

# Dynamic sparsity on dynamic regression models

Hedibert Lopes<sup>a</sup>  and Paloma Uribe<sup>b</sup>

Inspere Institute of Education and Research, São Paulo, Brazil, <sup>a</sup>[hedibertFL@insper.edu.br](mailto:hedibertFL@insper.edu.br),  
<sup>b</sup>[PalomaVU@insper.edu.br](mailto:PalomaVU@insper.edu.br)

**Abstract.** In the present work, we consider variable selection and shrinkage for the Gaussian dynamic linear regression within a Bayesian framework. In particular, we propose a novel method that allows for time-varying sparsity, based on an extension of spike-and-slab priors for dynamic models. This is done by assigning appropriate Markov switching priors for the time-varying coefficients' variances, extending the previous work of (*The Annals of Statistics* **33** (2005) 730–773). Furthermore, we investigate different priors, including the common Inverted gamma prior for the process variances, and other mixture prior distributions such as Gamma priors for both the spike and the slab, which leads to a mixture of Normal-Gamma priors (*Bayesian Analysis* **5** (2010) 171–188) for the coefficients. In this sense, our prior can be viewed as a dynamic variable selection prior, which induces either smoothness (through the slab) or shrinkage towards zero (through the spike) at each time point. The MCMC method used for posterior computation uses Markov latent variables that can assume binary regimes at each time point to generate the coefficients' variances. In that way, our model is a dynamic mixture model; thus, we could use the algorithm of (*Journal of the American Statistical Association* **95** (2000) 819–828) to generate the latent processes without conditioning on the states. Finally, our approach is exemplified through simulated examples and a real data application.

## 1 Introduction

Over the past few decades, advances in computational processing have encouraged the proliferation of massive datasets, bringing new challenges to statistical research due to high-dimensionality issue. In this sense, regularization and variable selection techniques have become even more relevant to prevent overfitting and to solve ill-posed problems by inducing sparsity and/or shrinkage.

The advantage of regularization was highlighted a few years ago when [Hastie, Tibshirani and Friedman \(2001\)](#) coined the informal *Bet on Sparsity* principle. The principle encourages the use of procedures that do well in sparse problems for high-dimensional problems, since no procedure does well in dense problems, according to them. Indeed, they have shown that for a dense problem, where all the numerous coefficients are different from zero, and/or there is a high *Noise-to-Signal Ratio* (NSR), both the former *ridge regression* procedure of [Hoerl and Kennard \(1970\)](#) and the *least absolute shrinkage and selection operator* (lasso) from [Tibshirani \(1996\)](#) do poorly in terms of prediction.

Consider the Gaussian linear model defined by

$$y = \beta_0 \mathbf{1} + X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I), \quad (1)$$

where  $y$  is an  $n$ -dimensional vector of continuous responses,  $\beta_0$  is the intercept,  $\beta$  is a  $q$ -dimensional vector of regression coefficients associated with covariates,  $X$  is a  $(n \times q)$  design matrix with each column representing a covariate.

---

*Key words and phrases.* Cholesky decomposition, dynamic models, Normal-Gamma prior, spike-and-slab priors, high-dimensional data, scale mixture of Normals.

Received June 2025; accepted October 2025.

It is well known that the *ordinary least squares* (OLS) or *maximum likelihood estimator* (MLE)  $\hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'y$  often does poorly on both interpretation and prediction accuracy. While interpretation stands for the preference for parsimony, in the sense that simpler models put more light on the relationship between the response and covariates, prediction accuracy is related to the bias-variance trade-off. Although the OLS/ MLE estimator has the smallest variance among all linear unbiased estimators according to the Gauss–Markov theorem, an estimator with slight bias but smaller variance could be preferable, leading to a substantial decrease in prediction error.

Modern statistics addresses this trade-off between bias and variance through regularization and variable selection methods, which encourages simpler models because the space of values of estimators  $\hat{\beta}$  considered is smaller. In general terms, the notion of regularization summarizes approaches that allow to solve ill-posed problems, such as those which arise when  $p \gg n$ , or to prevent overfitting. The problem of variable selection arises when there is some unknown subset of the predictors with regression coefficients so small that it would be preferable to ignore them.

There has been great interest in regularizing the coefficients within problems in which the parameters vary over time and a few methods were proposed. A first wave of contributions are from Belmonte, Koop and Korobilis (2014), Kalli and Griffin (2014) and Bitto and Frühwirth-Schnatter (2019). In linear regression models with a large number of predictors, it is common to assume that only a subset of them is important for prediction. In the context of dynamic regression, it is reasonable to assume that these relevant subsets change over time.

Actually, we can define two sources of sparsity in dynamic regression problems: the *vertical sparsity* or *dynamic sparsity*, which stands for time-varying subsets of relevant predictors, and the *horizontal sparsity*, which allows for intermittent zeros for when each individual predictor is not relevant at all times  $t$ .

Consider the Gaussian dynamic linear regression model defined by

$$\begin{aligned} y_t &= F_t' \beta_t + v_t, & v_t &\sim \mathcal{N}(0, \sigma_t^2), \\ \beta_t &= G_t \beta_{t-1} + \omega_t, & \omega_t &\sim \mathcal{N}(0, W_t), \end{aligned} \quad (2)$$

for  $t = 1, \dots, T$ , where  $F_t' = X_t$  is the  $(1 \times q)$  vector of regressors,  $\beta_t = (\beta_{1t}, \dots, \beta_{qt})$  is the  $(q \times 1)$  vector of coefficients and  $v_t$  and  $\omega_t$  are two independent sequences of independent Gaussian errors with mean zero and variances  $\sigma_t^2$  and  $W_t$ , respectively. Note that setting  $W_t = 0$  for all  $t$  is equivalent to  $\beta_t = \beta$ , that is, the static regression. Usually, we may define  $G_t = G = I_q$ .

The problem of shrinking in dynamic regression problems has been addressed by Belmonte, Koop and Korobilis (2014) and Bitto and Frühwirth-Schnatter (2019) in a similar strategy. See also Huber, Koop and Onorante (2021) for recent comparative study. Their basic approach was to rewrite the states  $\beta_{j,t}$  from (2) in terms of the scaled states as  $\tilde{\beta}_{j,t} = \beta_{j,t}/\omega_j$ . Then the shrinkage of the time-varying coefficients  $\beta_{j,t}$  was done by assigning priors to their standard deviations  $\omega_j$ . While Belmonte, Koop and Korobilis (2014) used the Laplace prior for shrinking the standard deviations, Bitto and Frühwirth-Schnatter (2019) used the Normal-Gamma prior, which is more general since the Laplace prior is a special case of the Normal-Gamma prior where the shrinking parameter equals one.

In both approaches, the standard deviation  $\omega_j$  plays the role of relevance of the  $j$ th predictor: small values of  $\omega_j$  leads to greater shrinkage of the coefficient  $\beta_{j,t}$  for all times  $t$ . That is, because the standard deviation  $\omega_j$  is taken as fixed for all times  $t$ , if it is pulled toward zero, then the time-varying effect of the covariate  $X_j$  is nonsignificant. In this sense, both tackles horizontal sparsity, as the shrinkage effect of the prior for  $\omega_j$  is equal over all times  $t$ .

On the other hand, Kalli and Griffin (2014) developed an extension of the Normal-Gamma prior discussed by Griffin, Brown et al. (2010) for dynamic regression where both the values

of the regression coefficients and the importance of the variables are allowed to change over time. Time-varying sparsity is allowed by giving independent Normal-Gamma autoregressive (NGAR) process priors to the time series of regression coefficients. For details on the model specification, please refer to [Kalli and Griffin \(2014\)](#). See also [Lopes, McCulloch and Tsay \(2022\)](#), who propose a customized four-component mixture prior for the elements of  $\beta_t$ . The prior induces sparsity by either flatlining the coefficient to a constant, possibly zero, or letting move around freely according the dynamic structure when the state vector is of very high dimensions, say five thousand equations. All these methods deal with horizontal sparsities.

Vertical sparsity in dynamic models has become an increasingly interesting research topic. Related literature include [Nakajima and West \(2013\)](#), [Kalli and Griffin \(2014\)](#), [Rocková and McAlinn \(2021\)](#) and [Kowal, Matteson and Ruppert \(2019\)](#) and, more recently, [Koop and Korobilis \(2023\)](#) who uses variational Bayes ideas.

In the present work, we consider variable selection and shrinkage for the Gaussian dynamic linear regression model within a Bayesian framework. In particular, we propose a method that allows for time-varying sparsity, based on an extension of spike-and-slab priors for dynamic models using Markov switching auxiliary variables. The paper is organized as follows: Section 2 reviews the Bayesian approach for regularization and variable selection, emphasizing the main shrinking priors and giving a unifying approach for the spike-and-slab priors. Section 3 introduces the dynamic spike-and-slab prior and describes the full Bayesian model for dynamic regression with time-varying sparsity, the posterior inference and the Markov chain Monte Carlo (MCMC) method for sampling the parameters. Section 4 applies the proposed model to simulated data and considers empirical studies in inflation modeling. Section 5 summarizes our findings and conclusions.

## 2 Shrinking and variable selection priors: A brief review

Classical regularization is done by maximizing the likelihood subject to a penalty function. While the ridge regression estimate is a penalized least squares method imposing a  $\ell_2$  penalty on the regression coefficients, the lasso uses the  $\ell_1$  norm instead, leading not only to shrunken coefficients but also to sparse solutions. In that way, the lasso is also considered a variable selection method.

In contrast, Bayesian approaches for both variable selection and regularization in the Gaussian linear model stated by (1) can be formalized through the conditional distribution  $p(y|\beta, \phi)$ , where  $\phi$  is a parameter vector, comprising, for example, the error variance  $\sigma^2$ . The regularization is achieved by specifying appropriate informative priors  $p(\beta|\theta)$ , where the hyperparameter vector  $\theta$  includes parameters controlling shrinkage properties. The model is completed by assuming hyperpriors  $p(\theta)$  and  $p(\phi)$  and the inference is based on the posterior  $p(\beta, \phi, \theta, y) \propto p(y|\beta, \phi)p(\beta|\theta)p(\theta)p(\phi)$ .

