(M-1)M}^{-1} + b_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ b_{K1} & b_{K1}u_{12}^{-1} + b_{K2} & \cdots & b_{K1}u_{1M}^{-1} + \cdots + b_{K(M-1)}u_{(M-1)M}^{-1} + b_{KM} \end{pmatrix}. \tag{D.2}$$

Table 6
Variables used in the empirical application.

Name	FRED MNEMONIC	Transformation
National Income and Product Accounts		
Real Gross Domestic Product	GDPC1	$\Delta \log(x_t)$
Real Personal Consumption Expenditures	PCECC96	$\Delta \log(x_t)$
Real Gross Private Domestic Investment	GPDIC1	$\Delta \log(x_t)$
Real private fixed investment: Residential	PRFIx	$\Delta \log(x_t)$
Industrial Production		
Industrial Production Index	INDPRO	$\Delta \log(x_t)$
Capacity Utilization: Manufacturing	CUMFNS	$\Delta \log(x_t)$
Employment		
All employees: Service-Providing Industries	SRVPRD	$\Delta \log(x_t)$
Civilian Employment	CE16OV	$\Delta \log(x_t)$
Average Weekly Hours of Production and Nonsupervisory Employees: Manufacturing	AWHMAN	$\Delta \log(x_t)$
Prices		
Personal Consumption Expenditures: Chain-type Price Index	PCECTPI	$\Delta \log(x_t)$
Gross Domestic Product: Chain-type Price Index	GDPCTPI	$\Delta \log(x_t)$
Gross Private Domestic Investment: Chain-type Price Index	GPDICTPI	$\Delta \log(x_t)$
Consumer Price Index for All Urban Consumers: All Items	CPIAUCSL	$\Delta \log(x_t)$
Earnings and Productivity		
Real Average Hourly Earnings of Production and Nonsupervisory Employees: Construction	CES2000000008x	$\Delta \log(x_t)$
Interest Rates		
Effective Federal Funds Rate	FEDFUNDS	no transformation
1-Year Treasury Constant Maturity Rate	GS1	no transformation
10-Year Treasury Constant Maturity Rate	GS10	no transformation
Money		
Real M2 Money Stock	M2REAL	$\Delta \log(x_t)$
Exchange Rates		
U.S./U.K. Foreign Exchange Rate	EXUSUKx	$\Delta \log(x_t)$
Other		
University of Michigan: Consumer Sentiment	UMCSENTx	$\Delta \log(x_t)$
Stock Markets		
S&P's Common Stock Price Index: Composite	S&P 500	$\Delta \log(x_t)$

It is clear that the induced prior on the reduced-form coefficients now depends on $\boldsymbol{U}$. Moreover, a potential drawback of using the parametrization in form of $\boldsymbol{B} = \boldsymbol{\Phi U}$ is that the posterior is not permutation invariant: The triangular form of $\boldsymbol{U}$ causes results to depend on the order of the vector $\boldsymbol{y}_t$. Especially in high-dimensional settings, justifying a specific order of the endogenous variables is implausible.[11]

To get a rough feeling of the induced prior distributions on $\boldsymbol{\Phi}$, we simulate from i.i.d. Gaussian priors for the elements in $\boldsymbol{B}$ and $\boldsymbol{u}$. Fig. 9 depicts quantile–quantile plots and scatter plots of the recovered reduced-form coefficients. Except for the coefficients of the first columns (where $\boldsymbol{\Phi}$ and $\boldsymbol{B}$ are identical), the induced prior distributions have much heavier tails. The tails get heavier from column to column. Moreover, the distributions are of no well-known forms.

Obviously, reparametrization can heavily influence results in the Bayesian paradigm; one must consider that a good prior for $\boldsymbol{\Phi}$ is not necessarily a good prior for $\boldsymbol{B}$, and vice versa. Unfortunately, this issue is often ignored. E.g., Cross et al. (2020) find that their implementation of a hierarchical Minnesota prior (SHM) generates better forecasts of macroeconomic data than various global–local priors (GL), which seems contradictory to the results in Huber and Feldkircher (2019). Consequently, they claim that their hierarchical setup of SHM is better, disregarding that they use different parametrizations: Whereas Huber and Feldkircher (2019) place the priors on $\boldsymbol{\Phi}$, Cross et al. (2020) place them on $\boldsymbol{B}$. Since neither the same data (same time span, same transformations, etc.) nor exactly the same evaluation metrics are used, it is unclear which full model specification (i.e., a specific prior on a specific form of the VAR) yields better forecasts. It seems plausible that the reparametrization of the VAR makes a bigger difference than some subtleties in the prior configuration. Indeed, in an empirical application predicting stock returns, Bernardi et al. (2024) find that various shrinkage priors lead to better out-of sample forecasts when they are imposed on the reduced-form coefficients instead of the structural-form coefficients. Another example can be found in Chan (2021a), where different SV specifications of the errors are examined. For one SV specification, the prior for the coefficients is placed on $\boldsymbol{\Phi}$, whereas for another SV

