Journal of Econometrics 210 (2019) 75–97

Contents lists available at ScienceDirect

Journal of Econometrics

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

Achieving shrinkage in a time-varying parameter model framework

Angela Bitto, Sylvia Frühwirth-Schnatter [*]

Institute for Statistics and Mathematics, Department of Finance, Accounting and Statistics, WU Vienna University of Economics and Business, Vienna, Austria

ARTICLE INFO

Article history:
Available online 12 November 2018

Keywords:
Bayesian inference
Bayesian Lasso
Double gamma prior
Hierarchical priors
Kalman filter
Log predictive density scores
Normal–gamma prior
Sparsity
State space model

ABSTRACT

Shrinkage for time-varying parameter (TVP) models is investigated within a Bayesian framework, with the aim to automatically reduce time-varying parameters to static ones, if the model is overfitting. This is achieved through placing the double gamma shrinkage prior on the process variances. An efficient Markov chain Monte Carlo scheme is developed, exploiting boosting based on the ancillarity-sufficiency interweaving strategy. The method is applicable both to TVP models for univariate as well as multivariate time series. Applications include a TVP generalized Phillips curve for EU area inflation modeling and a multivariate TVP Cholesky stochastic volatility model for joint modeling of the returns from the DAX-30 index.

1. Introduction

Time-varying parameter (TVP) models are widely used in time series analysis to deal with processes which gradually change over time and provide an interesting alternative to models that allow multiple change points as considered, for instance, in Geweke and Jiang (2011). A variety of interesting econometric applications of TVP models appeared in recent years; for example, Primiceri (2005) used time-varying structural VAR models in a monetary policy application, Dangl and Halling (2012) used TVP models for equity return prediction and Belmonte et al. (2014) used a TVP model to model EU-area inflation.

A huge advantage of TVP models is their flexibility in capturing gradual changes. However, the risk of overfitting increases with a growing number of coefficients, as many of them might in reality be constant over the entire observation period. This will be exemplified in the present paper for a TVP Cholesky stochastic volatility (SV) model (Lopes et al., 2016) for a time series of returns from the DAX-30 index, where out of 406 potentially time-varying coefficients only a small fraction actually changes over time. Allowing static coefficients to be time-varying leads to a considerable loss of statistical efficiency compared to a model, where coefficients are constant apriori.

Identifying fixed coefficients in a TVP model amounts to a *variance selection* problem, involving a decision whether the variances of the shocks driving the dynamics of a time-varying parameter are equal to zero. Variance selection in latent variable models is known to be a non-regular problem within the framework of classical statistical hypothesis testing (Harvey, 1989). The introduction of shrinkage priors for variances within a Bayesian framework has proven to be an attractive alternative both for random effects models (Frühwirth-Schnatter and Tüchler, 2008; Frühwirth-Schnatter and Wagner, 2011) as well as state space models (Frühwirth-Schnatter, 2004; Frühwirth-Schnatter and Wagner, 2010; Nakajima and West, 2013; Belmonte et al., 2014; Kalli and Griffin, 2014). For TVP models, shrinkage priors can automatically reduce time-varying coefficients to static ones, if the model is overfitting.

[*] Corresponding author.
E-mail addresses: angela.bitto@wu.ac.at (A. Bitto), sfruehwi@wu.ac.at (S. Frühwirth-Schnatter).

https://doi.org/10.1016/j.jeconom.2018.11.006
0304-4076/© 2018 Elsevier B.V. All rights reserved.

The literature on variance selection in TVP models is still rather slender, despite this pioneering work, compared to the vast literature on *variable selection* using shrinkage priors to shrink coefficients toward zero in a common regression framework. This class includes mixture priors such as spike-and-slab priors which assign positive probability to zero values (Mitchell and Beauchamp, 1988) and stochastic search variable selection (SSVS) priors (George and McCulloch, 1993) as well as continuous shrinkage priors with a pronounced spike at zero, well-known examples being the Bayesian Lasso prior (Park and Casella, 2008), the normal–gamma prior (Griffin and Brown, 2010; Caron and Doucet, 2008) and the horseshoe prior (Carvalho et al., 2010), among many others; see Fahrmeir et al. (2010) and Polson and Scott (2011) for a review.

One of the main contributions of Frühwirth-Schnatter and Wagner (2010) has been to recast the *variance selection* problem for state space models as a *variable selection* problem in the so-called non-centered parameterization of the state space model. This established the possibility to extend shrinkage priors from standard regression analysis to this more general framework to define a "sparse" state space model. To this aim, Frühwirth-Schnatter and Wagner (2010) employed spike-and-slab priors, whereas Belmonte et al. (2014) relied on the Bayesian Lasso prior for variance selection in TVP models. However, other shrinkage priors might be useful and overcome limitations of these priors, such as computational issues for the spike-and-slab prior and the risk of overshrinking coefficients for the Bayesian Lasso prior.

The present paper makes several contributions in the context of sparse state space models. We develop a new continuous shrinkage prior for process variances by introducing the normal–gamma prior in the non-centered parameterization. This leads to a gamma–gamma (called double gamma) prior for the process variances, which has many attractive properties compared to the popular inverted gamma prior (Petris et al., 2009). We show that the double gamma prior is more flexible than the Bayesian Lasso prior (which is a special case of the double gamma) and yields posterior distributions with a pronounced spike at zero for coefficients which are not time-varying, while at the same time overshrinkage is avoided for time-varying coefficients. A second shrinkage prior allows to shrink static coefficients to coefficients which are not significant over the entire observation period. As a result, we are able to discriminate between time-varying coefficients, coefficients which are significant, but static and insignificant coefficients. We compare different prior settings using log predictive density scores (Geweke and Amisano, 2010) and discuss an accurate approximation of the one-step ahead predictive density.

Based on these priors, we define a very general class of sparse TVP models, both for univariate and multivariate times series, and allow for homoscedastic error variances as well as error variances following a stochastic volatility (SV) model (Jacquier et al., 1994). The later model has proven to be useful in various applications, because neglecting time-varying volatilities might lead to overstating the role of time-varying coefficients in explaining structural changes in the dynamics of macroeconomic variables, as exemplified by Sims (2001) and Nakajima (2011).

Finally, we develop a new Markov chain Monte Carlo (MCMC) scheme for Bayesian inference in sparse TVP models. Using the scale-mixture representation of the normal–gamma prior allows us to implement full conditional Gibbs sampling, thus avoiding Metropolis–Hastings steps which are often used to implement MCMC methods for non-Gaussian state space models, see e.g. Geweke and Tanizaki (1999). To improve MCMC performance, we exploit the ancillarity-sufficiency interweaving strategy of Yu and Meng (2011).

The rest of the paper is structured as follows. Section 2 discusses our novel shrinkage method in the context of sparse TVP models. In Section 3, we present the MCMC scheme. Section 4 discusses evaluation of various priors using log predictive density scores. In Section 5, we extend our method to a multivariate framework. Section 6 presents a simulated data example and Section 7 exemplifies our approach through EU area inflation modeling based on the generalized Phillips curve as well as estimating a time-varying covariance matrix based on a TVP Cholesky SV model for a multivariate time series of returns of the DAX-30 index. Section 8 concludes.

2. Sparse time-varying parameter models

2.1. Bayesian inference for time-varying parameter models

Starting point is the well known state space model, which has been studied in many fields, see e.g. West and Harrison (1997) for a comprehensive review. For the ease of exposition, we consider in this section a univariate time series y_t, observed for T time points $t = 1, \dots, T$, whereas multivariate time series are discussed in Section 5. In a state space model, the distribution of y_t is driven by a latent d-dimensional state vector $\boldsymbol{\beta}_t$ which we are unable to observe. The time-varying parameter (TVP) model is a special case of a state space model and can be regarded as a regression model with time-varying regression coefficients $\boldsymbol{\beta}_t$ following a random walk:

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim \mathcal{N}_d(\mathbf{0}, \mathbf{Q}), \tag{1}$$

$$y_t = \mathbf{x}_t \boldsymbol{\beta}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left(0, \sigma_t^2\right), \tag{2}$$

where $\mathbf{x}_t = (x_{t1}, x_{t2}, \dots, x_{td})$ is a d-dimensional row vector, containing the regressors of the model, one of them being a constant (e.g. $x_{t1} \equiv 1$). To avoid any scaling issues, we assume that all covariates except the intercept are standardized such that for each j the average of x_{tj} over t is equal to zero and the sample variance is equal to 1. The unknown initial value $\boldsymbol{\beta}_0$ is assumed to follow a normal prior distribution,

$$\boldsymbol{\beta}_0 | \boldsymbol{\beta}, \mathbf{Q} \sim \mathcal{N}_d(\boldsymbol{\beta}, \mathbf{P}_0\mathbf{Q}), \tag{3}$$

with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)'$ being unknown fixed regression coefficients and $\mathbf{P}_0 = \text{Diag}\left(P_{0,11}, \ldots, P_{0,dd}\right)$ being a diagonal matrix. Furthermore, $\boldsymbol{\beta}_0$ is independent of the innovations (ε_t) and $(\boldsymbol{\omega}_t)$, which are independent Gaussian white noise processes.

We assume that $\mathbf{Q} = \text{Diag}\,(\theta_1, \ldots, \theta_d)$ is a diagonal matrix, hence each element β_{jt} of $\boldsymbol{\beta}_t = (\beta_{1t}, \ldots, \beta_{dt})'$ follows a random walk for $j = 1, \ldots, d$:

$$\beta_{jt} = \beta_{j,t-1} + \omega_{jt}, \quad \omega_{jt} \sim \mathcal{N}\left(0, \theta_j\right), \tag{4}$$

with initial value $\beta_{j0}|\beta_j, \theta_j, P_{0,jj} \sim \mathcal{N}\left(\beta_j, \theta_j P_{0,jj}\right)$. Hence, θ_j is the process variance governing the dynamics of the time-varying coefficient β_{jt}.[1]

Concerning the error variances in the observation equation (2), we consider the homoscedastic case ($\sigma_t^2 \equiv \sigma^2$ for all $t = 1, \ldots, T$) as well as a more flexible model specification, where σ_t^2 is time-dependent. To capture heteroscedasticity, we use a stochastic volatility (SV) specification as in Jacquier et al. (1994) where $\sigma_t^2 = \mathrm{e}^{h_t}$ and the log volatility h_t follows an AR(1) process:

$$h_t|h_{t-1}, \mu, \phi, \sigma_\eta^2 \sim \mathcal{N}\left(\mu + \phi(h_{t-1} - \mu), \sigma_\eta^2\right). \tag{5}$$

In this setup, the latent volatility process $\mathbf{h} = (h_0, \ldots, h_T)$ is not observed and the initial state h_0 is assumed to follow the stationary distribution of the autoregressive process, i.e. $h_0|\mu, \phi, \sigma_\eta^2 \sim \mathcal{N}\left(\mu, \sigma_\eta^2/(1 - \phi^2)\right)$.

We perform Bayesian inference for the TVP model based on a new family of shrinkage priors for the unknown model parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)'$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)'$ to be introduced in Section 2.2. A shrinkage prior for the process variance θ_j allows to pull the jth time-varying regression coefficient $\{\beta_{j0}, \beta_{j1}, \ldots, \beta_{jT}\}$ toward the fixed regression coefficient β_j, if the model is overfitting and the effect of the jth covariate x_{tj} is, in fact, not changing over time. This requires the definition of priors on the process variances θ_j that are able to shrink θ_j toward the boundary value 0. At the same time, these priors are flexible enough to avoid overshrinking for regression coefficients that are, actually, changing over time t and are characterized by a non-zero process variance $\theta_j \neq 0$.

Concerning the remaining priors, we assume that the scaling factor $P_{0,jj}$ in the initial distribution $\beta_{j0}|\beta_j, \theta_j, P_{0,jj} \sim \mathcal{N}\left(\beta_j, \theta_j P_{0,jj}\right)$ is unknown, following the prior $P_{0,jj} \sim \mathcal{G}^{-1}\left(\nu_P, (\nu_P - 1)c_P\right)$ with hyperparameters $c_P = 1$ and $\nu_P = 20$, implying that no prior moments exist. We employ commonly used priors for the parameters of the error distribution in Eq. (1), namely a hierarchical prior for the homoscedastic case,

$$\sigma^2|C_0 \sim \mathcal{G}^{-1}\left(c_0, C_0\right), \qquad C_0 \sim \mathcal{G}\left(g_0, G_0\right), \tag{6}$$

with hyperparameters c_0, g_0, and G_0. In our practical applications, $c_0 = 2.5, g_0 = 5$, and $G_0 = g_0/\mathrm{E}(\sigma^2)(c_0 - 1)$, with $\mathrm{E}(\sigma^2)$ being a prior guess of σ^2.

In the SV framework (5), unknown parameters are the level μ, the persistence ϕ, and the volatility of volatility σ_η^2. The priors are chosen as in Kastner and Frühwirth-Schnatter (2014), assuming prior independence, i.e. $p(\mu, \phi, \sigma_\eta^2) = p(\mu)p(\phi)p(\sigma_\eta^2)$, with $\mu \sim \mathcal{N}\left(b_\mu, B_\mu\right)$, $(\phi+1)/2 \sim \mathcal{B}\left(a_0, b_0\right)$, and $\sigma_\eta^2 \sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{2B_\sigma}\right)$, with hyperparameters $b_\mu = 0, B_\mu = 100$, $a_0 = 20, b_0 = 1.5$, and $B_\sigma = 1$.

An important building block of our approach is a non-centered parameterization of the TVP model in the vein of Frühwirth-Schnatter and Wagner (2010). First, we define d independent random walk processes $\tilde{\beta}_{jt}, j = 1, \ldots, d$, with standard normal independent increments, i.e.

$$\tilde{\beta}_{jt} = \tilde{\beta}_{j,t-1} + \tilde{\omega}_{jt}, \quad \tilde{\omega}_{jt} \sim \mathcal{N}\left(0, 1\right), \tag{7}$$

and initial value $\tilde{\beta}_{j0}|P_{0,jj} \sim \mathcal{N}\left(0, P_{0,jj}\right)$. Using the transformation

$$\beta_{jt} = \beta_j + \sqrt{\theta_j}\tilde{\beta}_{jt}, \qquad t = 0, \ldots, T, \tag{8}$$

we rewrite the state space model (2) and (4) by combining the d state equations for $\tilde{\beta}_{jt}$ given in (7) with following observation equation:

$$y_t = \mathbf{x}_t\boldsymbol{\beta} + \mathbf{x}_t\text{Diag}(\sqrt{\theta_1}, \ldots, \sqrt{\theta_d})\tilde{\boldsymbol{\beta}}_t + \varepsilon_t. \tag{9}$$

The resulting state space model with state vector $\tilde{\boldsymbol{\beta}}_t = (\tilde{\beta}_{1t}, \ldots, \tilde{\beta}_{dt})'$ is an alternative parameterization of the TVP model, where the observation equation (9) contains all unknown parameters, i.e. the fixed regression coefficients $\beta_1, \ldots, \beta_d$, as well as the (square roots of the) unknown process variances $\theta_1, \ldots, \theta_d$, whereas the state equations (7) are independent of any parameter. Such a parameterization is called non-centered in the spirit of Papaspiliopoulos et al. (2007), whereas the original parameterization (2) and (4) is called centered. Note that the initial state in the non-centered parameterization follows $\tilde{\boldsymbol{\beta}}_0|\mathbf{P}_0 \sim \mathcal{N}_d\left(0, \mathbf{P}_0\right)$ with $\mathbf{P}_0 = \text{Diag}\left(P_{0,11}, \ldots, P_{0,dd}\right)$.

[1] Eisenstat et al. (2014) discuss an extension where the covariance matrix $\mathbf{Q}$ in the state equation (1) is a full matrix instead of a diagonal matrix.

2.2. Shrinking process variances through the double gamma prior

A popular prior choice for the process variance θ_j is the inverted gamma distribution, which is the conjugate prior for θ_j in the centered parameterization (4), see e.g. Petris et al. (2009):

$$\theta_j \sim \mathcal{G}^{-1}\left(s_0, S_0\right). \tag{10}$$

However, as shown by Frühwirth-Schnatter and Wagner (2010), this prior fails to introduce shrinkage as it is bounded away from zero. Frühwirth-Schnatter (2004) introduced a shrinkage prior for the process variance in a univariate TVP model (that is $d = 1$) through the scale parameter in the non-centered parameterization (9) and Frühwirth-Schnatter and Wagner (2010) extended this idea to state space models with $d > 1$. The scale parameter $\sqrt{\theta}_j \in \mathbb{R}$ is defined as the positive and the negative root of θ_j and is allowed to take on positive and negative values. Since the conjugate prior for $\sqrt{\theta}_j$ in the non-centered parameterization (9) is the normal distribution, $\sqrt{\theta}_j$ is assumed to be Gaussian with zero mean and scale parameters ξ_j^2:

$$\sqrt{\theta}_j|\xi_j^2 \sim \mathcal{N}\left(0, \xi_j^2\right) \quad \Leftrightarrow \quad \theta_j|\xi_j^2 \sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{2\xi_j^2}\right). \tag{11}$$

Shrinking θ_j toward the boundary value is achieved by shrinking $\sqrt{\theta}_j$ toward 0 (which is an interior point of the parameter space in the non-centered parameterization). For a sparse state space model, prior (11) substitutes the inverted gamma prior (10) by a gamma prior.[2]

To discriminate between static and time-varying components, Frühwirth-Schnatter and Wagner (2010) introduced spike-and-slab priors, where $\xi_j^2 = 0$ with positive prior probability and ξ_j^2 is fixed, otherwise (e.g. $\xi_j^2 = 10$). Instead of using spike-and-slab priors, Belmonte et al. (2014) extended prior (11) by adding two levels of hierarchy to define a hierarchical Bayesian Lasso prior, where ξ_j^2 follows an exponential distribution.