It can be shown that, if  $\phi$  and  $\theta$  are fixed, then the posterior mode or the maximum a posteriori (MAP)

$$\arg \max_{\beta} \{p(y|\beta, \phi)p(\beta|\theta)\}$$

is equivalent to penalizing the log-likelihood  $\log p(y|\beta, \phi)$  with penalty equal to the (minus) log prior  $\log p(\beta|\theta)$ .

### 2.1 Shrinking priors

In general, practically all shrinking priors are defined hierarchically as a *Scale-Mixture of Normals* (SMN) (see, e.g., [West \(1987\)](#)). Considering the Gaussian linear model as in (1), the

SMN has the following general structure:

$$\beta_j | \psi_j \stackrel{\text{ind}}{\sim} \mathcal{N}(0, \psi_j), \quad \psi_j | \theta \sim p(\psi_j | \theta), \quad (3)$$

where  $\beta_i$  and  $\beta_j$  are independent for any  $i, j \in \{1, \dots, q\}$  and  $\psi_j$  depends on the vector of hyperparameters  $\theta$ . Note that the marginal distribution  $p(\beta_j | \theta) = \int p(\beta_j | \psi_j) p(\psi_j | \theta) d\psi_j$  is non Gaussian, and can assume many forms depending on the mixing distribution  $p(\psi_j | \theta)$ . A famous form arises when the mixing distribution  $p(\psi_j | \theta)$  from (3) is Exponential, that is,

$$\psi_j | \lambda \sim \mathcal{E}(\lambda^2/2),$$

for  $j = 1, \dots, q$ , where  $\mathcal{E}(\alpha)$  denotes the Exponential distribution with mean  $1/\alpha$ . Thus, marginally,  $\beta_j$  follows a Laplace distribution with null location and scale  $1/\lambda$ , so that  $p(\beta_j) \propto \exp(-\lambda|\beta_j|)$ . The Laplace prior is also known as the Bayesian lasso from [Park and Casella \(2008\)](#).

The frequentist ridge regression also has a Bayesian analogue and can be represented by (3) if we consider an Inverted-Gamma mixing distribution  $p(\psi_j | \theta)$ . That is, assuming

$$\psi_j | a, b \sim \mathcal{IG}(a, b),$$

for  $j = 1, \dots, q$ , then  $\beta_j$  follows a scaled  $t$  distribution with  $2a$  degrees of freedom and scale parameter  $\sqrt{a/b}$ , marginally. Hence, the ridge prior leads to weaker penalization of large coefficients as long as the  $t$  distribution has heavier tails than the Gaussian distribution.

## 2.2 The Normal-Gamma prior

Even more shrinkage than the Bayesian lasso and the ridge prior can be achieved by using a Gamma mixing distribution in (3) as

$$\psi_j | \lambda, \gamma^2 \sim \mathcal{G}(\lambda, 1/(2\gamma^2)),$$

for  $j = 1, \dots, q$ , where  $\mathcal{G}(\lambda, 1/(2\gamma^2))$  denotes the Gamma distribution with shape parameter  $\lambda$  and mean  $2\lambda\gamma^2$ . This structure leads to the Normal-Gamma density, which was applied to regression problems in [Griffin, Brown et al. \(2010\)](#) and can be expressed in closed form as

$$p(\beta_j) = \frac{1}{\sqrt{\pi} 2^{\lambda-1/2} \gamma^{\lambda+1/2} \Gamma(\lambda)} |\beta_j|^{\lambda-1/2} K_{\lambda-1/2}(|\beta_j|/\gamma), \quad (4)$$

where  $K$  is the modified Bessel function of the third kind. Taking this parametrization, the variance of  $\beta_j$  is  $2\lambda\gamma^2$  and the excess kurtosis is  $3/\lambda$ . As the shape parameter  $\lambda$  of the Gamma distribution decreases these include distributions that place a lot of mass close to zero but at the same time have heavy tails. Thus, the effect of the parameter  $\lambda$  is related to shrinking, while lower values of  $\lambda$  are associated with more shrinkage as more mass is placed to zero. Thus, the Normal-Gamma prior is more general than the Bayesian lasso, which corresponds to  $\lambda = 1$ .

One could fix the shrinking parameter  $\lambda$  such as in [Frühwirth-Schnatter and Wagner \(2010\)](#) and in [Kastner \(2019\)](#)<sup>1</sup> or adopt a fully Bayesian approach by assigning hyperpriors to both  $\lambda$  and  $\gamma^2$  as did [Griffin, Brown et al. \(2010\)](#). A prior, which seemed to work well in the simulations, is taking  $\lambda$  to be an exponential distribution with mean 1, which offers variability around the Bayesian lasso.

<sup>1</sup>In [Frühwirth-Schnatter and Wagner \(2010\)](#), they set  $\lambda = 1/2$ , in which case  $\psi_j | \gamma^2 \sim \gamma^2 \chi_1^2$ , or equivalently,  $\sqrt{\psi_j} \sim N(0, \gamma^2)$ . In [Kastner \(2019\)](#), it was assumed that  $\lambda = 0.1$ .

### 2.3 Variable selection: Spike-and-slab priors

Bayesian variable selection is commonly based on spike-and-slab priors for regression coefficients. The basic idea is that each component  $\beta_j$  from the coefficients' vector  $\beta$  is modeled as having come either from a distribution with most (or all) of its mass concentrated around zero (the *spike*), or from a comparably diffuse distribution with mass spread out over a large range of values (the *slab*), that is,

$$\beta_j | J_j \sim J_j p_{\text{slab}}(\beta_j | \theta) + (1 - J_j) p_{\text{spike}}(\beta_j | \theta), \tag{5}$$

where  $\theta$  is a vector of parameters,  $p_{\text{spike}}(\beta_j | \theta)$  is the spike distribution,  $p_{\text{slab}}(\beta_j | \theta)$  is the slab distribution and  $J_j \in \{0, 1\}$  is a binary random variable with  $p(J_j = 1) = 1 - P(J_j = 0) = \omega$ .

Famous seminal *stochastic search variable selection* (SSVS) method of [George and McCulloch \(1993\)](#) used Gaussian distributions for both the spike and the slab as follows:

$$\beta_j | J_j \sim (1 - J_j) \mathcal{N}(0, \tau_j^2) + J_j \mathcal{N}(0, c_j^2 \tau_j^2), \tag{6}$$

for  $j = 1, \dots, q$ , where  $c_j > 1$  is a large scalar and  $\tau_j > 0$  is a small scalar. Note that  $c_j^2$  is the ratio of variances between the slab and the spike distributions.

Indeed, one can achieve (5) by formulating appropriate spike-and-slab priors to the component variances  $(\psi_j | J_j = 0) = \text{Var}_{\text{spike}}(\beta_j | \theta)$  and  $(\psi_j | J_j = 1) = \text{Var}_{\text{slab}}(\beta_j | \theta)$ , what was first proposed by [Ishwaran and Rao \(2005\)](#). Their idea is that, considering the hierarchical SMN representation from (3), if we assume absolutely continuous priors for the component variances, that is, a mixture prior for  $\psi_j$ , then we reach the spike-and-slab structure for the coefficient  $\beta_j$  as in (5).

Actually, even the SSVS approach given by (6), where the coefficients are given a mixture of Normals prior, can be represented in that way. Letting  $c_j^2 = 1/r = \text{Var}_{\text{slab}}(\beta_j | \theta) / \text{Var}_{\text{spike}}(\beta_j | \theta)$  and  $\tau_j^2 = r$  from (6) and assuming a two-point mixture prior to the variance  $\psi_j$  in (3),

$$\psi_j | \omega, Q, r \sim (1 - \omega) \delta_{rQ}(\cdot) + \omega \delta_Q(\cdot),$$

where  $\delta_v(\cdot)$  is a discrete measure concentrated at value  $v$ , we get the original formulation of the SSVS prior of [George and McCulloch \(1993\)](#).

Nevertheless, it can be difficult to set the hyperparameters  $r$ ,  $Q$  and  $\omega$ , used in the two-point mixture prior for the variance  $\psi_j$ , as noted by [Ishwaran and Rao \(2005\)](#). Hence, their approach is to place absolutely continuous priors for  $\psi_j$ . In particular, they chose Inverted-Gamma densities for both the spike and the slab variances as

$$\psi_j | J_j = 0 \sim \mathcal{IG}(v, rQ), \quad \psi_j | J_j = 1 \sim \mathcal{IG}(v, Q),$$

where  $\theta = (v, r, Q)$  is the vector of hyperparameters that define the conditional (on  $J_j$ ) variances' densities. Thus, each  $\beta_j$  has the marginal distribution

$$\beta_j | \omega, r, Q \sim \omega t_{2v}(0, Q/v) + (1 - \omega) t_{2v}(0, rQ/v), \tag{7}$$

where  $t_{\xi}(0, s)$  denotes the Student's t distribution with zero location, scale  $\sqrt{s}$  and  $\xi$  degrees of freedom. Because of the SMN representation, (7) is also known as *Normal mixture of Inverse-Gamma* (NMIG) prior. Although it allows discrimination or variable selection, it does not encourage shrinkage in the sense that the resulting marginal distribution of each coefficient  $\beta_j$  is a two component mixture of scaled Student's t distributions. Hence, it makes sense to choose other component specific distributions, besides the Inverse Gamma, which could actually induce shrinkage. A prior that solves this is assuming Exponential densities for both the spike and the slab as

$$\psi_j | J_j = 0 \sim \mathcal{E}(1/2rQ), \quad \psi_j | J_j = 1 \sim \mathcal{E}(1/2Q),$$

where  $\mathcal{E}(\alpha)$  denotes the Exponential distribution with mean  $1/\alpha$ , which leads to

$$\beta_j | \omega, r, Q \sim \omega \text{Lap}(\sqrt{Q}) + (1 - \omega) \text{Lap}(\sqrt{rQ}), \quad (8)$$

where  $\text{Lap}(x)$  denotes the Laplace distribution with mean 0 and scale parameter  $x$  and the weight  $\omega$  is the prior probability of the slab, that is,  $\omega = p(J_j = 1)$ . That is, (8) is a mixture of Laplace densities for  $\beta_j$ . Finally, if we assume that  $\psi_j$  is a mixture of Gammas, that is,

$$\psi_j | J_j = 0 \sim \mathcal{G}(a, 1/2rQ), \quad \psi_j | J_j = 1 \sim \mathcal{G}(a, 1/2Q),$$

then the marginal distribution of  $\beta_j$  is a mixture of Normal-Gamma densities as showed in (4), that is,

$$\beta_j | \omega, r, Q \sim \omega \mathcal{NG}(\beta_j | a, Q) + (1 - \omega) \mathcal{NG}(\beta_j | a, r, Q). \quad (9)$$

Hence, assuming either (7), (8) or (9), we adopt a unifying approach for spike-and-slab priors as discussed in Frühwirth-Schnatter and Wagner (2011). Taking this parametrization, we will always have

$$\text{Var}_{\text{spike}}(\beta_j | r, Q) = cQr, \quad \text{Var}_{\text{slab}}(\beta_j | Q) = cQ,$$

with  $c$  being a constant, which depends on the distribution assumption. Table 1 gives a summary for what as discussed through this section assuming the general form from (5) and viewing each prior as a scaled mixture of Normals (SMN).