[11] Our reduced-form results also depend on the ordering, because of the factorization of $\boldsymbol{\Sigma}_t$ in Eq. (2). Yet, the prior imposed on $\boldsymbol{\Phi}$ remains order-invariant. Using the structural form, the ordering affects both $\boldsymbol{U}$ and $\boldsymbol{\Phi}$, which might aggravate the problem. We refer to Kastner and Huber (2020) and Chan et al. (2024) for methods that are completely order-invariant.

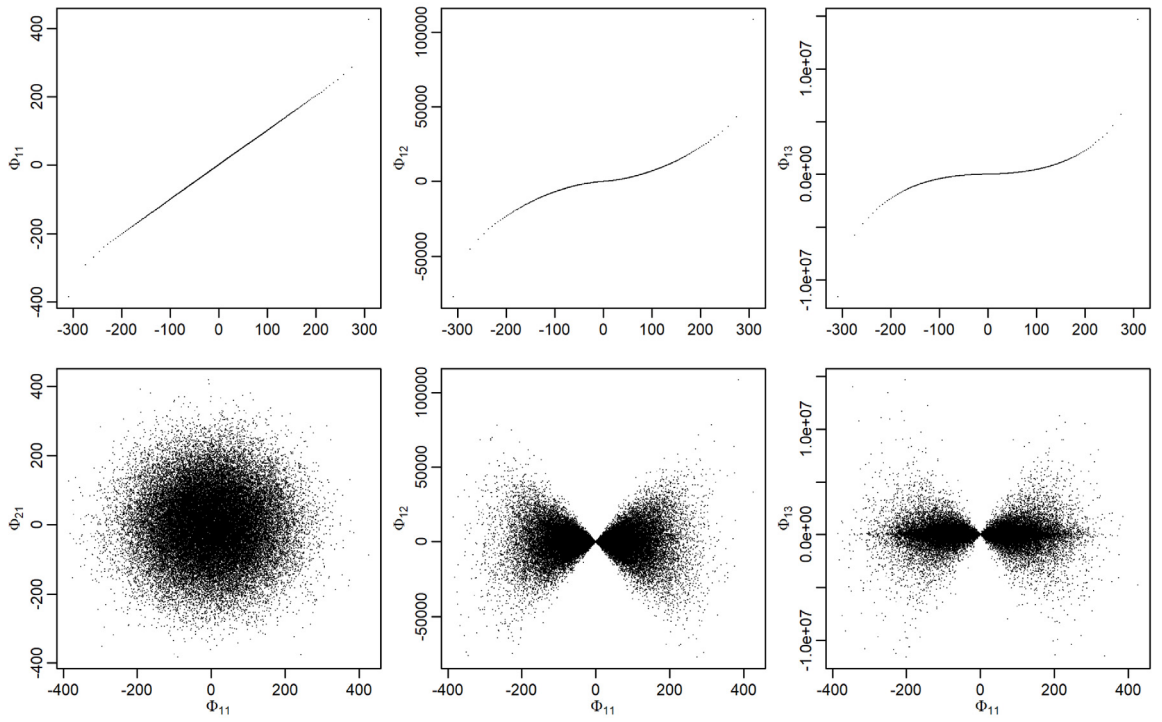

Fig. 9. Prior simulations: Induced prior distributions for $\boldsymbol{\Phi}^{3\times 3}$ resulting from i.i.d. normal priors (zero mean and variance 10) for the elements in $\boldsymbol{B}^{3\times 3}$ and $\boldsymbol{u}$. Top panel: quantile–quantile plots for Φ_{11}, Φ_{12} and Φ_{13} against theoretical quantiles of N(0,10). Bottom panel: scatter plots of the pairs (Φ_{11}, Φ_{21}), (Φ_{11}, Φ_{12}), and (Φ_{11}, Φ_{13}).

specification, the same prior is placed on $\boldsymbol{B}$. The question arises to which degree the results are driven by different SV specifications, and to which degree by different parametrizations of the VAR coefficients.

Data and code availability

Link to the reproducibility package: https://github.com/luisgruber/GK-IJF-BVAR.

References

Bańbura, M., Giannone, D., & Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, *25*(1), 71–92.

Bateman, H. (1953). *Higher transcendental functions, Volume I*. New York: McGraw-Hill Book Company.

