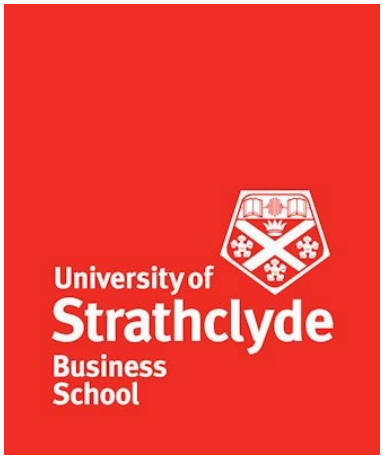
Strathclyde
Discussion Papers
in Economics



Dynamic Shrinkage Priors for Large Time-varying Parameter Regressions using Scalable Markov Chain Monte Carlo Methods*

NIKO HAUZENBERGER1, 2, FLORIAN HUBER1 , and GARY KOOP2

No. 23 – 5

1University of Salzburg
2University of Strathclyde

Dynamic Shrinkage Priors for Large Time-varying Parameter Regressions using Scalable Markov Chain Monte Carlo Methods*

NIKO HAUZENBERGER[1, 2], FLORIAN HUBER[1], and GARY KOOP[2]

[1] *University of Salzburg*
[2] *University of Strathclyde*

May 15, 2023

Abstract. Time-varying parameter (TVP) regression models can involve a huge number of coefficients. Careful prior elicitation is required to yield sensible posterior and predictive inferences. In addition, the computational demands of Markov Chain Monte Carlo (MCMC) methods mean their use is limited to the case where the number of predictors is not too large. In light of these two concerns, this paper proposes a new dynamic shrinkage prior which reflects the empirical regularity that TVPs are typically sparse (i.e., time variation may occur only episodically and only for some of the coefficients). A scalable MCMC algorithm is developed which is capable of handling very high dimensional TVP regressions or TVP Vector Autoregressions. In an exercise using artificial data we demonstrate the accuracy and computational efficiency of our methods. In an application involving the term structure of interest rates in the eurozone, we find our dynamic shrinkage prior to effectively pick out small amounts of parameter change and our methods to forecast well.

JEL: C11, C30, C50, E3, E43
Keywords: Time-varying parameter regression, dynamic shrinkage prior, global-local shrinkage prior, Bayesian variable selection, scalable Markov Chain Monte Carlo

*Corresponding author: Niko Hauzenberger. Department of Economics, University of Salzburg. Address: Mönchsberg 2A, 5020 Salzburg, Austria. Email: niko.hauzenberger@plus.ac.at. The first two authors gratefully acknowledge financial support by the Austrian Science Fund (FWF): ZK-35 and by funds of the Oesterreichische Nationalbank (Austrian Central Bank, Anniversary Fund, project no. 18127, 18763 and 18765).

1 Introduction

The increasing availability of large data sets in economics has led to interest in regressions involving large numbers of explanatory variables. Given the evidence of instability and parameter change in many macroeconomic variables, there is also an interest in time-varying parameter (TVP) regression models and multi-equation extensions such as time-varying parameter Vector Autoregressions (TVP-VARs). This combination of large numbers of explanatory variables with TVPs can lead to regressions with a huge number of parameters. But such regressions are often sparse, in the sense that most of these parameters are zero. In this context, Bayesian methods have proved particularly useful since Bayesian priors can be used to find and impose this sparsity, leading to more accurate inferences and forecasts. A range of priors have been suggested for high-dimensional regression models (see, among many others, Ishwaran and Rao, 2005; Park and Casella, 2008; Griffin and Brown, 2010; Carvalho *et al.*, 2010; Bhattacharya *et al.*, 2015). There is also a growing literature which extends these methods to the TVP case. Examples include Belmonte *et al.* (2014), Kalli and Griffin (2014), Eisenstat *et al.* (2016), Kowal *et al.* (2019), Petrova (2019), Kalli and Griffin (2019), Knaus *et al.* (2021), Chan *et al.* (2020), Hauzenberger *et al.* (2022) and Fischer *et al.* (2023).

Most of these papers assume particular forms of parameter change (e.g., it is common to assume parameters evolve according to random walks) and use computationally-demanding Markov Chain Monte Carlo (MCMC) methods. The former aspect can be problematic (e.g., if parameter change is rare and abrupt, then a model which assumes all parameters evolve gradually according to random walks is inappropriate). The latter aspect means these methods are not scalable (i.e., MCMC-based methods cannot handle models with huge numbers of coefficients).

The contributions of the present paper relate to issues of prior elicitation and computation in TVP regressions. With regards to prior elicitation, we develop novel dynamic shrinkage priors for TVP regressions. These modify recent approaches to dynamic shrinkage priors in papers such as Kowal *et al.* (2019). We work with the static representation of the TVP regression model which breaks the coefficients into two groups. One group

contains constant coefficients (we call these $\boldsymbol{\alpha}$). The other, which we call $\boldsymbol{\beta}$, are TVPs. In the static representation, the dimension of $\boldsymbol{\beta}$ can be enormous. Our dynamic global-local shrinkage priors are carefully designed to push unimportant elements in $\boldsymbol{\beta}$ to zero in a time-varying fashion. This is done using a global shrinkage parameter that varies over time as well as local shrinkage parameters. The global shrinkage parameter has an interpretation similar to a dynamic factor model with a single factor. This single factor can be used to find periods of time-variation in coefficients and periods when they are constant. Since the assumption of a common volatility factor hampers the use of standard stochastic volatility MCMC algorithms based on a mixture of Gaussians approximation (Kim *et al.*, 1998), we propose a simple approximation that works particularly well in high dimensional settings.

With regards to computation, we develop a scalable MCMC algorithm. This algorithm is suitable for cases where the posterior for $\boldsymbol{\beta}$, conditional on the other parameters in the model, is Gaussian. This occurs for a wide range of global-local shrinkage priors including the dynamic shrinkage priors used in this paper. In this case, the exact MCMC algorithm of Bhattacharya *et al.* (2016) is the state of the art.[1] However, even it is too computationally slow to handle the huge number of regressors that appear in the static representation of the TVP regression model. Recently, Johndrow *et al.* (2017) has proposed an approximate algorithm based on this exact algorithm which is computationally much more efficient in sparse models and, thus, is scalable.

In our paper, it is precisely this scalable MCMC algorithm which forms the basis of the algorithm we use. It involves a thresholding step (described below) which we implement in a different manner than Johndrow *et al.* (2017). In particular, as opposed to fixing the threshold to a small number, we set it adaptively. Since this would typically imply a number of thresholds that match the dimension of $\boldsymbol{\beta}$, we use a method called Signal Adaptive Variable Selection (SAVS), see Ray and Bhattacharya (2018), to determine the thresholds in a novel way. SAVS has the advantage of being computationally fast and easy to implement. Recent papers use SAVS for determining variable relevance (Hahn and Carvalho, 2015), portfolio applications (Puelz *et al.*, 2020) or improving macroeconomic

[1] Kastner and Huber (2020), Hauzenberger (2021) and Korobilis (2022), for example, use this exact algorithm in the context of large VARs to reduce the computational burden of estimating these models.

forecasts (Huber *et al.*, 2021). We solely use SAVS to identify which variables can be safely set to zero in order to construct an approximate posterior distribution for the TVPs. Thus, the use of SAVS in the context of the algorithm of Johndrow *et al.* (2017) provides two-fold benefits: computational improvements and more flexibility due to its adaptive nature.

We investigate the use of our methods in artificial and real data. The artificial data exercise demonstrates that our scalable algorithm is a good approximation to exact MCMC and that its computational benefits are substantial. Our application to the eurozone yield curve shows how our methods can effectively pick out small amounts of occasional parameter change in some parameters. Furthermore, allowing for such change in the coefficients improves forecasts.

The remainder of the paper is organized as follows. The second section defines the TVP regression and TVP-VAR models used in this paper. The third section discusses MCMC methods for the regression coefficients and introduces our computationally-efficient approximate method. Section 4 develops different dynamic shrinkage priors and discusses Bayesian estimation. This section also describes a novel method for drawing the volatilities in the context of a multivariate stochastic volatility process with a common factor. Sections 5 and 6 present our artificial data exercise and our empirical application, respectively. Section 7 summarizes and concludes.

2 Static Representation of a TVP Regression

2.1 A TVP regression

The static representation of a TVP regression model involving a T-dimensional dependent variable, $\boldsymbol{y}$, and a $T \times K$-dimensional matrix of predictors, $\boldsymbol{X}$ is:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{W}\boldsymbol{\beta} + \boldsymbol{L}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_T), \quad \boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_T')', \tag{1}$$

where $\boldsymbol{\alpha}$ is a K-dimensional vector of time-invariant coefficients, $\boldsymbol{\beta}_t$ is a $K \times 1$ vector of time-varying coefficients and $\boldsymbol{L} = \text{diag}(\sigma_1, \ldots, \sigma_T)$ with σ_t denoting time-varying error volatilities. The TVP part of this model arises through the $\boldsymbol{W}\boldsymbol{\beta}$ term. $\boldsymbol{W}$ is a $T \times k(= TK)$ matrix given by:

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{x}_1' & \boldsymbol{0}_{K\times 1}' & \ldots & \boldsymbol{0}_{K\times 1}' \\ \boldsymbol{0}_{K\times 1}' & \boldsymbol{x}_2' & \ldots & \boldsymbol{0}_{K\times 1}' \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0}_{K\times 1}' & \boldsymbol{0}_{K\times 1}' & \ldots & \boldsymbol{x}_T' \end{pmatrix}, \tag{2}$$

with $\boldsymbol{x}_t$ denoting a K-dimensional sub-vector of $\boldsymbol{X}$. Equation 1 is simply a regression which leads to the terminology *static representation*. But it is a regression with an enormous number of explanatory variables.

Note that (2) implies that the TVPs are mean zero and uncorrelated over time. However, extensions to other forms can be trivially done through a re-definition of $\boldsymbol{W}$. For instance, if we are interested in random walk-type behavior in the TVPs, we can set

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{x}_1' & \boldsymbol{0}_{K\times 1}' & \ldots & \boldsymbol{0}_{K\times 1}' \\ \boldsymbol{x}_2' & \boldsymbol{x}_2' & \ldots & \boldsymbol{0}_{K\times 1}' \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_T' & \boldsymbol{x}_T' & \ldots & \boldsymbol{x}_T' \end{pmatrix}. \tag{3}$$

This specification implies that $\boldsymbol{\beta}$ can be interpreted as the changes in the parameters and multiplication with $\boldsymbol{W}$ yields the cumulative sum over $\boldsymbol{\beta}$. In our empirical exercise, we

consider both of these specifications for $\boldsymbol{W}$ and refer to the former as the flexible (FLEX) and the latter as the random walk (RW) specification.

The existing literature using Bayesian shrinkage techniques typically uses MCMC methods. Exact MCMC sampling, however, quickly becomes computationally cumbersome since k is extremely large even for moderate values of T and K.

Various solutions to this have been proposed in the literature. The standard solution is simply not to work with the static representation, but instead make some parametric assumption about how the TVPs evolve (e.g., assume they follow random walks or a Markov switching process). Unless K is extremely large, exact MCMC methods are feasible. However, with macroeconomic data it is common to find strong evidence of changes in the conditional variance of a series, but much less evidence in favor of change in the conditional mean of a series, (see, e.g., Clark, 2011). When K is large, it is plausible to assume that only some of the predictors have time-varying coefficients and, even for these, coefficient change may only rarely happen. Common conventional approaches are not suited for data sets which exhibit such sparsity in the TVPs. If changes in the conditional mean of the parameters happen only rarely then a random walk assumption, which assumes change is continually happening, is not appropriate. If changes in the conditional mean only occur for a small sub-set of the K variables (or occur at different times for different variables), then a Markov switching model which assumes all coefficients change at the same time is not appropriate. These considerations motivate our use of the static representation and the development of a dynamic shrinkage prior suited for the case of TVP sparsity.

The literature has proposed a few ways of overcoming the computational hurdle that arises if the static representation is used. Korobilis (2021) uses message passing techniques to estimate large TVP regressions and shows that these large models outperform a range of competing models. Similarly, Huber *et al.* (2020) approximate the TVPs using message passing techniques based on a rotated model representation and sample from the full conditional posterior of $\boldsymbol{\alpha}$ using MCMC methods. Both approaches have the drawback that the quality of the approximation inherent in the use of message passing techniques might be questionable. In another recent paper, Hauzenberger *et al.* (2022) propose using

the singular value decomposition of $\boldsymbol{W}$ in combination with a conjugate shrinkage prior on $\boldsymbol{\beta}$ to ensure computational efficiency. However, this method has the potential drawback that conjugate priors might be too restrictive for discriminating signals and noise in high dimensional models.

In this paper, we develop another approach which should work particularly well when $\boldsymbol{\beta}$ is extremely sparse. This is the scalable MCMC method, based on posterior perturbations, of Johndrow *et al.* (2017).

2.2 Extension to a TVP-VAR

Before discussing the scalable MCMC algorithm, we note that methods developed for the TVP regression can also be used for the TVP-VAR if it is written in equation-by-equation form (see, for instance, Carriero *et al.*, 2019; Huber *et al.*, 2021). In particular, we can use the following structural representation of the TVP-VAR:

$$\boldsymbol{y}_t = \boldsymbol{c}_t + \boldsymbol{A}_{0t}\boldsymbol{y}_t + \sum_{p=1}^{P} \boldsymbol{A}_{pt}\boldsymbol{y}_{t-p} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \boldsymbol{\Sigma}_t), \tag{4}$$

with $\boldsymbol{y}_t$ being an M-dimensional vector of endogenous variables, $\boldsymbol{c}_t$ denoting an M-dimensional vector of intercepts, $\boldsymbol{A}_{pt}$, for $p = 1, \ldots, P$, denoting an $M \times M$-dimensional time-varying coefficient matrix that may be stacked in a matrix $\boldsymbol{A}_t = (\boldsymbol{A}_{1t}, \ldots, \boldsymbol{A}_{Pt})$. Furthermore, $\boldsymbol{\epsilon}_t$ is an M-dimensional vector of errors and $\boldsymbol{\Sigma}_t = \text{diag}\,(\sigma_{1t}^2, \ldots, \sigma_{Mt}^2)$ refers to its diagonal time-varying covariance matrix. Finally, $\boldsymbol{A}_{0t}$ defines contemporaneous relationships between the elements of $\boldsymbol{y}_t$ and is lower-triangular with zeros on the diagonal.