In this section, we have discussed several shrinkage and sparsity inducing priors independently assigned to static coefficients from the Gaussian linear model. Now we turn attention to the case where the coefficients from regression are time varying, the so-called time-varying parameter (TVP) regression models. The next section describes the proposed model for shrinking and selecting subsets of relevant variables dynamically.

### 3 Our approach for sparsity in TVP models

The aim of the method is allowing for vertical sparsity, so that in each snapshot of time  $t$  we can have different subsets of relevant predictors as well enabling shrinkage of the time-varying coefficients. This is accomplished by extending the approach of spike-and-slab priors over the variances discussed in Section 2.3 for dynamic models. In our approach, at each time point, an auxiliary Markov switching variable can assume two regimes—the spike or the slab—driving the evolution of the variance  $\psi_{j,t}$ , assuming that  $\beta_{j,t} | \psi_{j,t} \sim \mathcal{N}(0, \psi_{j,t})$ .

#### 3.1 Model specification

The observation and the state equations considered are given below. We work with univariate time series responses, although extending the framework to multivariate time series is straightforward. We assume the Gaussian dynamic linear regression model

$$y_t = X_t \beta_t + v_t, \quad v_t \sim \mathcal{N}(0, \sigma_t^2), \quad (10)$$

for  $t = 1, \dots, T$ , where  $X_t$  is a  $(T \times q)$  matrix of regressors,  $\beta_t$  is a  $(q \times 1)$  vector of coefficients with the following evolution equation for the scaled states  $\tilde{\beta}_{1:T} = (\tilde{\beta}_1, \dots, \tilde{\beta}_T)$ :

$$\tilde{\beta}_t = G_t \tilde{\beta}_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, W_t), \quad (11)$$

for  $t = 2, \dots, T$ , with

$$\tilde{\beta}_t = (\beta_{1,t}/\sqrt{\psi_{1,t}}, \dots, \beta_{q,t}/\sqrt{\psi_{q,t}}),$$

$$G_t = \text{diag}(\phi_1, \dots, \phi_q),$$

$$W_t = \text{diag}((1 - \phi_1^2), \dots, (1 - \phi_q^2)),$$

**Table 1** Summary table of an unifying approach for spike-and-slab mixture priors. Depending on the assumption for the mixture prior of  $\psi_j$ , the constant  $c$  changes so that we can compare the different priors, fixing the value of the component variances  $\text{Var}_{\text{spike}}(\beta_j|r, Q)$  and  $\text{Var}_{\text{slab}}(\beta_j|Q)$

Prior	Spike $\psi J=0$	Slab $\psi J=1$	Marginal $\beta \omega$	Constant $c$
SSVS	$\psi J=0 = \delta_{rQ}(\cdot)$	$\psi J=1 = \delta_Q(\cdot)$	$\omega\mathcal{N}(0, Q) + (1-\omega)\mathcal{N}(0, rQ)$	1
NMIG	$\mathcal{IG}(v, rQ)$	$\mathcal{IG}(v, Q)$	$\omega t_{2v}(0, Q/v) + (1-\omega)t_{2v}(0, rQ/v)$	$1/(v-1)$
Mixture of Laplaces	$\mathcal{E}(1/2rQ)$	$\mathcal{E}(1/2Q)$	$\omega \text{Lap}(\sqrt{Q}) + (1-\omega) \text{Lap}(\sqrt{rQ})$	2
Mixture of Normal-Gammas	$\mathcal{G}(a, 1/2rQ)$	$\mathcal{G}(a, 1/2Q)$	$\omega\mathcal{NG}(\beta_j a, Q) + (1-\omega)\mathcal{NG}(\beta_j a, r, Q)$	$2a$
Laplace-t	$\mathcal{E}(1/2rQ)$	$\mathcal{IG}(v, Q)$	$\omega t_{2v}(0, Q/v) + (1-\omega) \text{Lap}(\sqrt{rQ})$	$c_1 = 2, c_2 = 1/(v-1)$

where the initial condition for the scaled states is  $\tilde{\beta}_1 \sim \mathcal{N}(0, I)$ . The  $G_t$  matrix specify (conditionally independent) autoregressive models for the coefficients  $\tilde{\beta}_t$ . The parameters  $\phi$  appear in  $G_t$ , as well as in  $W_t$  in order to accommodate a collapsed state when  $\phi$  approaches the unity. The definition below specifies the generic dynamic spike-and-slab prior that can be assigned to the coefficients' variances  $\psi_{j,1:T}$  in order to induce shrinkage and/or variable selection.

**Definition 3.1 (Dynamic spike-and-slab prior).** Consider that  $\psi_{j,t} = K_{j,t} \tau_j^2$ . The dynamic spike-and-slab prior for  $\beta_{j,1:T}$  is defined by (10), (11) and

$$\begin{aligned} \tau_j^2 &\stackrel{\text{i.i.d.}}{\sim} p(\tau_j^2 | \theta), \\ (K_{j,t} | K_{j,t-1} = v_i) &\stackrel{\text{ind}}{\sim} \omega_{j,1,i} \delta_1(\cdot) + (1 - \omega_{j,1,i}) \delta_r(\cdot), \\ \omega_{j,1,i} &= p(K_{j,t} = 1 | K_{j,t-1} = v_i), \end{aligned} \tag{12}$$

for  $i \in \{0, 1\}$ ,  $j = 1, \dots, q$ ,  $t = 2, \dots, T$ , where  $\delta_x(\cdot)$  is a discrete measure concentrated at value  $x$ ,  $v_i \in \{r, 1\}$ ,  $p(K_{j,1} = r) = p(K_{j,1} = 1) = 1/2$  and  $p(\tau_j^2 | \theta)$  can be one of the mixing distributions from Table 1. As in Section 2.3, we assume that  $r = \text{Var}_{\text{spike}}(\beta_j | \theta) / \text{Var}_{\text{slab}}(\beta_j | \theta) \ll 1$  and that  $v_1 = 1$ .

Because now we are talking about dynamic models and dynamic sparsity, we have a time-varying scale parameter  $\psi_{j,t}$ , which is taken to be  $\psi_{j,t} = K_{j,t} \tau_j^2$ . That is, the time-varying pattern for the scale parameter is driven by the latent variable  $K_{j,t}$ , which evolves as a Markov switching process of order 1 and can assume two values  $K_{j,t} = 1$  or  $K_{j,t} = r$  according to a transition matrix. Thus, we assume a finite mixture prior for  $\psi_{j,t}$  as

$$\psi_{j,t} | K_{j,t-1} = v_i \sim \omega_{j,1,i} p_{\text{slab}}(\psi_j | Q_j) + (1 - \omega_{j,1,i}) p_{\text{spike}}(\psi_j | r, Q_j),$$

where  $\omega_{j,1,i}$  is the transition probability of the first-order Markov process  $K_{j,t}$  to regime  $K_{j,t} = 1$  (the slab) given that  $K_{j,t-1} = v_i \in \{r, 1\}$ . Thus, by adopting a regime switching model, the process  $\psi_{j,t}$  can switch between the spike and the slab variances' distributions according to the following transition probabilities:

$$\mathcal{P}_j = \begin{bmatrix} \omega_{j,0,0} & \omega_{j,0,1} \\ \omega_{j,1,0} & \omega_{j,1,1} \end{bmatrix}$$

where  $\omega_{j,k,i} = P(K_{j,t} = v_k | K_{j,t-1} = v_i)$  denotes the probability of  $K_{j,t}$  changing to regime  $v_k$  from regime  $v_i$ ,  $k, i \in \{0, 1\}$ . Note that  $\omega_{j,0,1} = (1 - \omega_{j,1,1})$  and  $\omega_{j,1,0} = (1 - \omega_{j,0,0})$ .

For the other component, a prior distribution (Inverse-Gamma, Gamma or Exponential) is placed on  $\tau_j^2$ . Together with the variable  $K_{j,t}$ , this results in a spike-and-slab prior for  $\psi_{j,t}$ , which shrinks the coefficients whenever  $\sigma_j^2$  takes a small value through the spike component of the mixture prior.

For instance, if  $p(\tau_j^2 | \theta) \sim \mathcal{IG}(v, Q_j)$ , then we have a mixture of scaled-t for each  $\beta_{j,t}$  as

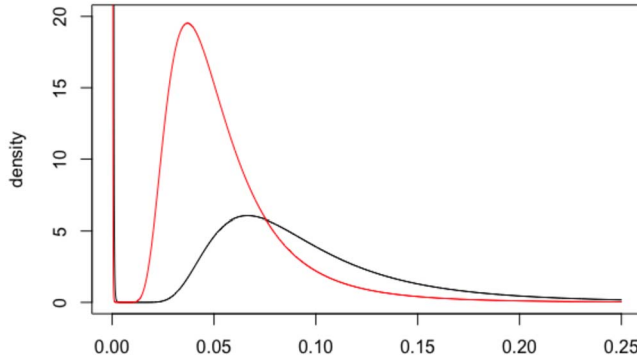
$$\beta_{j,t} | \omega_{j,1,i} \sim \omega_{j,1,i} t_{2v}(0, Q_j/v) + (1 - \omega_{j,1,i}) t_{2v}(0, r Q_j/v). \tag{13}$$

If  $p(\tau_j^2 | \theta) \sim \mathcal{G}(a_\tau, 1/2 Q_j)$ , then we have a mixture of Normal-Gamma densities for each  $\beta_{j,t}$  as

$$\beta_{j,t} | \omega_{j,1,i} \sim \omega_{j,1,i} \mathcal{NG}(\beta_{j,t} | a_\tau, Q_j) + (1 - \omega_{j,1,i}) \mathcal{NG}(\beta_{j,t} | a_\tau, r Q_j). \tag{14}$$

Finally, if  $p(\tau_j^2 | \theta) \sim \mathcal{E}(1/2 Q_j)$ , then we have a mixture of Laplace densities for each  $\beta_{j,t}$  as

$$\beta_{j,t} | \omega_{j,1,i} \sim \omega_{j,1,i} \text{Lap}(\sqrt{Q_j}) + (1 - \omega_{j,1,i}) \text{Lap}(\sqrt{r Q_j}). \tag{15}$$



**Figure 1** Mixture prior for  $\psi_{j,t}$  assuming the NMIG structure from (13) with the following hyperparameters  $\nu = 5, r = 0.0025, c_\psi = 2, C_\psi = 0.05, \omega_{j,1,i} = 0.5$  (black) and  $\omega_{j,1,i} = 0.9$  (red).