Bernaciak, D., & Griffin, J. E. (2024). A loss discounting framework for model averaging and selection in time series models. *International Journal of Forecasting*, *40*(4), 1721–1733.

Bernardi, M., Bianchi, D., & Bianco, N. (2024). Variational inference for large Bayesian vector autoregressions. *Journal of Business & Economic Statistics*, *42*(3), 1066–1082.

Bhattacharya, A., Chakraborty, A., & Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, *103*(4), 985–991.

Bhattacharya, A., Pati, D., Pillai, N. S., & Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, *110*(512), 1479–1490.

Bitto, A., & Frühwirth-Schnatter, S. (2019). Achieving shrinkage in a time-varying parameter model framework. *Journal of Econometrics*, *210*(1), 75–97.

Brown, P. J., & Griffin, J. E. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, *5*(1), 171–188.

Carriero, A., Chan, J., Clark, T. E., & Marcellino, M. (2022). Corrigendum to "Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors" [J. Econometrics 212 (1) (2019) 137–154]. *Journal of Econometrics*, *227*(2), 506–512.

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.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480.

Chan, J. C. C. (2021a). Comparing stochastic volatility specifications for large Bayesian VARs. https://joshuachan.org/papers/ML-LargeVARSV.pdf. (Accessed 19 May 2021).

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

Chan, J. C. C. (2022). Asymmetric conjugate priors for large Bayesian VARs. *Quantitative Economics*, *13*(3), 1145–1169.

Chan, J. C. C. (2023). Comparing stochastic volatility specifications for large Bayesian VARs. *Journal of Econometrics*, *235*(2), 1419–1446.