The i^{th} $(i = 2, \ldots, M)$ equation of $\boldsymbol{y}_t$ can be written as a standard TVP regression model:

$$y_{it} = \boldsymbol{x}_{it}' \underbrace{(\boldsymbol{\alpha}_i + \boldsymbol{\beta}_{it})}_{\boldsymbol{\gamma}_{it}} + \sigma_{it}\epsilon_{it}, \quad \epsilon_{it} \sim \mathcal{N}(0, 1).$$

Here, $\boldsymbol{x}_{it}$ is a $K_i(= MP+i)$-dimensional vector of covariates with $\boldsymbol{x}_{it} = (1, \{y_{jt}\}_{j=1}^{i-1}, \boldsymbol{y}_{t-1}', \ldots, \boldsymbol{y}_{t-P}')'$, $\boldsymbol{\gamma}_{it} = (\boldsymbol{\alpha}_i + \boldsymbol{\beta}_{it}) = (c_{it}, \{a_{ij,0t}\}_{j=1}^{i-1}, \boldsymbol{A}_{i\bullet,t})'$ denotes a K_i-dimensional vector of time-varying coefficients, with c_{it} referring to the i^{th} element in $\boldsymbol{c}_t$, $a_{ij,0t}$ denoting the $(i,j)^{th}$ element

of $\boldsymbol{A}_{0t}$ and $\boldsymbol{A}_{i\bullet,t}$ referring to the i^{th} row of $\boldsymbol{A}_t$. For $i = 1$, $\boldsymbol{x}_{1t} = (1, \boldsymbol{y}'_{t-1}, \ldots, \boldsymbol{y}'_{t-p})'$ and $\boldsymbol{\gamma}_{1t} = (c_{1t}, \boldsymbol{A}_{1\bullet,t})'$. Thus, the TVP-VAR can be written as a set of M independent TVP regressions which can be estimated separately using the MCMC methods described in the following section. An additional computational advantage arises in that the M equations can be estimated in parallel using multiple CPUs.

Depending on the particular choice of $\boldsymbol{W}$, this model nests a variety of commonly used specifications in the literature. For instance, if $\boldsymbol{W}$ implies a random walk behavior of the latent states we arrive at a TVP-VAR closely related to the one proposed in Primiceri (2005). As we will show below, the main difference is that we have a more flexible state equation by allowing for heteroskedasticity in the shocks to the states through dynamic shrinkage priors. Another model that is closely related to ours is the one proposed in Cogley *et al.* (2010). This model assumes that the variances of the state innovations evolve according to independent stochastic volatility models.

3 Scalable MCMC Algorithm for a Large TVP Model

In this section, we explain the MCMC algorithm of Johndrow *et al.* (2017) and Johndrow *et al.* (2020) and discuss how we adapt it for our TVP regression model. The parameters in the static representation are $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Since $\boldsymbol{\alpha}$ is typically of moderate size and potentially non-sparse, we use conventional (exact) MCMC methods for it. It is $\boldsymbol{\beta}$ which is high-dimensional and potentially sparse, characteristics the algorithm of Johndrow *et al.* (2017) is perfectly suited for. Thus, we use this algorithm for $\boldsymbol{\beta}$. Every model used in the empirical application also includes stochastic volatility.

In the following section, we develop an MCMC algorithm to produce draws of $\boldsymbol{L}$. Since there is nothing new in our MCMC algorithm for $\boldsymbol{\alpha}$ and our algorithm for drawing $\boldsymbol{L}$ is discussed later, in this section we will proceed conditionally on them and work with the transformed regression involving dependent variable $\tilde{\boldsymbol{y}} = \boldsymbol{L}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha})$ and explanatory variables $\tilde{\boldsymbol{W}} = \boldsymbol{L}^{-1}\boldsymbol{W}$. The appendix provides full details of our MCMC algorithm. In this section, we will also assume that the prior on $\boldsymbol{\beta}$ is (conditional on other parameters) Gaussian with mean zero and a diagonal prior covariance matrix $\boldsymbol{D}_0 = \text{diag}(d_1, \ldots, d_k)$.

Many different global-local shrinkage priors have this general form and, in the following section, we will suggest several different choices likely to be well-suited to TVP regressions.

The exact MCMC algorithm of Bhattacharya *et al.* (2016) for drawing $\boldsymbol{\beta}$ proceeds as follows:

1. Draw a k-dimensional vector $\boldsymbol{v} \sim \mathcal{N}(\mathbf{0}_k, \boldsymbol{D}_0)$,
2. Sample a T-dimensional vector $\boldsymbol{q} \sim \mathcal{N}(\mathbf{0}_T, \boldsymbol{I}_T)$,
3. Define $\boldsymbol{w} = \tilde{\boldsymbol{W}}\boldsymbol{v} + \boldsymbol{q}$
4. Solve $(\tilde{\boldsymbol{y}} - \boldsymbol{w}) = (\boldsymbol{I}_T + \tilde{\boldsymbol{W}}\boldsymbol{D}_0\tilde{\boldsymbol{W}}')\boldsymbol{u}$ for $\boldsymbol{u}$,
5. Set $\boldsymbol{\beta} = (\boldsymbol{D}_0\tilde{\boldsymbol{W}}'\boldsymbol{u}) + \boldsymbol{v}$.

Bhattacharya *et al.* (2016) show that this algorithm is fast compared to existing approaches which involve taking the Cholesky factorization of the posterior covariance matrix. However, it can still be slow when k is very large. The computational bottleneck lies in the calculation of $\boldsymbol{\Gamma} = \tilde{\boldsymbol{W}}\boldsymbol{D}_0\tilde{\boldsymbol{W}}'$ which has computational complexity of order $\mathcal{O}(T^2k)$. In macroeconomic or financial applications involving hundreds of observations, $T^2k = T^3K$ can be enormous.

Johndrow *et al.* (2017) and Johndrow *et al.* (2020) propose an approximation to the algorithm of Bhattacharya *et al.* (2016) which, in sparse contexts, will be much faster and, thus, scalable to huge dimensions. The basic idea of the algorithm is to approximate the high-dimensional matrix $\boldsymbol{\Gamma}$ by dropping irrelevant columns of $\tilde{\boldsymbol{W}}$ so as to speed up computation. To be precise, Steps 4 and 5 of the algorithm are replaced with

4* Solve $(\tilde{\boldsymbol{y}} - \boldsymbol{w}) = (\boldsymbol{I}_T + \hat{\boldsymbol{\Gamma}})\boldsymbol{u}$ for $\boldsymbol{u}$, with $\hat{\boldsymbol{\Gamma}} = \tilde{\boldsymbol{W}}_S\boldsymbol{D}_{0,S}\tilde{\boldsymbol{W}}_S'$,

5* Set $\boldsymbol{\beta} = (\boldsymbol{D}_{0,S}\tilde{\boldsymbol{W}}_S'\boldsymbol{u}) + \boldsymbol{v}$.

Here, $\tilde{\boldsymbol{W}}_S$ denotes a $T \times s$-dimensional sub-matrix of $\tilde{\boldsymbol{W}}$ that consists of columns defined by a set S and $\boldsymbol{D}_{0,S}$ is constructed by taking the diagonal elements of $\boldsymbol{D}_0$ also defined by S. Let $S = \{j : \delta_j = 1\}$ denote an index set with δ_j being the j^{th} element of a k-dimensional selection vector $\boldsymbol{\delta}$ with elements $\delta_j = 1$ with probability p_j and $\delta_j = 0$ with probability

$(1 - p_j)$. Johndrow *et al.* (2017) approximates δ_j by setting $\hat{\delta}_j = 0$ if $d_j \in (0, \xi]$ for ξ being a small threshold. Computational complexity is reduced from $\mathcal{O}(T^2k)$ to $\mathcal{O}(T^2s)$, where $s = \sum_{j=1}^{k} \delta_j$ is the cardinality of the set S or equivalently the number of non-zero parameters in $\boldsymbol{\beta}$. Step 5* yields a draw from the approximate posterior $\hat{p}(\boldsymbol{\beta}|\bullet)$ with the $\bullet$ notation indicating that we condition on the data and the remaining parameters in the model.

The algorithm requires a choice of a threshold for constructing $\boldsymbol{\delta}$. Johndrow *et al.* (2017) suggest simple thresholding rules that seem to work well in their work with artificial data (e.g., recommendations include setting the threshold to 0.01 when explanatory variables are largely uncorrelated, but 10^{-4} when they are more highly correlated). However, choosing the threshold might be problematic for real data applications and can require a significant amount of tuning in practice. Instead we propose to choose the thresholds in a different way using SAVS.

To explain what SAVS is and how we use it in practice, note first that papers such as Hahn and Carvalho (2015) recommend separating out shrinkage (i.e., use of a Bayesian prior to shrink coefficients towards zero) and sparsification (i.e., setting the coefficents on de-selected variables to be precisely zero so as to remove them from the model) into different steps. First, MCMC output from a standard model (e.g., a regression with global-local shrinkage prior) is produced. Secondly, this MCMC output is then sparsified by choosing a sparse coefficient vector that minimizes the distance between the predictive distribution of the shrunk model and the predictive density of a model based on this sparse coefficient vector plus an additional penalty term for non-zero coefficients. This assumption is critically based on assuming normally distributed shocks. The optimal solution, $\tilde{\boldsymbol{\beta}}$, is then a sparse vector which can be used to construct $\boldsymbol{\delta}$.

The advantages of this shrink-then-sparsify approach are discussed in Hahn and Carvalho (2015) and, in the context of TVP regressions, in Huber *et al.* (2021). One important advantage is that estimation error is removed for the sparsified coefficients. When using global shrinkage priors in high dimensional contexts with huge numbers of parameters, small amounts of estimation error can build up and have a deleterious impact on forecasts. By sparsifying, estimation error in the small coefficients is eliminated, thus improving fore-

casts. This paper differs from the aforementioned papers by using SAVS to approximate the indicators $\boldsymbol{\delta}$ which is then used in our approximate MCMC algorithm.

The SAVS algorithm, developed in Ray and Bhattacharya (2018), is a fast method for solving the optimization problem outlined above, making it feasible to sparsify each draw from the posterior of $\boldsymbol{\beta}$. In the present context, our contention is that a strategy which uses SAVS to shrink-then-sparsify our coefficients can be used to provide a sensible estimate of $\boldsymbol{\delta}$ that does not lead to a deterioration in forecast accuracy. Using SAVS, we first produce a sparsified draws $\tilde{\boldsymbol{\beta}}$.[2] For each draw $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_k)'$, we then set

$$\hat{\delta}_j = I(\tilde{\beta}_j^* \neq 0).$$

Each draw of $\hat{\delta}_j$ is used in the construction of $\hat{\boldsymbol{\Gamma}}$ in the MCMC algorithm of Johndrow *et al.* (2017) described above. We will refer to this algorithm as being approximate to distinguish it from the exact algorithm of Ray and Bhattacharya (2018).

4 Bayesian Estimation and Inference

4.1 Dynamic global-local shrinkage priors

For the time-invariant coefficients, $\boldsymbol{\alpha}$, we use a horseshoe shrinkage prior (Carvalho *et al.*, 2010). Since the properties of this prior are familiar and posterior simulation methods for this prior are standard, we do not discuss it further here. See the appendix for additional details.

The important contribution of the present paper lies in the development of a dynamic extension of the horseshoe prior for $\boldsymbol{\beta}$. We modify methods outlined in Kowal *et al.* (2019) to design a prior which reflects our beliefs about what kinds of parameter change are commonly found in macroeconomic applications. In particular, we want to allow for a high degree of sparsity in the TVPs. That is, we want a prior that allows for the possibility that parameter change is rare and may occur for only some coefficients in the regression. There may be periods of instability when parameters change and times of stability when

[2]Precise details for how SAVS works in TVP regressions, along with additional motivation for the approach, are provided in Huber *et al.* (2021).

they do not. A dynamic global-local shrinkage prior which has these properties is:

$$p(\boldsymbol{\beta}_t) = \prod_{j=1}^{K} \mathcal{N}(\beta_{jt}|0, \tau\lambda_t\phi_{jt}^2), \quad \phi_{jt} \sim \mathcal{C}^+(0,1), \tag{5}$$

where $\boldsymbol{\beta}_t = (\beta_{1t}, \ldots, \beta_{Kt})'$ denotes the coefficients at time t, τ denotes a global shrinkage parameter that pushes all elements in $\boldsymbol{\beta}$ towards zero, λ_t is a time-specific shrinkage factor that pushes all elements in $\boldsymbol{\beta}_t$ towards zero and ϕ_{jt} is a coefficient and time-specific shrinkage term that follows a half-Cauchy distribution.

Thus, the prior covariance matrix of $\boldsymbol{\beta}_t$ is given by:

$$\boldsymbol{\Omega}_t = \tau\lambda_t \times \text{diag}(\phi_{1t}^2, \ldots, \phi_{Kt}^2),$$

which implies that λ_t acts as a common factor that aims to detect periods characterized by substantive amounts of time variation.

The main innovation of this paper lies in our treatment of this common factor. Before we discuss the precise specifications for λ_t, it is worth summarizing the key innovation of this prior. As opposed to the dynamic horseshoe of Kowal *et al.* (2019), we only introduce persistence in the common shrinkage factor λ_t. The key point to note here is that, as opposed to assuming a dynamic law of motion for the coefficient-specific prior scaling parameters, we borrow strength from the cross-sectional dimension and by doing this we substantially reduce the computational burden necessary.