In the present paper, we introduce a more general family of shrinkage priors derived from the normal–gamma prior, introduced by Griffin and Brown (2010) for variable selection in standard regression models and applied in Caron and Doucet (2008) to multivariate regression models. The main idea is to use the normal–gamma prior as a prior for $\sqrt{\theta}_j$ in the non-centered state space model, extending (11). The normal–gamma prior is a scale mixture of normal distributions with following hierarchical representation:

$$\sqrt{\theta}_j|\xi_j^2 \sim \mathcal{N}\left(0, \xi_j^2\right), \qquad \xi_j^2|a^\xi, \kappa^2 \sim \mathcal{G}\left(a^\xi, a^\xi\kappa^2/2\right). \tag{12}$$

In terms of the process variances θ_j, (12) implies that θ_j follows a "double gamma" prior:

$$\theta_j|\xi_j^2 \sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{2\xi_j^2}\right), \qquad \xi_j^2|a^\xi, \kappa^2 \sim \mathcal{G}\left(a^\xi, a^\xi\kappa^2/2\right). \tag{13}$$

For $a^\xi = 1$, $\xi_j^2|a^\xi, \kappa^2$ reduces to an exponential distribution and the Bayesian Lasso prior considered by Belmonte et al. (2014) results as a special case of the double gamma prior.

Marginalizing over ξ_j^2 yields closed form expressions for $p(\sqrt{\theta}_j|a^\xi, \kappa^2)$ and $p(\theta_j|a^\xi, \kappa^2)$[3]:

$$\begin{aligned} p(\sqrt{\theta}_j|a^\xi, \kappa^2) &= \frac{(\sqrt{a^\xi\kappa^2})^{a^\xi+1/2}}{\sqrt{\pi}2^{a^\xi-1/2}\Gamma(a^\xi)}|\sqrt{\theta}_j|^{a^\xi-1/2}K_{a^\xi-1/2}(\sqrt{a^\xi\kappa^2}|\sqrt{\theta}_j|), \\ p(\theta_j|a^\xi, \kappa^2) &= \frac{(\sqrt{a^\xi\kappa^2})^{a^\xi+1/2}}{\sqrt{\pi}2^{a^\xi-1/2}\Gamma(a^\xi)}(\theta_j)^{a^\xi/2-3/4}K_{a^\xi-1/2}(\sqrt{a^\xi\kappa^2\theta_j}), \end{aligned} \tag{14}$$

where $K_p(\cdot)$ is the modified Bessel function of the second kind with index p. The display of $\log p(\sqrt{\theta}_j|a^\xi, \kappa^2)$ for different values of κ^2 in Fig. 1 shows that the double gamma prior with $a^\xi \leq 1$ is an example of a global–local shrinkage prior (Polson and Scott, 2011). A pronounced spike at zero is present and the mass placed close to zero strongly depends on the global parameter κ^2. From representation (13) we obtain that, marginally, $\mathrm{E}(\theta_j) = 2/\kappa^2$, whereas

$$\mathrm{V}(\theta_j) = \mathrm{E}(\theta_j^2) - \mathrm{E}(\theta_j)^2 = 3\mathrm{E}((\xi_j^2)^2) - \frac{4}{\kappa^4} = \frac{12}{a^\xi\kappa^4} + \frac{8}{\kappa^4} = \mathrm{E}(\theta_j)^2(2 + 3/a^\xi).$$

Hence, independently of a^ξ, the hyperparameter κ^2 controls the global level of shrinkage, which is the stronger, the larger κ^2. At the same time, also $\mathrm{V}(\theta_j)$ decreases, as κ increases. Therefore, the larger κ^2, the more mass is placed close to zero. On the other hand, the term $3/a^\xi$ – which is equal to the excess kurtosis of $\sqrt{\theta}_j$ – controls local adaption to the global level of

[2] We use the parameterization of the $\mathcal{G}(\alpha, \beta)$ distribution with pdf given by $f(y) = \beta^\alpha y^{\alpha-1}e^{-\beta y}/\Gamma(\alpha)$.

[3] Note that $F_{\theta_j}(c) = \Pr(\theta_j \leq c) = \Pr(-\sqrt{c} \leq \sqrt{\theta}_j \leq \sqrt{c}) = 2F_{\sqrt{\theta}_j}(\sqrt{c})$, where $F_{\theta_j}(\cdot)$ is the cdf of the random variable $\sqrt{\theta}_j$. Therefore, $p(\theta_j|a^\xi, \kappa^2) = p(\sqrt{\theta}_j|a^\xi, \kappa^2)/\sqrt{\theta}_j$.

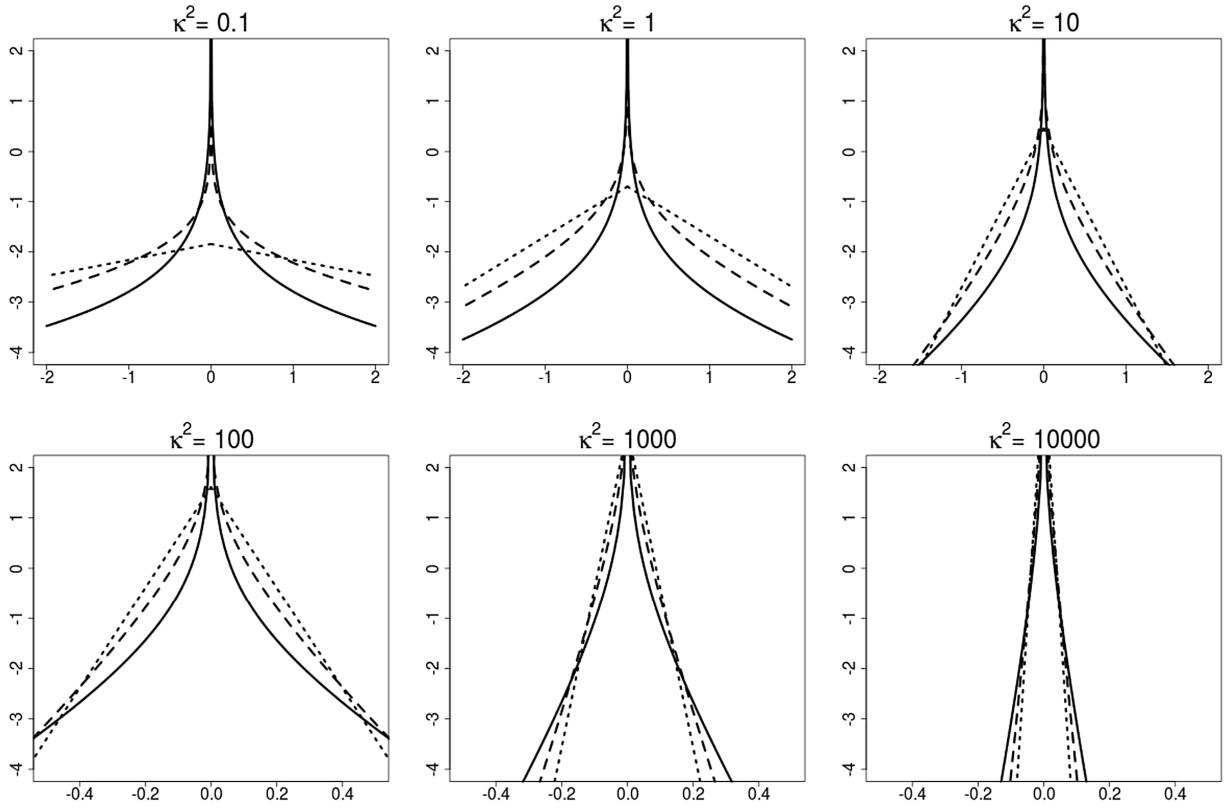

Fig. 1. Log $p(\sqrt{\theta}_j|a^\xi, \kappa^2)$ of the double gamma prior for different values of κ^2 and $a^\xi = 0.1$ (solid line), $a^\xi = 1/3$ (dashed line) and $a^\xi = 1$ (dotted line).

shrinkage, with more local adaption, the smaller a^ξ. As a^ξ decreases, the excess kurtosis of $\sqrt{\theta}_j$ increases and the tails of $p(\sqrt{\theta}_j|a^\xi, \kappa^2)$ become thicker.

It is also illuminating to investigate the joint marginal prior distribution of $(\theta_1, \ldots, \theta_d)$ or (equivalently) of $(\sqrt{\theta}_1, \ldots, \sqrt{\theta}_d)$ given a^ξ and κ^2. Since the random prior variances ξ_j^2 in (13) are drawn independently, also marginally the double gamma prior is characterized by prior conditional independence of $(\theta_1, \ldots, \theta_d)$ given fixed values of a^ξ and κ^2: $p(\theta_1, \ldots, \theta_d|a^\xi, \kappa^2) = \prod_{j=1}^d p(\theta_j|a^\xi, \kappa^2)$.

For illustration, Fig. 2 shows simulations from the joint prior $p(\sqrt{\theta}_1, \sqrt{\theta}_2|a^\xi, \kappa^2)$ for $d = 2$ for various values of a^ξ and κ^2. Not surprisingly from the previous discussions, for the same value of κ^2, the double gamma with $a^\xi = 0.1$ has a pronounced spike at 0 with fat tails in both directions of $\sqrt{\theta}_1$ and $\sqrt{\theta}_2$ and provides more flexible shrinkage compared to the Bayesian Lasso prior ($a^\xi = 1$). For the Bayesian Lasso prior, large values of κ^2 (e.g. $\kappa^2 = 200$) are needed to introduce strong shrinkage toward 0.

To infer appropriate values of a^ξ and κ^2 from the data, hierarchical priors are employed. We assume that κ^2 follows a gamma distribution with fixed hyperparameters d_1 and d_2:

$$\kappa^2 \sim \mathcal{G}(d_1, d_2)\,. \tag{15}$$

For $a^\xi = 1$ this corresponds to the hierarchical Bayesian Lasso prior considered by Belmonte et al. (2014). In addition, we assume that the shrinkage parameter a^ξ follows an exponential distribution as in Griffin and Brown (2010),

$$a^\xi \sim \mathcal{E}(b^\xi), \tag{16}$$

with a fixed hyperparameter $b^\xi \geq 1$. Combining (13) with (15) and (16) defines the hierarchical double gamma prior. Given the hyperparameters d_1, d_2, and b^ξ, this hierarchical prior introduces prior dependence among $(\theta_1, \ldots, \theta_d)$ which is advantageous in a shrinkage framework, as recently shown by Griffin and Brown (2017). Prior dependence is desirable in situations, where only a few variances are expected to be different from 0. In this case, whether a certain process variance is shrunken toward 0 depends on how close the other process variances are to 0.

Prior dependence also exists between β_{j0} and θ_j, as the size (but not the sign) of $\beta_{j0} - \beta_j$ depends on θ_j through $V(\beta_{j0} - \beta_j|\theta_j) = \theta_j P_{0,jj}$. If θ_j is shrunken toward 0, then β_{j0} and all subsequent values β_{jt} are pulled toward β_j for covariate x_{tj}. In high dimensions, where many coefficients are expected to be static, it is of interest to allow a practically constant coefficient

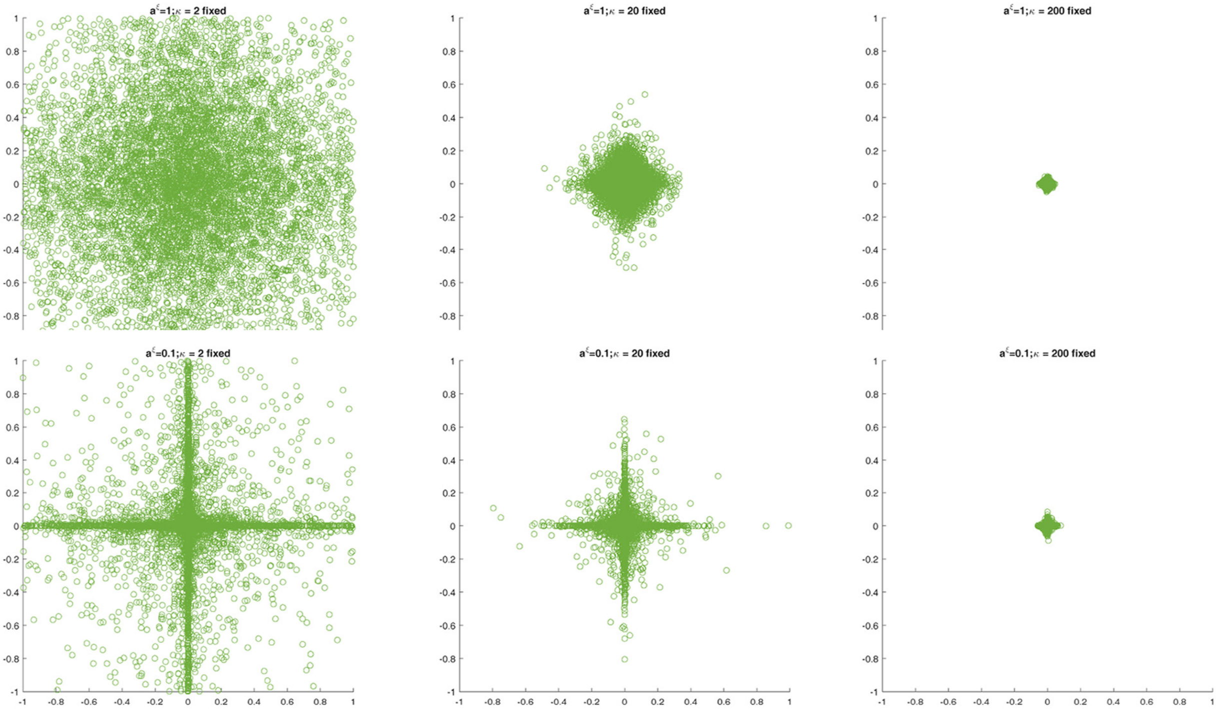

Fig. 2. Simulations from the double gamma prior $p(\sqrt{\theta}_1, \sqrt{\theta}_2|a^\xi, \kappa^2)$ for $a^\xi = 1$ (top) and $a^\xi = 0.1$ (bottom) for different values of κ^2 (left-hand side: $\kappa^2 = 2$, middle: $\kappa^2 = 20$, right-hand side: $\kappa^2 = 200$). The plots at the top correspond to the Bayesian Lasso prior.

β_{jt} to be insignificant throughout the entire observation period. As these coefficients are characterized by a parameter setting where both θ_j and β_j are close to 0, a second normal–gamma prior is employed as a shrinkage prior for β_j to allow shrinkage of β_j toward 0[4]:

$$\beta_j|\tau_j^2 \sim \mathcal{N}\left(0, \tau_j^2\right), \qquad \tau_j^2|a^\tau, \lambda^2 \sim \mathcal{G}\left(a^\tau, a^\tau\lambda^2/2\right). \tag{17}$$

In this case, any (practically constant) coefficient β_{jt} is insignificant, whenever the corresponding fixed regression effect β_j is zero.[5] Similarly as for θ_j, another layer of hierarchy is added, by assuming that $\lambda^2 \sim \mathcal{G}(e_1, e_2)$ and $a^\tau \sim \mathcal{E}(b^\tau)$ with fixed hyperparameters e_1, e_2 and $b^\tau \geq 1$.

3. MCMC estimation

To carry out Bayesian inference for a sparse TVP model under the shrinkage priors introduced in Section 2, we develop an efficient scheme for Markov chain Monte Carlo (MCMC) sampling, given all hyperparameters, i.e. $e_1, e_2, b^\tau, d_1, d_2, b^\xi$ in the priors for $\boldsymbol{\beta}$ and $\mathbf{Q}$, c_P, ν_P in the prior of $P_{0,11}, \ldots, P_{0,dd}$, as well as c_0, g_0, G_0 for homoscedastic variances σ^2 and $b_\mu, B_\mu, a_0, b_0, B_\sigma$ for parameters of the SV model (5). Bayesian inference operates in the latent variable formulation of the TVP model and relies on data augmentation of the latent processes $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T)$ for the centered and $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_T)$ for the non-centered parameterization. For the SV model, the log volatilities $\mathbf{h} = (h_0, \ldots, h_T)$ are introduced as additional latent variables.

For the centered parameterization under the common inverted gamma prior (10) for the process variances θ_j, Gibbs sampling is totally standard, see e.g. Petris et al. (2009). However, if some of the process variances are small, then this MCMC scheme suffers from slow convergence and poor mixing of the sampler. As shown by Frühwirth-Schnatter and Wagner (2010), MCMC estimation based on the non-centered parameterization proves to be useful, in particular if process variances are close to 0.