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

Clark, T. E., & Ravazzolo, F. (2015). Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *Journal of Applied Econometrics*, *30*(4), 551–575.
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.
Cross, J. L., Hou, C., & Poon, A. (2020). Macroeconomic forecasting with large Bayesian VARs: Global-local priors and the illusion of sparsity. *International Journal of Forecasting*, *36*(3), 899–915.
Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18.
Eddelbuettel, D., & Sanderson, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics & Data Analysis*, *71*, 1054–1063.
Fava, B., & Lopes, H. F. (2021). The illusion of the illusion of sparsity: An exercise in prior sensitivity. *Brazilian Journal of Probability and Statistics*, *35*(4), 699–720.
Feldkircher, M., Gruber, L., Huber, F., & Kastner, G. (2024). Sophisticated and small versus simple and sizeable: When does it pay off to introduce drifting coefficients in Bayesian vector autoregressions? *Journal of Forecasting*, *43*(6), 2126–2145.
Follett, L., & Yu, C. (2019). Achieving parsimony in Bayesian vector autoregressions with the horseshoe prior. *Econometrics and Statistics*, *11*, 130–144.
Frühwirth-Schnatter-Schnatter, S., & Knaus, P. (2021). Sparse Bayesian state-space and time-varying parameter models. In M. G. Tadesse, & M. Vannucci (Eds.), *Handbook of Bayesian variable selection* (pp. 297–326). Chapman & Hall.
George, E. I., Sun, D., & Ni, S. (2008). Bayesian stochastic search for VAR model restrictions. *Journal of Econometrics*, *142*(1), 553–580.
Geweke, J., & Amisano, G. (2010). Comparing and evaluating Bayesian predictive distributions of asset returns. *International Journal of Forecasting*, *26*(2), 216–230.
Giannone, D., Lenza, M., & Primiceri, G. E. (2015). Prior selection for vector autoregressions. *The Review of Economics and Statistics*, *97*(2), 436–451.
Giannone, D., Lenza, M., & Primiceri, G. E. (2019). Priors for the long run. *Journal of the American Statistical Association*, *114*(526), 565–580.
Giannone, D., Lenza, M., & Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, *89*(5), 2409–2437.
Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, *29*(3), 411–422.
Hosszejni, D., & Kastner, G. (2021). Modeling univariate and multivariate stochastic volatility in R with stochvol and factorstochvol. *Journal of Statistical Software*, *100*, 1–34.
Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, *5*, 1457–1469.
Huber, F., & Feldkircher, M. (2019). Adaptive shrinkage in Bayesian vector autoregressive models. *Journal of Business & Economic Statistics*, *37*(1), 27–39.
Huber, F., Kastner, G., & Feldkircher, M. (2019). Should I stay or should I go? A latent threshold approach to large-scale mixture innovation models. *Journal of Applied Econometrics*, *34*(5), 621–640.
Iacopini, M., Ravazzolo, F., & Rossini, L. (2023). Proper scoring rules for evaluating density forecasts with asymmetric loss functions. *Journal of Business & Economic Statistics*, *41*(2), 482–496.
Jacquier, E., Polson, N. G., & Rossi, P. E. (1994). Bayesian analysis of stochastic volatility models. *Journal of Business & Economic Statistics*, *12*(4), 371–389.
Jarociński, M., & Maćkowiak, B. (2017). Granger causal priority and choice of variables in vector autoregressions. *The Review of Economics and Statistics*, *99*(2), 319–329.
Kastner, G. (2019). Sparse Bayesian time-varying covariance estimation in many dimensions. *Journal of Econometrics*, *210*(1), 98–115.
Kastner, G., & Frühwirth-Schnatter, S. (2014). Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models. *Computational Statistics & Data Analysis*, *76*, 408–423.
Kastner, G., & Huber, F. (2020). Sparse Bayesian vector autoregressions in huge dimensions. *Journal of Forecasting*, *39*(7), 1142–1165.
Kim, S., Shephard, N., & Chib, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies*, *65*(3), 361–393.
Koop, G. M. (2013). Forecasting with medium and large Bayesian VARs. *Journal of Applied Econometrics*, *28*(2), 177–203.
Koop, G., & Korobilis, D. (2010). Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends® in Econometrics*, *3*(4), 267–358.
Koop, G., & Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, *53*(3), 867–886.
Koop, G., & Korobilis, D. (2013). Large time-varying parameter VARs. *Journal of Econometrics*, *177*(2), 185–198.
Leydold, J., & Hörmann, W. (2022). GIGrvg: Random variate generator for the GIG distribution. R package version 0.7.
Litterman, R. B. (1986). Forecasting with Bayesian vector autoregressions: Five years of experience. *Journal of Business & Economic Statistics*, *4*(1), 25–38.
Luo, Y., & Griffin, J. E. (2025). Bayesian inference of vector autoregressions with tensor decompositions. *Journal of Business & Economic Statistics*. (in press).
Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Berlin, Heidelberg: Springer.
Makalic, E., & Schmidt, D. F. (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, *23*(1), 179–182.
Mathai, A. M., Saxena, R. K., & Haubold, H. J. (2009). *The H-function: theory and applications*. Springer Science & Business Media.
Mavroeidis, S. (2021). Identification at the zero lower bound. *Econometrica*, *89*(6), 2855–2885.
McCracken, M. W., & Ng, S. (2021). FRED-QD: A quarterly database for macroeconomic research. *Federal Reserve Bank of St. Louis Review*, *103*(1), 1–44.
Onorante, L., & Raftery, A. E. (2016). Dynamic model averaging in large model spaces using dynamic Occam's window. *European Economic Review*, *81*, 2–14.
Polson, N. G., & Scott, J. G. (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, & M. West (Eds.), *Bayesian Statistics 9: Proceedings of the Ninth Valencia International Meeting* (pp. 501–538). Oxford, UK: Oxford University Press.
Raftery, A. E., Kárný, M., & Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, *52*(1), 52–66.
Rossini, L., Villa, C., Prevenas, S., & McCrea, R. (2024). Loss-based prior for the degrees of freedom of the Wishart distribution. *Econometrics and Statistics*, (in press).
Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, *48*(1), 1–48.
Sims, C. A., & Zha, T. (1998). Bayesian methods for dynamic multivariate models. *International Economic Review*, *39*(4), 949–968.
Stock, J. H., & Watson, M. W. (2012). Disentangling the channels of the 2007–09 recession. *Brookings Papers on Economic Activity*, 81–156.
Team, R. C. (2023). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Villani, M. (2009). Steady-state priors for vector autoregressions. *Journal of Applied Econometrics*, *24*(4), 630–650.
West, M. (2020). Bayesian forecasting of multivariate time series: scalability, structure uncertainty and decisions. *Annals of the Institute of Statistical Mathematics*, *72*(1).

Wu, P., & Koop, G. (2022). *Fast, order-invariant Bayesian inference in VARs using the eigendecomposition of the error covariance matrix*: *Working Papers 2310*, University of Strathclyde Business School, Department of Economics.
Yu, Y., & Meng, X.-L. (2011). To center or not to center: That is not the question—An ancillarity–sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. *Journal of Computational and Graphical Statistics*, *20*(3), 531–570.
Zhang, Y. D., Naughton, B. P., Bondell, H. D., & Reich, B. J. (2022). Bayesian regression using a prior on the model fit: The R2-D2 shrinkage prior. *Journal of the American Statistical Association*, *117*(538), 862–874.
Zhou, M., & Carin, L. (2012). Negative binomial process count and mixture modeling. arXiv:1209.3442v1 [stat.ME].
Zwillinger, D., & Jeffrey, A. (2007). *Table of integrals, series, and products*. Elsevier.