For the global shrinkage parameter we consider four different laws of motion. The first and second of these involve setting $g_t = \log(\tau\lambda_t)$ and assuming it follows an AR(1) process:

$$g_t = \mu + \rho(g_{t-1} - \mu) + \nu_t,$$

with $\mu = \log\tau$. We consider two possible distributions for ν_t. In the first of these it follows a four parameter Z-distribution, $\mathcal{Z}(1/2, 1/2, 0, 0)$, leading to a variant of the dynamic horseshoe prior proposed in Kowal *et al.* (2019) (henceforth labeled `dHS svol-Z`). The second of these follows a Gaussian distribution, leading to a standard stochastic

volatility model for this prior variance (labeled `dHS svol-N`). This model resembles the one stipulated in Cogley *et al.* (2010) but with a single dynamic volatility process. Both of these processes imply a gradual evolution of g_t and thus a smooth transition from times of rapid parameter change to times of less parameter change.

The third and fourth specifications allow for more abrupt change between times of stability and times of instability. They assume that λ_t is a regime switching process with:

$$\lambda_t = \kappa_0^2(1 - d_t) + \kappa_1^2 d_t, \tag{6}$$

Here, d_t denotes an indicator that either follows a Markov switching model (labeled `dHS MS`) or a mixture specification (labeled `dHS Mix`) and κ_0, κ_1 denote prior variances with the property that $\kappa_1 \gg \kappa_0$. For the Markov switching model, we assume that d_t is driven by a (2×2)-dimensional transition probability matrix P with transition probabilities from state i to j denoted by p_{ij} (with $p_{ii} \sim \mathcal{B}(a_{i,MS}, b_{i,MS})$, for $i = 0, 1$, following a Beta distribution a priori). The mixture model assumes that $p(d_t = 1) = \underline{p}$, with $\underline{p} \sim \mathcal{B}(a_{Mix}, b_{Mix})$. In the empirical application we specify $\kappa_1 = 100/K$, $\kappa_0 = 0.01/K$, $a_{Mix} = a_{1,MS} = b_{0,MS} = 3$ and $b_{Mix} = a_{0,MS} = b_{1,MS} = 30$.

We also include a fifth specification by setting $\lambda_t = 1$ for all t. We refer to this setup as the static horseshoe prior (abbreviated as `sHS`). For these last three specifications (i.e., the ones that do not assume λ_t to evolve according to an AR(1) process), we use a half-Cauchy prior on $\sqrt{\tau} \sim \mathcal{C}^+(0, 1)$.

4.2 Markov Chain Monte Carlo (MCMC) algorithm

For all of these models, Bayesian estimation and prediction can be done using MCMC methods. In this sub-section we mainly focus on how to sample λ_t under the assumption that it evolves according to an AR(1) process. For this step we propose a simple and accurate approximation that renders the corresponding hierarchical model linear and conditionally Gaussian. We only briefly discuss the remaining steps since most of them are standard in the literature.

For the time varying regression coefficients, the scalable algorithms (with or without sparsification) of the preceding section, based on Johndrow *et al.* (2017), can be used. The only modification is that we construct $\boldsymbol{D}_0$ as follows:

$$\boldsymbol{D}_0 = \text{diag}(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_T),$$

with λ_t depending on the specific law of motion adopted. Most of the prior hyperparameters introduced in this section have posterior conditionals of standard forms. These are given in the appendix.

Sampling λ_t for the specifications that assume it to be binary is also straightforward and can be carried out using standard algorithms. To sample from the posterior of λ_t under the assumption that it evolves according to an AR(1) process, the algorithm proposed in Jacquier *et al.* (1995) can be used. However, since this algorithm simulates the λ_t's one at a time mixing is often an issue. A second option would be to view the prior (after squaring each element of $\boldsymbol{\beta}_t$ and taking logs) as the observation equation of a dynamic factor model. This strategy, however, would be computationally challenging for moderate to large values of K. As a solution, we propose a new algorithm that is straightforward to implement and, if K is large, has good properties.

Let $\hat{\boldsymbol{\beta}}_t$ be a K-dimensional vector of normalized TVPs with typical element $\hat{\beta}_{jt} = \beta_{jt}/(\phi_{jt}\tau^{1/2})$. Using (5) and squaring yields:

$$b_t = (\hat{\boldsymbol{\beta}}_t{}'\hat{\boldsymbol{\beta}}_t) = \lambda_t \nu_t, \tag{7}$$

with $\nu_t = \boldsymbol{v}_t'\boldsymbol{v}_t$ for $\boldsymbol{v}_t \sim \mathcal{N}(\boldsymbol{0}_K, \boldsymbol{I}_K)$. Notice that ν_t follows a χ^2 distribution with K degrees of freedom, denoted by χ^2_K. This implies that sampling algorithms that rely on the Gaussian mixture approximation proposed in Kim *et al.* (1998) cannot be used. Instead we approximate the χ^2_K using a well-known limit theorem that implies, as $K \to \infty$,

$$\frac{\nu_t - K}{\sqrt{2K}} \xrightarrow{d} \mathcal{N}(0,1) \quad \Leftrightarrow \quad \nu_t \approx \hat{\nu}_t = \sqrt{2K} q_t + K, \quad q_t \sim \mathcal{N}(0,1).$$

This approximation works if K is large. In our case, K is often large. For instance, in the

largest TVP-VAR model we consider, K is around 100. Since we estimate the TVP-VAR one equation at a time, values of this order of magnitude hold in each equation and the approximation is likely to be good. But if one were to do full system estimation of the TVP-VAR, there are on the order of MK VAR coefficients at each point in time and the approximation would be even better.

Substituting the Gaussian approximation into (7) and taking logs yields:

$$\log b_t = \log \lambda_t + \log \hat{\nu}_t. \tag{8}$$

Finally, under the assumption that $(\sqrt{2K}q_t + K) > 0$ and by using a Taylor series expansion,[3] we approximate $\log \hat{v}_t$ with a $\mathcal{N}(\log(K) - 1/K, 2/K)$ to render (8) conditionally Gaussian. This implies that any of the standard algorithms proposed in the literature on Gaussian linear state space models can be used. In this paper, we simulate $\log \lambda_t$ using the precision sampler outlined, for example, in Chan and Jeliazkov (2009) and McCausland *et al.* (2011).

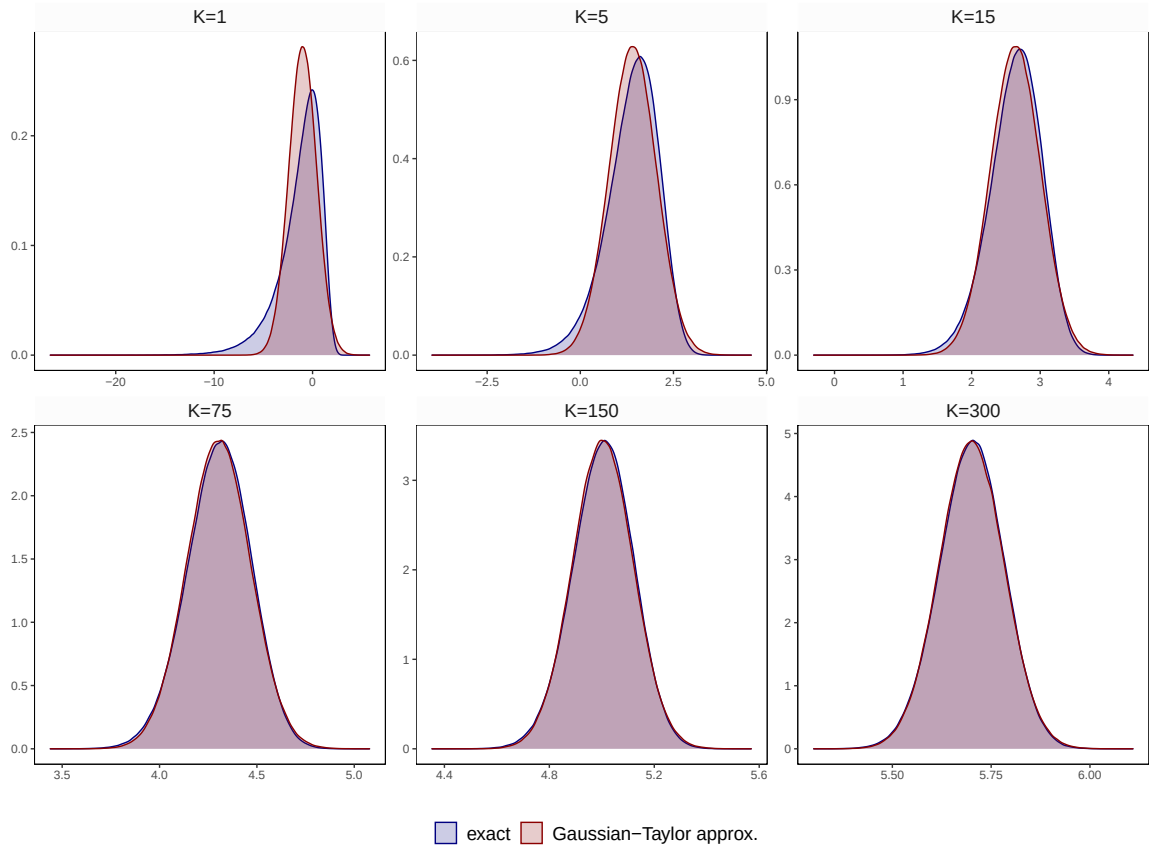
The accuracy of this approximation for different values of K is illustrated in Figure 1. From this figure it is clearly visible that, if K is greater than 5, our approximation works extremely well. In these cases, there is hardly any difference visible between the $\log \chi^2_K$ and the single-component Gaussian distribution. For $K = 1$ (the most extreme case) and $K = 5$, some differences arise which mainly relate to the left tail of the distribution. However, already for $K = 5$ these differences are so small that we do not expect them to have any serious consequences on our estimates of λ_t, even for small values of K.

5 Illustration Using Artificial Data

In this section we illustrate the merits of our approach using synthetic data.

[3]More precisely, we compute the mean and variance of $\log \hat{\nu}_t$ using a second and first order Taylor series expansion of $\mathrm{E}(\log(K + \hat{\nu}_t - K))$ and $\mathrm{Var}(\log(K + \hat{\nu}_t - K))$ around K, respectively.

Figure 1: Approximation error of a single-component Gaussian used to approximate a $\log \chi^2_K$ distribution.



Notes: This figure illustrates the approximation error resulting from approximating the error distribution (which is $\log \chi^2_K$) with a single-component Gaussian with mean $\log(K) - 1/K$ and variance $2/K$. For different values of K, the blue shaded areas show the exact error distribution, while the red shaded areas indicate the approximate error distribution.

5.1 How does our algorithm compare to exact MCMC?

We start by showing that using our approximate (sparsified) algorithm yields estimates that are close to the exact ones in terms of precision. This is achieved by considering five different data generating processes (DGPs). These are all based on Equation (1) but make different assumptions about the density and nature of parameter change. Dense DGPs are characterized by having time-variation in a large number of parameters (with sparse DGPs being the opposite of dense). The nature of parameter change can be gradual (e.g., characterized by constant evolution of the parameters) or abrupt. For each of the five DGPs, we simulate a time series of length $T = 250$ and with $K = 50$.

The different DGPs assume that the states evolve as follows:

- *dense gradual:* $\boldsymbol{\beta}_t \sim \mathcal{N}(\boldsymbol{\beta}_{t-1}, \frac{1}{100} \times \boldsymbol{I}_K)$,

- *dense mixed:* $\boldsymbol{\beta}_t \sim \mathcal{N}\left(\boldsymbol{\beta}_{t-1}, \left(d_t + \frac{(1-d_t t)}{100}\right) \times \boldsymbol{I}_K\right)$ with $Prob(d_t = 1) = 0.1$,
- *medium-dense gradual:* $\boldsymbol{\beta}_t \sim \mathcal{N}(\boldsymbol{\beta}_{t-1}, \frac{d_t}{100} \times \boldsymbol{I}_K)$ with $Prob(d_t = 1) = 0.3$,
- *sparse abrupt:* $\boldsymbol{\beta}_t \sim \mathcal{N}(\boldsymbol{\beta}_{t-1}, \boldsymbol{I}_K)$ with $Prob(d_t = 1) = 0.02$,
- *no TVPs:* $\boldsymbol{\beta}_t = \mathbf{0}_{K\times 1}$ for all t.

The remaining parameters are set as follows: $\boldsymbol{\beta}_0 = \mathbf{0}$, $\boldsymbol{L} = 0.01 \times \boldsymbol{I}_T$, $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_K)$ and $\boldsymbol{X}_j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_T)$ for $j = 1, \ldots, K$. Based on these, we use the true path of the parameters $\boldsymbol{\beta}_t$ to obtain a realization of y_t. In all simulation experiments and for all models considered we simulate $2,500$ draws from the joint posterior of the parameters and latent states and discard the first 500 draws as burn-in.

We investigate the accuracy of our scalable approximate MCMC methods relative to the exact MCMC algorithm of Bhattacharya *et al.* (2016) (i.e., it is the version of our algorithm which imposes $\delta_j = 1$ for all j). Table 1 shows the ratio of mean absolute errors (MAEs), computed using the posterior mean of $\{\boldsymbol{\beta}_t\}_{t=1}^T$ and the true parameters, for the approximate relative to the exact approach for the five priors averaged over the five DGPs. With one exception, MAE ratios are essentially one indicating that the approximate and exact algorithms are producing almost identical results. The one exception is for the DGP which does not have any TVPs. For this case, the approximate algorithm is substantially better than the exact one. This is because our approximate algorithm uses SAVS which (correctly for this DGP) can set the TVPs to be precisely zero. In this case, draws from the posterior will coincide with draws from the prior that induce heavy shrinkage. Hence, compared to the exact model, the likelihoood does not influence the prior and more shrinkage can be achieved.