Frühwirth-Schnatter (2004) discusses the relationship between the various parametrizations for a simple TVP model and the computational efficiency of the resulting MCMC samplers, see also Papaspiliopoulos et al. (2007). For TVP models with

[4] A closed form expression, comparable to (14), is available for $p(\beta_j|a^\tau, \lambda^2)$, with expectation $\mathrm{E}(|\beta_j|) = \sqrt{\frac{4}{\pi a^\tau \lambda^2}} \frac{\Gamma(a^\tau+1/2)}{\Gamma(a^\tau)}$, $\mathrm{V}(\beta_j) = \frac{2}{\lambda^2}$, while the excess kurtosis is given by $\frac{3}{a^\tau}$.

[5] It should be noted that the data are not informative about β_j, if $\theta_j > 0$, but they are always informative about the initial regression coefficient β_{j0}. For $\theta_j = 0$, β_{j0} and β_j coincide.

$d > 1$, MCMC estimation in the centered parameterization is preferable for all coefficients that are actually time-varying, whereas the non-centered parameterization is preferable for (nearly) constant coefficients. For practical time series analysis, both types of coefficients are likely to be present and choosing a computationally efficient parameterization in advance is not possible.

We show how these two data augmentation schemes can be combined through the *ancillarity-sufficiency interweaving strategy* (ASIS) introduced by Yu and Meng (2011) to obtain an efficient sampler combining the "best of both worlds". ASIS provides a principled way of interweaving different data augmentation schemes by re-sampling certain parameters conditional on the latent variables in the alternative parameterization of the model. This strategy has been successfully employed to univariate SV models (Kastner and Frühwirth-Schnatter, 2014), multivariate factor SV models (Kastner et al., 2017) and dynamic linear state space models (Simpson et al., 2017). In the present paper, ASIS is applied to interweave the centered and the non-centered parameterization of a TVP model. More specifically, we use the non-centered parameterization as baseline, and interweave into the centered parameterization. This leads to the MCMC sampling scheme outlined in Algorithm 1 which increases posterior sampling efficiency considerably compared to conventional Gibbs sampling for either of the two parameterizations.

Algorithm 1. Choose starting values for $\boldsymbol{\beta}, \mathbf{Q}, \boldsymbol{\tau} = (\tau_1, \dots, \tau_d), \boldsymbol{\xi} = (\xi_1, \dots, \xi_d), a^\tau, \lambda^2, a^\xi, \kappa^2, \mathbf{P}_0$, and (for homoscedastic variances) σ^2 and C_0 and repeat the following steps:

(a) Sample the states $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_0, \dots, \tilde{\boldsymbol{\beta}}_T)$ in the non-centered parameterization from the multivariate Gaussian posterior $\tilde{\boldsymbol{\beta}}|\boldsymbol{\beta}, \mathbf{Q}, \mathbf{P}_0, \sigma^2 \sim \mathcal{N}_{(T+1)d}\left(\boldsymbol{\Omega}^{-1}\mathbf{c}, \boldsymbol{\Omega}^{-1}\right)$ given in (A.1).

(b) Joint sampling of $\boldsymbol{\alpha} = (\beta_1, \dots, \beta_d, \sqrt{\theta_1}, \dots, \sqrt{\theta_d})'$ from the multivariate Gaussian posterior $p(\boldsymbol{\alpha}|\tilde{\boldsymbol{\beta}}, \boldsymbol{\tau}, \boldsymbol{\xi}, \sigma^2, \mathbf{y})$ given in (A.3).

(c) For each $j = 1, \dots, d$, redraw the constant coefficient β_j and the square root of the process variance $\sqrt{\theta_j}$ through interweaving into the state equation of the centered parameterization:

(c-1) Use the transformation (8) to match the draws of the latent process $\tilde{\beta}_{j0}, \dots, \tilde{\beta}_{jT}$ in the non-centered to the latent process $\beta_{j0}, \dots, \beta_{jT}$ in the centered parameterization and store the sign of $\sqrt{\theta_j}$.

(c-2) Update β_j and θ_j in the centered parameterization by sampling $\theta_j{}^{\text{new}}$ from the generalized inverse Gaussian posterior $\theta_j|\beta_{j0}, \dots, \beta_{jT}, \beta_j, \xi_j^2, P_{0,jj}$, given in (18), and $\beta_j{}^{\text{new}}$ from the Gaussian posterior $\beta_j|\beta_{j0}, \theta_j{}^{\text{new}}, \tau_j^2, P_{0,jj}$, given in (19).

(c-3) Determine $\sqrt{\theta_j^{\text{new}}}$ using the same sign as the old value $\sqrt{\theta_j}$. Based on $\sqrt{\theta_j^{\text{new}}}$ and $\beta_j{}^{\text{new}}$, the state process $\tilde{\beta}_{jt}$ in the non-centered parameterization is updated in a deterministic manner through the inverse of the transformation (8):

$$\tilde{\beta}_{jt}{}^{\text{new}} = (\beta_{jt} - \beta_j^{\text{new}})/\sqrt{\theta_j^{\text{new}}}, \qquad t = 0, \dots, T.$$

(d) Sample from $a^\tau|\beta_1, \dots, \beta_d, \lambda^2$ and $a^\xi|\sqrt{\theta_1}, \dots, \sqrt{\theta_d}, \kappa^2$ using a random walk Metropolis–Hastings (MH) step based on proposing $\log a^{\tau,\text{new}} \sim \mathcal{N}\left(\log a^\tau, c_\tau^2\right)$ and $\log a^{\xi,\text{new}} \sim \mathcal{N}\left(\log a^\xi, c_\xi^2\right)$.

(e) Sample the prior variances $\tau_j|\beta_j, a^\tau, \lambda^2$ and $\xi_j|\theta_j, a^\xi, \kappa^2$, for $j = 1, \dots, d$, from conditionally independent generalized inverse Gaussian distributions given in (A.4) and (A.5), respectively, and update the hyperparameters $\lambda^2|a^\tau, \boldsymbol{\tau}$ and $\kappa^2|a^\xi, \boldsymbol{\xi}$ from the gamma distributions given in (A.6) and (A.7).

(f) Sample $\sigma^2|\tilde{\boldsymbol{\beta}}, \boldsymbol{\alpha}, C_0, \mathbf{y}$ from the following inverted gamma distribution

$$\sigma^2|\tilde{\boldsymbol{\beta}}, \boldsymbol{\alpha}, C_0, \mathbf{y} \sim \mathcal{G}^{-1}\left(c_0 + \frac{T}{2}, C_0 + \frac{1}{2}\sum_{t=1}^{T}(y_t - \mathbf{z}_t\boldsymbol{\alpha})^2\right),$$

where $\mathbf{z}_t$ is defined in (A.2), and sample C_0 from $C_0|\sigma^2 \sim \mathcal{G}\left(g_0 + c_0, G_0 + \frac{1}{\sigma^2}\right)$.

(g) Sample the scale parameters of the initial distribution for each $j = 1, \dots, d$, from $P_{0,jj}|\tilde{\beta}_{j0} \sim \mathcal{G}^{-1}\left(\nu_P + \frac{1}{2}, (\nu_P - 1)c_P + \frac{1}{2}\tilde{\beta}_{j0}^2\right)$.

After discarding a certain amount of initial draws (the *burn-in*), the full conditional sampler iterating Steps (a) to (g) of Algorithm 1 yields draws from the joint posterior distribution $p(\tilde{\boldsymbol{\beta}}, \beta_1, \dots, \beta_d, \sqrt{\theta_1}, \dots, \sqrt{\theta_d}, \boldsymbol{\tau}, \boldsymbol{\xi}, a^\tau, \lambda^2, a^\xi, \kappa^2, \mathbf{P}_0, \sigma^2, C_0, |\mathbf{y})$ under the hierarchical shrinkage priors outlined in Section 2.2.

In Step (a), we sample the latent states $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_0, \dots, \tilde{\boldsymbol{\beta}}_T)$ in the non-centered parameterization conditional on known parameters $\boldsymbol{\beta}, \mathbf{Q}, \mathbf{P}_0$ and known error variances σ^2. As an alternative to the commonly used *Forward Filtering Backward Sampling* (Frühwirth-Schnatter, 1994; Carter and Kohn, 1994), we implemented a multi-move sampling algorithm in the spirit of McCausland et al. (2011) which allows to sample the entire state process $\tilde{\boldsymbol{\beta}}$ *all without a loop* (AWOL; Kastner and Frühwirth-Schnatter (2014)). Full details are provided in Appendix A.1.1.1.

In Step (b), conditional on the latent states $\tilde{\boldsymbol{\beta}}$, a regression type model results from the observation equation (9) of the non-centered state space model. Based on the Gaussian priors appearing in the hierarchical representations of the shrinkage priors (12) and (17), we sample the parameters $\beta_1, \dots, \beta_d$ and $\sqrt{\theta_1}, \dots, \sqrt{\theta_d}$ jointly from the conditionally Gaussian posterior given in (A.3); see Appendix A.1.1.2 for details. One major advantage of working with the square root of the process variance $\sqrt{\theta_j}$, instead of θ_j, is that we avoid boundary space problems for small variances, resulting in better mixing behavior of the sampler.

The interweaving Step (c) turns out to be instrumental for an efficient implementation of the hierarchical shrinkage priors introduced in Section 2.2. In this step, we temporarily move from the non-centered to the centered parameterization to resample β_j and θ_j. To ensure that the posterior distributions obtained with and without interweaving are identical, the priors between the non-centered and the centered parameterization are matched. Whereas the Gaussian prior $\beta_j|\tau_j^2$ for the initial value β_j is the same for both parameterizations, we transform the Gaussian prior for $\sqrt{\theta_j}|\xi_j^2$ to the corresponding gamma prior for $\theta_j|\xi_j^2$ in the centered parameterization, see (11). In Step (c-2), the posteriors of θ_j and β_j in the centered parameterization, conditional on the state process $\beta_{j0}, \dots, \beta_{jT}$, are easily obtained. First, the conditional posterior

$$p(\theta_j|\beta_{j0}, \dots, \beta_{jT}, \beta_j, \xi_j^2, P_{0,jj}) \propto p(\theta_j|\xi_j^2)p(\beta_{j0}|\beta_j, \theta_j, P_{0,jj}) \prod_{t=1}^{T} p(\beta_{jt}|\beta_{j,t-1}, \theta_j),$$

where $\beta_{j0}|\beta_j, \theta_j, P_{0,jj} \sim \mathcal{N}\left(\beta_j, \theta_j P_{0,jj}\right)$ and $\beta_{jt}|\beta_{j,t-1}, \theta_j \sim \mathcal{N}\left(\beta_{j,t-1}, \theta_j\right)$, is the density of a generalized inverse Gaussian distribution (GIG) with following parameters:

$$\theta_j|\beta_{j0}, \dots, \beta_{jT}, \beta_j, \xi_j^2, P_{0,jj} \sim \mathcal{GIG}\left(-\frac{T}{2}, \frac{1}{\xi_j^2}, \sum_{t=1}^{T}(\beta_{jt} - \beta_{j,t-1})^2 + \frac{(\beta_{j0} - \beta_j)^2}{P_{0,jj}}\right). \tag{18}$$

Note that sampling the process variance θ_j from this GIG posterior[6] deviates from the usual MCMC inference for the centered state space model, since the conditionally conjugate inverted gamma prior (10) is substituted by a prior from the gamma distribution. Second, the posterior $p(\beta_j|\beta_{j0}, \theta_j, \tau_j^2, P_{0,jj})$ is a Gaussian distribution, obtained by combining the prior $\beta_j|\tau_j^2 \sim \mathcal{N}\left(0, \tau_j^2\right)$ with the conditional likelihood $\beta_{j0}|\beta_j, \theta_j, P_{0,jj} \sim \mathcal{N}\left(\beta_j, \theta_j P_{0,jj}\right)$:

$$\beta_j|\beta_{j0}, \theta_j, \tau_j^2, P_{0,jj} \propto \mathcal{N}\left(\frac{\beta_{j0}\tau_j^2}{\tau_j^2 + \theta_j P_{0,jj}}, \frac{\tau_j^2 \theta_j P_{0,jj}}{\tau_j^2 + \theta_j P_{0,jj}}\right). \tag{19}$$

Sampling the parameters a^τ and a^ξ in Step (d) is performed without conditioning on $\tau_1, \dots, \tau_d$ and $\xi_1, \dots, \xi_d$. The acceptance probability for $a^{\xi,\text{new}}$ reads:

$$\min\left\{1, \frac{p(a^{\xi,\text{new}})a^{\xi,\text{new}}}{p(a^\xi)a^\xi} \prod_{j=1}^{d} \frac{p(\sqrt{\theta_j}|a^{\xi,\text{new}}, \kappa^2)}{p(\sqrt{\theta_j}|a^\xi, \kappa^2)}\right\},$$

based on the marginal prior (14). A similar acceptance probability holds for $a^{\tau,\text{new}}$.

Sampling the latent prior variances τ_j^2 and ξ_j^2 of the hierarchical shrinkage priors (17) and (12) for β_j and $\sqrt{\theta_j}$ in Step (e) is less standard and we briefly discuss sampling ξ_j^2 (full details are given in Appendix A.1.1.3). The conditionally normal prior $\sqrt{\theta_j}|\xi_j^2$ in (12) leads to a likelihood for ξ_j^2 which is the kernel of an inverted gamma density. In combination with the gamma prior for $\xi_j^2|a^\xi, \kappa^2$, this leads to a posterior distribution arising from a generalized inverse Gaussian (GIG) distribution: $\xi_j^2|\theta_j, a^\xi, \kappa^2 \sim \mathcal{GIG}\left(a^\xi - 1/2, a^\xi\kappa^2, \theta_j\right)$.

Finally, Step (f) has to be modified for the SV model defined in (5). To sample $(h_0, \dots, h_T)$ as well as μ, ϕ, and σ_η^2, we rely on Kastner and Frühwirth-Schnatter (2014) who developed an interweaving strategy for boosting MCMC estimation of SV models.[7]

4. Comparing shrinkage priors through log predictive density scores

Log predictive density scores (LPDS) are an often used scoring rule to compare models; see, e.g., Gneiting and Raftery (2007). Geweke and Keane (2007) introduced LPDS for model comparison of econometric models, see also Geweke and Amisano (2010) for an excellent review of Bayesian predictive analysis. In the present paper, we use log predictive density scores as a means of evaluating and comparing different shrinkage priors.

As common in this framework, the first t_0 time series observations $\mathbf{y}^{\text{tr}} = (y_1, \dots, y_{t_0})$ are used as a "training sample", while evaluation is performed for the remaining time series observations $y_{t_0+1}, \dots, y_T$, based on the log predictive density:

$$\text{LPDS} = \log p(y_{t_0+1}, \dots, y_T|\mathbf{y}^{\text{tr}}) = \sum_{t=t_0+1}^{T} \log p(y_t|\mathbf{y}^{t-1}) = \sum_{t=t_0+1}^{T} \text{LPDS}_t^\star. \tag{20}$$

[6] To sample from the GIG distribution, we use a method proposed by Hörmann and Leydold (2014) which is implemented in the R-package `GIGrvg` (Hörmann and Leydold, 2015). This method is especially reliable for TVP models where the scale parameters of the GIG distribution can be extremely small due to shrinkage and other samplers tend to fail.

[7] This step is easily incorporated into Algorithm 1 using the R-package `stochvol` (Kastner, 2016).

In (20), $p(y_t|\mathbf{y}^{t-1})$ is the one-step ahead predictive density for time t given $\mathbf{y}^{t-1} = (y_1, \ldots, y_{t-1})$ which is evaluated at the observed value y_t. The (individual) log predictive density scores $\mathrm{LPDS}_t^{\star} = \log p(y_t|\mathbf{y}^{t-1})$ provide a tool to analyze performance separately for each observation y_t, whereas LPDS is an aggregated measure of performance for the entire time series.

As shown by Frühwirth-Schnatter (1995) in the context of selecting time-varying and fixed components for a basic structural state space model, LPDS can be interpreted as a log marginal likelihood based on the training sample prior $p(\boldsymbol{\vartheta}|\mathbf{y}^{\mathrm{tr}})$, since

$$p(y_{t_0+1}, \ldots, y_T|\mathbf{y}^{\mathrm{tr}}) = \int p(y_{t_0+1}, \ldots, y_T|\mathbf{y}^{\mathrm{tr}}, \boldsymbol{\vartheta})p(\boldsymbol{\vartheta}|\mathbf{y}^{\mathrm{tr}})d\,\boldsymbol{\vartheta},$$

where $\boldsymbol{\vartheta}$ summarizes the unknown model parameters, e.g. $\boldsymbol{\vartheta} = (\beta_1, \ldots, \beta_d, \sqrt{\theta_1}, \ldots, \sqrt{\theta_d}, \sigma^2)$ for the homoscedastic state space model. This provides a sound and coherent foundation for using the log predictive density score for model – or, in our context, rather prior – comparison.