Furthermore, when assuming that  $\psi_{j,t} \sim \mathcal{IG}(c_\psi, C_\psi)$ , it can be shown that the distribution of  $Q_j$  is Inverse-Gamma:

$$Q_j | \omega_{j,1,i} \stackrel{\text{ind}}{\sim} \mathcal{IG}(c_\psi, C_\psi / f^*(w)), \tag{16}$$

for  $j = 1, \dots, q$ , where  $f^*(w) = c[(1 - \omega_{j,1,i})r + \omega_{j,1,i}]$  depends on the value of on the distribution constant  $c$  from Table 1 specified for  $\tau_j^2 | \theta$  and on the value of  $\omega_{j,1,i} = p(K_{j,t} = \nu_1 | K_{j,t-1} = \nu_i)$ . By defining the hyperparameters  $c_\psi$  and  $C_\psi$  appropriately, we can learn about  $Q_j$  and, therefore, about  $\tau_j^2$ . For instance, if we assume that  $\nu = 5, r = 0.0025, c_\psi = 2, C_\psi = 0.05$ , we have the following mixture densities for  $\psi_{j,t}$  in Figure 1, considering two values of  $\omega_{j,1,i} = 0.5, 0.9$  and the NMIG structure.

The assumption that  $Q_j$  is given a prior as defined in (16) makes  $\tau_j^2$  also indirectly depend on the previous value of the Markov latent variable  $K_{j,t-1} = \nu_i \in \{r, 1\}$ . The purpose of this is to keep the variance of the coefficients  $\beta_{j,t}$  at each time point  $t$  constant across the priors defined by equations (13), (14) and (15), and thus, comparable.

For a simpler specification, from now on we assume that the observation variance is constant over time such that  $\sigma_t^2 = \sigma^2$ . Extending the model to accommodate stochastic volatility is straightforward. In order to complete the specification, we shall assign prior distributions to parameters  $\sigma^2, \phi$  and to the transition probabilities  $\mathcal{P}$  in a fully Bayesian strategy. For the observation variance  $\sigma^2$ , we assume the conjugate traditional prior

$$\sigma^2 \sim \mathcal{IG}(a_\sigma, b_\sigma).$$

For the AR parameters  $\phi$ , we assume that each  $\phi_j$  are independent from each other and distributed as

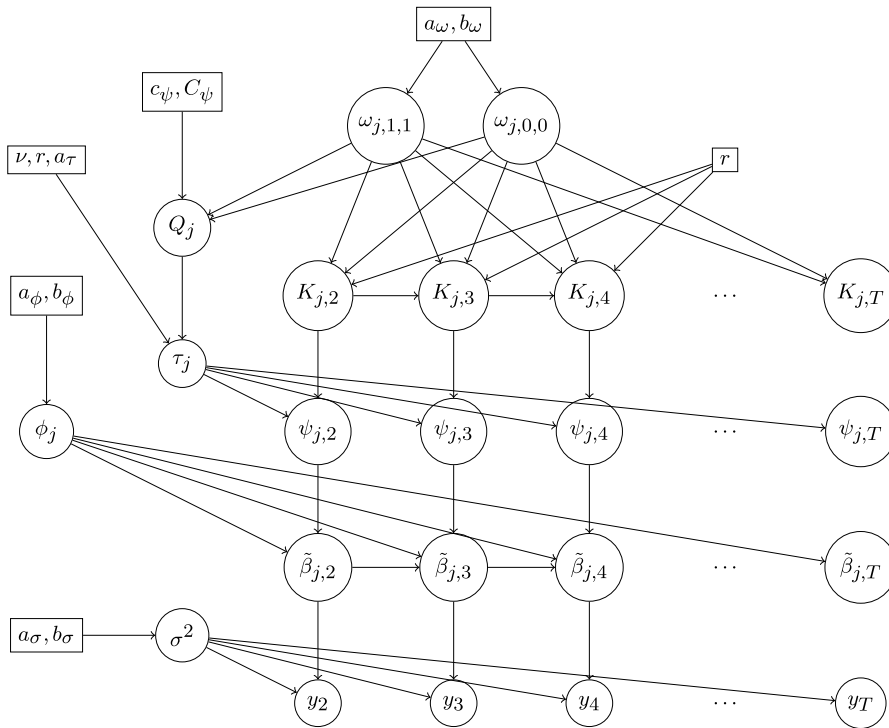
$$\phi_j \sim \mathcal{B}(a_\phi, b_\phi),$$

for  $j = 1, \dots, q$ , where we are not considering the case  $-1 < \phi_j < 0$ . Finally, for the transition probabilities  $\mathcal{P}_j$  we assign independent Beta distributions as

$$\omega_{j,i,i} \sim \mathcal{B}(a_\omega, b_\omega),$$

for  $j = 1, \dots, q$ , with  $i \in \{0, 1\}$  denoting the spike ( $i = 0$ ) or the slab ( $i = 1$ ) and  $\omega_{j,k,i} = (1 - \omega_{j,i,i}), k \neq i, k, i \in \{0, 1\}$ .

The directed acyclic graph (DAG) that summarizes the dependencies of the proposed model is shown in Figure 2. We can see that the observations  $y_1, \dots, y_T$  are conditionally independent given the scaled coefficients  $\hat{\beta}_{j,1}, \dots, \hat{\beta}_{j,T}$ , which in turn depend on the AR pa-



**Figure 2** Dependence structure for dynamic spike-and-slab model.

parameter  $\phi_j$  and on the scale  $\psi_{j,t}$ . The latter is driven by the Markov switching variable  $K_{j,t}$ , which evolves accordingly to the transition probabilities  $\omega_{j,1,1}$  and  $\omega_{j,0,0}$  and with  $\tau_j^2$ , whose distribution governs the prior choice for  $\beta_{j,t}$  (e.g., mixture of Laplaces, Normal-Gammas or scaled-t).

### 3.2 Posterior inference

The posterior distribution of the parameters can be drawn using a hybrid Gibbs sampler with an additional Metropolis–Hastings update. The scaled states  $\tilde{\beta}_1, \dots, \tilde{\beta}_T$  can be updated using the FFBS algorithm (due to Carter and Kohn (1994) and Frühwirth-Schnatter (1994)) within the Gibbs sampler, while the process  $K = (K_1, \dots, K_T)$  is updated using the algorithm proposed by Gerlach, Carter and Kohn (2000). At each snapshot of time  $t$ , the probability of  $K_t = 1$  depends only on the previous value observed  $K_{t-1} = v_t \in \{r, 1\}$ . Thus, the sequence  $(K_1, \dots, K_T)$  is a sequence of random variables that are Markovian as for  $t = 2, \dots, T$ . We have  $p(K_t | K_{1:t-1}) = p(K_t | K_{t-1})$ .

This feature was discussed by Gerlach, Carter and Kohn (2000) in their work about *dynamic mixture models*. These models add to dynamic linear model equations (2) the assumption that the system matrices  $F_t$ ,  $G_t$  and  $W_t$  and the variance  $\sigma_t^2$  are determined, up to a set of unknown parameters, by the value of  $K_t$ . The algorithm and the lemmas associated are detailed in the Appendix.

In summary, the MCMC scheme is given below:

1. Draw  $\tilde{\beta}$  jointly using Forward Filtering Backward Sampling (FFBS).
2. Draw  $K$  jointly using the algorithm of Gerlach, Carter and Kohn (2000).

3. Draw  $\sigma^2$  by its full conditional

$$(\sigma^2 | \theta_{\setminus \sigma^2}, y) \sim \mathcal{IG}\left(a_\sigma + \frac{T}{2}, b_\sigma + \frac{1}{2} \sum_{t=1}^T (y_t - X_t \beta_t)^2\right),$$

where  $\theta_{\setminus \sigma^2}$  denotes all the parameters to be sampled except from  $\sigma^2$ .