Thus, Table 1 shows that, where there is substantial time variation in parameters, the approximation inherent in our scalable MCMC algorithm is an excellent one, yielding results that are virtually identical to the slower exact algorithm. The table also shows the usefulness of SAVS in cases of very sparse DGPs.

Table 1: Mean absolute errors of the TVPs relative to exact estimation.

Specification	MAE ratios: different forms of TVPs				
	dense gradual	dense mixed	medium-dense gradual	sparse abrupt	no TVPs
dHS Mix	1.001	1.003	1.001	1.002	0.755
dHS MS	0.998	0.999	1.000	0.999	0.558
dHS svol-N	1.000	1.000	1.000	1.000	0.817
dHS svol-Z	1.001	1.001	1.000	1.000	0.696
sHS	0.999	1.000	1.000	1.001	0.653

Notes: Numbers are averages based on 20 replications from each of the DGPs.

5.2 How big are the computational gains of our algorithm?

Our second artificial data experiment is designed to investigate the computational gains of our algorithm relative to exact MCMC for various choices of K, T, degrees of sparsity and data configurations. Since we are only interested in computation time we just generate one artificial data set for each of two different ways of specifying $\boldsymbol{W}$. The random numbers refered to below are drawn from the standard Gaussian distribution.

For $K = 1, ..., 400$ and $T \in \{100, 200\}$ we randomly draw a $\boldsymbol{y}$ and an $\boldsymbol{X}$. The $\boldsymbol{W}$ is drawn in two ways which correspond to the flexible and random walk specifications of equations (2) and (3), respectively.

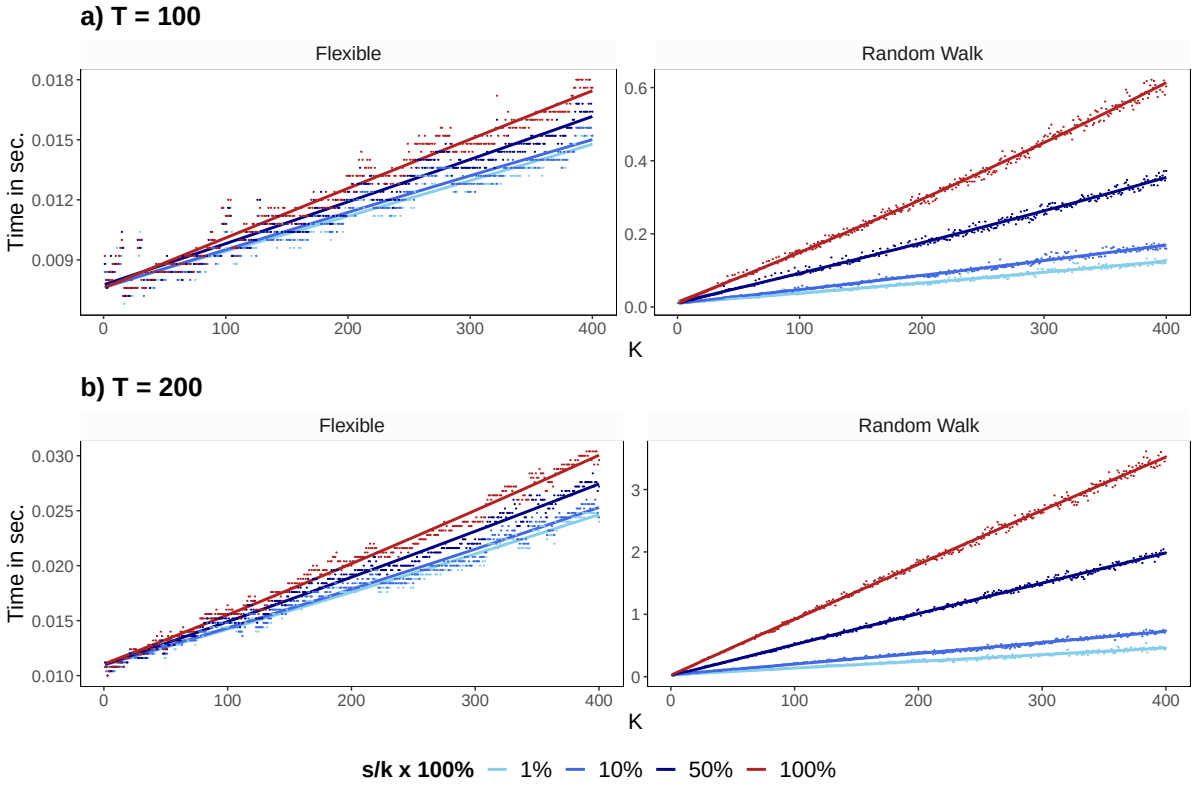
In terms of sparsity, we consider four scenarios based on how we choose $\tilde{\boldsymbol{W}}_S$:

- 100% dense: $\tilde{\boldsymbol{W}}_S = \boldsymbol{W}$. This is the exact algorithm.
- 50% dense: $\tilde{\boldsymbol{W}}_S$ contains 50% of the columns of $\boldsymbol{W}$ (i.e., $s = 0.5k$).
- 10% dense: $\tilde{\boldsymbol{W}}_S$ contains 10% of the columns of $\boldsymbol{W}$ (i.e., $s = 0.1k$).
- 1% dense: $\tilde{\boldsymbol{W}}_S$ contains 1% of the columns of $\boldsymbol{W}$ (i.e., $s = 0.01k$).

Figure 2 depicts the computational advantages of our approximate MCMC algorithm relative to the exact algorithm of Bhattacharya *et al.* (2016). It shows the time necessary to obtain a draw of $\boldsymbol{\beta}$. It can be seen that when the TVPs are highly correlated over time as with the random walk specification, then our scalable algorithm has substantial computational advantages relative to the exact algorithm particularly for large K and in sparse data sets. When the TVPs are uncorrelated the computational advantages of our approach relative to the exact algorithm are smaller, but still appreciable.[4]

[4]The relatively good performance of the exact algorithm in this case is partly due to the fact that

Figure 2: Time necessary to obtain a draw for K time-varying coefficients.



Notes: The figure shows the estimation time in seconds required to obtain a draw for K time-varying coefficients for different degrees of overall sparsity (i.e., $1\%, 10\%, 50\%$, and 100% dense). The dots refer to the empirical run times for which we fit a nonlinear trend (indicated by the solid lines). The red colored dots and red solid lines indicate run times of the exact algorithm (100% dense, with $s = k$).

6 Empirical Application using Eurozone Yield Data

6.1 Data overview and specification issues

We illustrate our methods using a monthly data set of 30 government bond yields in the euro area (EA). As opposed to forecasting standard US macroeconomic time series such as output, inflation and unemployment rates, forecasting EA government bond yields is challenging due to, at least, three reasons. The first is that the researcher has to decide on the segment of yield curve she is interested in or use techniques that allow for analyzing the full term structure of government bond yields. Following the latter approach leads to overfitting issues whereas the former approach might suffer from omitted variable bias. The second challenge is that these time series are often subject to outliers as well as sharp shifts in the conditional variance. The final reason is that the time series we consider are

we are coding using sparse algorithms. In the flexible specification for $\boldsymbol{W}$, the underlying matrices are block-diagonal and thus exact sampling is already quite fast.

rather short and in such circumstances TVP-VARs risk overfitting if the estimates of the TVPs are not regularized sufficiently. We expect that the techniques proposed in this paper are capable of handling both issues well.

We use monthly yield curve data obtained from Eurostat. This dataset includes the yield to maturity of a (hypothetical) zero coupon bond on AAA-rated government bonds of eurozone countries for 30 different maturities. These maturities range from one-year to 30-years and span the period from 2005:01 to 2019:12.

If we wish to model all 30 yields jointly we have to estimate a TVP-VAR with $M = 30$ equations, a challenging statistical and computational task which we will take on in the next sub-section. Since the parameter space of such a model is vast and difficult to interpret, in this sub-section where we present some in-sample results, we will use a small-scale example. This model is based on the Nelson-Siegel three factor model (see, e.g., Nelson and Siegel, 1987; Diebold *et al.*, 2006) and assumes that the yield on a security with maturity $\mathfrak{t}$, labeled $r_t(\mathfrak{t})$, features a factor structure:

$$r_t(\mathfrak{t}) = L_t + S_t \left(\frac{1 - e^{-\zeta \mathfrak{t}}}{\zeta \mathfrak{t}} \right) + C_t \left(\frac{1 - e^{-\zeta \mathfrak{t}}}{\zeta \mathfrak{t}} - e^{-\zeta \mathfrak{t}} \right) + \eta_t(\mathfrak{t}), \quad \eta_t(\mathfrak{t}) \sim \mathcal{N}\left(0, \sigma_\eta^2(\mathfrak{t})\right). \quad (9)$$

Here, L_t, S_t and C_t refer to the level, slope and curvature factor, respectively, while $\eta_t(\mathfrak{t})$ denotes maturity-specific measurement errors which are independent across maturities and feature variance $\sigma_\eta^2(\mathfrak{t})$. ζ denotes a parameter that controls the shape of the factor loadings. Following Diebold *et al.* (2006), we set $\zeta = 0.7308$ (12×0.0609). Since the loading of the level factor is one for all maturities and does not feature a discount factor, it defines the behavior at the long end of the yield curve. Moreover, the slope factor mainly shapes the short end of the yield curve and the curvature factor defines the middle part of the curve. The latent yield curve factors are obtained by running OLS on a t-by-t basis. These estimates are then consequently used as our endogenous variables by setting $\boldsymbol{y}_t = (L_t, S_t, C_t)'$ and estimating the TVP-VAR defined in (4). We use the flexible specification for $\boldsymbol{W}$ in (2) and the approximate algorithm to estimate the model. In addition, we set the lag length to two. After obtaining forecasts for $\boldsymbol{y}_t$, we use (9) to map the factors back to the observed yields. It is worth noting that (9) constitutes an

observation equation which links the observed yields to the latent Nelson-Siegel factors. To compute predictive densities, we also take the corresponding measurement errors into account by estimating the measurement error variance independently for each observed series.

6.2 In-sample results

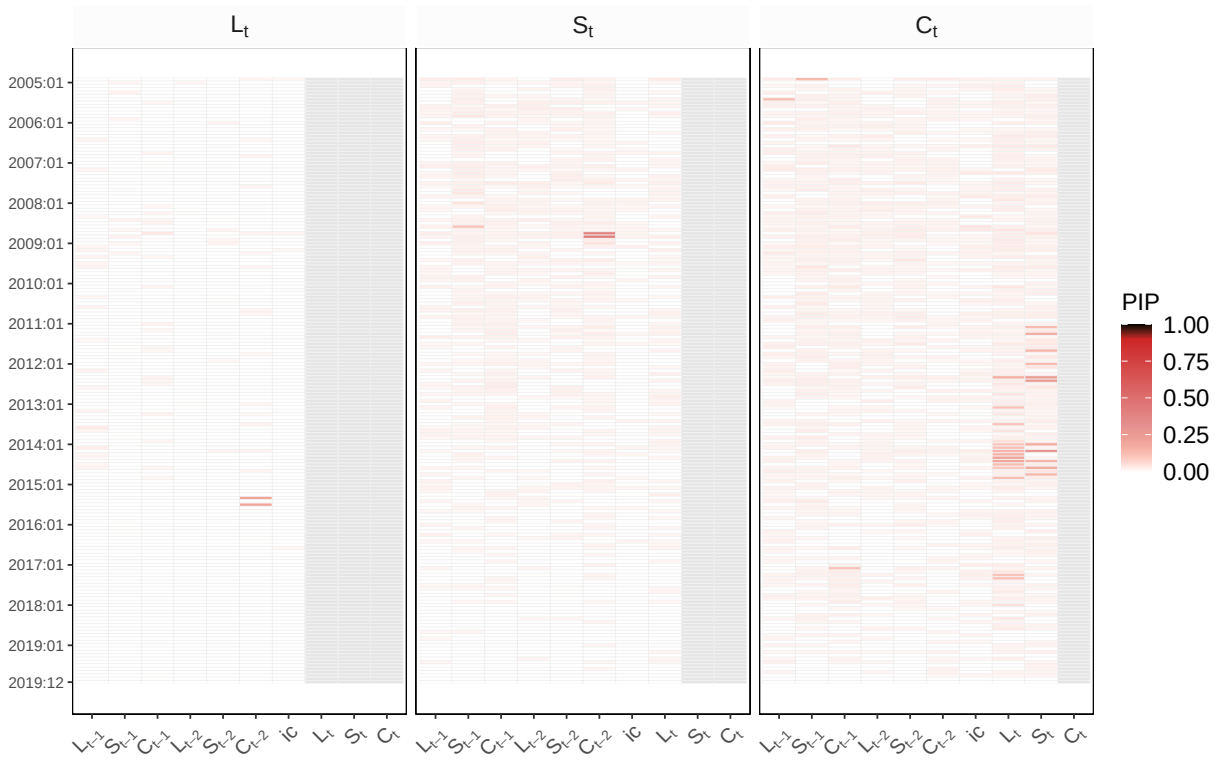
To provide some information on the amount of time variation, Figures 3 and 4 depicts heatmaps of the posterior inclusion probability (PIPs) for a Nelson-Siegel model with panels a) to d) referring to the four different dynamic priors for λ_t. These PIPs are the posterior means of the elements of $\boldsymbol{\delta}$.

The main impression provided by Figure 3 and 4 is that there is little evidence of strong time-variation in the parameters when using this data set. However, there does seem to be some in the sense that there are many variables and time periods where the PIPs are appreciably above zero. That is, even though the figures contain a lot of white (PIPs essentially zero) and just a handful of deep reds (PIPs above one half), there is a great deal of pink of various shades (e.g., PIPs 20%-30%). This is consistent with time-variation being small, episodic and only occurring in some coefficients.