To approximate the one-step ahead predictive density $p(y_t|\mathbf{y}^{t-1})$, we use Gaussian sum approximations, which are derived from the MCMC draws $(\boldsymbol{\vartheta}^{(m)}, m = 1, \ldots, M)$ from the posterior distribution $p(\boldsymbol{\vartheta}|\mathbf{y}^{t-1})$ given information up to $\mathbf{y}^{t-1}$, i.e:

$$\mathrm{LPDS}_t^* = \log\, p(y_t|\mathbf{y}^{t-1}) = \log \int p(y_t|\mathbf{y}^{t-1}, \boldsymbol{\vartheta})p(\boldsymbol{\vartheta}|\mathbf{y}^{t-1})d\,\boldsymbol{\vartheta} \approx \log \left(\frac{1}{M} \sum_{m=1}^{M} p(y_t|\mathbf{y}^{t-1}, \boldsymbol{\vartheta}^{(m)}) \right), \tag{21}$$

where the one-step ahead predictive density $p(y_t|\mathbf{y}^{t-1}, \boldsymbol{\vartheta})$ is Gaussian conditional on knowing $\boldsymbol{\vartheta}$.

We derive an approximation, called the *conditionally optimal Kalman mixture approximation*, which exploits the fact that the TVP model is a conditionally Gaussian state space model given $\boldsymbol{\vartheta} = (\beta_1, \ldots, \beta_d, \sqrt{\theta_1}, \ldots, \sqrt{\theta_d}, \sigma_t^2)$.[8] For each draw $\boldsymbol{\vartheta}^{(m)} = (\beta_1^{(m)}, \ldots, \beta_d^{(m)}, \sqrt{\theta_1}^{(m)}, \ldots, \sqrt{\theta_d}^{(m)}, \sigma_t^{2(m)})$ from the posterior $p(\boldsymbol{\vartheta}|\mathbf{y}^t)$, we determine the *exact* predictive density $p(y_t|\mathbf{y}^{t-1}, \boldsymbol{\vartheta}^{(m)})$ given by the normal distribution $y_t|\mathbf{y}^{t-1}, \boldsymbol{\vartheta}^{(m)} \sim \mathcal{N}_d\left(\hat{y}_t^{(m)}, S_t^{(m)}\right)$, where $\hat{y}_t^{(m)}$ and $S_t^{(m)}$ are obtained from the prediction step of the Kalman filter (see Appendix A.1.2.1), based on the filtering density $\tilde{\boldsymbol{\beta}}_{t-1}|\mathbf{y}^{t-1}, \boldsymbol{\vartheta}^{(m)} \sim \mathcal{N}_d\left(\mathbf{m}_{t-1}^{(m)}, \mathbf{C}_{t-1}^{(m)}\right)$:

$$\begin{aligned}\hat{y}_t^{(m)} &= \mathbf{x}_t\boldsymbol{\beta}^{(m)} + \mathbf{F}_t^{(m)}\mathbf{m}_{t-1}^{(m)},\\ S_t^{(m)} &= \mathbf{F}_t^{(m)}(\mathbf{C}_{t-1}^{(m)} + \mathbf{I}_d)\mathbf{F}_t^{\prime(m)} + \sigma_t^{2(m)},\end{aligned}$$

where $\mathbf{F}_t^{(m)} = \mathbf{x}_t\mathrm{Diag}\left(\sqrt{\theta_1}^{(m)}, \ldots, \sqrt{\theta_d}^{(m)}\right)$ and $\mathbf{I}_d$ is the $d \times d$ identity matrix. This yields the following Gaussian mixture approximation for $p(y_t|\mathbf{y}^{t-1})$:

$$p(y_t|\mathbf{y}^{t-1}) \approx \frac{1}{M} \sum_{m=1}^{M} f_N\left(y_t; \hat{y}_t^{(m)}, S_t^{(m)}\right). \tag{22}$$

Draws from $p(\boldsymbol{\vartheta}|\mathbf{y}^{t-1})$ are obtained by running the Gibbs sampler outlined in Algorithm 1 for the reduced sample $\mathbf{y}^{t-1} = (y_1, y_2, \ldots, y_{t-1})$. For a homoscedastic error specification, $\sigma_t^{2(m)} \equiv \sigma^{2(m)}$, whereas $\sigma_t^{2(m)}$ is forecasted in the following way for the SV model (5). Given the posterior draw $h_{t-1}^{(m)}$, we simulate $h_t^{(m)}$ from a conditional normal distribution with mean $\mu^{(m)} + \phi^{(m)}(h_{t-1}^{(m)} - \mu^{(m)})$ and variance $\sigma_\eta^{2(m)}$ and define $\sigma_t^{2(m)} = \mathrm{e}^{h_t^{(m)}}$.

5. Extension to multivariate time series

5.1. Sparse TVP models for multivariate time series

The methods introduced in the previous sections are easily extended to TVP models for multivariate time series, such as time-varying parameter VARs, see e.g. Eisenstat et al. (2014) who analyze the response of macro variables to fiscal shocks, and time-varying structural VARs, see e.g. Primiceri (2005) for a monetary policy application. Consider, as illustration, the following TVP model for an r-dimensional time series $\mathbf{y}_t$,

$$\mathbf{y}_t = \mathbf{B}_t\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}_r\left(\mathbf{0}, \boldsymbol{\Sigma}_t\right), \tag{23}$$

where $\mathbf{x}_t$ is a *column* vector of d regressors, and $\mathbf{B}_t$ is a time-varying $(r \times d)$ matrix with coefficient $\beta_{ij,t}$ in row i and column j, potentially containing structural zeros or constant values such that $\beta_{ij,t} \equiv c$ apriori. The (apriori) unconstrained time-varying coefficients $\beta_{ij,t}$ are assumed to follow independent random walks as in the univariate case:

$$\beta_{ij,t} = \beta_{ij,t-1} + \omega_{ij,t}, \quad \omega_{ij,t} \sim \mathcal{N}\left(0, \theta_{ij}\right), \tag{24}$$

[8] Alternative approximations are discussed in Appendix A.1.2.2.

with initial value $\beta_{ij,0} \sim \mathcal{N}\left(\beta_{ij}, \theta_{ij} P_{0,ijj}\right)$, where $P_{0,ijj} \sim \mathcal{G}^{-1}\left(\nu_P, (\nu_P - 1)c_P\right)$ as before. Both the fixed regression coefficients β_{ij} as well as the process variances θ_{ij} are assumed to be unknown.

Each of the apriori unconstrained coefficients $\beta_{ij,t}$ is potentially constant, with the corresponding process variance θ_{ij} being 0. A constant coefficient $\beta_{ij,t} \equiv \beta_{ij}$ is potentially insignificant, in which case $\beta_{ij} = 0$. Hence, shrinkage priors as introduced in Section 2.2 for the univariate case, are imposed on the θ_{ij}s and β_{ij}s to define a sparse TVP model for identifying which of these scenarios holds for each coefficient $\beta_{ij,t}$.

For $i = 1, \ldots, r$, the hierarchical double gamma prior for the process variances θ_{ij} of the coefficients in the ith row of a multivariate TVP model reads:

$$\theta_{ij}|\xi^2_{ij} \sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{2\xi^2_{ij}}\right), \quad \xi^2_{ij}|a^\xi_i, \kappa^2_i \sim \mathcal{G}\left(a^\xi_i, a^\xi_i \kappa^2_i/2\right), \quad \kappa^2_i \sim \mathcal{G}\left(d_1, d_2\right), \quad a^\xi_i \sim \mathcal{E}(b^\xi), \tag{25}$$

with prior expectation ξ^2_{ij} for each process variance θ_{ij}. Similarly, an individual prior variance τ^2_{ij} is introduced for each fixed regression coefficient β_{ij} as in (17):

$$\beta_{ij}|\tau^2_{ij} \sim \mathcal{N}\left(0, \tau^2_{ij}\right), \quad \tau^2_{ij}|a^\tau_i, \lambda^2_i \sim \mathcal{G}\left(a^\tau_i, a^\tau_i \lambda^2_i/2\right), \quad \lambda^2_i \sim \mathcal{G}\left(e_1, e_2\right), \quad a^\tau_i \sim \mathcal{E}(b^\tau). \tag{26}$$

By choosing $a^\tau_i = a^\xi_i = 1$ in (25) and (26), a hierarchical Bayesian Lasso prior for multivariate TVP models results.

We assume row specific hyperparameters a^τ_i, λ^2_i and a^ξ_i, κ^2_i, drawn from common hyperpriors with fixed hyperparameters e_1, e_2, b^τ and d_1, d_2, b^ξ. This leads to prior independence across the r rows of the observation equation (23) and is advantageous for computational reasons, in particular, if the errors $\boldsymbol{\varepsilon}_t$ are uncorrelated, i.e. $\boldsymbol{\Sigma}_t$ is a diagonal matrix. In this case, the multivariate TVP model has a representation as r independent univariate TVP models as in Section 2.1 and MCMC estimation using Algorithm 1 can be performed independently for each of the r rows of the system, e.g. in a parallel computing environment.

If $\boldsymbol{\Sigma}_t$ is a full covariance matrix, then the rows are not independent, because of the correlation among the various components in $\boldsymbol{\varepsilon}_t$. However, as shown by Lopes et al. (2016), a Cholesky decomposition of $\boldsymbol{\Sigma}_t$ leads to such a representation, see also Eisenstat et al. (2014) and Zhao et al. (2016). Further details are provided in the next subsection.

5.2. The sparse TVP Cholesky SV model

Lopes et al. (2016) demonstrate how a multivariate time series $\mathbf{y}_t \sim \mathcal{N}_r\left(\mathbf{0}, \boldsymbol{\Sigma}_t\right)$ with time-varying covariance matrix $\boldsymbol{\Sigma}_t$ can be transformed into a system of r independent equations using the time-varying Cholesky decomposition $\boldsymbol{\Sigma}_t = \mathbf{A}_t\mathbf{D}_t\mathbf{A}'_t$, where $\mathbf{A}_t\mathbf{D}^{1/2}_t$ is the lower triangular Cholesky decomposition of $\boldsymbol{\Sigma}_t$. $\mathbf{A}_t$ is lower triangular with ones on the main diagonal, while $\mathbf{D}_t$ is a time-varying diagonal matrix. It follows that $\mathbf{A}^{-1}_t\mathbf{y}_t \sim \mathcal{N}_r\left(\mathbf{0}, \mathbf{D}_t\right)$. Denoting the elements of $\mathbf{A}^{-1}_t$ as $\Phi_{ij,t}$, for $j < i$, this can be expressed as

$$\begin{pmatrix} 1 & \ldots & & & 0 \\ \Phi_{21,t} & 1 & & & 0 \\ & & \ddots & & 0 \\ \vdots & & & 1 & 0 \\ \Phi_{r1,t} & \Phi_{r2,t} & \ldots & \Phi_{r,r-1,t} & 1 \end{pmatrix} \begin{pmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{rt} \end{pmatrix} \sim \mathcal{N}_r\left(\mathbf{0}, \mathbf{D}_t\right),$$

which can be written as in (23):

$$\mathbf{y}_t \sim \mathcal{N}_r\left(\mathbf{B}_t\mathbf{x}_t, \mathbf{D}_t\right), \tag{27}$$

where $\mathbf{B}_t$ is a $r \times (r-1)$ matrix with elements $\beta_{ij,t} = -\Phi_{ij,t}$, $\mathbf{D}_t$ is a diagonal matrix and the $(r-1)$-dimensional vector $\mathbf{x}_t = (y_{1t}, \ldots, y_{r-1,t})'$ is a regressor derived from $\mathbf{y}_t$. Thus the distribution of $\mathbf{y}_t$ can be represented by a system of r independent TVP models as in Section 5.1, where each time-varying coefficient $\beta_{ij,t}, j < i, i = 1, \ldots, r$, follows a random walk as in (24). Employing the prior (26) for β_{ij} and (25) for θ_{ij} yields the sparse TVP Cholesky SV model.

To capture conditional heteroscedasticity, the matrix $\mathbf{D}_t = \text{Diag}\left(\mathrm{e}^{h_{1t}}, \ldots, \mathrm{e}^{h_{rt}}\right)$ is assumed to be time-varying, where for each row $i = 1, \ldots, r$, the log volatility h_{it} is assumed to follow an individual SV model as in (5), with row specific parameters μ_i, ϕ_i, and $\sigma^2_{\eta,i}$:

$$h_{it}|h_{i,t-1}, \mu_i, \phi_i, \sigma^2_{\eta,i} \sim \mathcal{N}\left(\mu_i + \phi_i(h_{i,t-1} - \mu_i), \sigma^2_{\eta,i}\right).$$

For $r = 3$, for instance, the TVP Cholesky SV model reads:

$$\begin{aligned} y_{1t} &= \varepsilon_{1t}, & \varepsilon_{1t} &\sim \mathcal{N}\left(0, \mathrm{e}^{h_{1t}}\right), \\ y_{2t} &= \beta_{21,t}y_{1t} + \varepsilon_{2t}, & \varepsilon_{2t} &\sim \mathcal{N}\left(0, \mathrm{e}^{h_{2t}}\right), \\ y_{3t} &= \beta_{31,t}y_{1t} + \beta_{32,t}y_{2t} + \varepsilon_{3t}, & \varepsilon_{3t} &\sim \mathcal{N}\left(0, \mathrm{e}^{h_{3t}}\right). \end{aligned}$$

Table 1
Simulated data. Average mean squared error ($avMSE$), average variance ($avVAR$), and average squared bias ($avBIAS^2$) over 100 simulated time series for the hierarchical double gamma prior with $a^\tau \sim \mathcal{E}(10)$ and $a^\xi \sim \mathcal{E}(10)$ and the hierarchical Bayesian Lasso prior with $a^\tau = a^\xi = 1$.

	$a^\tau \sim \mathcal{E}(10), a^\xi \sim \mathcal{E}(10)$			$a^\tau = a^\xi = 1$		
	$avMSE$	$avVAR$	$avBIAS^2$	$avMSE$	$avVAR$	$avBIAS^2$
β_1	3.30E−01	1.67E−01	1.63E−01	3.60E−01	1.57E−01	2.03E−01
β_2	8.18E−03	8.11E−03	6.47E−05	1.56E−02	1.55E−02	1.77E−04
β_3	2.10E−03	2.10E−03	1.36E−06	1.14E−02	1.13E−02	1.31E−04
$\lvert\sqrt{\theta_1}\rvert$	1.81E−03	1.79E−03	2.50E−05	1.61E−03	1.56E−03	5.32E−05
$\lvert\sqrt{\theta_2}\rvert$	1.14E−04	9.33E−05	2.11E−05	5.02E−04	2.47E−04	2.55E−04
$\lvert\sqrt{\theta_3}\rvert$	4.33E−05	3.53E−05	7.97E−06	3.10E−04	1.44E−04	1.66E−04

No intercept is present in these TVP models. For the TVP model in the first row, no regressors are present and only the time-varying volatilities h_{1t} have to be estimated. In the ith equation, $i-1$ regressors are present and $d = i-1$ time-varying regression coefficients $\beta_{ij,t}$ as well as the time-varying volatilities h_{it} need to be estimated. Each of these equations is transformed into a non-centered TVP model and the MCMC scheme in Algorithm 1 is applied to perform Bayesian inference independently for each row i.

6. Illustrative application to simulated data

To illustrate our methodology for simulated data, we generated 100 univariate time series of length $T = 200$ from a TVP model where $d = 3$, $\{x_{1t}\} \equiv 1$, $\{x_{jt}\} \sim \mathcal{N}(0,1)$ for $j = 2,3$, $\sigma^2 = 1$, $(\beta_1, \beta_2, \beta_3) = (1.5, -0.3, 0)$ and $(\theta_1, \theta_2, \theta_3) = (0.02, 0, 0)$. For each time series, β_{1t} is a strongly time-varying coefficient, β_{2t} is a constant, but significant coefficient, and β_{3t} is an insignificant coefficient. As shrinkage priors on β_j and $\sqrt{\theta_j}$, we consider the hierarchical double gamma prior with $a^\tau \sim \mathcal{E}(10)$ and $a^\xi \sim \mathcal{E}(10)$ and the hierarchical Bayesian Lasso prior (that is $a^\tau = a^\xi = 1$) under the hyperparameter setting $d_1 = d_2 = e_1 = e_2 = 0.001$. For each of the 100 simulated time series, MCMC estimation is based on Algorithm 1 by drawing $M = 30{,}000$ samples after a burn-in of length 30,000.[9]

In Fig. 3 we compare the posterior densities for β_j and $\sqrt{\theta_j}$ for one such time series under both shrinkage priors. In general, we want to distinguish three types of coefficients: time-varying, static but significant, and insignificant. One way to achieve a classification is by visual inspection of the posterior distributions of β_j and $\sqrt{\theta_j}$. The posterior density of the scale parameter $\sqrt{\theta_j}$ is symmetric around zero by definition. Thus, if the unknown variance θ_j is different from zero, then the posterior density of $\sqrt{\theta_j}$ is likely to be bimodal. If we find that the posterior density of $\sqrt{\theta_j}$ is unimodal, then the unknown variance is likely to be zero.