4. Draw each  $\tau_j^2$  by its full conditional. Assuming the NMIG prior structure from (13),

$$(\tau_j^2 | \theta_{\setminus \tau_j^2}, y) \sim \mathcal{IG}\left(v + \frac{T}{2}, Q_j + \frac{1}{2} \sum_{t=1}^T \frac{(\beta_{j,t} - \sqrt{\frac{K_{j,t}}{K_{j,t-1}}} \phi_j \beta_{j,t-1})^2}{K_{j,t}(1 - \phi_j^2)}\right).$$

Assuming the mixture of Laplaces from (14) or the mixture of Normal-Gammas from (15),

$$(\tau_j^2 | \theta_{\setminus \tau_j^2}, y) \sim \mathcal{GIG}(p, g, h),$$

where

$$g = 1/Q_j, \quad h = \sum_{t=1}^T \frac{(\beta_{j,t} - \sqrt{\frac{K_{j,t}}{K_{j,t-1}}} \phi_j \beta_{j,t-1})^2}{K_{j,t}(1 - \phi_j^2)}, \quad p = a_\tau - T/2.$$

5. Draw each  $\phi_j$  using Metropolis–Hastings algorithm since the full conditional

$$\begin{aligned} p(\phi_j | \theta_{\setminus \phi_j}, y) &\propto p(\phi_j | a_\phi, b_\phi) p(\beta_j | K_j, \sigma^2, \tau_j^2) \\ &\propto \phi_j^{(a_\phi - 1)} (1 - \phi_j)^{(b_\phi - 1)} \exp\left\{-\sum_{t=1}^T \frac{(\beta_{j,t} - \sqrt{\frac{\psi_{j,t}}{\psi_{j,t-1}}} \phi_j \beta_{j,t-1})^2}{2\psi_{j,t}(1 - \phi_j^2)}\right\}, \end{aligned}$$

has no closed form. We use a Beta proposal density  $q(\phi_j^* | \phi_j^{(m-1)})$  as

$$\phi_j^* \sim \mathcal{B}(\alpha, \xi(\phi_j^{(m-1)})), \quad \xi(\phi_j^{(m-1)}) = \alpha \left( \frac{1 - \phi_j^{(m-1)}}{\phi_j^{(m-1)}} \right),$$

where  $\alpha$  is a tuning parameter and the acceptance distribution is

$$\mathcal{A}(\phi_j^* | \phi_j^{(m-1)}) = \min\left\{1, \frac{f(\phi_j^*) q(\phi_j^{(m-1)} | \phi_j^*)}{f(\phi_j^{(m-1)}) q(\phi_j^* | \phi_j^{(m-1)})}\right\}.$$

6. Update the transition probabilities from the latent Markov process by their full conditionals

$$\begin{aligned} (\omega_{1,1} | \theta_{\setminus \omega_{1,1}}, y) &\sim \mathcal{B}(a_\omega + \#\{t : v_1 \rightarrow v_1\}, b_\omega + \#\{t : v_1 \rightarrow v_0\}), \\ (\omega_{0,0} | \theta_{\setminus \omega_{0,0}}, y) &\sim \mathcal{B}(a_\omega + \#\{t : v_0 \rightarrow v_0\}, b_\omega + \#\{t : v_0 \rightarrow v_1\}), \end{aligned}$$

with  $v_0 = r$  and  $v_1 = 1$ .

7. Draw each  $Q_j$  by its full conditional, which depends on the mixing distribution assumed for the spike-and-slab process as follows:

- NMIG prior:  $\mathcal{GIG}(p, g, h)$ , with  $p = v - c_\psi$ ,  $g = 2\tau_j^{-2}$  and  $h = 2[C_\psi / f^*(w)]$ ,
- Mixture of Normal-Gammas:  $\mathcal{IG}(c_\psi + a_\tau, \tau_j^2/2 + [C_\psi / f^*(w)])$ ,
- Mixture of Laplaces:  $\mathcal{IG}(c_\psi + a_\tau, \tau_j^2/2 + [C_\psi / f^*(w)])$ , with  $a_\tau = 1$ .

## 4 Synthetic and real data analyses

In this section, we present two simulated examples where some coefficients are relevant in some periods of time and negligible in others. The first example is a singular equation model with five coefficients with four possible patterns and the second example is an application of the modified Cholesky decomposition where we simulate time-varying coefficients that compose the Cholesky factor and then apply the spike-and-slab priors on each recursive regression. In our empirical application, we use the inflation data obtained from Griffin's research page.<sup>2</sup> Inflation forecasting is a frequent topic within the shrinkage in time-varying parameter models literature and was also the main subject of Belmonte, Koop and Korobilis (2014).

### 4.1 First simulation example

We generated the data using equation (10) with  $q = 5$  predictors,  $T = 200$  and constant observational variance  $\sigma_t^2 = \sigma^2 = 1$ , where  $X_t \sim N(0, I)$  and  $X_{j,1}, \dots, X_{j,T}$  are independent. We simulate the five regression coefficients as follows:

1. The first coefficient  $\beta_{1,t}$  follows a stationary AR(1) process with AR parameter 0.97 and a Normal stationary distribution with mean 2 and variance 0.25. The initial value was drawn from its stationarity distribution  $\beta_{1,1} \sim \mathcal{N}(2, 0.25)$ .
2. The second coefficient  $\beta_{2,t}$  also follows an AR(1) process with autocorrelation parameter 0.97 and a Normal marginal distribution with mean 0 and variance 0.25, but only until the half of the sample, that is,

$$\beta_{2,t} = \begin{cases} 0.97\beta_{2,t-1} + \epsilon_{2,t}, & t \leq 100, \\ 0, & t > 100, \end{cases}$$

with the initial value drawn as  $\beta_{2,1} \sim \mathcal{N}(2, 0.25)$ .

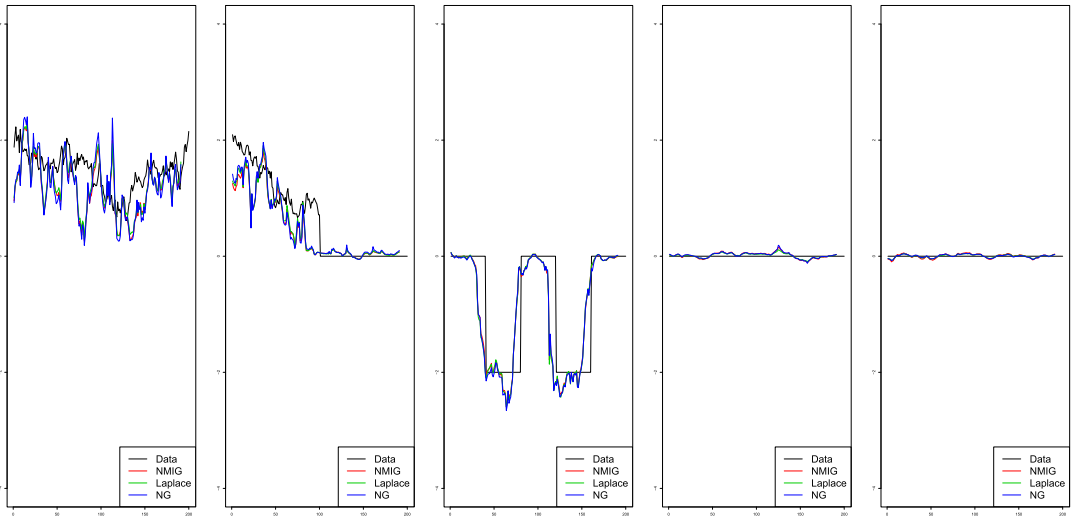
3. The third coefficient is always zero, except from two short periods when it equals  $-2$ :

$$\beta_{3,t} = \begin{cases} 0, & t \leq 20; 51 \leq t \leq 120; 151 \leq t \leq 200 \\ -2, & 21 \leq t \leq 50; 121 \leq t \leq 150. \end{cases}$$

4. The fourth coefficient is  $\beta_{4,t} = 0, \forall t$ .
5. The fifth coefficient  $\beta_{5,t} = 0, \forall t$ .

We generate 5 replications of the data and then sample from the posterior distribution using the three mentioned priors for  $\beta_{j,t}$  with the the following hyperparameters settings:  $\nu_0 = r = 0.005$ ,  $\nu_1 = 1$ ,  $a_\tau = 0.5$  (for the NG prior),  $\nu = 5$ ,  $c_0 = 51$ ,  $C_0 = 5$ ,  $a_\sigma = 0.0001$ ,  $b_\sigma = 0.0001$  (improper prior) and  $\alpha = 1000$  (tuning parameter for Metropolis). We have used the standard prior specification illustrated in Figure 1. The main contribution of (only) 5 scenarios is to illustrate the flexibility of the currently modeling when the state is truly a random walk, random walk plus static, regime-shifting and purely static. The MCMC algorithm was run for 10,000 iterations with half discarded as a burn-in. The prior for the autoregressive parameter is  $\phi_j \sim \mathcal{B}(77.6, 2.4)$  for  $j = 1, \dots, 5$ , so that it has mean 0.97. The same choice was made for the transition probabilities  $\omega_{j,0,0}$  and  $\omega_{j,1,1}$ . The informative prior choice is to assure that  $\beta_{j,t}|\beta_{j,t-1}$  and  $\psi_{j,t}|\psi_{j,t-1}$  evolves smoothly and  $K_t$  does not switch regimes so rapidly. Figure 3 shows the posterior medians of the coefficients  $\beta_{j,t}$ , comparing them to the real values, while Table 2 shows the root mean square error (RMSE) of the three priors

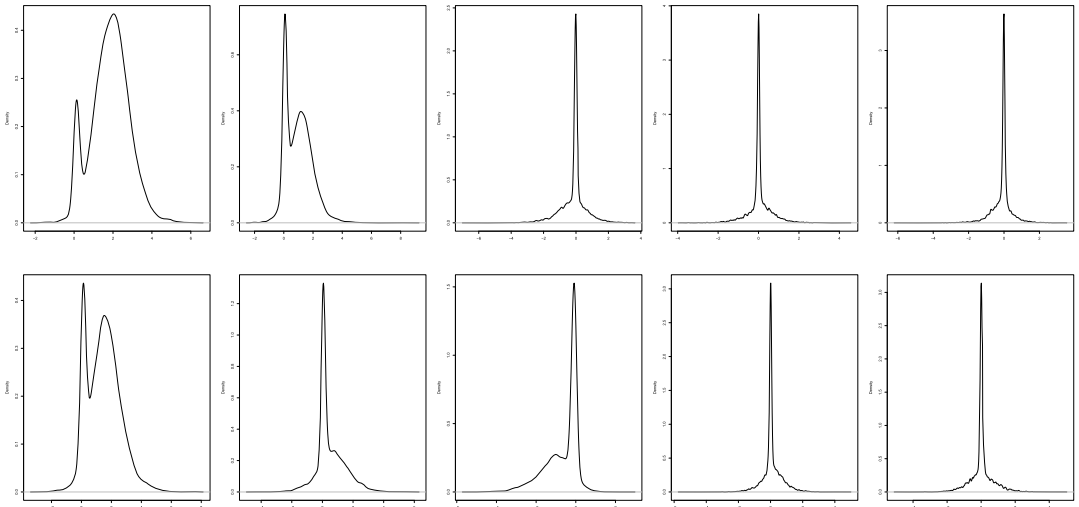
<sup>2</sup>Available in <https://www.kent.ac.uk/smsas/personal/jeg28/index.htm>.