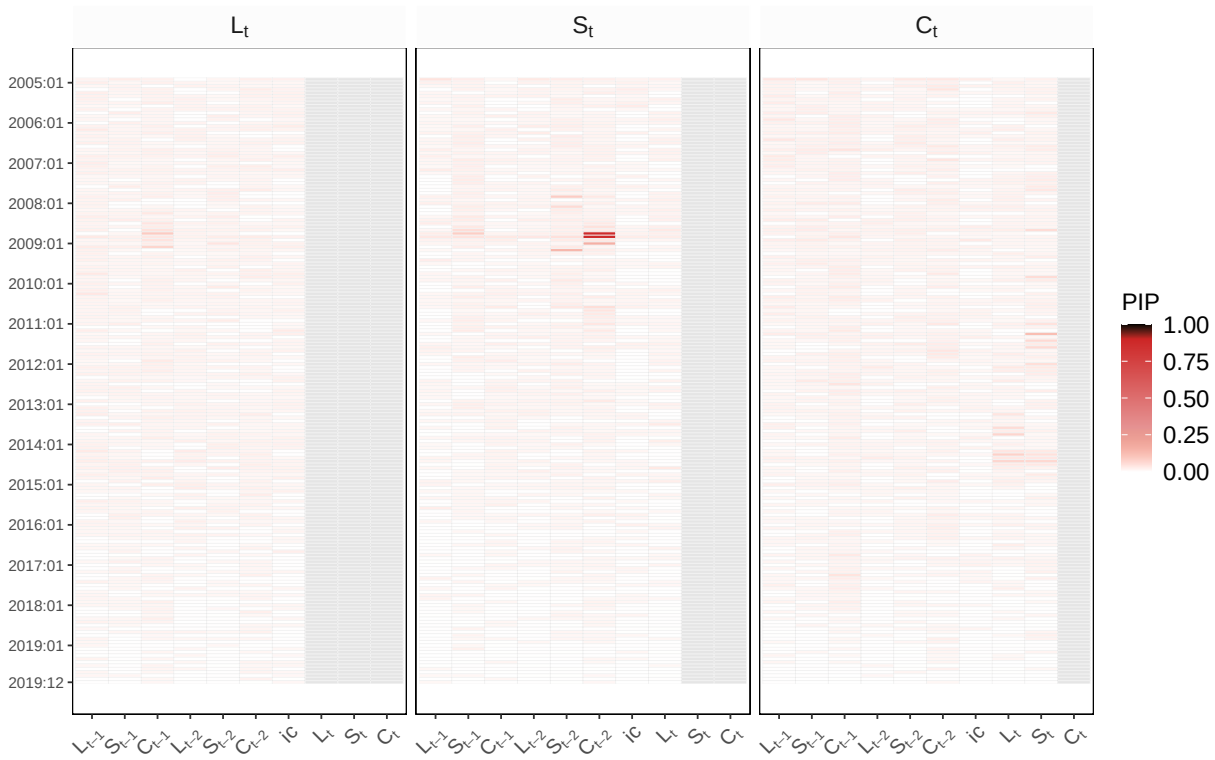
Results for our four different dynamic horseshoe priors are slightly different indicating the dynamic prior choice can have an impact on results. A clear pattern emerges only for the dynamic horseshoe prior with a mixture specification. It is finding that small amounts of time-variation occur only for the coefficients on the curvature factor. If the mixture part of the prior is replaced by a Markov switching specification, we tend to find short-lived periods where a small amount of time-variation occurs for all of the coefficients in an equation. But, interestingly, `dHS MS` finds that different equations have time-variation occuring at different periods of time. Evidence for TVPs is the least when we use stochastic volatility specifications in the dynamic horseshoe priors. For these priors, tiny amounts of time variation (i.e., tiny PIPs) are spread much more widely throughout the sample and across variables.

Figure 3: Heatmaps of posterior inclusion probability (PIPs) for time-variation in structural TVP-VAR coefficients with a gradually changing common shrinkage factor.

a) *dHS svol-Z*



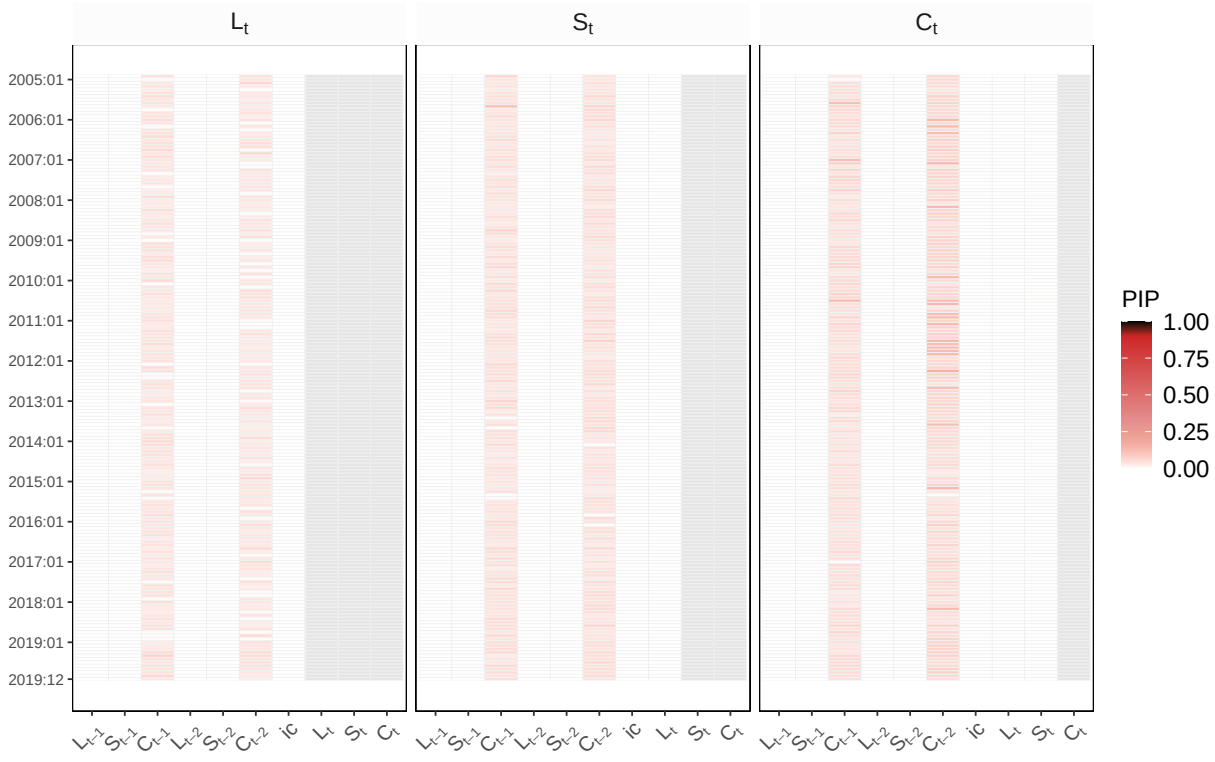
b) *dHS svol-N*



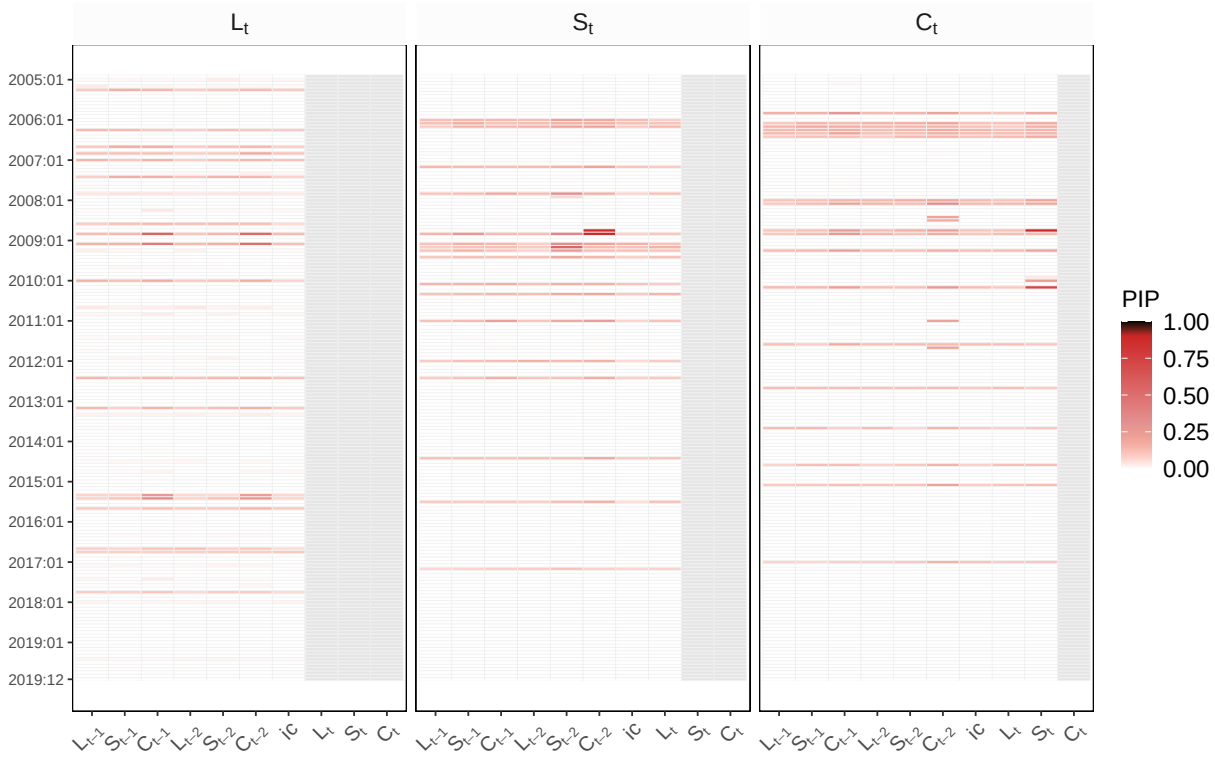
Notes: Grey shaded areas indicate coefficients which do not appear in the model due to the lower triangularity of $\boldsymbol{A}_{0t}$.

Figure 4: Heatmaps of posterior inclusion probability (PIPs) for time-variation in structural TVP-VAR coefficients with a regime-switching common shrinkage factor.

a) *dHS Mix*



b) *dHS MS*



Notes: Grey shaded areas indicate coefficients which do not appear in the model due to the lower triangularity of $\boldsymbol{A}_{0t}$.

6.3 Forecast exercise

The dataset covers the entire yield curve and includes yields from one-year to thirty-year bonds in one-year steps. We choose {1y, 3y, 5y, 7y, 10y, 15y, 30y} maturities as our target variables that we wish to forecast and consider one-month and one-quarter ahead as forecast horizons. We use a range of competing models that differ in terms of how they model time-variation in coefficients and the number of endogenous variables they have. All models feature stochastic volatility in the measurement errors and have two lags. We also offer comparison between the two MCMC algorithms: exact and approximate.

In terms of VAR dimension, we have large TVP-VARs and VARs with all 30 maturities ($M = 30$) as well as the three factor Nelson-Siegel model described in the previous sub-section ($M = 3$).

In terms of time variation specified through the likelihood function (i.e., through the definition of $\boldsymbol{W}$), we consider the flexible (FLEX) and random walk (RW) specifications defined in (2) and (3). In terms of time variation specified through the prior, we consider the five global-local shrinkage priors (four dynamic and one static) given in Sub-section 4.1. In addition, we consider as a competitor the conventional TVP-VAR setup of Primiceri (2005). We estimate the TVP-VAR only for the Nelson-Siegel model since the original prior overfits in higher dimensions.[5]

We also have VAR models where coefficients are constant over time. For these we do two versions, one with a Minnesota prior (MIN) and the other a horseshoe prior (HS). These models are estimated by setting $\boldsymbol{\beta} = \mathbf{0}$ and then using the sampling steps for $\boldsymbol{\alpha}$ detailed in the appendix. For the Minnesota prior, we use a non-conjugate version that allows for asymmetric shrinkage patterns and integrate out the corresponding hyperparameters within MCMC.

To evaluate one-month and one-quarter ahead forecasts, we use a recursive prediction design and split the sample into an initial estimation period that ranges from 2005:01 to 2008:12 and a forecast evaluation period from 2009:01 to 2019:12. We use Root Mean Squared Forecast Errors (RMSEs) as the measure of performance for our point forecasts

[5]The priors of the conventional Primiceri (2005) TVP-VAR are informed by OLS estimates using an initial training sample (in our case the initial first 18 observations). Such an empirical Bayesian calibration strategy is only sensible for models that feature a small number of endogenous variables.

and Continuous Ranked Probability Scores (CRPSs, Gneiting and Raftery, 2007) as the measure of performance of our density forecasts. Both are presented in ratio form relative to the benchmark model which is the large VAR with Minnesota prior. Values less than one indicate an approach is beating the benchmark.

We present our forecasting results in two tables. Table 2 shows the one-month ahead forecast performance of the different models while Table B.1 in the appendix shows the one-quarter ahead forecasting results. Our focus on one-step ahead forecasts is predicated by the fact that the density forecast measures based on proper scoring rules (such as CRPSs) can be viewed as a training sample marginal likelihood and thus enables model comparison (see Gneiting and Raftery, 2007).

Overall, the evidence in Table 2 (and Table B.1) is mixed, with no single approach being dominant. In principle, one robust pattern is that models with TVPs tend to produce more accurate forecasts than the large VAR with stochastic volatility benchmark. These gains range from being rather small (particularly at the short-end of the yield curve) to appreciable (when the focus is on the long-end of the yield curve). This is consistent with recent findings in Fischer *et al.* (2023) who document that flexible models work well for this particular dataset when longer maturities are considered.