While such a bimodal structure of $p(\sqrt{\theta_j}|\mathbf{y})$ is well pronounced for the first coefficient where $\sqrt{\theta}_1 = 0.141$, $p(\sqrt{\theta_j}|\mathbf{y})$ is indeed shrunken toward zero for the two coefficients with zero variances $\theta_2 = \theta_3 = 0$. For the third coefficient, where in addition $\beta_3 = 0$, also the posterior $p(\beta_3|\mathbf{y})$ is shrunken toward zero. Further, we show the posterior paths of β_{jt} in Fig. 4. Evidently, shrinkage priors are able to detect the time-varying coefficient β_{1t}, the constant but significant coefficient β_{2t} and the insignificant coefficient β_{3t}. In both figures, the advantage of the double gamma prior compared to the Bayesian Lasso prior is reflected by increased efficiency in identifying coefficients that are not time-varying.

Table 1 summarizes the average mean squared error ($avMSE$), the average squared bias ($avBIAS^2$) and the average variance ($avVAR$) for the parameters β_1, β_2, β_3, $|\sqrt{\theta}_1|$, $|\sqrt{\theta}_2|$, and $|\sqrt{\theta}_3|$ over the 100 simulated time series.[10] Heavier shrinkage introduced by the hierarchical double gamma prior leads to reduced $avMSE$ compared to the hierarchical Bayesian Lasso prior, in particular for the two coefficients which are not time-varying.

7. Applications in economics and finance

7.1. Modeling EU area inflation

As a first application, we reconsider EU-area inflation data analyzed in Belmonte et al. (2014) and consider the generalized Phillips curve specification, where inflation π_t depends on (typically $p = 12$) lags of inflation and other predictors $\mathbf{z}_t$:

$$\pi_{t+h} = \sum_{j=0}^{p-1} \phi_{jt}\pi_{t-j} + \mathbf{z}_t\boldsymbol{\gamma}_t + \varepsilon_{t+h}, \quad \varepsilon_{t+h} \sim \mathcal{N}\left(0, \sigma^2\right). \tag{28}$$

[9] The Bayesian Lasso prior is combined with the ASIS strategy by fixing $a^\tau = a^\xi = 1$ and skipping Step (d).

[10] Given M draws $\vartheta^{(i1)}, \ldots, \vartheta^{(iM)}$, of a parameter ϑ for each time series i, these measures are defined as $avMSE = avVAR + avBIAS^2$, where $avVAR = \frac{1}{100}\sum_{i=1}^{100} \mathrm{V}_i$ and $avBIAS^2 = \frac{1}{100}\sum_{i=1}^{100}(\mathrm{E}_i - \vartheta^{\mathrm{true}})^2$ with $\mathrm{E}_i = \frac{1}{M}\sum_{m=1}^{M}\vartheta^{(im)}$ and $\mathrm{V}_i = \frac{1}{M}\sum_{m=1}^{M}(\vartheta^{(im)} - \mathrm{E}_i)^2$.

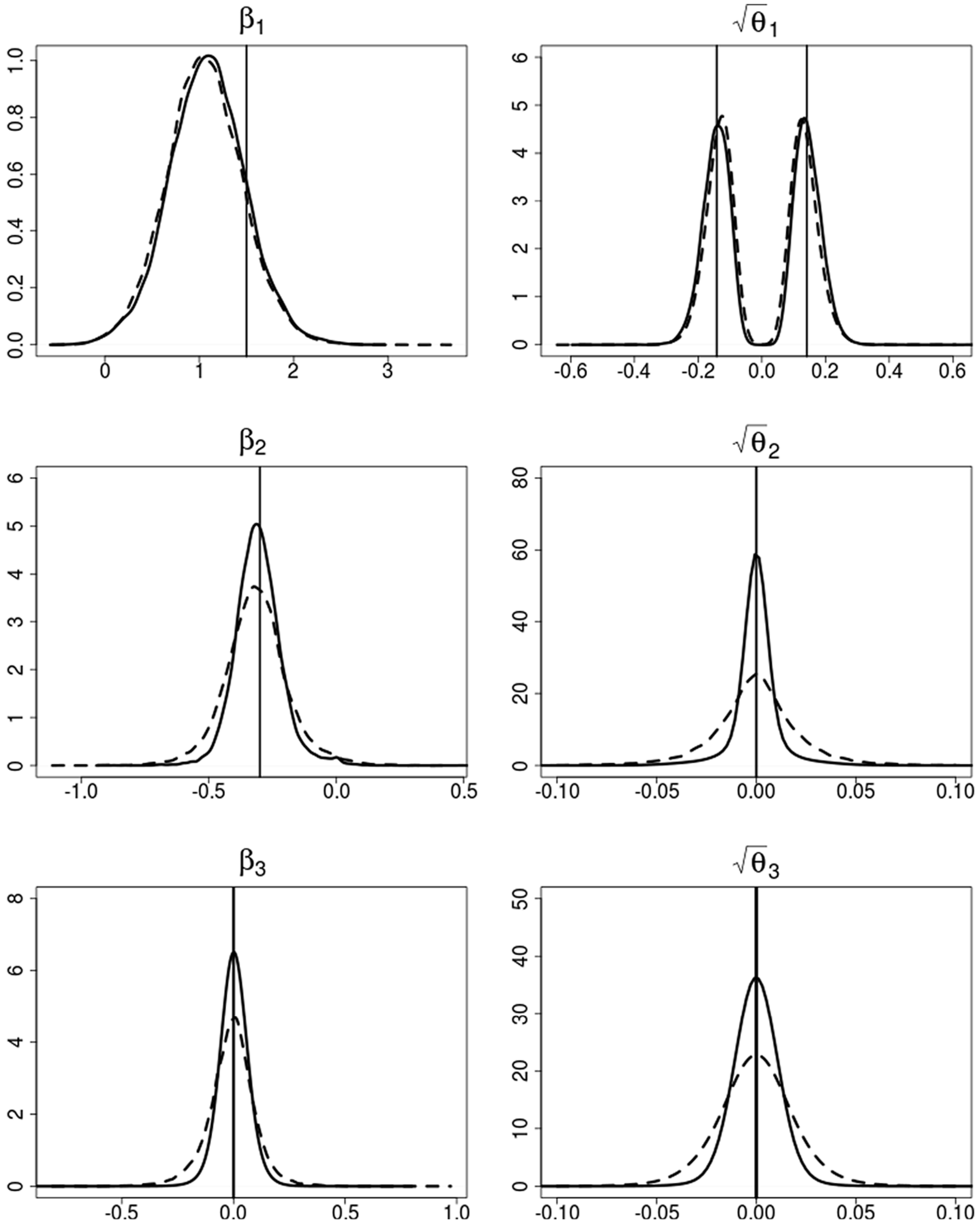

Fig. 3. Simulated data. Posterior densities of β_j (left-hand side) and $\sqrt{\theta_j}$ (right-hand side) together with the true values (indicated by the vertical lines), based on the hierarchical double gamma prior with $a^\tau \sim \mathcal{E}(10)$ and $a^\xi \sim \mathcal{E}(10)$ (solid line) and the hierarchical Bayesian Lasso prior (dashed line).

This set-up has been discussed by Stock and Watson (2012), among others, for forecasting the annual inflation rate, that is $h = 12$. Data are monthly and range from February 1994 until November 2010, i.e. $T = 190$. We list precise definitions of all variables in Appendix A.2.1. As the time series are not seasonally adjusted, we include monthly dummy variables as covariates in (28) to account for seasonal patterns. Thus we are estimating in total $d = 37$ possibly time-varying coefficients, consisting of the intercept, 13 regressors like the *unemployment rate* and the *1-month interest rate*, 12 lagged values of inflation and 11 seasonal dummies.

As shrinkage priors on β_j and $\sqrt{\theta_j}$, we consider the hierarchical double gamma prior with $a^\tau \sim \mathcal{E}(b^\tau)$ and $a^\xi \sim \mathcal{E}(b^\xi)$ under the hyperparameter setting $d_1 = d_2 = e_1 = e_2 = 0.001$ and compare it with the hierarchical Bayesian Lasso prior (that is $a^\tau = a^\xi = 1$) applied by Belmonte et al. (2014). For each prior, MCMC inference is based on Algorithm 1 with $M = 100{,}000$

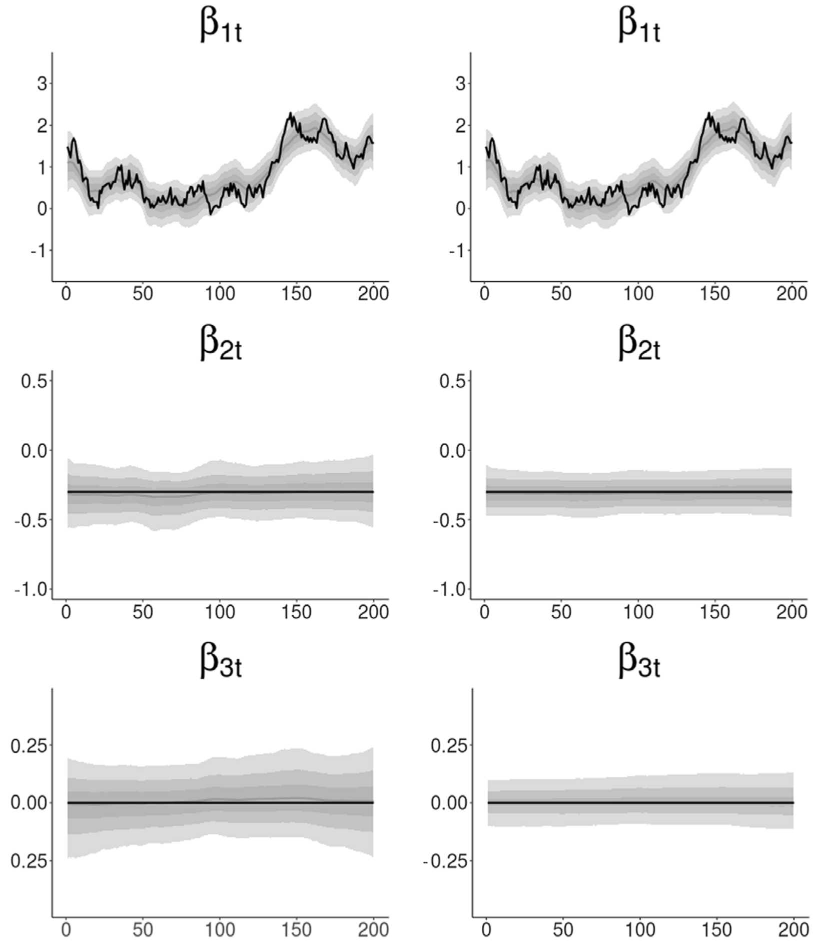

Fig. 4. Simulated data. Pointwise (0.025, 0.25, 0.5, 0.75, 0.975)-quantiles of the posterior paths $\beta_{jt} = \beta_j + \sqrt{\theta_j}\tilde{\beta}_{jt}$ in the centered parameterization in comparison to the true paths (thick black line) for one of the simulated time series; left-hand side: hierarchical Bayesian Lasso prior, right-hand side: hierarchical double gamma prior with $a^\tau \sim \mathcal{E}(10)$ and $a^\xi \sim \mathcal{E}(10)$.

Table 2
ECB data. Posterior summaries for $p(a^\tau|\mathbf{y})$ and $p(a^\xi|\mathbf{y})$ for various values $b^\tau = b^\xi$.

$b^\tau = b^\xi$	$p(a^\tau\|\mathbf{y})$				$p(a^\xi\|\mathbf{y})$			
	1st Qu.	Median	Mean	3rd Qu.	1st Qu.	Median	Mean	3rd Qu.
1	0.128	0.153	0.158	0.181	0.078	0.091	0.094	0.107
2	0.127	0.151	0.158	0.181	0.078	0.092	0.096	0.109
5	0.124	0.147	0.153	0.174	0.076	0.090	0.094	0.107
10	0.122	0.145	0.150	0.172	0.074	0.088	0.090	0.103

draws after a burn-in of the same size. For the hierarchical double gamma prior, we considered various hyperparameters b^τ and b^ξ and the corresponding posterior densities of a^τ and a^ξ are shown in Fig. 5, with posterior summaries being provided in Table 2. The acceptance probability for the random walk MH algorithm in Step (d) of Algorithm 1 lies in the range of 0.24 to 0.26. For these data, the posteriors of a^τ and a^ξ clearly point at the double gamma prior rather than the Bayesian Lasso prior. The choice of the hyperparameters b^τ and b^ξ does not play a significant role and the following results are presented for $b^\tau = b^\xi = 10$.

Summary statistics of $p(\beta_j|\mathbf{y})$ and $p(\sqrt{\theta_j}|\mathbf{y})$ for the hierarchical double gamma prior with $a^\tau \sim \mathcal{E}(10)$ and $a^\xi \sim \mathcal{E}(10)$ are given in Table A.3 in Appendix A.3.1. The easiest combination to spot is the case where both parameters β_j and $\sqrt{\theta_j}$ are shrunken toward zero and the corresponding posterior densities exhibit peaks at zero. This is the case for most of the 37 covariates. The posterior median of $\sqrt{|\theta_j|}$ in Table A.3 is smaller than 10^{-3} for 34 regression coefficients, among them the lagged values of inflation and the monthly dummy variables. In addition, for these coefficients the 95%-confidence regions for β_j obtained from $p(\beta_j|\mathbf{y})$ are reported in Table A.3 and show that none of these variables is “significant”.

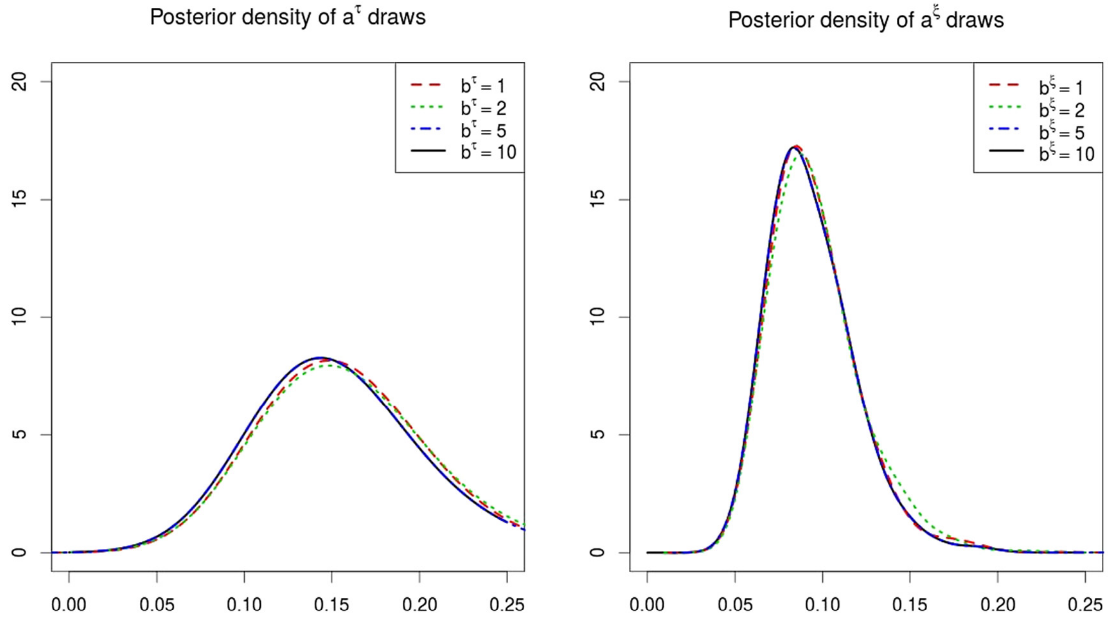

Fig. 5. ECB data. Posterior density of a^τ (left-hand side) and a^ξ (right-hand side).

Table 3
ECB data. Inefficiency factors of MCMC posterior draws of selected parameters, obtained from Algorithm 1 with and without interweaving under the hierarchical double gamma prior with $a^\tau \sim \mathcal{E}(10)$ and $a^\xi \sim \mathcal{E}(10)$ and the hierarchical Bayesian Lasso prior.

j	$a^\tau \sim \mathcal{E}(10)$, $a^\xi \sim \mathcal{E}(10)$				$a^\tau = a^\xi = 1$			
	No ASIS		ASIS		No ASIS		ASIS	
	β_j	$\lvert\sqrt{\theta_j}\rvert$	β_j	$\lvert\sqrt{\theta_j}\rvert$	β_j	$\lvert\sqrt{\theta_j}\rvert$	β_j	$\lvert\sqrt{\theta_j}\rvert$
1	4368	271	86	105	1464	185	72	71
14	535	367	77	231	53	57	28	57
15	116	280	109	245	143	197	45	76
22	1192	328	110	136	64	95	45	56
26	450	672	231	441	266	203	79	100

The four variables in Table A.3 with a posterior mean of $\sqrt{|\theta_j|}$ larger than 10^{-2} are the *1-month interest rate* ($j = 14$), the *1-year interest rate* ($j = 15$), *M3* ($j = 22$), and the *unemployment rate* ($j = 26$). For illustration, we present the corresponding posterior densities of β_j and $\sqrt{\theta_j}$ in Fig. 6 under the hierarchical double gamma prior with $a^\tau \sim \mathcal{E}(10)$ and $a^\xi \sim \mathcal{E}(10)$ and the hierarchical Bayesian Lasso prior with $a^\tau = a^\xi = 1$. The corresponding posterior paths $\beta_{jt} = \beta_j + \sqrt{\theta_j}\tilde{\beta}_{jt}$ are reported in Fig. 7 for the hierarchical double gamma prior. A time-varying behavior is visible for *M3* and the *unemployment rate*. The path of the *1-month interest rate* is significantly different from zero, but the posterior density of $\sqrt{\theta_j}$ exhibits a peak at zero and indicates a constant coefficient. The *1-year interest rate* is basically shrunken toward zero and can be regarded as insignificant.