**Figure 3** Fitting of the models using posterior medians of the sampled coefficients.

**Table 2** Mean of the RMSEs of the five replications for the dynamic spike-and-slab priors using the mean and the median of the sampled coefficients—simulated example 1

Prior	RMSE (mean)	RMSE (median)
NMIG	0.3522	0.3543
NG	0.3636	0.3641
Laplace	0.3425	0.3441



**Figure 4** Posterior densities of the sampled coefficients using the dynamic Laplace prior in  $t = 10$  (first line) and  $t = 40$  (second line).

considering the 5 replications. Figure 4 shows the posterior densities of the sampled coefficients with the Laplace prior in time points  $t = 10$  and  $t = 40$ . We can note the change between these two time points: the third coefficient is equal  $-2$  in  $t = 40$ , so that its posterior densities shows more mass near this value.

We note that the dynamic Laplace prior was slightly superior in terms of RMSE than the other two priors, but the difference is tiny. The dynamic NG with  $a_\tau = 0.5$  have some issues: they are much more volatile than the NMIG prior and the Laplace prior.

### 4.2 Second simulation example

In the Cholesky decomposition, each variable is regressed on its predecessors in a dynamic regression problem, that is,

$$y_{i,t} = \sum_{j=1}^{i-1} \beta_{i,j,t} y_{j,t} + \varepsilon_{i,t}$$

for  $i = 2, \dots, q$ , with  $y_{1,t} = \varepsilon_{1,t}$ . The Cholesky factor is then

$$T_t = (I - B_t),$$

where  $B_t$  is the lower triangular matrix of coefficients for each time  $t$  with zeros in the diagonal, that is, the matrix with entries  $\beta_{2,1,t}, \beta_{3,1,t}, \beta_{3,2,t}, \dots, \beta_{q,1,t}, \dots, \beta_{q,q-1,t}$ . Thus, we have  $q(q - 1)T$  parameters to be estimated.

In this second example, we simulate time-varying coefficients that compose the Cholesky factor  $T_t$  and then apply the spike-and-slab priors on each recursive regression. The simulation is done as follows. We define that the number of time points  $T = 240$  and the number of ordered variables that compose the vector  $y$  is  $q = 10$ . We sample from four possible processes for the time-varying coefficients with the same probability of occurrence. They are:

1. A stationary AR(1) process with autoregressive coefficient  $\phi = 0.98$  and with fixed variance  $\sigma^2 = (1 - \phi)0.15$ , without an intercept term, that is,

$$\beta_{i,j,t} = \phi\beta_{i,j,t-1} + v_{i,j,t},$$

with  $v_{i,j,t} \sim \mathcal{N}(0, \sigma^2)$ .

2. A stationary AR(1) process with autoregressive coefficient  $\phi = 0.98$  and with fixed variance  $\sigma^2 = (1 - \phi)0.15$  until the half of the time points. Then the coefficient is set to zero.
3. A fixed interval process similar to the third coefficient from the first simulated example as follows:

$$\beta_{i,j,t} = \begin{cases} 0, & t \leq T/8; 3T/8 < t \leq 5T/8; t > 7T/8 \\ -0.5, & T/8 < t < 3T/8; 5T/8 < t \leq 7T/8. \end{cases}$$

4. A constant coefficient equal to zero.

In this manner, we want to give a structure to the Cholesky factor, but now allowing for time-varying coefficients. Each coefficient  $\beta_{i,j,t}$  follows one of the four processes: (1) AR(1), (2) AR(1) with zeros, (3) fixed intervals or (4) zeros, which are sampled using equal probabilities. Then we build the 10 time series  $y_1, \dots, y_{10}$  as

$$\begin{aligned} y_{1,t} &= \varepsilon_{1,t} \\ y_{2,t} &= \beta_{2,1,t}y_{1,t} + \varepsilon_{2,t} \\ &\dots \\ y_{10,t} &= \sum_{j=1}^9 \beta_{10,j,t}y_{j,t} + \varepsilon_{10,t}, \end{aligned}$$

for  $t = 1, \dots, 240$  and where  $\varepsilon_{i,t} \sim \mathcal{N}(0, 0.0625), \forall j = 1, \dots, 10$ .

**Table 3** RMSE for the dynamic spike-and-slab priors using the mean and the median of the sampled coefficients—simulated example 2

Prior	RMSE (mean)	RMSE (median)
NMIG	0.2472	0.2863
NG	0.2398	0.2820
Laplace	0.2401	0.2842

The results for the RMSE are shown in Table 3. Once again, we have used the standard prior specification illustrated in Figure 1. The MCMC scheme uses 10,000 simulations with 5000 discarded as burn-in. The hyperparameters were set as follows:  $\nu_0 = r = 0.005$ ,  $\nu_1 = 1$ ,  $a_\tau = 0.5$  (for the NG prior),  $\nu = 25$ ,  $c_0 = 50$ ,  $C_0 = 1.5$ ,  $a_\sigma = 5$ ,  $b_\sigma = 1.5$  and  $\alpha = 1000$  (tuning parameter for Metropolis).

### 4.3 Predicting inflation

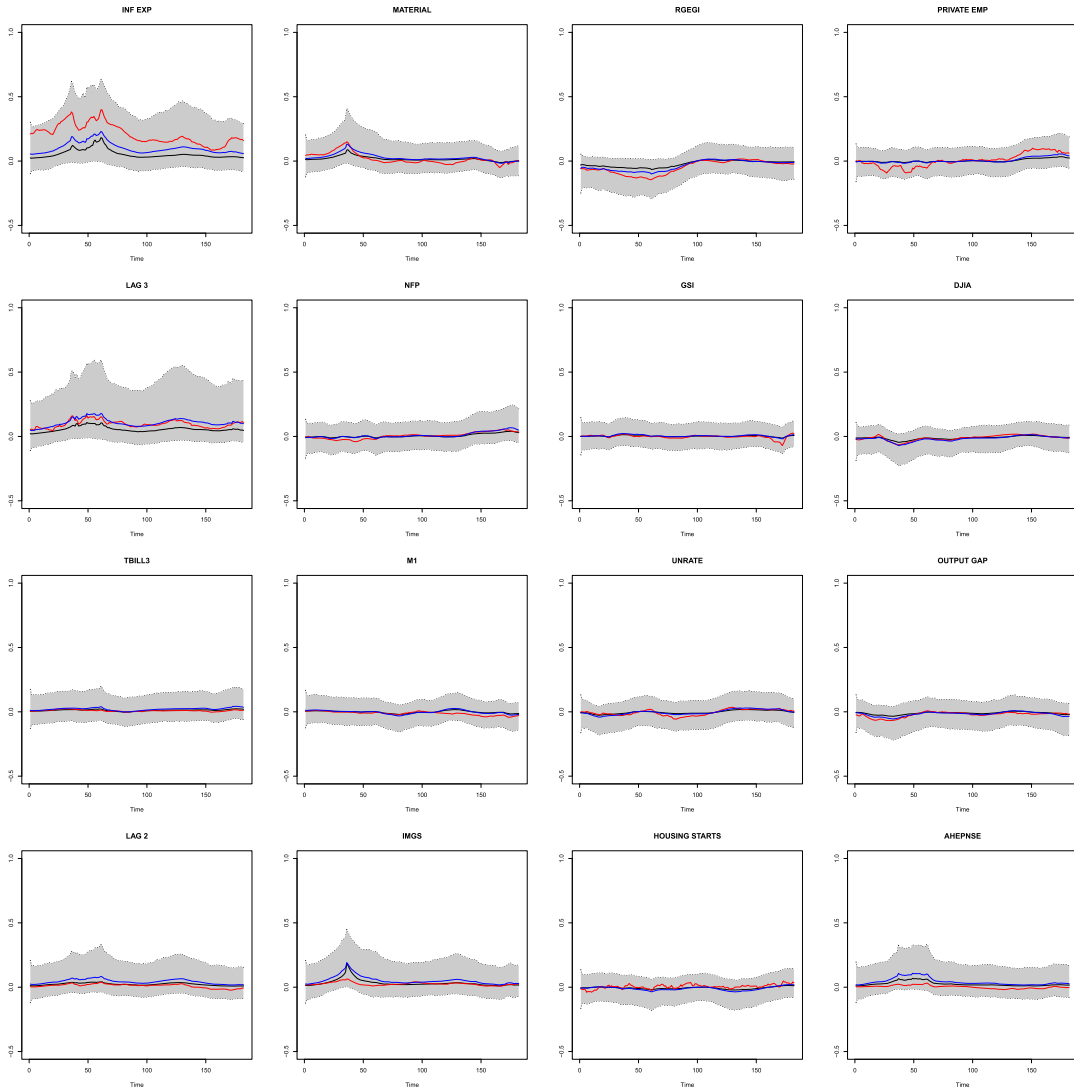
The empirical application uses inflation data obtained from Professor Griffin's research page.<sup>3</sup> We use the inflation data collected from his research page with the independent variable as the US quarterly inflation measure based on the Gross Domestic Product (GDP). The data was obtained from FRED database, Federal Reserve Bank of St. Louis, University of Michigan Consumer Survey database, Federal Reserve Bank of Philadelphia and Institute of Supply Management. The data set includes 31 predictors, from activity and term structure variables to survey forecasts and previous lags. A full description of the 31 explanatory variables can be found in Table 4 of Appendix B. The sample period is from the second quarter of 1965 to first quarter of 2011 with  $T = 182$  observations.

Inflation forecasting is a frequent topic within the shrinkage in time varying parameter models literature and was also the main subject of Belmonte, Koop and Korobilis (2014). The size of the set of potential variables to forecast inflation is huge and, as noted by Kalli and Griffin (2014), this is usually split into four subsets: past inflation forecasts, where the explanatory variables are previous lags of inflation; Phillips curve forecasts, which involve activity variables, such as economic growth rate or output gap, unemployment rate and lagged inflation; forecasts based on variables which are themselves forecasts of asset prices (combination indices), term structures of nominal debt and consumer surveys; and forecasts based on other exogenous variables such as government investment, the number of new private houses.