If we compare results for the large TVP-VARs to results for the smaller TVP-VARs based on the Nelson-Siegel factors reveals that both specifications produce forecasts of similar quality. When forecasting one-month ahead and focusing on the CRPS as a measure of forecast performance, the best average forecast performance is produced by one of the large TVP-VARs. But when we focus on point forecasting performance, one of the NS models emerges as the best forecasting model. This finding indicates that using more information in an unrestricted manner seems to exert benign effect on higher order moments of the predictive density whereas for the first moment the effect is negligible (or even negative). Interestingly, this finding only holds for one-month ahead predictive densities. When we focus on one-quarter ahead forecasts (see Table B.1 in the appendix), this result is reversed with CRPSs indicating one of the NS models is forecasting best and RMSEs indicating one of the large TVP-VARs is forecasting best.

The comparison of the different choices for $\boldsymbol{W}$ also yields a mixed pattern of results.

At the short end of the yield curve the RW specification tends to forecast better, but at the longer end the FLEX specification does better. It is interesting to note, however, that the good performance for RW occurs with a large TVP-VAR whereas for the FLEX specification it occurs for a Nelson-Siegel version of the model.

In terms of which of our dynamic horseshoe priors forecasts best, it does seem to be the priors which assume λ_t to exhibit rapid change between values forecast better than the gradual change of the stochastic volatility specifications. That is, the Markov switching or mixture versions of the prior, `dHS MS` and `dHS Mix`, tend to forecast better than `dHS svol-Z` or `dHS svol-N`. Although there are several exceptions to this pattern. At this point it is also worth highlighting that the original Primiceri (2005) model is outperformed by our shrinkage specifications in all segments of the yield curve. This suggests that using proper shrinkage priors on the state innovation variances and allowing for dynamic shrinkage pays off.

Thus, overall (and with several exceptions) we have a story where, in this data set, time variation in the regression coefficients is present and there are gains to be made from capturing them. This can be seen by noting that the constant parameter VARs with stochastic volatility are never the best performing specifications across the different maturities and also for both time horizons we consider. As we have shown in the previous sub-section, this time variation is episodic (rather than gradually evolving) and only occurs occasionally and for some of the coefficients. However, ignoring this time variation and using constant parameter models leads to a deterioration in forecasts in almost all situations.

In terms of computation, our scalable algorithm does seem to work well. If we compare results from the exact MCMC algorithm to our approximate (non-sparsified) algorithm, it can be seen that using the computationally-faster approximation is not leading to a deterioration in forecast performance. In fact, there are some cases where the approximate forecasts are better than their exact counterparts.

Table 2: One-month ahead forecast performance for EA central government bond yields at different maturities using non-sparsified models.

Specification	One-month-ahead							
	Avg.	1y	3y	5y	7y	10y	15y	30y
VAR with constant coefficients								
Large with MIN	0.99	0.70	0.84	0.90	0.96	1.04	1.16	1.23
	(0.51)	(0.34)	(0.43)	(0.50)	(0.54)	(0.57)	(0.60)	(0.60)
Large VAR with HS prior	0.93	0.98	0.96	0.97	0.98	0.96	0.90	0.82
	(0.96)	(0.98)	(0.97)	(0.98)	(0.98)	(0.97)	(0.95)	(0.90)
Nelson-Siegel VAR with HS prior	0.91	1.02	0.98	0.97	0.97	0.94	0.88	0.78
	(0.96)	(1.01)	(1.01)	(0.99)	(0.98)	(0.96)	(0.94)	(0.87)
Nelson-Siegel with MIN prior	0.92	1.01	0.99	0.98	0.97	0.94	0.88	0.78
	(0.97)	(1.03)	(1.03)	(1.01)	(0.99)	(0.96)	(0.94)	(0.87)
Large TVP-VAR with the random walk specification for W								
dHS Mix	0.98	1.00	0.97	1.01	1.04	1.03	0.96	0.92
	(1.00)	(0.99)	(0.98)	(0.99)	(1.01)	(1.01)	(1.00)	(0.98)
dHS Mix (approx.)	0.93	0.99	0.98	0.99	0.99	0.95	0.90	0.83
	(0.97)	(0.97)	(0.98)	(0.98)	(0.98)	(0.97)	(0.97)	(0.94)
dHS MS	0.97	0.97	0.97	1.00	1.02	1.01	0.96	0.87
	(0.99)	(0.98)	(0.98)	(1.00)	(1.01)	(1.01)	(1.00)	(0.95)
dHS MS (approx.)	0.91	0.97	0.95	**0.96**	**0.96**	0.94	0.88	0.82
	(**0.95**)	(0.98)	(**0.96**)	(**0.96**)	(**0.97**)	(0.96)	(0.95)	(0.92)
dHS svol-N	0.92	0.99	0.98	0.98	0.98	0.94	0.88	0.81
	(0.96)	(0.98)	(0.99)	(0.99)	(0.98)	(0.97)	(0.95)	(0.89)
dHS svol-N (approx.)	0.93	0.98	0.97	0.99	0.99	0.96	0.90	0.82
	(1.01)	(0.98)	(0.98)	(0.99)	(0.99)	(1.03)	(1.06)	(1.02)
dHS svol-Z	0.94	0.98	0.96	0.98	0.99	0.97	0.92	0.83
	(0.99)	(0.97)	(0.97)	(0.99)	(0.99)	(0.99)	(0.97)	(1.02)
dHS svol-Z (approx.)	0.93	0.98	0.97	0.99	0.99	0.97	0.90	0.83
	(0.97)	(0.98)	(0.98)	(0.99)	(0.99)	(0.98)	(0.96)	(0.92)
sHS	0.96	1.00	0.96	0.99	1.02	1.01	0.95	0.88
	(1.04)	(0.99)	(0.97)	(0.99)	(1.02)	(1.02)	(1.13)	(1.09)
sHS (approx.)	0.93	0.99	**0.94**	0.97	0.98	0.96	0.91	0.82
	(0.97)	(0.98)	(0.97)	(0.98)	(0.98)	(0.98)	(0.97)	(0.92)
Large TVP-VAR with the flexible specification for W								
dHS Mix	1.05	0.99	0.96	1.01	1.07	1.09	1.07	1.05
	(1.04)	(0.98)	(0.98)	(1.02)	(1.05)	(1.07)	(1.08)	(1.06)
dHS Mix (approx.)	0.91	0.98	0.96	0.97	0.97	0.94	0.89	0.78
	(1.01)	(0.98)	(0.97)	(0.98)	(0.98)	(0.98)	(0.99)	(1.18)
dHS MS	1.13	1.00	1.03	1.12	1.17	1.17	1.17	1.14
	(1.18)	(1.04)	(1.10)	(1.16)	(1.20)	(1.21)	(1.22)	(1.26)
dHS MS (approx.)	0.93	0.98	0.97	0.99	0.99	0.95	0.89	0.82
	(1.01)	(0.98)	(0.98)	(0.99)	(0.99)	(0.97)	(0.96)	(1.15)
dHS svol-N	0.92	0.98	0.95	0.96	0.97	0.95	0.89	0.85
	(0.96)	(0.98)	(0.97)	(0.97)	(0.98)	(0.97)	(0.95)	(0.91)
dHS svol-N (approx.)	0.92	0.98	0.95	0.97	0.97	0.95	0.91	0.83
	(1.00)	(0.98)	(0.97)	(0.97)	(0.97)	(0.97)	(0.96)	(1.15)
dHS svol-Z	0.95	0.98	0.96	0.98	1.00	0.98	0.92	0.88
	(0.98)	(0.98)	(0.98)	(0.99)	(1.00)	(1.00)	(0.99)	(0.95)
dHS svol-Z (approx.)	0.93	0.97	0.96	0.98	0.98	0.96	0.92	0.83
	(0.96)	(0.98)	(0.97)	(0.98)	(0.98)	(0.97)	(0.96)	(0.91)
sHS	0.92	0.98	0.96	0.97	0.97	0.94	0.89	0.82
	(0.95)	(0.98)	(0.97)	(0.98)	(0.97)	(0.96)	(0.95)	(0.89)
sHS (approx.)	0.93	0.96	0.95	0.97	0.98	0.96	0.91	0.83
	(0.96)	(**0.96**)	(0.97)	(0.98)	(0.98)	(0.97)	(0.95)	(0.92)

Table 2 continued

Specification	One-month-ahead							
	Avg.	1y	3y	5y	7y	10y	15y	30y
Nelson-Siegel TVP-VAR with the random walk specification for W								
dHS Mix	1.11	1.14	1.21	1.19	1.18	1.15	1.08	0.98
	(1.12)	(1.10)	(1.16)	(1.14)	(1.13)	(1.13)	(1.12)	(1.06)
dHS Mix (approx.)	1.10	1.27	1.06	1.11	1.14	1.13	1.08	1.00
	(1.04)	(1.09)	(1.05)	(1.06)	(1.05)	(1.04)	(1.04)	(0.98)
dHS MS	1.06	1.11	1.11	1.14	1.13	1.11	1.03	0.92
	(1.07)	(1.09)	(1.12)	(1.10)	(1.09)	(1.08)	(1.07)	(1.00)
dHS MS (approx.)	0.98	1.01	1.06	1.08	1.07	1.02	0.94	0.81
	(0.99)	(1.02)	(1.05)	(1.03)	(1.01)	(0.99)	(0.97)	(0.88)
dHS svol-N	1.07	1.14	1.15	1.14	1.14	1.11	1.04	0.92
	(1.07)	(1.09)	(1.11)	(1.09)	(1.09)	(1.08)	(1.07)	(0.99)
dHS svol-N (approx.)	0.92	0.99	0.98	0.99	0.99	0.96	0.89	0.78
	(0.97)	(1.00)	(1.02)	(1.00)	(0.99)	(0.97)	(0.95)	(0.87)
dHS svol-Z	1.11	1.19	1.21	1.19	1.18	1.14	1.07	0.96
	(1.10)	(1.11)	(1.14)	(1.12)	(1.11)	(1.11)	(1.10)	(1.03)
dHS svol-Z (approx.)	0.92	1.03	1.00	0.99	0.98	0.94	0.88	0.78
	(0.97)	(1.02)	(1.02)	(1.00)	(0.99)	(0.96)	(0.94)	(0.87)
sHS	1.09	1.09	1.15	1.15	1.15	1.13	1.08	0.96
	(1.13)	(1.10)	(1.16)	(1.15)	(1.14)	(1.14)	(1.14)	(1.07)
sHS (approx.)	0.91	1.00	0.97	0.98	0.97	0.94	0.88	0.78
	(0.96)	(1.01)	(1.02)	(1.00)	(0.98)	(0.96)	(0.94)	(0.86)
Nelson-Siegel TVP-VAR with the flexible specification for W								
dHS Mix	1.19	1.25	1.38	1.36	1.28	1.20	1.10	0.99
	(1.33)	(1.32)	(1.43)	(1.40)	(1.35)	(1.31)	(1.29)	(1.27)
dHS Mix (approx.)	**0.90**	1.02	0.98	0.98	0.97	**0.93**	**0.85**	**0.75**
	(0.95)	(1.01)	(1.03)	(1.00)	(0.98)	(**0.95**)	(**0.92**)	(**0.85**)
dHS MS	1.00	**0.93**	1.13	1.13	1.10	1.04	0.95	0.83
	(1.01)	(0.99)	(1.06)	(1.05)	(1.04)	(1.01)	(0.99)	(0.91)
dHS MS (approx.)	1.06	1.20	1.20	1.18	1.16	1.09	0.98	0.85
	(1.01)	(1.05)	(1.07)	(1.06)	(1.04)	(1.02)	(0.98)	(0.90)
dHS svol-N	1.11	1.26	1.09	1.11	1.14	1.14	1.08	1.02
	(1.11)	(1.18)	(1.14)	(1.13)	(1.13)	(1.11)	(1.10)	(1.05)
dHS svol-N (approx.)	0.92	1.01	0.98	0.99	0.98	0.95	0.88	0.78
	(0.97)	(1.01)	(1.02)	(1.01)	(0.99)	(0.97)	(0.95)	(0.87)
dHS svol-Z	1.07	1.20	1.18	1.17	1.14	1.09	1.01	0.92
	(1.23)	(1.23)	(1.29)	(1.28)	(1.26)	(1.22)	(1.20)	(1.16)
dHS svol-Z (approx.)	0.93	1.09	1.07	1.03	1.00	0.94	0.86	0.75
	(0.97)	(1.03)	(1.04)	(1.01)	(0.99)	(0.97)	(0.94)	(0.86)
sHS	0.93	1.11	0.97	0.98	0.99	0.96	0.90	0.80
	(0.98)	(1.03)	(1.02)	(1.01)	(1.00)	(0.98)	(0.97)	(0.89)
sHS (approx.)	0.91	1.01	0.99	0.98	0.98	0.94	0.87	0.77
	(0.96)	(1.02)	(1.03)	(1.00)	(0.98)	(0.96)	(0.94)	(0.86)
Nelson-Siegel TVP-VAR with the conventional Primiceri (2005) setup								
	0.99	1.05	1.14	1.09	1.06	1.01	0.93	0.82
	(1.02)	(1.06)	(1.12)	(1.07)	(1.03)	(1.01)	(0.99)	(0.92)

Notes: This table displays the one-step ahead forecast performance for non-sparsified models. We focus on seven maturities (1y, 3y, 5y, 7y, 10y, 15y, and 30y) as our target variables and use a hold-out period from 2009:01 to 2019:12. Point forecast performance is measured by relative root mean square errors (RMSEs), while density forecast performance (shown in parentheses) by relative continuous ranked probability scores (CRPSs). We consider two different models in terms of the dimension of the (TVP-)VARs: a large model including all 30 maturities ($M = 30$) and a small model specified as a three factor Nelson-Siegel model ($M = 3$). For the main TVP-VARs, we consider a flexible and a RW specification of $\boldsymbol{W}$, each with five different global-local shrinkage priors (four dynamic and one static). These TVP-VARs are estimated with two different algorithms: our proposed approximate approach and an exact algorithm. In addition, we consider the conventional TVP-VAR setup of Primiceri (2005) for the Nelson-Siegel model and a set of VARs with constant coefficients. For the VARs with constant parameters, we adopt either a Minnesota or a horseshoe (HS) shrinkage prior. As overall benchmark model we choose a large VAR with constant parameters and a Minnesota prior. The red shaded rows correspond to the actual RMSE and CRPS values of this benchmark model, while the grey shaded rows correspond to models for which we use our approximate (but non-sparsified) MCMC algorithm. The best performing specification is in bold.