For this data set, full conditional MCMC sampling turned out to be extremely inefficient and motivated us to include the interweaving step in the Gibbs sampler outlined in Algorithm 1. For illustration, Fig. 8 shows MCMC paths obtained for β_1 with and without interweaving. As illustrated for selected parameters in Table 3, adding the interweaving step leads to substantial improvement of the mixing behavior of MCMC sampling, with considerably reduced inefficiency factors.[11]

Finally, as discussed in Section 4, we use log predictive density scores (LPDS) to evaluate the various shrinkage priors. Fig. 9 shows cumulative LPDS over the last 100 time points, using the conditionally optimal Kalman mixture approximation derived in Section 4.[12] Evidently, for this time series, the hierarchical double gamma prior is clearly preferable to the hierarchical Bayesian Lasso prior applied by Belmonte et al. (2014).

[11] Inefficiency factors were computed using the function `effectiveSize` from the R package `coda` (Plummer et al., 2006).

[12] For numerical reasons, it is essential to use the conditionally optimal Kalman mixture approximation rather than the naive approximation to approximate the predictive density, see Appendix A.3.2 for details. See Frühwirth-Schnatter (1992) for an earlier discussion of that problem.

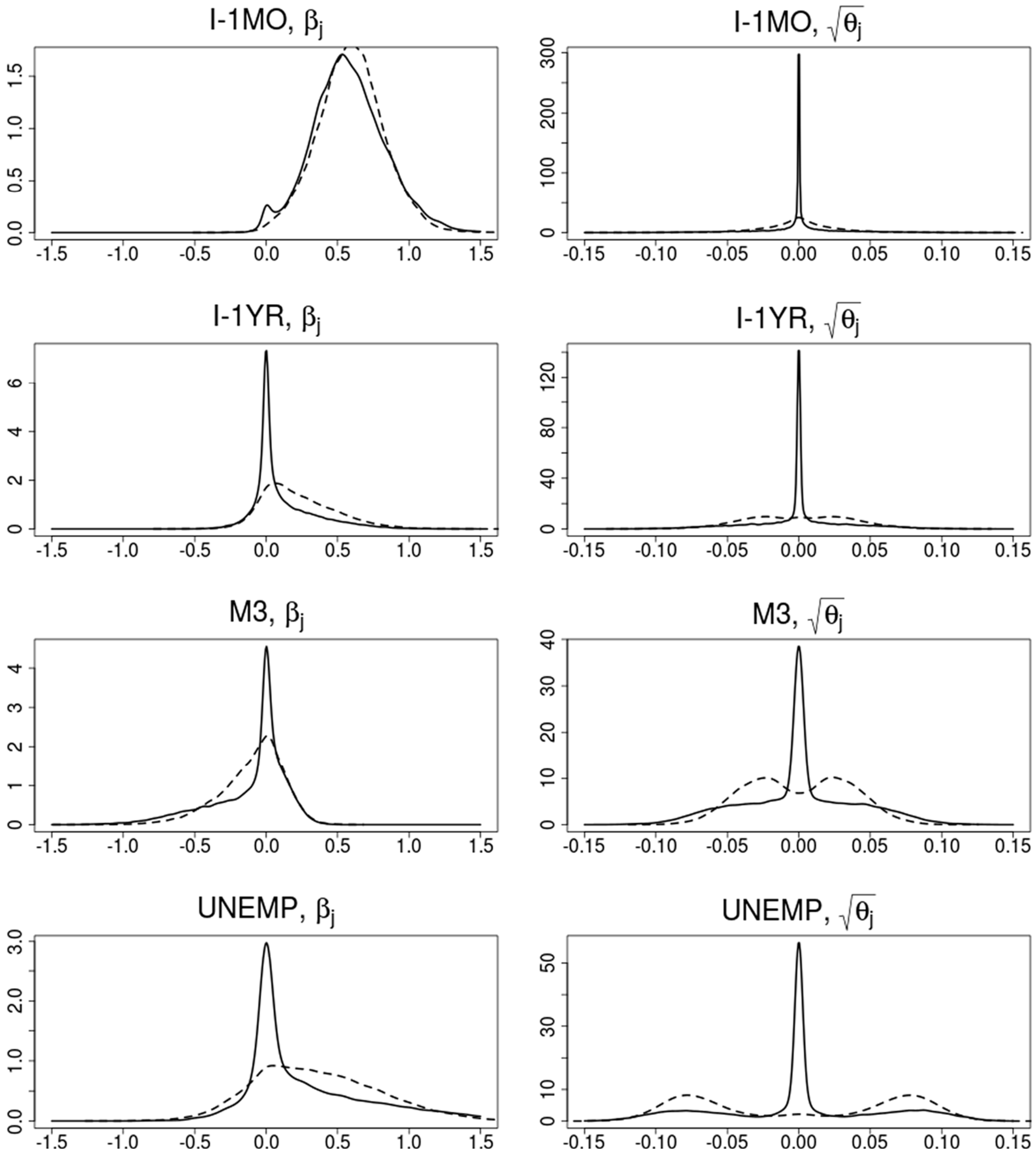

Fig. 6. ECB data. Posterior densities of β_j (left-hand side) and $\sqrt{\theta_j}$ (right-hand side), based on the hierarchical double gamma prior with $a^\tau \sim \mathcal{E}(10)$ and $a^\xi \sim \mathcal{E}(10)$ (solid line) and the hierarchical Bayesian Lasso prior (dashed line) for following predictors (from top to bottom): *1-month interest rate*, *1-year interest rate*, *M3*, and *unemployment rate*.

7.2. Sparse TVP Cholesky SV modeling of DAX returns

As a second real world data application, we fit the sparse TVP Cholesky SV model introduced in Section 5.2 to 29 indices from the German Stock Index DAX, see Appendix A.2.2 for more details on the data. The ordering of the indices is alphabetical and our data set spans roughly 2500 daily stock returns from September 4th, 2001 until August 31st, 2011.[13]

Due to the nature of the TVP Cholesky SV model, a representation of the multivariate model in terms of 29 independent equations exists. Exploiting representation (27), we estimate a pure stochastic volatility model for the first index and 28

[13] As any model based on a Cholesky decomposition, inference is not invariant with respect to reordering the indices. While inference for the elements of $\mathbf{\Sigma}_t$ was fairly robust, we observed sensitivity to the ordering of the data for functionals of $\mathbf{\Sigma}_t^{-1}$, e.g. the time-varying global minimum variance portfolio weights derived from $\mathbf{\Sigma}_t^{-1}$.

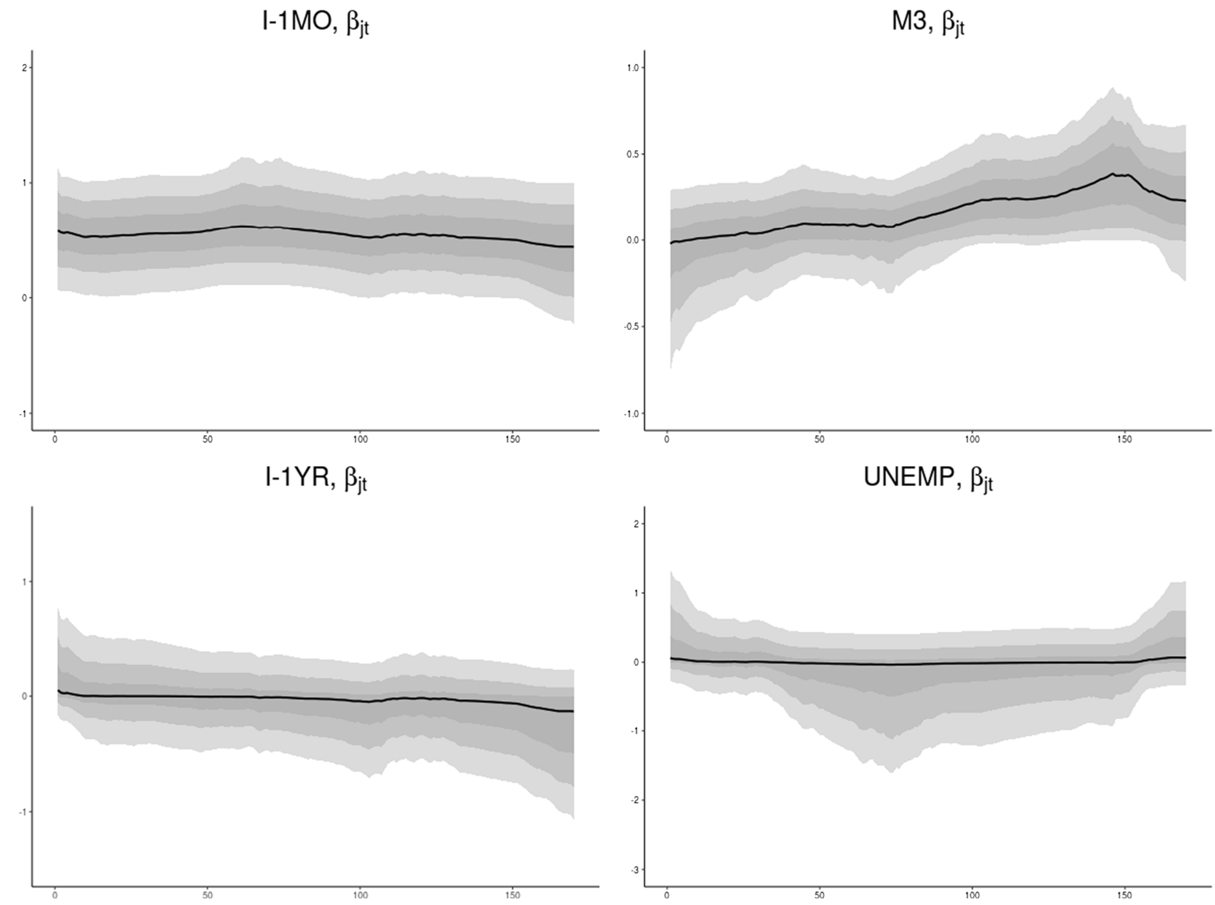

Fig. 7. ECB data. Pointwise (0.025, 0.25, 0.5, 0.75, 0.975)-quantiles of the posterior paths of $\beta_{jt} = \beta_j + \sqrt{\theta_j}\tilde{\beta}_{jt}$, based on the hierarchical double gamma prior with $a^\tau \sim \mathcal{E}(10)$ and $a^\xi \sim \mathcal{E}(10)$; left-hand side: *1-month interest rate* (top) and *1-year interest rate* (bottom); right-hand side: *M3* (top) and *unemployment rate* (bottom).

TVP models with SV error specification for the remaining indices, with the dimension d increasing from 1 to 28. To estimate the resulting 406 potentially time-varying coefficients $\beta_{ij,t}$ in an efficient manner, we apply the hierarchical double gamma priors introduced in (25) and (26) with $a_i^\tau \sim \mathcal{E}(10)$ and $a_i^\xi \sim \mathcal{E}(10)$ as well as the hierarchical Bayesian Lasso prior (that is $a_i^\tau = a_i^\xi = 1$) under the hyperparameter setting $d_1 = d_2 = e_1 = e_2 = 0.001$. In addition to these shrinkage priors, we apply the usual conditionally conjugate prior, i.e. $\theta_{ij} \sim \mathcal{IG}(s_0, S_0)$ for all process variances θ_{ij} and $\beta_{ij} \sim \mathcal{N}(0, A_0)$ for all fixed regression coefficients β_{ij}, with prior setting as in Petris et al. (2009), namely $s_0 = 0.1$, $S_0 = 0.001$ and $A_0 = 10$.

For all TVP models and all priors, MCMC inference is performed using Algorithm 1 with $M = 50{,}000$ draws after a burn-in of 50,000.[14] The acceptance probability for the MH algorithm in Step (d) lies in the range of 0.24 to 0.26. Posterior densities of a_i^τ and a_i^ξ under the hierarchical double gamma prior are provided in Fig. 10. The posterior medians of a_i^τ are in the range of 0.11 to 0.37, whereas the posterior medians of a_i^ξ lie between 0.07 and 0.15.

Exemplarily, detailed results are presented for the tenth time-varying regression in Fig. 11, where we compare the posterior densities of β_{ij} and $\sqrt{\theta}_{ij}$, for $i = 10$ and $j = 1, \dots, 9$, obtained under the different priors. As expected, under the inverted gamma prior all posteriors distributions of $\sqrt{\theta}_{ij}$ are bounded away from 0, with the position of the symmetric posterior modes (roughly ±0.015) being more or less the same for all coefficients.[15] As opposed to this, both shrinkage priors allow the posterior distribution of $\sqrt{\theta}_{ij}$ to concentrate at 0, if appropriate, and in this way allows to distinguish between coefficients that are time-varying ($j = 1, 2, 7$) and the remaining coefficients which turn out to be static. When comparing

[14] MCMC estimation under the inverted gamma prior requires a minor modification of Algorithm 1. We sample θ_{ij} only in the centered parameterization from the conditional posterior $\theta_{ij}|\boldsymbol{\beta} \sim \mathcal{IG}\left(s_0 + \frac{T+1}{2}, S_0 + \frac{1}{2}\sum_{t=1}^{T}(\beta_{ij,t} - \beta_{ij,t-1})^2 + \frac{(\beta_{ij,0}-\beta_{ij})^2}{2P_{0,ijj}}\right)$.

[15] The location of the posterior distribution is mainly driven by the prior – for the alternative hyperparameters $s_0 = 0.5$ and $S_0 = 0.2275$ (not shown in the figure) the posterior modes shift to around ±0.1.

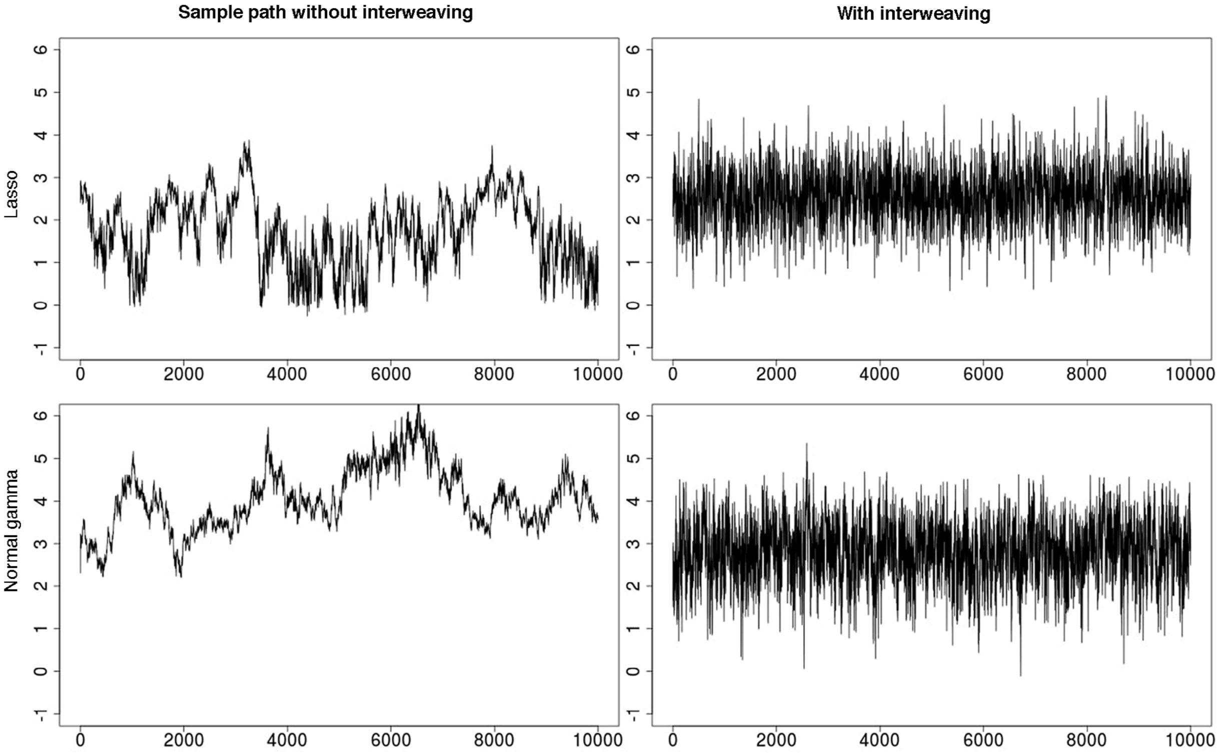

Fig. 8. ECB data. Sample paths of β_1 comparing the MCMC schemes without interweaving (left-hand side) and with interweaving (right-hand side) for $a^\tau = a^\xi = 1$ (top row) and $a^\tau \sim \mathcal{E}(10)$, $a^\xi \sim \mathcal{E}(10)$ (bottom row). $M = 100{,}000$ draws, only every tenth draw is shown.

both shrinkage priors, the influence of the increased shrinkage introduced by the double gamma prior is evident for static coefficients, with the posterior of $\sqrt{\theta}_{ij}$ showing a much more pronounced spike at 0 than the Bayesian Lasso prior.