We applied the three variable selection priors (13), (14) and (15) to the GDP deflator data with the following hyperparameter settings:  $\nu_0 = r = 0.05$ ,  $\nu_1 = 1$ ,  $a_\tau = 0.5$  (for the NG prior),  $\nu = 50$ ,  $c_0 = 50$ ,  $C_0 = .05$ ,  $a_\sigma = 31$ ,  $b_\sigma = 30\hat{\sigma}^2 = 4.22$ , with  $\hat{\sigma}^2 = 0.14$  being the sum of square residuals of the OLS estimate divided by  $(T - 1)$  and  $\alpha = 1000$  (tuning parameter for Metropolis). The previous Beta priors, that is,  $\phi_j \sim \mathcal{B}(77.6, 2.4)$  and for the transition probabilities  $\omega_{j,0,0} \sim \mathcal{B}(77.6, 2.4)$  and  $\omega_{j,1,1} \sim \mathcal{B}(77.6, 2.4)$  were maintained. We ran a total of 20,000 iterations of the MCMC scheme and we discarded 10,000 as a burn-in.

The results (the mean of the coefficients  $\beta_{j,t}$ ) were compared to results from the NGAR process defined Kalli and Griffin (2014); see Figure 5 below. We used the MATLAB code provided by Professor Griffin in his website for the GDP inflation data after standardizing both the response and the predictors in the same way as done by the authors. As Figure 5 indicates, the results are quite similar. Despite its simplicity our proposed method, which is computationally faster, produces competitive results.

<sup>3</sup>Available in <https://www.kent.ac.uk/smsas/personal/jeg28/index.htm>.



**Figure 5** Comparison between the mean of the NGAR model and the dynamic NMIG prior relevances. Mean NGAR=red line; Mean NMIG=blue line; Median NMIG=black line; 95% confidence intervals=grey area.

## 5 Conclusions

We introduce a novel strategy to allow dynamic sparsity, or vertical sparsity, in dynamic regression models, with particular interest in the time-varying Cholesky decomposition set up by Lopes, McCulloch and Tsay (2022). Our scheme allows for time-varying sparsity, based on an extension of spike-and-slab priors for dynamic models using Markov switching auxiliary variables. It is simpler than the existing ones, such as Nakajima and West (2013) and Kalli and Griffin (2014), while maintaining scalability. It also leads to a more efficient MCMC as time-varying latent variables  $\tilde{\beta} = (\tilde{\beta}_1, \dots, \tilde{\beta}_T)$  and  $K = (K_1, \dots, K_T)$  can be sampled jointly by combining a marginal Gibbs step for  $K$  (Gerlach, Carter and Kohn (2000)) with a conditional FFBS step for  $\tilde{\beta}|K$  (Frühwirth-Schnatter (1994)). The alternative ones sample from full conditional distributions, which is notoriously known to lead to slow mixing in dynamic systems.

## Appendix A: Efficient Bayesian inference for dynamic mixtures

We present here in details the dynamic mixture model approach proposed by [Gerlach, Carter and Kohn \(2000\)](#). Consider the univariate Gaussian DLM as

$$\begin{aligned} y_t &= f_t + F_t' \theta_t + \gamma_t u_t, & u_t &\sim \mathcal{N}(0, 1) \\ \theta_t &= G_t \theta_{t-1} + \Gamma v_t, & v_t &\sim \mathcal{N}(0, I), \end{aligned} \quad (17)$$

for  $t = 1, \dots, n$ , where  $\theta_t$  is a  $q$ -dimensional vector of states,  $u_t$  and  $v_t$  are independent and standard Normally distributed, and  $f_t$ ,  $F_t'$ ,  $\gamma_t$ ,  $G_t$ ,  $G_t$  and  $\Gamma_t$  may all depend on the vector Markov  $K_t$  and on a vector of parameters  $\Phi$ . This makes observations  $y_t$  mixture of normals.

Note that  $K_{1:n} = (K_1, \dots, K_n)$  is a sequence of random vectors that are Markov:

$$p(K_t | K_{1:t-1}) = p(K_t | K_{t-1}), \quad t = 2, \dots, n.$$

The sampling scheme proposed generates  $K_t$  from density  $p(K_t | y_{1:n}, K_{s \neq t})$  for  $t = 1, \dots, n$  without conditioning on the states  $\theta_{1:n}$ . The crucial thing is to notice that

$$\begin{aligned} p(K_t | y_{1:n}, K_{s \neq t}) &\propto p(y_{1:n} | K_{1:n}) p(K_t | K_{s \neq t}) \\ &\propto p(y_{t+1:n} | y_{1:t}, K_{1:n}) p(y_t | y_{1:t-1}, K_{1:t}) p(K_t | K_{s \neq t}), \end{aligned} \quad (18)$$

where the dependence on the parameters  $\Phi$  has been suppressed for convenience.

For each value of  $K_t$ , the right-hand side of (18) is evaluated as follows. The term  $p(K_t | K_{s \neq t})$  is obtained from the prior. The term  $p(y_t | y_{1:t-1}, K_{1:t})$  is obtained from  $p(\theta_{t-1} | y_{1:t-1}, K_{1:t-1})$ , that is, from the filtering distribution, using one step of the Kalman filter.

Obtaining the term  $p(y_{t+1:n} | y_{1:t}, K_{1:n})$  is the crucial innovation of the algorithm of [Gerlach, Carter and Kohn \(2000\)](#). See algorithm 1 below for our version of their algorithm. Traditional sampling algorithms use  $n - t + 1$  steps of the Kalman filter given the current values of  $K_{t,n}$  to obtain the term  $p(y_{t+1:n} | y_{1:t}, K_{1:n})$ . Therefore, it requires  $\mathcal{O}(n)$  operations to generate each  $K_t$ , and hence  $\mathcal{O}(n^2)$  operations to generate  $K_{1:n}$ . Nevertheless, in the proposed algorithm the term  $p(y_{t+1:n} | y_{1:t}, K_{1:n})$  is obtained in one step after an initial set of backward recursions. This reduces the number of operations required to generate the complete vector  $K_{1:n}$  to  $\mathcal{O}(n)$ .

Before giving the efficient method for generating  $K_{1:n}$ , we are going to state several preliminary lemmas, whose proofs can be found in the Appendix of [Gerlach, Carter and Kohn \(2000\)](#). All of the lemmas refer to the univariate Gaussian DLM (17).

**Lemma A.1.** *Let  $r_{t+1} = \text{Var}(y_{t+1} | \theta_t, K_{1:t+1})$ . Then the following hold:*

$$\begin{aligned} \mathbb{E}(y_{t+1} | \theta_t, K_{1:t+1}) &= f_{t+1} + F_{t+1}' (G_{t+1} + G_{t+1} \theta_t), \\ r_{t+1} &= F_{t+1}' \Gamma_{t+1} \Gamma_{t+1}' F_{t+1} + \gamma_{t+1}^2, \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}(\theta_{t+1} | \theta_t, y_{t+1}, K_{1:n}) &= a_{t+1} + A_{t+1} \theta_t + B_{t+1} y_{t+1}, \\ \text{Var}(\theta_{t+1} | \theta_t, y_{t+1}, K_{1:n}) &= C_{t+1} C_{t+1}', \end{aligned}$$

where

$$\begin{aligned} a_{t+1} &= (I - B_{t+1} F_{t+1}') G_t - B_{t+1} f_t, \\ A_{t+1} &= (I - B_{t+1} F_{t+1}') G_{t+1}, \\ B_{t+1} &= \Gamma_{t+1} \Gamma_{t+1}' F_{t+1} r_{t+1}^{-1}, \\ C_{t+1} C_{t+1}' &= \Gamma_{t+1} (I - \Gamma_{t+1}' F_{t+1} r_{t+1}^{-1} F_{t+1}' \Gamma_{t+1}) \Gamma_{t+1}, \end{aligned}$$

It is straightforward to factor the expression on the right-hand side of the last equality to get a matrix  $C_{t+1}$  that either is null or has full column rank. Then we can write

$$\theta_{t+1} = a_{t+1} + A_{t+1} + B_{t+1}y_{t+1} + C_{t+1}\xi_{t+1},$$

where  $\xi_{t+1} \sim \mathcal{N}(0, I)$  and is independent of  $\theta_t$  and  $y_{t+1}$ , conditional on  $K_{1:n}$ .

**Lemma A.2.** For  $t = 1, \dots, n-1$ , the density  $p(y_{t+1:n}|\theta_t, K_{1:n})$  is independent of  $K_{1:t}$  and can be expressed as

$$p(y_{t+1:n}|\theta_t, K_{1:n}) \propto \exp\left\{-\frac{1}{2}(\theta_t \Omega_t (\theta_t - 2\mu'_t \theta_t))\right\},$$

where the terms  $\Omega_t$  and  $\mu_t$  are computed recursively starting from

$$\Omega_n = 0, \quad \mu_n = 0.$$

Updating backward, we obtain

$$\Omega_t = A'_{t+1}(\Omega_{t+1} - \Omega_{t+1}C_{t+1}D_{t+1}^{-1}C'_{t+1}\Omega_{t+1})A_{t+1} + G'_{t+1}F_{t+1}r_{t+1}^{-1}F'_{t+1}G_{t+1},$$

$$\begin{aligned} \mu_t &= A'_{t+1}(I - \Omega_{t+1}C_{t+1}D_{t+1}^{-1}C'_{t+1})(\mu_{t+1} - \Omega_{t+1}(a_{t+1} + B_{t+1}y_{t+1})) \\ &\quad + G'_{t+1}F_{t+1}r_{t+1}^{-1}(y_{t+1} - f_{t+1} - F'_{t+1}G_{t+1}), \end{aligned}$$

$$D_{t+1} = C'_{t+1}\Omega_{t+1}C_{t+1} + I.$$

**Lemma A.3.** Let  $m_t = \mathbb{E}(\theta_t|y_{1:t}, K_{1:n})$ ,  $v_t = \text{Var}(\theta_t|y_{1:t}, K_{1:n})$  and  $R_t = \text{Var}(y_t|y_{1:t-1}, K_{1:n})$ . The Kalman filter for the model (17) is given by

$$\begin{aligned} R_t &= F'_t G_t v_{t-1} G'_t F_t + F'_t \Gamma_t \Gamma'_t F_t + \gamma_t^2, \\ m_t &= (I - J_t F'_t)(G_t + G_t m_{t-1}) + J_t (y_t - f_t), \\ v_t &= G_t v_{t-1} G'_t + \Gamma_t \Gamma'_t - J_t J'_t R_t, \end{aligned}$$

where

$$J_t = [G_t v_{t-1} G'_t F_t + \Gamma_t \Gamma'_t F_t] / R_t.$$

The conditional density  $p(y_t|y_{1:t-1}, K_{1:t})$  is such that

$$p(y_t|y_{1:t-1}, K_{1:t}) \propto R_t^{-1} \exp\left\{-\frac{1}{2R_t}(y_t - f_t - F'_t(G_t + G_t m_{t-1}))^2\right\}.$$

We can write  $v_t = T_t T'_t$ , where the matrix  $T_t$  either has full column rank if  $v_t \neq 0$  or is null if  $v_t = 0$ . Conditional on  $K_{1:n}$ , we can express  $\theta_t$  as

$$\theta_t = m_t + T_t \xi_t,$$

where  $\xi_t \sim \mathcal{N}(0, I)$  and is independent of  $y_{1:t}$ .