7 Closing remarks

VARs modelled with many macroeconomic and financial data sets exhibit parameter change and structural breaks. Typically, most parameter change is found in the error covariance matrix. But there can be small amounts of time-variation in VAR coefficients where only some coefficients change and even they only change at points in time. The problem is how to uncover TVPs of this sort. Simply working with a model where all VAR coefficients change can lead to over-fitting and poor forecast performance. In light of this situation, one contribution of this paper lies in our development of several dynamic horseshoe priors which are designed for picking up the kind of parameter change that often occurs in practice. In an application involving eurozone yield data our methods find small amounts of time variation in parameters. In a forecasting exercise we find that appropropriately modeling this time variation leads to forecast improvements.

The second contribution of this paper lies in computation. The approximate MCMC algorithm developed in this paper is scalable in a manner that exact MCMC algorithms are not. Thus, we have developed an algorithm which can be used in the huge dimensional models that are increasingly being used by economists. Finally, we have developed an MCMC algorithm for common stochastic volatility specifications which is particularly well-suited for large k applications such as the one considered in this paper.

References

Belmonte M, Koop G, and Korobilis D (2014), "Hierarchical shrinkage in time-varying coefficient models," *Journal of Forecasting* **33**(1), 80–94.

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

Bhattacharya A, Pati D, Pillai NS, and Dunson DB (2015), "Dirichlet–Laplace priors for optimal shrinkage," *Journal of the American Statistical Association* **110**(512), 1479–1490.

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

Carvalho CM, Polson NG, and Scott JG (2010), "The horseshoe estimator for sparse signals," *Biometrika* **97**(2), 465–480.

Chan JC, Eisenstat E, and Strachan RW (2020), "Reducing the state space dimension in a large TVP-VAR," *Journal of Econometrics* **218**(1), 105–118.

Chan JC, and Jeliazkov I (2009), "Efficient simulation and integrated likelihood estimation in state space models," *International Journal of Mathematical Modelling and Numerical Optimisation* **1**(1-2), 101–120.

Clark TE (2011), "Real-Time Density Forecasts From Bayesian Vector Autoregressions With Stochastic Volatility," *Journal of Business & Economic Statistics* **29**(3), 327–341.

Cogley T, Primiceri GE, and Sargent TJ (2010), "Inflation-gap persistence in the US," *American Economic Journal: Macroeconomics* **2**(1), 43–69.

Diebold FX, Rudebusch GD, and Aruoba SB (2006), "The macroeconomy and the yield curve: a dynamic latent factor approach," *Journal of Econometrics* **131**(1-2), 309–338.

Eisenstat E, Chan JC, and Strachan RW (2016), "Stochastic model specification search for time-varying parameter VARs," *Econometric Reviews* **35**(8-10), 1638–1665.

Fischer MM, Hauzenberger N, Huber F, and Pfarrhofer M (2023), "General Bayesian time-varying parameter vector autoregressions for modeling government bond yields," *Journal of Applied Econometrics* **38**(1), 69–87.

Gneiting T, and Raftery AE (2007), "Strictly proper scoring rules, prediction, and estimation," *Journal of the American statistical Association* **102**(477), 359–378.

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

Hahn PR, and Carvalho CM (2015), "Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective," *Journal of the American Statistical Association* **110**(509), 435–448.

Hauzenberger N (2021), "Flexible mixture priors for large time-varying parameter models," *Econometrics and Statistics* **20**, 87–108.

Hauzenberger N, Huber F, Koop G, and Onorante L (2022), "Fast and flexible Bayesian inference in time-varying parameter regression models," *Journal of Business & Economic Statistics* **40**(4), 1904–1918.

Huber F, Koop G, and Onorante L (2021), "Inducing sparsity and shrinkage in time-varying parameter models," *Journal of Business & Economic Statistics* **39**(3), 669–683.

Huber F, Koop G, and Pfarrhofer M (2020), "Bayesian Inference in High-Dimensional Time-varying Parameter Models using Integrated Rotated Gaussian Approximations," *arXiv preprint arXiv:2002.10274* .

Ishwaran H, and Rao JS (2005), "Spike and slab variable selection: Frequentist and Bayesian strategies," *The Annals of Statistics* **33**(2), 730–773.

Jacquier E, Polson N, and Rossi P (1995), "Models and Priors for Multivariate Stochastic Volatility Models," Technical report, Technical Report, University of Chicago, Graduate School of Business.

Johndrow J, Orenstein P, and Bhattacharya A (2020), "Scalable Approximate MCMC Algorithms for the Horseshoe Prior," *Journal of Machine Learning Research* **21**(73), 1–61.

Johndrow JE, Orenstein P, and Bhattacharya A (2017), "Bayes shrinkage at GWAS scale: Convergence and approximation theory of a scalable MCMC algorithm for the horseshoe prior," *arXiv preprint arXiv:1705.00841* .

Kalli M, and Griffin J (2014), "Time-varying sparsity in dynamic regression models," *Journal of Econometrics* **178**(2), 779 – 793.

——— (2019), "Bayesian nonparametric time varying vector autoregressive models," *Journal of Business & Economic Statistics* .

Kastner G, and Frühwirth-Schnatter S (2014), "Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models," *Compu-*

tational Statistics & Data Analysis **76**, 408–423.

Kastner G, and Huber F (2020), "Sparse Bayesian vector autoregressions in huge dimensions," *Journal of Forecasting* **39**(7), 1142–1165.

Kim CJ, and Nelson CR (1999a), "Has the US economy become more stable? A Bayesian approach based on a Markov-switching model of the business cycle," *Review of Economics and Statistics* **81**(4), 608–616.

——— (1999b), "State-space models with regime switching: classical and Gibbs-sampling approaches with applications," *MIT Press Books* **1**.

Kim S, Shephard N, and Chib S (1998), "Stochastic volatility: likelihood inference and comparison with ARCH models," *The Review of Economic Studies* **65**(3), 361–393.

Knaus P, Bitto-Nemling A, Cadonna A, and Frühwirth-Schnatter S (2021), "Shrinkage in the time-varying parameter model framework using the R package shrinkTVP," *Journal of Statistical Software* **100**(13), 1–32.

Korobilis D (2021), "High-dimensional macroeconomic forecasting using message passing algorithms," *Journal of Business & Economic Statistics* **39**(2), 493–504.

——— (2022), "A new algorithm for structural restrictions in Bayesian vector autoregressions," *European Economic Review* **148**, 104241.

Kowal DR, Matteson DS, and Ruppert D (2019), "Dynamic shrinkage processes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **81**(4), 781–804.

Makalic E, and Schmidt DF (2015), "A simple sampler for the horseshoe estimator," *IEEE Signal Processing Letters* **23**(1), 179–182.

McCausland WJ, Miller S, and Pelletier D (2011), "Simulation smoothing for state–space models: A computational efficiency analysis," *Computational Statistics & Data Analysis* **55**(1), 199–212.

Nelson CR, and Siegel AF (1987), "Parsimonious modeling of yield curves," *Journal of Business* 473–489.

Park T, and Casella G (2008), "The Bayesian Lasso," *Journal of the American Statistical Association* **103**(482), 681–686.

Petrova K (2019), "A quasi-Bayesian local likelihood approach to time varying parameter VAR models," *Journal of Econometrics* **212**(1), 286–306.

Primiceri G (2005), "Time varying structural autoregressions and monetary policy," *Oxford University Press* **72**(3), 821–852.

Puelz D, Hahn PR, and Carvalho CM (2020), "Portfolio selection for individual passive investing," *Applied Stochastic Models in Business and Industry* **36**(1), 124–142.

Ray P, and Bhattacharya A (2018), "Signal Adaptive Variable Selector for the Horseshoe Prior," *arXiv preprint arXiv:1810.09004* .

A Details of the MCMC Algorithm

A.1 Sampling the Log-Volatilities

We assume a stochastic volatility process of the following form for $h_t = log(\sigma_t^2)$:

$$h_t = \mu_h + \rho_h(h_{t-1} - \mu_h) + \sigma_h v_t, \quad v_t \sim \mathcal{N}(0,1), \quad h_0 \sim \mathcal{N}\left(\mu, \frac{\sigma_h^2}{1-\rho_h^2}\right).$$

Following Kastner and Frühwirth-Schnatter (2014) we make the prior assumptions that $\mu_h \sim \mathcal{N}(0, 10)$, $\frac{\rho_h+1}{2} \sim \mathcal{B}(5, 1.5)$ and $\sigma_h^2 \sim \mathcal{G}(1/2, 1/2)$ where $\mathcal{B}$ and $\mathcal{G}$ denote the Beta and Gamma distributions, respectively. We use the algorithm of Kastner and Frühwirth-Schnatter (2014) to take draws of h_t.

A.2 Sampling the Time-Invariant Regression Coefficients

Most of the conditional posterior distributions take a simple and well-known form. Here we briefly summarize these and provide some information on the relevant literature.

The time-invariant coefficients $\boldsymbol{\alpha}$ follow a K-dimensional multivariate Gaussian posterior given by

$$\begin{aligned}
\boldsymbol{\alpha}|\bullet &\sim \mathcal{N}(\overline{\boldsymbol{\alpha}}, \overline{\boldsymbol{V}}_\alpha), \\
\overline{\boldsymbol{V}}_\alpha &= \left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} + \boldsymbol{D}_\alpha^{-1}\right)^{-1}, \\
\overline{\boldsymbol{\alpha}} &= \overline{\boldsymbol{V}}_\alpha \tilde{\boldsymbol{X}}\hat{\boldsymbol{y}},
\end{aligned}$$

with $\tilde{\boldsymbol{X}} = \boldsymbol{L}^{-1}\boldsymbol{X}$, $\hat{\boldsymbol{y}} = \boldsymbol{L}^{-1}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\beta})$ and $\boldsymbol{D}_\alpha = \tau_\alpha \text{ diag}(\psi_1^2, \ldots, \psi_K^2)$ denoting a $K \times K$-dimensional prior variance-covariance matrix with ψ_j $(j = 1, \ldots, K)$ and $\sqrt{\tau}_\alpha$ following a half-Cauchy distribution, respectively.

A.3 Sampling the horseshoe Prior on the Constant and the Time-varying Parameters

Makalic and Schmidt (2015) show that one can simulate from the posterior distribution of ψ_j using standard distributions only. This is achieved by introducing additional auxiliary quantities

ϱ_j $(j = 1, \ldots, K)$. Using these, the posterior of ψ_j follows an inverted Gamma distribution:

$$\psi_j^2 | \bullet \sim \mathcal{G}^{-1}\left(1, \frac{1}{\varrho_j} + \frac{\alpha_j^2}{2\tau_\alpha}\right)$$

where α_j denotes the j^{th} element of $\boldsymbol{\alpha}$. The posterior of ϱ_j is also inverse Gamma distributed with $\varrho_j | \bullet \sim \mathcal{G}^{-1}(1, 1 + \psi_j^{-2})$.

For the global shrinkage parameter, we introduce yet another auxiliary quantity ϖ_α. This enables us to derive a conditional posterior for τ_α which is also inverse Gamma distributed:

$$\tau_\alpha | \bullet \sim \mathcal{G}^{-1}\left(\frac{K+1}{2}, \frac{1}{\varpi_\alpha} + \sum_{j=1}^{K} \frac{\alpha_j^2}{2\psi_j^2}\right)$$

and the posterior of ϖ_α being given by:

$$\varpi_\alpha | \bullet \sim \mathcal{G}^{-1}(1, 1 + \tau_\alpha^{-1}).$$

The local shrinkage parameters ϕ_{jt} can be simulated conditionally on τ and $\{\lambda_t\}_{t=1}^T$ similarly to the ψ_j's. Specificially, the posterior distribution of ϕ_{jt}^2 follows an inverse Gamma:

$$\phi_{jt}^2 | \bullet \sim \mathcal{G}^{-1}\left(1, \frac{1}{\vartheta_{jt}} + \frac{\beta_{jt}^2}{2\tau\lambda_t}\right)$$

with ϑ_{jt} denoting yet another scaling parameter that follows a inverse Gamma posterior distribution: $\vartheta_{jt} | \bullet \sim \mathcal{G}^{-1}(1, 1 + \phi_{jt}^{-2})$.

If we do not assume λ_t to evolve according to an AR(1) process, we sample the global shrinkage parameter τ similar to τ_α. The conditional posterior of τ also follows an inverse Gamma:

$$\tau | \bullet \sim \mathcal{G}^{-1}\left(\frac{k+1}{2}, \frac{1}{\varpi} + \sum_{t=1}^{T}\sum_{j=1}^{K} \frac{\beta_{jt}^2}{2\lambda_t\phi_{jt}^2}\right)$$

with the posterior of the auxiliary variable ϖ given by:

$$\varpi | \bullet \sim \mathcal{G}^{-1}(1, 1 + \tau^{-1}).$$

A.4 Sampling the Dynamic Shrinkage Parameters

As stated in Sub-section 4.1, the full history of λ_t in the case that it follows a mixture or Markov switching specification can be easily obtained through standard techniques. More precisely, if d_t in (6) follows a Markov switching model, we adopt the algorithm discussed in, e.g., Kim and Nelson (1999b,a). The posterior of the transition probabilities is Beta distributed:

$$p_{ii}|\bullet \sim \mathcal{B}(a_{i,MS} + T_{i0}, b_{i,MS} + T_{i1}),$$

whereby T_{ij} denotes the number of times a transition from state i to j has been observed in the full history of d_t.