For the static coefficients, the posterior distributions of β_{ij} indicate that some coefficients are significant, in particular when $j = 3$ and $j = 9$, whereas others are clearly insignificant, e.g. when $j = 6$. These findings are confirmed by the corresponding posterior paths of $\beta_{ij,t} = \beta_{ij} + \sqrt{\theta_{ij}}\tilde{\beta}_{ij,t}$ displayed in Fig. 12 under the hierarchical double gamma prior and the inverted gamma prior. For the double gamma prior, the coefficients $\beta_{i1,t}$, $\beta_{i2,t}$, and $\beta_{i7,t}$ are the only ones that are time-varying, whereas $\beta_{i3,t}$ and $\beta_{i9,t}$ are constant, but shifted away from 0. Fig. 12 also demonstrates a dramatic gain in statistical efficiency, in terms of dispersion of the posterior distribution of $\beta_{ij,t}$ for each point in time, compared to the inverted gamma prior. This holds in particular for coefficients which are static, but significant such as $\beta_{i3,t}$ and $\beta_{i9,t}$. In addition, the estimated paths are much smoother under the double gamma prior, which facilitates the interpretation of the time-varying components $\beta_{i1,t}$, $\beta_{i2,t}$, and $\beta_{i7,t}$. The coefficient $\beta_{i2,t}$, for instance, shows a trending behavior, which is not apparent under the inverted gamma prior.

Similar impact of our shrinkage method can be observed for the remaining 27 equations in the TVP model. Overall, we investigated all 406 posterior paths $\beta_{ij,t}$, together with the corresponding posterior distributions of β_{ij} and $\sqrt{\theta}_{ij}$, and found that a large fraction of these coefficients is not significant. For illustration, we display in Fig. 13 one (out of 2500) heat maps of the posterior median of the 29×28 Cholesky factor matrix $\mathbf{B}_t$ at $t = 1150$. Whereas the majority of the estimated coefficients $\hat{\beta}_{ij,t}$ is different from zero for the inverted gamma prior, only a small part is significantly different from zero for the double gamma prior.

Finally, we compare the various priors using LPDS for the last 500 returns, with the first 400 observations serving as training sample. Very conveniently, the triangular structure of the model allows to decompose the 29-dimensional predictive density as $p(\mathbf{y}_t|\mathbf{y}^{t-1}) = \prod_{i=1}^{r} p(y_{i,t}|\mathbf{y}^{t-1})$. Hence, the overall log predictive density score $\text{LPDS}^\star_t$ at time t results as the sum of the individual log predictive density scores $\text{LPDS}^\star_{i,t} = \log p(y_{i,t}|\mathbf{y}^{t-1})$, derived independently for each of the $r = 29$ TVP models:

$$\text{LPDS}^\star_t = \log p(\mathbf{y}_t|\mathbf{y}^{t-1}) = \sum_{i=1}^{r} \log p(y_{i,t}|\mathbf{y}^{t-1}) = \sum_{i=1}^{r} \text{LPDS}^\star_{i,t}. \tag{29}$$

The individual log predictive density scores $\text{LPDS}^\star_{i,t}$ are approximated using the conditionally optimal Kalman mixture approximation introduced in Section 4 and the cumulative log predictive scores are shown in Fig. 14 for the various priors. We find overwhelming evidence in favor of using shrinkage priors instead of the popular inverted gamma prior. For the later,

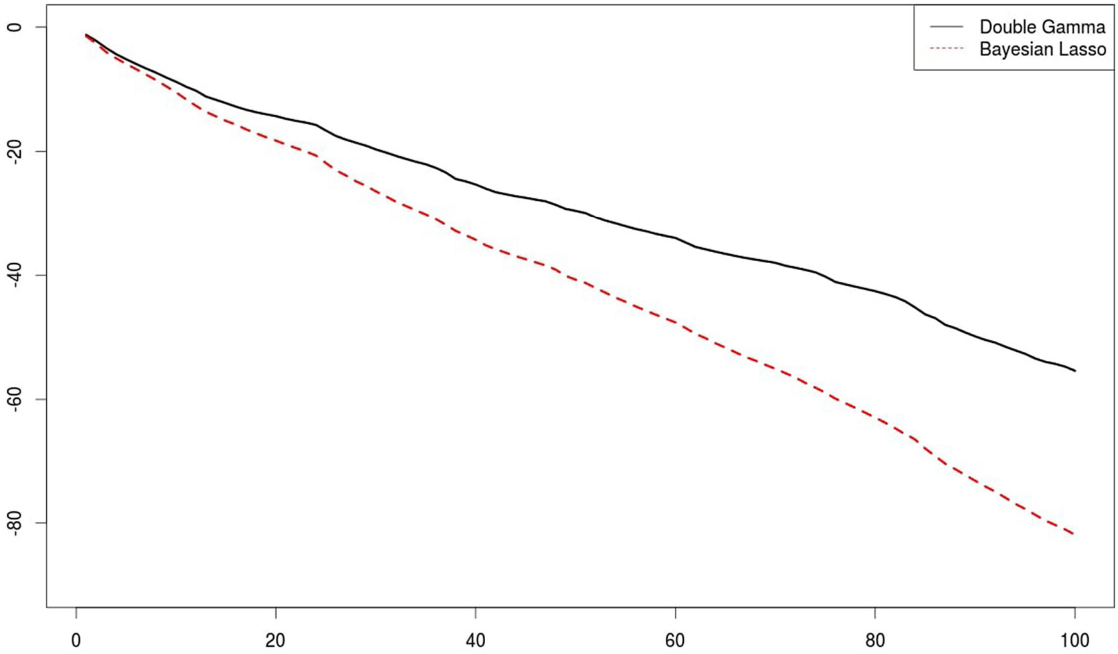

Fig. 9. ECB data. Cumulative log predictive scores for the last 100 time point (labeled with time index $t - t_0$, where $t_0 = 90$) under the hierarchical double gamma prior with $a^\tau \sim \mathcal{E}(10)$ and $a^\xi \sim \mathcal{E}(10)$ (solid line) and the hierarchical Bayesian Lasso prior (dashed line).

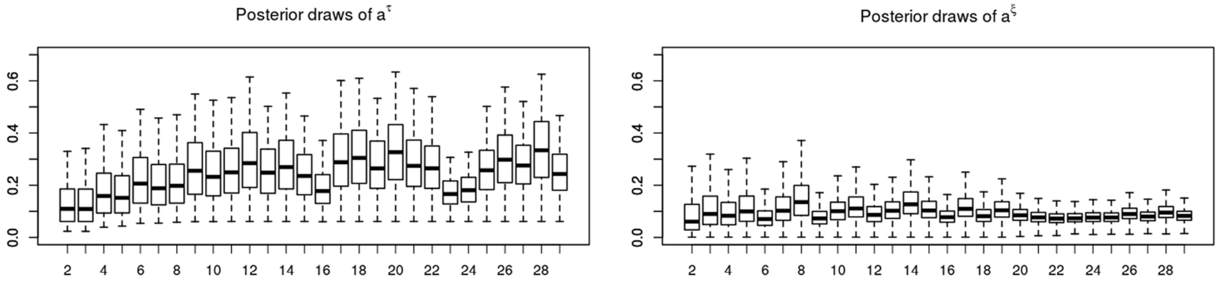

Fig. 10. DAX data. Posterior densities of a_i^τ (left-hand side) and a_i^ξ (right-hand side) under a hierarchical double gamma prior with $a_i^\tau \sim \mathcal{E}(10)$ and $a_i^\xi \sim \mathcal{E}(10)$, represented by box plots for $i = 2, \ldots, 29$. Whiskers correspond to the 0.05 and the 0.95 quantile.

the choice of the hyperparameters (see $\theta_{ij} \sim \mathcal{IG}(0.1, 0.001)$ versus $\theta_{ij} \sim \mathcal{IG}(0.5, 0.2275)$) exercises tremendous influence on the log predictive density scores. Fig. 14 also compares the hierarchical Bayesian Lasso prior with the hierarchical double gamma prior with fixed values $a^\tau = a^\xi = 0.1$. Although the posteriors of a^τ and a^ξ in Fig. 10 clearly are bounded away from the values $a^\tau = a^\xi = 1$ corresponding to the Bayesian Lasso prior, the log predictive density scores are very similar for both shrinkage priors.[16] Evidently, the major predictive gain comes from substituting the popular inverted gamma prior for the process variances by a sensible shrinkage prior that allows posterior concentration of the process variances at zero (see again Fig. 11). As long as these priors behave sensibly at zero, the data contain little information to discriminate between them due to the small signal-to-noise ratio inherent in financial time series.

8. Conclusion

In the present paper, shrinkage for time-varying parameter (TVP) models was investigated within a Bayesian framework both for univariate and multivariate time series, with the aim to automatically reduce time-varying parameters to static ones, if the model is overfitting. This goal was achieved by formulating shrinkage priors for the process variances based on the normal–gamma prior (Griffin and Brown, 2010), extending previous work using spike-and-slab priors (Frühwirth-Schnatter and Wagner, 2010) and the Bayesian Lasso prior (Belmonte et al., 2014). As a major computational contribution,

[16] The posteriors in Fig. 10 are based on the entire time series, but similar figures result for the last 500 observations.

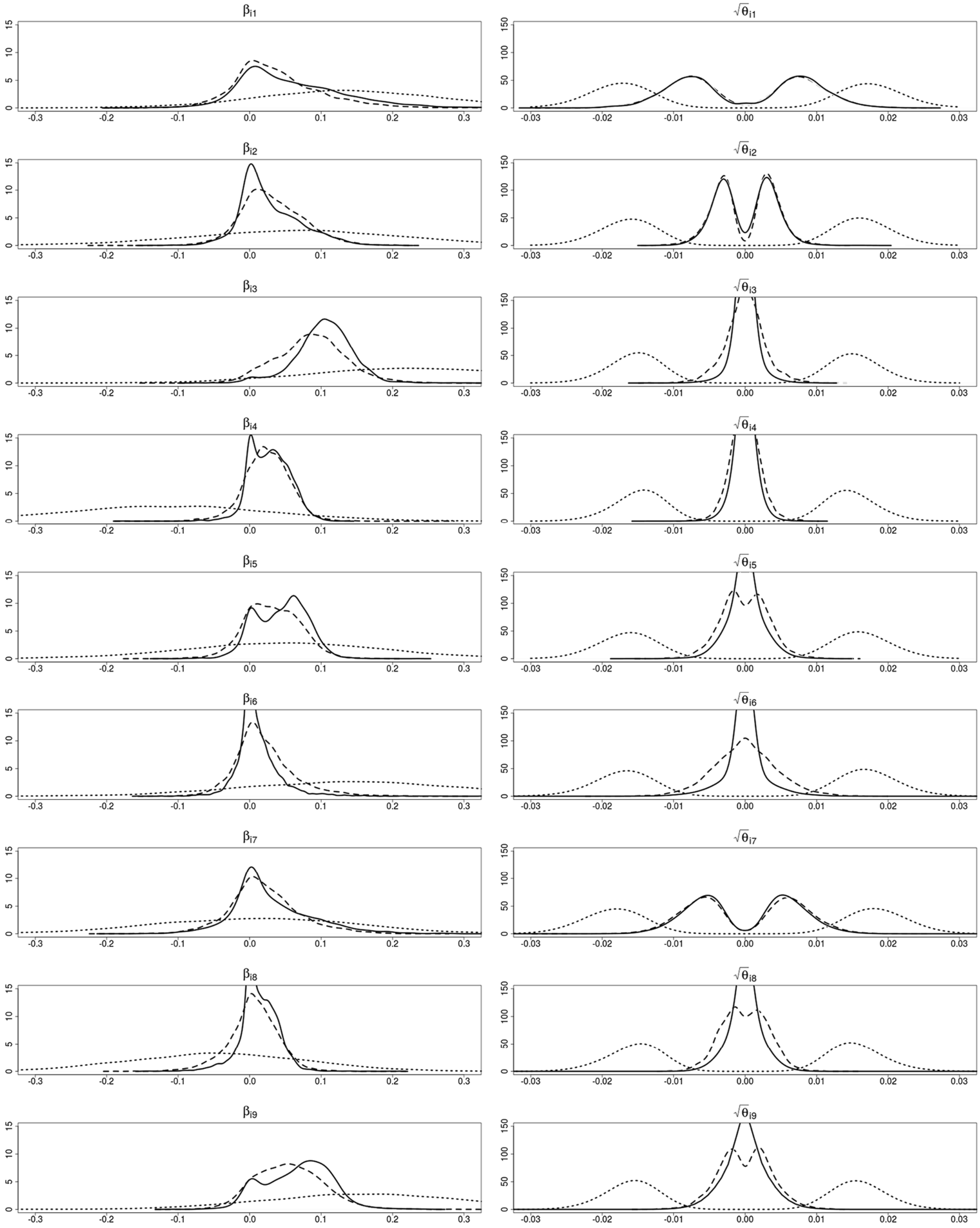

Fig. 11. DAX data. Posterior densities of β_{ij} (left-hand side) and $\sqrt{\theta_{ij}}$ (right-hand side) for $i = 10$ and $j = 1, \dots, 9$ (from top to bottom), derived under the conditionally conjugate prior $\beta_{ij} \sim \mathcal{N}(0, 10)$ and $\theta_{ij} \sim \mathcal{IG}(0.1, 0.001)$ (dotted line), the hierarchical Bayesian Lasso prior with $a_i^\tau = a_i^\xi = 1$ (dashed line) and a hierarchical double gamma prior with $a_i^\tau \sim \mathcal{E}(10)$ and $a_i^\xi \sim \mathcal{E}(10)$ (solid line).

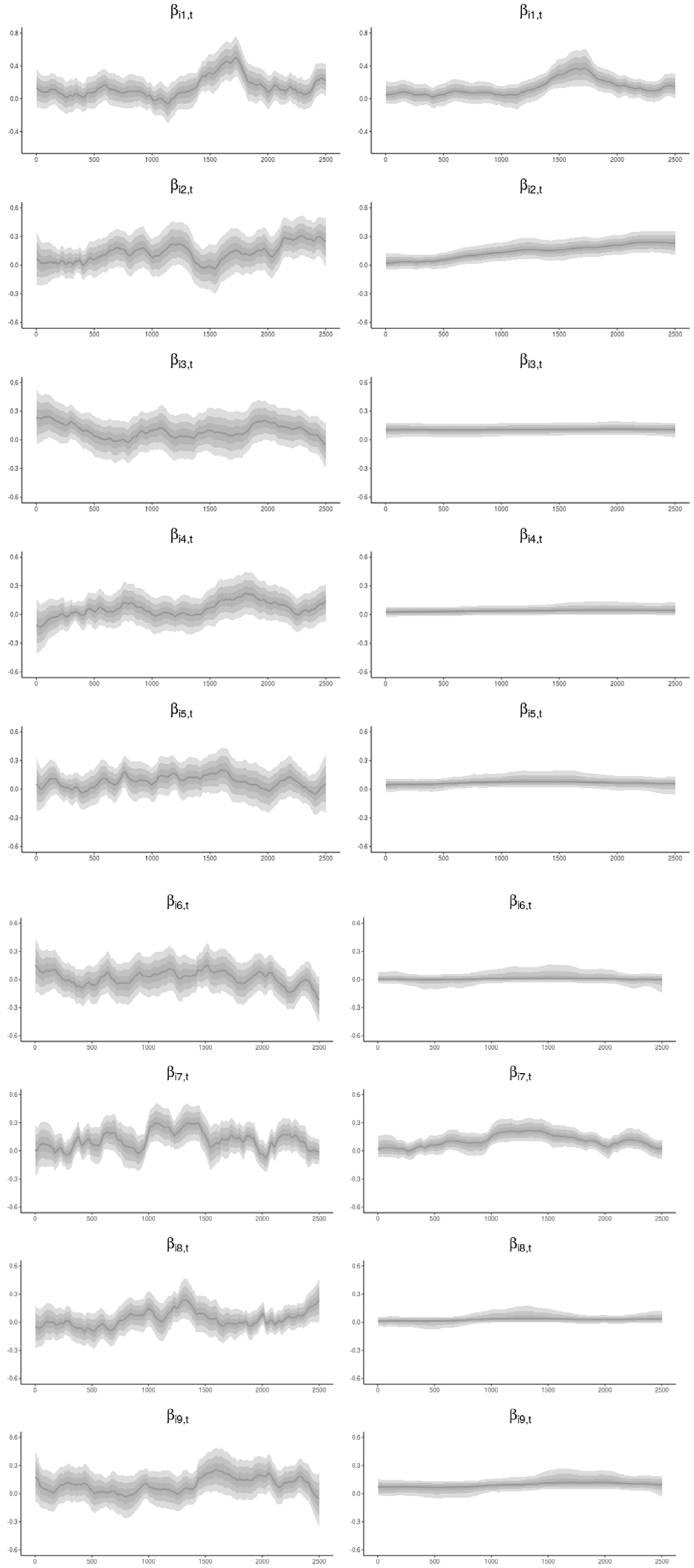

Fig. 12. DAX data. Pointwise (0.025, 0.25, 0.5, 0.75, 0.975)-quantiles of the posterior paths $\beta_{ij,t} = \beta_{ij} + \sqrt{\theta_{ij}}\tilde{\beta}_{ij,t}$ for $i = 10$ and $j = 1, \ldots, 9$ (from top to bottom); derived under the conditionally conjugate prior $\beta_{ij} \sim \mathcal{N}(0, 10)$ and $\theta_{ij} \sim \mathcal{IG}(0.1, 0.001)$ (left-hand side) and a hierarchical double gamma prior with $a^{\tau} \sim \mathcal{E}(10)$ and $a^{\xi} \sim \mathcal{E}(10)$ (right-hand side).