The next lemma uses Lemma A.3 to efficiently evaluate the factor  $p(y_{t+1:n}|y_{1:t}, K_{1:n})$ .

**Lemma A.4.** Using the results of Lemma A.3, it follows that

$$\begin{aligned} &p(y_{t+1:n}|y_{1:t}, K_{1:n}) \\ &= \int p(y_{t+1:n}|\theta_t, K_{t+1:n})p(\xi_t|K_{1:t})d\xi_t \\ &\propto |T'_t \Omega_t T_t + I|^{-1/2} \exp\left\{-\frac{1}{2}(m'_t \Omega_t m_t - 2\mu'_t m_t - \Phi'_t (T'_t \Omega_t T_t + I)^{-1} \Phi_t)\right\}, \end{aligned}$$

where  $\Phi_t = T'_t(\mu_t - \Omega_t m_t)$ .

The recursion for generating  $K_{1:n}$  in  $\mathcal{O}(n)$  operations is now given.

**Algorithm 1:** The algorithm of Gerlach, Carter and Kohn (2000) for dynamic mixture models:

1. Given the current value of  $K_{1:n}$ , calculate  $\Omega_t$  and  $\mu_t$  for  $t = n - 1, \dots, 1$ , using the recursions in Lemma A.2.
2. Given  $\mathbb{E}(\theta_0)$  and  $\text{Var}(\theta_0)$ , perform the following for  $t = 1, \dots, n$ :
- 3 (a) Obtain  $R_t$ ,  $m_t$  and  $v_t$  from  $m_{t-1}$  and  $v_{t-1}$  as in Lemma A.3;
- 4 (b) Obtain  $p(y_t|y_{1:t-1}, K_{1:t})$  as in Lemma A.3 and  $p(y_{t+1:n}|y_{1:t}, K_{1:n})$  as in Lemma A.4;
- 5 (c) Obtain  $p(K_t|y_{1:n}, K_{s \neq t})$  for all values of  $K_t$  by normalization of
 
$$p(K_t|y_{1:n}, K_{s \neq t}) \propto p(y_{1:n}|K_{1:n})p(K_t|K_{s \neq t})$$

$$\propto p(y_{t+1:n}|y_{1:t}, K_{1:n})p(y_t|y_{1:t-1}, K_{1:t})p(K_t|K_{s \neq t}).$$
- 6 Then draw  $K_t$ .
- 7 (d) Update  $m_t$  and  $v_t$  as in Lemma A.3, using the generated value of  $K_t$ .

## Appendix B: Inflation data

**Table 4** *Inflation Data. Sources: FRED database, Federal Reserve Bank of St. Louis, University of Michigan Consumer Survey database, Federal Reserve Bank of Philadelphia and Institute of Supply Management*

Name	Description
GDP	Difference in logs of real gross domestic product
PCE	Difference in logs of real personal consumption expenditure
GPI	Difference in logs of real gross private investment
RGEGI	Difference in logs of real government consumption expenditure and gross investment
IMGS	Difference in logs of imports of goods and services
NFP	Difference in logs nonfarm payroll
M2	Difference in logs M2 (commercial bank money)
ENERGY	Difference in logs of oil price index
FOOD	Difference in logs of food price index
MATERIALS	Difference in logs of producer price index (PPI) industrial commodities
OUTPUT GAP	Difference in logs of potential GDP level
GS10	Difference in logs of 10 yr Treasury constant maturity rate
GS5	Difference in logs of 5 yr Treasury constant maturity rate
GS3	Difference in logs 3 yr Treasury constant maturity rate
GS1	Difference in logs 1 yr Treasury constant maturity rate
PRIVATE EMPLOYMENT	Log difference in total private employment
PMI MANU	Log difference in PMI-manufacturing index
AHEPNSE	Log difference in average hourly earnings of private nonmanagement employees
DJIA	Log difference in Dow Jones Industrial Average Returns
M1	Log difference in M1 (narrow-commercial bank money)
ISM SDI	Institute for Supply Management (ISM) Supplier Deliveries Inventory
CONSUMER	University of Michigan consumer sentiment (level)
UNRATE	Log of the unemployment rate
TBILL3 3m	Treasury bill rate
TBILL SPREAD	Difference between GS10 and TBILL3
HOUSING STARTS	Private housing (in thousands of units)
INF EXP	University of Michigan inflation expectations (level)
LAG1, LAG2, LAG3, LAG4	The first, second, third and fourth lag

## Acknowledgments

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper. Hedibert Lopes is the corresponding author.

## Funding

The first author was partially supported by FAPESP Grants 2013/00506-1, 2018/04654-9 and 2023/02538-0.

## References

- Belmonte, M. A., Koop, G. and Korobilis, D. (2014). Hierarchical shrinkage in time-varying parameter models. *Journal of Forecasting* **33**, 80–94. [MR3148281 https://doi.org/10.1002/for.2276](https://doi.org/10.1002/for.2276)
- Bitto, A. and Frühwirth-Schnatter, S. (2019). Achieving shrinkage in a time-varying parameter model framework. *Journal of Econometrics* **210**, 75–97. arXiv preprint. Available at [arXiv:1611.01310](https://arxiv.org/abs/1611.01310).
- Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81**, 541–553. [MR1311096 https://doi.org/10.1093/biomet/81.3.541](https://doi.org/10.1093/biomet/81.3.541)
- Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis* **15**, 183–202. [MR1263889 https://doi.org/10.1111/j.1467-9892.1994.tb00184.x](https://doi.org/10.1111/j.1467-9892.1994.tb00184.x)
- Frühwirth-Schnatter, S. and Wagner, H. (2010). Stochastic model specification search for Gaussian and partial non-Gaussian state space models. *Journal of Econometrics* **154**, 85–100. [MR2558953 https://doi.org/10.1016/j.jeconom.2009.07.003](https://doi.org/10.1016/j.jeconom.2009.07.003)
- Frühwirth-Schnatter, S. and Wagner, H. (2011). Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data. *Bayesian Statistics* **9**, 165. [MR3204006 https://doi.org/10.1093/acprof:oso/9780199694587.003.0006](https://doi.org/10.1093/acprof:oso/9780199694587.003.0006)
- George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.
- Gerlach, R., Carter, C. and Kohn, R. (2000). Efficient Bayesian inference for dynamic mixture models. *Journal of the American Statistical Association* **95**, 819–828. [MR1804440 https://doi.org/10.2307/2669465](https://doi.org/10.2307/2669465)
- Griffin, J. E., Brown, P. J., et al (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* **5**, 171–188. [MR2596440 https://doi.org/10.1214/10-BA507](https://doi.org/10.1214/10-BA507)
- Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning 1*. Springer Series in Statistics. Berlin: Springer. [MR1851606 https://doi.org/10.1007/978-0-387-21606-5](https://doi.org/10.1007/978-0-387-21606-5)
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67. [MR4165988 https://doi.org/10.1080/00401706.2020.1742207](https://doi.org/10.1080/00401706.2020.1742207)
- Huber, F., Koop, G. and Onorante, L. (2021). Inducing sparsity and shrinkage in time-varying parameter models. *Journal of Business & Economic Statistics* **39**, 669–683. [MR4272927 https://doi.org/10.1080/07350015.2020.1713796](https://doi.org/10.1080/07350015.2020.1713796)
- Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics* **33**, 730–773. [MR2163158 https://doi.org/10.1214/009053604000001147](https://doi.org/10.1214/009053604000001147)
- Kalli, M. and Griffin, J. E. (2014). Time-varying sparsity in dynamic regression models. *Journal of Econometrics* **178**, 779–793. [MR3144682 https://doi.org/10.1016/j.jeconom.2013.10.012](https://doi.org/10.1016/j.jeconom.2013.10.012)
- Kastner, G. (2019). Sparse Bayesian time-varying covariance estimation in many dimensions. *Journal of Econometrics* **210**, 98–115. [MR3944765 https://doi.org/10.1016/j.jeconom.2018.11.007](https://doi.org/10.1016/j.jeconom.2018.11.007)
- Koop, G. and Korobilis, D. (2023). Bayesian dynamic variable selection in high dimensions. *International Economic Review* **64**, 1047–1074.
- Kowal, D. R., Matteson, D. S. and Ruppert, D. (2019). Dynamic shrinkage processes. *Journal of the Royal Statistical Society, Series B (Methodological)* **81**, 781–804. [MR3997101 https://doi.org/10.1111/rssb.12325](https://doi.org/10.1111/rssb.12325)
- Lopes, H. F., McCulloch, R. E. and Tsay, R. S. (2022). Parsimony inducing priors for large scale state-space models. *Journal of Econometrics* **230**, 39–61. [MR4436489 https://doi.org/10.1016/j.jeconom.2021.11.005](https://doi.org/10.1016/j.jeconom.2021.11.005)
- Nakajima, J. and West, M. (2013). Bayesian analysis of latent threshold dynamic models. *Journal of Business & Economic Statistics* **31**, 151–164. [MR3055329 https://doi.org/10.1080/07350015.2012.747847](https://doi.org/10.1080/07350015.2012.747847)
- Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association* **103**, 681–686. [MR2524001 https://doi.org/10.1198/016214508000000337](https://doi.org/10.1198/016214508000000337)

- Rocková, V. and McAlinn, K. (2021). Dynamic variable selection with spike-and-slab process priors. *Bayesian Analysis* **16**, 233–269. [MR4194280](#) <https://doi.org/10.1214/20-BA1199>
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)* **58**, 267–288. [MR1379242](#)
- West, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74**, 646–648. [MR0909372](#) <https://doi.org/10.1093/biomet/74.3.646>