In the case of the mixture model, the posterior distribution of d_t follows a Bernoulli distribution for each t:

$$Prob(d_t = 1|\bullet) = Ber(\overline{p}_t)$$

with $\overline{p}_t$ given by:

$$\overline{p}_t = \frac{\kappa_1^{-K/2} \exp\left(-\frac{\sum_{j=1}^{K} \hat{\beta}_{jt}}{2\kappa_1^2}\right) \times \underline{p}}{\kappa_1^{-K/2} \exp\left(-\frac{\sum_{j=1}^{K} \hat{\beta}_{jt}}{2\kappa_1^2}\right) \times \underline{p} + \kappa_0^{-K/2} \exp\left(-\frac{\sum_{j=1}^{K} \hat{\beta}_{jt}}{2\kappa_0^2}\right) \times (1 - \underline{p})}.$$

and the posterior of $\underline{p}$ follows a Beta distribution $\underline{p}|\bullet \sim \mathcal{B}\left(\sum_{t=1}^{T} d_t + a_{Mix}, 1 - \sum_{t=1}^{T} d_t + b_{Mix}\right)$.

Finally, in the case that λ_t evolves according to an AR(1) process with Gaussian shocks, we use precisely the same algorithm as Kastner and Frühwirth-Schnatter (2014) for simulating μ and ρ. In the case that we use Z-distributed shocks, the algorithm proposed in Kowal *et al.* (2019) is adopted. This implies that we use Polya-Gamma (PG) auxiliary random variables to approximate the Z-distribution using a scale-mixture of Gaussians. Essentially, the main implication is that conditional on the T PG random variates, the parameters of the state evolution equation can be estimated similarly to the Gaussian case after normalizing everything by rendering the AR(1) conditionally homoscedastic. For more details, see Kowal *et al.* (2019).

B Higher-order forecast performance

Table B.1: One-quarter ahead forecast performance for EA central government bond yields at different maturities using non-sparsified models.

Specification	One-quarter-ahead							
	Avg.	1y	3y	5y	7y	10y	15y	30y
VAR with constant coefficients								
Large VAR with MIN prior	0.97	0.74	0.83	0.95	1.03	1.08	1.10	0.99
	(0.53)	(0.36)	(0.45)	(0.53)	(0.58)	(0.61)	(0.62)	(0.56)
Large VAR with HS prior	0.95	0.92	0.94	0.94	0.94	0.94	0.96	1.00
	(0.95)	(0.95)	(0.96)	(0.96)	(0.95)	(0.95)	(0.95)	(0.96)
Nelson-Siegel VAR with HS prior	0.95	1.02	0.97	0.94	0.92	0.92	0.94	0.98
	(0.95)	(1.03)	(1.01)	(0.96)	(0.93)	(0.92)	(0.93)	(0.94)
Nelson-Siegel VAR with MIN prior	0.96	1.02	1.00	0.95	0.93	0.93	0.95	0.99
	(0.97)	(1.05)	(1.03)	(0.97)	(0.95)	(0.94)	(0.94)	(0.95)
Large TVP-VAR with random walk specification for W								
dHS Mix	1.06	0.94	0.96	0.99	1.01	1.04	1.09	1.25
	(1.01)	(0.96)	(0.97)	(0.99)	(1.00)	(1.02)	(1.03)	(1.07)
dHS Mix (approx.)	**0.94**	**0.91**	0.94	0.93	0.91	**0.91**	0.94	1.00
	(0.96)	(0.95)	(**0.94**)	(**0.93**)	(0.93)	(0.94)	(0.97)	(1.09)
dHS MS	0.97	0.93	0.95	0.94	0.93	0.95	0.99	1.08
	(0.97)	(0.95)	(0.97)	(0.96)	(0.96)	(0.96)	(0.98)	(1.01)
dHS MS (approx.)	0.94	0.95	0.93	0.92	0.92	0.93	0.94	0.99
	(0.96)	(0.95)	(0.95)	(0.95)	(0.95)	(0.95)	(0.96)	(0.99)
dHS svol-N	0.95	0.92	0.94	0.94	0.93	0.94	0.95	1.01
	(0.97)	(0.95)	(0.96)	(0.96)	(0.95)	(0.96)	(0.97)	(1.04)
dHS svol-N (approx.)	0.95	0.94	0.94	0.94	0.93	0.94	0.95	1.00
	(1.10)	(1.16)	(1.13)	(1.10)	(1.08)	(1.07)	(1.07)	(1.11)
dHS svol-Z	0.95	0.93	0.94	0.93	0.93	0.94	0.96	1.01
	(1.09)	(1.15)	(1.12)	(1.09)	(1.07)	(1.06)	(1.07)	(1.10)
dHS svol-Z (approx.)	0.96	0.93	0.96	0.94	0.93	0.94	0.96	1.02
	(1.03)	(0.95)	(0.97)	(0.97)	(0.96)	(1.08)	(1.08)	(1.13)
sHS	1.00	0.97	0.96	0.96	0.96	0.99	1.02	1.07
	(1.13)	(1.18)	(1.14)	(1.12)	(1.11)	(1.12)	(1.12)	(1.16)
sHS (approx.)	0.94	0.93	0.93	0.93	0.92	0.92	0.94	0.99
	(1.02)	(0.96)	(0.96)	(0.96)	(0.95)	(1.07)	(1.08)	(1.12)
Large TVP-VAR with the flexible specification for W								
dHS Mix	1.11	0.96	0.93	0.95	1.01	1.11	1.22	1.36
	(1.10)	(0.98)	(0.99)	(1.03)	(1.07)	(1.13)	(1.18)	(1.25)
dHS Mix (approx.)	0.99	0.97	0.99	0.98	0.96	0.95	0.97	1.11
	(1.52)	(1.06)	(1.20)	(1.26)	(1.27)	(1.37)	(1.60)	(2.68)
dHS MS	2.11	0.95	1.66	2.06	2.10	2.22	2.30	2.49
	(2.58)	(1.45)	(2.02)	(2.37)	(2.53)	(2.71)	(2.90)	(3.50)
dHS MS (approx.)	0.96	0.94	0.96	0.95	0.94	0.95	0.97	1.01
	(1.26)	(1.39)	(1.31)	(1.25)	(1.22)	(1.21)	(1.21)	(1.29)
dHS svol-N	0.96	0.93	0.93	0.93	0.93	0.94	0.98	1.03
	(0.96)	(0.95)	(0.95)	(0.95)	(0.95)	(0.95)	(0.96)	(0.97)
dHS svol-N (approx.)	0.94	0.91	**0.92**	**0.92**	0.92	0.93	0.95	0.99
	(1.23)	(1.36)	(1.28)	(1.23)	(1.20)	(1.19)	(1.19)	(1.22)
dHS svol-Z	0.98	0.91	0.93	0.94	0.94	0.97	1.01	1.09
	(0.98)	(0.95)	(0.96)	(0.97)	(0.96)	(0.98)	(1.01)	(1.03)
dHS svol-Z (approx.)	0.95	0.97	0.96	0.95	0.93	0.93	0.95	0.99
	(0.95)	(0.96)	(0.96)	(0.96)	(0.95)	(0.95)	(0.95)	(0.95)
sHS	0.95	0.93	0.95	0.94	0.93	0.93	0.95	0.99
	(0.95)	(0.95)	(0.96)	(0.95)	(0.95)	(0.94)	(0.95)	(0.94)
sHS (approx.)	0.94	0.93	0.94	0.93	0.92	0.92	0.94	0.97
	(0.95)	(**0.94**)	(0.96)	(0.96)	(0.95)	(0.95)	(0.94)	(0.95)

Table B.1 continued

Specification	One-quarter-ahead							
	Avg.	1y	3y	5y	7y	10y	15y	30y
Nelson-Siegel TVP-VAR with random walk specification for W								
dHS Mix	1.02	1.02	1.07	1.03	1.00	0.99	1.01	1.05
	(1.08)	(1.13)	(1.17)	(1.10)	(1.05)	(1.03)	(1.03)	(1.08)
dHS Mix (approx.)	1.37	1.93	1.47	1.34	1.27	1.24	1.24	1.33
	(1.12)	(1.35)	(1.21)	(1.12)	(1.08)	(1.06)	(1.07)	(1.10)
dHS MS	0.98	0.96	1.00	0.97	0.96	0.96	0.98	1.04
	(1.05)	(1.09)	(1.12)	(1.06)	(1.02)	(1.01)	(1.02)	(1.06)
dHS MS (approx.)	0.94	0.95	0.98	0.94	0.92	0.92	0.94	0.98
	(0.96)	(1.01)	(1.02)	(0.97)	(0.94)	(0.94)	(0.94)	(0.95)
dHS svol-N	0.98	0.98	1.02	0.99	0.97	0.96	0.97	1.02
	(1.02)	(1.08)	(1.10)	(1.03)	(1.00)	(0.98)	(0.99)	(1.02)
dHS svol-N (approx.)	0.97	1.23	0.96	0.94	0.92	0.93	0.95	0.98
	(0.96)	(1.08)	(1.01)	(0.96)	(0.94)	(0.93)	(0.94)	(0.94)
dHS svol-Z	1.00	0.97	1.04	1.00	0.98	0.98	0.99	1.03
	(1.04)	(1.10)	(1.13)	(1.06)	(1.02)	(1.00)	(1.01)	(1.05)
dHS svol-Z (approx.)	1.00	1.53	0.98	0.94	0.92	0.92	0.94	0.97
	(0.96)	(1.15)	(1.01)	(0.96)	(0.93)	(0.92)	(0.93)	(0.93)
sHS	1.28	1.54	1.15	1.17	1.22	1.28	1.28	1.36
	(1.21)	(1.32)	(1.27)	(1.21)	(1.17)	(1.16)	(1.18)	(1.24)
sHS (approx.)	0.98	1.04	1.09	1.05	0.97	0.93	**0.92**	**0.96**
	(0.97)	(1.05)	(1.05)	(1.00)	(0.96)	(0.93)	(0.93)	(0.93)
Nelson-Siegel TVP-VAR with the flexible specification for W								
dHS Mix	1.90	2.10	2.49	1.89	1.64	1.59	1.70	2.15
	(2.13)	(2.24)	(2.55)	(2.23)	(2.01)	(1.92)	(1.94)	(2.15)
dHS Mix (approx.)	1.22	1.51	1.42	1.11	1.26	1.19	1.12	1.06
	(1.06)	(1.18)	(1.16)	(1.09)	(1.06)	(1.03)	(1.01)	(0.97)
dHS MS	1.03	1.04	1.28	1.10	1.00	0.96	0.95	0.99
	(1.00)	(1.05)	(1.09)	(1.02)	(0.98)	(0.97)	(0.97)	(0.98)
dHS MS (approx.)	0.96	0.99	1.07	0.97	0.92	0.92	0.93	0.97
	(0.98)	(1.07)	(1.06)	(0.99)	(0.96)	(0.95)	(0.96)	(0.96)
dHS svol-N	1.37	1.23	1.63	1.44	1.33	1.30	1.32	1.37
	(1.33)	(1.36)	(1.51)	(1.38)	(1.30)	(1.26)	(1.26)	(1.29)
dHS svol-N (approx.)	0.94	0.97	0.97	0.93	0.91	0.91	0.93	0.97
	(**0.95**)	(1.01)	(1.00)	(0.95)	(**0.92**)	(**0.92**)	(0.93)	(0.94)
dHS svol-Z	1.61	1.50	2.00	1.70	1.52	1.46	1.49	1.68
	(1.73)	(1.70)	(2.04)	(1.83)	(1.68)	(1.61)	(1.62)	(1.75)
dHS svol-Z (approx.)	0.94	1.00	0.95	0.92	**0.91**	0.91	0.94	0.98
	(0.95)	(1.02)	(1.00)	(0.95)	(0.92)	(0.92)	(0.93)	(0.93)
sHS	1.00	1.07	1.11	1.01	0.95	0.94	0.96	1.01
	(1.00)	(1.06)	(1.09)	(1.02)	(0.97)	(0.96)	(0.97)	(0.97)
sHS (approx.)	0.95	1.04	0.98	0.94	0.92	0.92	0.94	0.97
	(0.95)	(1.03)	(1.01)	(0.96)	(0.93)	(0.92)	(**0.93**)	(**0.93**)
Nelson-Siegel TVP-VAR with the conventional Primiceri (2005) setup								
	1.00	0.98	1.02	1.00	0.98	0.98	0.99	1.04
	(1.02)	(1.07)	(1.11)	(1.04)	(1.01)	(0.99)	(0.99)	(1.00)

Notes: This table displays the three-step ahead forecast performance for non-sparsified models. We focus on seven maturities (1y, 3y, 5y, 7y, 10y, 15y, and 30y) as our target variables and use a hold-out period from 2009:01 to 2019:12. Point forecast performance is measured by relative root mean square errors (RMSEs), while density forecast performance (shown in parentheses) by relative continuous ranked probability scores (CRPSs). We consider two different models in terms of the dimension of the (TVP-)VARs: a large model including all 30 maturities ($M = 30$) and a small model specified as a three factor Nelson-Siegel model ($M = 3$). For the main TVP-VARs, we consider a flexible and a RW specification of $\boldsymbol{W}$, each with five different global-local shrinkage priors (four dynamic and one static). These TVP-VARs are estimated with two different algorithms: our proposed approximate approach and an exact algorithm. In addition, we consider the conventional TVP-VAR setup of Primiceri (2005) for the Nelson-Siegel model and a set of VARs with constant coefficients. For the VARs with constant parameters, we adopt either a Minnesota or a horseshoe (HS) shrinkage prior. As overall benchmark model we choose a large VAR with constant parameters and a Minnesota prior. The red shaded rows correspond to the actual RMSE and CRPS values of this benchmark model, while the grey shaded rows correspond to models for which we use our approximate (but non-sparsified) MCMC algorithm. The best performing specification is in bold.