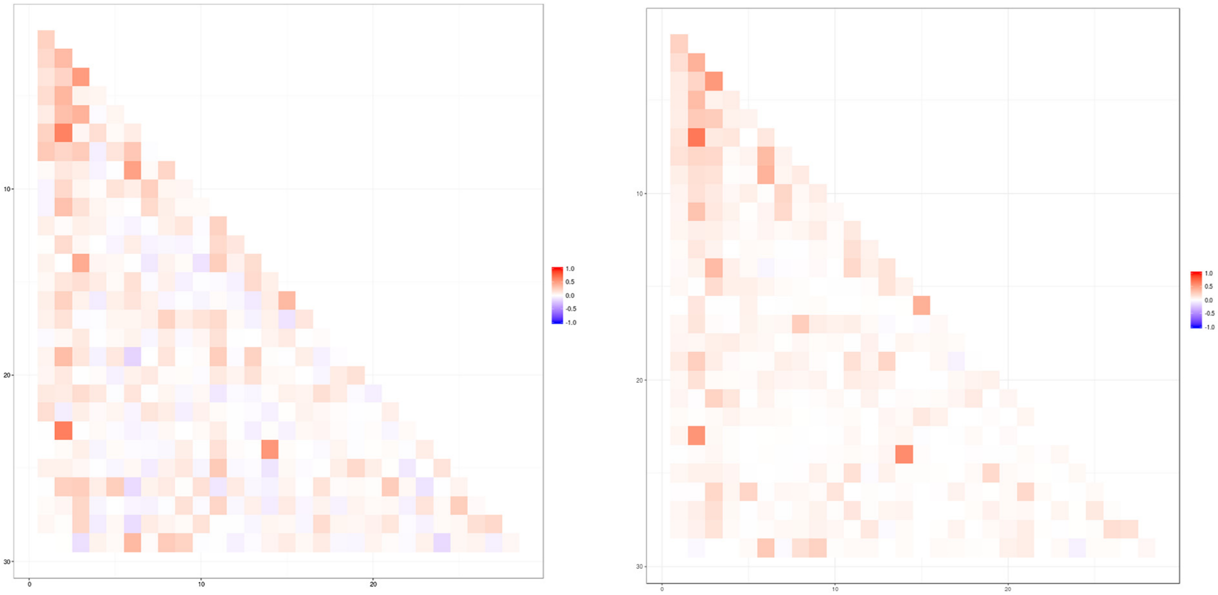

Fig. 13. DAX data. Heat plot of the posterior median of the 29 × 28 Cholesky factor matrix $\mathbf{B}_t$ at $t = 1150$, derived under the conditionally conjugate prior $\beta_{ij} \sim \mathcal{N}(0, 10)$ and $\theta_{ij} \sim \mathcal{IG}(0.1, 0.001)$ (left-hand side) and a hierarchical double gamma prior with $a^{\tau} \sim \mathcal{E}(10)$ and $a^{\xi} \sim \mathcal{E}(10)$ (right-hand side). Values shrunken to zero are white.

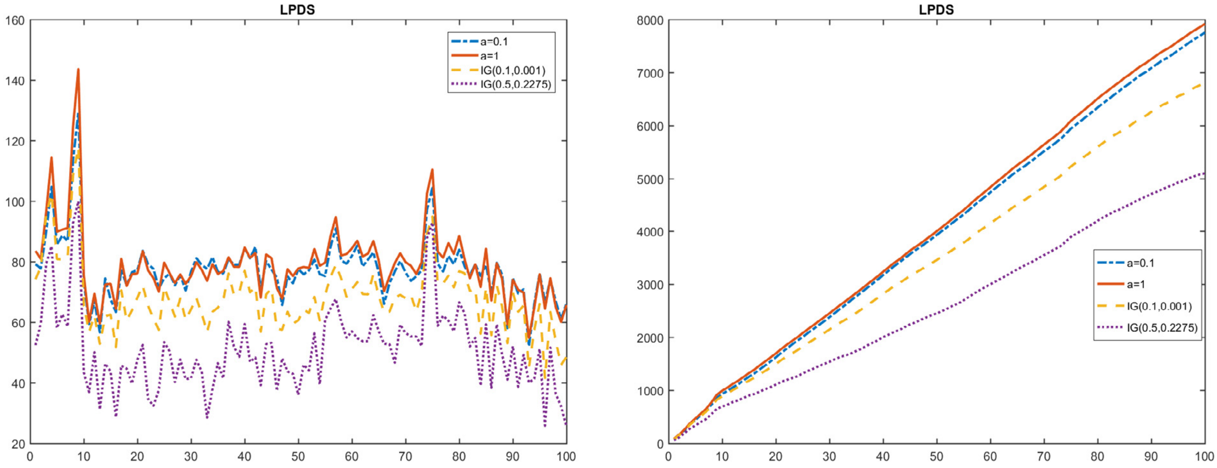

Fig. 14. DAX data. Individual (left-hand side) and cumulative (right-hand side) log predictive density scores for the last 100 time points using the last 400 observations as training sample. Shrinkage prior with $a^{\tau} = a^{\xi} = 1$ (full line) and $a^{\tau} = a^{\xi} = 0.1$ (dash-dotted line) in comparison to the inverted gamma priors $\theta_{ij} \sim \mathcal{IG}(0.1, 0.001)$ (dashed line) and $\theta_{ij} \sim \mathcal{IG}(0.5, 0.2275)$ (dotted line).

an efficient MCMC estimation scheme was developed, exploiting the ancillarity-sufficiency interweaving strategy of Yu and Meng (2011).

Our applications included EU area inflation modeling based on a TVP generalized Phillips curve and estimating a time-varying covariance matrix based on a sparse TVP Cholesky SV model for a multivariate time series of returns of the DAX-30 index. We investigated different prior settings, including the popular inverted gamma prior for the process variances, using log predictive density scores. Overall, our findings suggest that the family of double gamma priors introduced in this paper for sparse TVP models is successful in avoiding overfitting, if coefficients are, indeed, static or even insignificant. The framework developed in this paper is very general and holds the promise to be useful for introducing sparsity in other TVP and state space models in many different settings. In particular, sparse time-varying parameter VAR models result by straightforward extensions of the methods discussed in this paper.

The underlying strategy of using the non-centered parameterization of a state space model to extend shrinkage priors introduced for variable selection in regression models to variance selection in a state space model is very generic and many alternative shrinkage priors for variance selection seem worth to be investigated. As pointed out by a reviewer, extending the normal–gamma–gamma prior, introduced recently for highly structured regression models (Griffin and Brown, 2017), to variance selection is a particularly promising venue for future research. This strategy leads to the following "triple gamma prior" in the context of variance selection for state space models for univariate time series:

$$\theta_j|\xi_j^2 \sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{2\xi_j^2}\right), \quad \xi_j^2|a^\xi, \kappa_j^2 \sim \mathcal{G}\left(a^\xi, \kappa_j^2/2\right), \quad \kappa_j^2 \sim \mathcal{G}\left(c^\xi, d^\xi\right),$$

with three hyperparameters a^ξ, c^ξ, and d^ξ. The special case where $a^\xi = c^\xi = 1/2$ is of particular interest, as it extends the horseshoe prior (Carvalho et al., 2010) to variance selection for state space models which is very popular in regression analysis for its outstanding properties, see e.g. Bhadra et al. (2017).

Acknowledgments

We are extremely grateful to the Guest Editors Sylvia Kaufmann and Herman van Dijk and two anonymous referees for numerous valuable comments and suggestions that improved the paper considerably. We acknowledge inspiring comments from participants of ESOBE Meetings (Vienna 2012, Oslo 2013, Gerzensee 2015) and NBER-NSF Time Series Conferences (St. Louis 2014, Vienna 2015, New York 2016). We owe special thanks to Mauro Bernardi, discussant of the paper at the 2016 Bolzano Workshop on Forecasting in Finance and Macroeconomics. We are grateful for comments on preliminary versions by participants of CFE 2011, the 1st EFaB Workshop (2013), the 7th Rimini Bayesian Econometrics Workshop (2013), the 1st Vienna Workshop on High Dimensional Time Series in Macroeconomics and Finance (2013), and the 12th ISBA World Meeting (2014).

Appendix A. Achieving shrinkage in a time-varying parameter model framework webappendix

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2018.11.006.

References

Belmonte, M.A.G., Koop, G., Korobolis, D., 2014. Hierarchical shrinkage in time-varying parameter models. J. Forecast. 33, 80–94.
Bhadra, A., Datta, J., Polson, N., Willard, B., 2017. Lasso meets horsheshoe. https://arxiv.org/abs/1706.10179.
Caron, F., Doucet, A., 2008. Sparse Bayesian nonparametric regression. In: McCallum, A., Roweis, S. (Eds.), Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008). Omnipress, pp. 88–95.
Carter, C.K., Kohn, R., 1994. On Gibbs sampling for state space models. Biometrika 81, 541–553.
Carvalho, C.M., Polson, N.G., Scott, J.G., 2010. The horseshoe estimator for sparse signals. Biometrika 97, 465–480.
Dangl, T., Halling, M., 2012. Predictive regressions with time-varying coefficients. J. Financ. Econ. 106, 157–181.
Eisenstat, E., Chan, J.C., Strachan, R.W., 2014. Stochastic model specification search for time-varying parameter VARs. SSRN Electron. J. 1.
Fahrmeir, L., Kneib, T., Konrath, S., 2010. Bayesian regularisation in structured additive regression: A unifying perspective on shrinkage, smoothing and predictor selection. Stat. Comput. 20, 203–219.
Frühwirth-Schnatter, S., 1992. Approximate predictive integrals for dynamic generalized linear models. In: Fahrmeir, L., Francis, B., Gilchrist, R., Tutz, G. (Eds.), Advances in GLIM and Statistical Modelling. In: Lecture Notes in Statistics, (vol. 78), Springer, New York, pp. 123–151.
Frühwirth-Schnatter, S., 1994. Data augmentation and dynamic linear models. J. Time Series Anal. 15, 183–202.
Frühwirth-Schnatter, S., 1995. Bayesian model discrimination and Bayes factors for linear Gaussian state space models. J. Roy. Statist. Soc. Ser. B 57, 237–246.
Frühwirth-Schnatter, S., 2004. Efficient Bayesian parameter estimation. In: Harvey, A., Koopman, S.J., Shephard, N. (Eds.), State Space and Unobserved Component Models: Theory and Applications. Cambridge University Press, Cambridge, pp. 123–151.
Frühwirth-Schnatter, S., Tüchler, R., 2008. Bayesian parsimonious covariance estimation for hierarchical linear mixed models. Stat. Comput. 18, 1–13.
Frühwirth-Schnatter, S., Wagner, H., 2010. Stochastic model specification search for Gaussian and partially Non-Gaussian state space models. J. Econometrics 154, 85–100.
Frühwirth-Schnatter, S., Wagner, H., 2011. Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data. In: Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., West, M. (Eds.), Bayesian Statistics 9. Oxford University Press, pp. 165–200.
George, E.I., McCulloch, R., 1993. Variable selection via Gibbs sampling. J. Amer. Statist. Assoc. 88, 881–889.
Geweke, J., Amisano, G., 2010. Comparing and evaluating Bayesian predictive distributions of asset returns. Int. J. Forecast. 26, 216–230.
Geweke, J., Jiang, Y., 2011. Inference and prediction in a multiple-structural-break model. J. Econometrics 163, 172–185.
Geweke, J., Keane, M., 2007. Smoothly mixing regressions. J. Econometrics 138, 252–291.
Geweke, J., Tanizaki, H., 1999. On Markov chain Monte Carlo methods for nonlinear and non-Gaussian state space models. Commun. Statist. Part B – Simul. Comput. 28, 867–894.
Gneiting, T., Raftery, A., 2007. Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. 102, 359–378.
Griffin, J.E., Brown, P.J., 2010. Inference with normal-gamma prior distributions in regression problems. Bayesian Anal. 5, 171–188.
Griffin, J.E., Brown, P.J., 2017. Hierarchical shrinkage priors for regression models. Bayesian Anal. 12, 135–159.
Harvey, A.C., 1989. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, Cambridge.
Hörmann, W., Leydold, J., 2014. Generating generalized inverse Gaussian random variates. Stat. Comput. 24, 547–557.
Hörmann, W., Leydold, J., 2015. GIGrvg: Random Variate Generator for the GIG Distribution. R package version 0.4. URL: http://CRAN.R-project.org/package=GIGrvg.
Jacquier, E., Polson, N.G., Rossi, P.E., 1994. Bayesian analysis of stochastic volatility models. J. Bus. Econ. Statist. 12, 371–417.
Kalli, M., Griffin, J.E., 2014. Time-varying sparsity in dynamic regression models. J. Econometrics 178 (2), 779–793.
Kastner, G., 2016. Dealing with stochastic volatility in time series using the R package stochvol. J. Stat. Softw. 69, 1–30.
Kastner, G., Frühwirth-Schnatter, S., 2014. Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models. Comput. Statist. Data Anal. 76, 408–423.
Kastner, G., Frühwirth-Schnatter, S., Lopes, H.F., 2017. Efficient Bayesian inference for multivariate factor stochastic volatility models. J. Comput. Graph. Statist. 26, 905–917.

Lopes, H.F., McCulloch, R.E., Tsay, R.S., 2016. Parsimony inducing priors for large scale state-space models. Bayesian Anal..

McCausland, W.J., Miller, S., Pelletier, D., 2011. Simulation smoothing for state space models: A computational efficiency analysis. Comput. Statist. Data Anal. 55, 199–212.

Mitchell, T.J., Beauchamp, J.J., 1988. Bayesian variable selection in linear regression. J. Amer. Statist. Assoc. 83, 1023–1036.

Nakajima, J., 2011. Time-varying parameter VAR model with stochastic volatility: An overview of methodology and empirical applications. Monetary Econ. Stud. 29, 107–142.

Nakajima, J., West, M., 2013. Bayesian analysis of latent threshold dynamic models. J. Bus. Econom. Statist. 31, 151–164.

Papaspiliopoulos, O., Roberts, G., Sköld, M., 2007. A general framework for the parameterization of hierarchical models. Statist. Sci. 22, 59–73.

Park, T., Casella, G., 2008. The Bayesian Lasso. J. Amer. Statist. Assoc. 103, 681–686.

Petris, G., Petrone, S., Campagnoli, P., 2009. Dynamic Linear Models with R. Springer, New York.

Plummer, M., Best, N., Cowles, K., Vines, K., 2006. CODA: Convergence diagnosis and output analysis for MCMC. R News 6 (1), 7–11.

Polson, N.G., Scott, J.G., 2011. Shrink globally, act locally: Sparse Bayesian regularization and prediction. In: Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., West, M. (Eds.), Bayesian Statistics 9. Oxford University Press, pp. 501–538.

Primiceri, G., 2005. Time varying structural vector autoregressions and monetary policy. Rev. Econom. Stud. 72, 821–852.

Simpson, M., Niemi, J., Roy, V., 2017. Interweaving Markov chain Monte Carlo strategies for efficient estimation of dynamic linear models. J. Comput. Graph. Statist. 26, 152–159.

Sims, C.A., 2001. [Evolving Post-World war II US inflation Dynamics]: Comment. NBER Macroecon. Annu. 16, 373–379.

Stock, J.H., Watson, M.W., 2012. Generalized shrinkage methods for forecasting using many predictors. J. Bus. Econom. Statist. 30 (4), 481–493.

West, M., Harrison, P.J., 1997. Bayesian Forecasting and Dynamic Models, second ed. Springer, New York.

Yu, Y., Meng, X.L., 2011. To center or not to center: that is not the question - an ancillarity-suffiency interweaving strategy (ASIS) for boosting MCMC efficiency. J. Comput. Graph. Statist. 20 (3), 531–570.

Zhao, Z.Y., Xie, M., West, M., 2016. Dynamic dependence networks: Financial time series forecasting and portfolio decisions. Appl. Stoch. Models Bus. Ind. 32, 311–332.