INTERNATIONAL ECONOMIC REVIEW
Vol. 64, No. 3, August 2023 DOI: 10.1111/iere.12623

BAYESIAN DYNAMIC VARIABLE SELECTION IN HIGH DIMENSIONS*

By Gary Koop and Dimitris Korobilis

University of Strathclyde, Glasgow, United Kingdom; University of Glasgow, Glasgow, United Kingdom

This article addresses the issue of inference in time-varying parameter regression models in the presence of many predictors and develops a novel dynamic variable selection strategy. The proposed variational Bayes dynamic variable selection algorithm allows for assessing at each time period in the sample which predictors are relevant (or not) for forecasting the dependent variable. The algorithm is used to forecast inflation using over 400 macroeconomic, financial, and global predictors, many of which are potentially irrelevant or short-lived. The new methodology is able to ensure parsimonious solutions to this high-dimensional estimation problem, which translate into excellent forecast performance.

1. INTRODUCTION

Regression models that incorporate stochastic variation in parameters have been used by economists at least since the work of Cooley and Prescott (1976). Thirty years later, Granger (2008) argued that time-varying parameter (TVP) models might become the norm in econometric inference due to the fact that (as he illustrated via *White's theorem*) time variation is able to approximate generic forms of nonlinearity in parameters. Indeed, initiated by the unprecedented shocks observed during and after the Global Recession of 2007–09, a large recent literature has established the importance of modeling time variation in the intercept, slopes, and variance of regressions for forecasting economic time series; see Stock and Watson (2007) for a representative example of a parsimonious model using only a stochastic intercept and stochastic volatilities. At the same time, the stylized fact that economic predictors are short-lived—that is, predictors are relevant for forecasting the dependent variable only in short periods of the sample—has emerged in various forecasting problems such as inflation (Koop and Korobilis, 2012), stock returns (Dangl and Halling, 2012), and exchange rates (Byrne et al., 2018). Following these observations, there is no shortage of recent econometric work on methods for penalized estimation of TVP models via shrinkage as well as variable selection methods; see, for example, Belmonte et al. (2014), Bauwens et al. (2015), Bitto and Frühwirth-Schnatter (2019), Giordani and Kohn (2008), Kalli and Griffin (2014), Chan et al. (2012), Korobilis (2021), Kowal et al. (2019), Nakajima and West (2013), Rockova and McAlinn (2021), and Uribe and Lopes (2017).[1]

In this article, we add to this literature by proposing a new dynamic variable selection (DVS) prior and a novel, for the field of economics, Bayesian estimation methodology. In particular, we propose to use variational Bayes (VB) inference to estimate TVP regressions using state-space methods. Variational inference has long been used in data science problems such as large-scale document analysis, computational neuroscience, and computer vision (Blei

*Manuscript received September 2020; revised November 2022.

Please address correspondence to: Dimitris Korobilis, Adam Smith Business School, University of Glasgow, G12 8QQ Glasgow, UK. E-mail: *Dimitris.Korobilis@glasgow.ac.uk*.

[1] In turn, all these papers built on a long tradition in flexible Bayesian modeling of state-space models; see, for example, Gerlach et al. (2000), McCulloch and Tsay (1993), and Shively and Kohn (1997).

et al., 2017). Nevertheless, it is only relatively recently that posterior consistency and other theoretical properties of these methods have been explored by mainstream statisticians (Wang and Blei, 2019). Variational inference is a comprehensive estimation methodology that shares similarities with the Gibbs sampler that many economists traditionally use to estimate TVP models (see, e.g., Stock and Watson, 2007). Similar to the Gibbs sampler, parameter updates are derived for one parameter at a time conditional on all other parameters using an iterative scheme. Unlike the Gibbs sampler, there is no repeated sampling involved and the output of VB is typically the first two moments of the posterior distribution of parameters. Our first task is to introduce this estimation scheme in the context of TVP regressions, and contrast it to existing estimation algorithms used in economics for capturing structural change via TVPs.

Our second contribution lies in the development of a DVS prior that is a conceptually straightforward extension of the static variable selection prior of George and McCulloch (1993). The dynamic extension of their popular variable selection prior tackles the nontrivial econometric problem of allowing some predictor variables to enter the TVP regression only in some time periods. With p predictors and a sample of T time periods, static variable selection involves choosing the "best" among 2^p possible models. In contrast, in DVS, the model space becomes 2^{Tp}. Therefore, for dimensions of p and T typically found in macroeconomic data, the DVS problem can only be approximated using machine learning methods—hence our proposal to adopt VB inference. Is there a need to consider such a high-dimensional problem instead of relying on more parsimonious modeling approaches? The answer is yes, as there is strong support in favor of DVS procedures. Recent empirical evidence suggests that different factors might be driving predictability of economic variables over time; see Rossi (2013) for a thorough review of this idea. By specifying our new prior within a VB framework, we are able to derive an algorithm that is numerically stable and can be extended to much larger p and T than was possible before, whereas DVS ensures that parsimonious solutions to the TVP problem are achieved.[2]

The purpose of the proposed algorithm is prediction and high-dimensional inference. Although we do show using synthetic data that the estimation algorithm is able to do a good job at recovering the true parameters and dynamic probabilities of inclusion of each predictor, it is important to note that VB is computationally faster than alternative algorithms (e.g., Markov chain Monte Carlo (MCMC)) at the cost of introducing some approximation. In general, it has been shown that VB may suffer errors in recovering the posterior variance of the model parameters (Giordano et al., 2018) or lead to general inferential inaccuracies (Frazier et al., 2022). However, there are two major arguments in favor of using VB, especially for forecasting. First, it is well established in macroeconomic forecasting (especially after observing events such as the 2007 global recession, or the 2020 pandemic) that model uncertainty and structural break uncertainty are far more important than parameter uncertainty in determining forecast accuracy. Therefore, even assuming that VB will always misestimate posterior variances, the algorithm we present here is the only one we know of, in all of econometrics, that allows inference with hundreds of TVPs and DVS/shrinkage.[3] Second, in the high-dimensional case, we are interested in (where we have many more predictors than observations) estimation uncertainty will be large even for the most accurate of estimation methods. In MCMC-based estimation, in particular, bias related to repeated sampling can become troubling.[4]

[2] In particular, many of the algorithms cited above, such as Koop and Korobilis (2012), Kalli and Griffin (2014), or Nakajima and West (2013), are unable to scale up to TVP regressions with hundreds of predictors.

[3] In certain fields, such as causal inference, accurate estimation of parameter uncertainty is of paramount importance for testing and prediction. In such settings, one can always combine the benefits of machine learning estimators for fitting high-dimensional data, with those of using an unbiased, efficient estimator. For example, Belloni and Chernozhukov (2013) suggest to use least-squares after performing high-dimensional model selection using a machine learning estimator, in a procedure they name "post-lasso." Similar procedures are trivial to devise using the DVS algorithm proposed in this article.

[4] When using MCMC methods, the bias due to initialization of the chain and the finite number of Monte Carlo samples collected ("transient bias") can be quite large in high-dimensional settings. This is because the larger the di-

Although our main contribution is methodological, on the empirical front, we specify a flexible forecasting model for inflation that incorporates the desirable empirical features of time variation and dynamic shrinkage of coefficients. In a thorough comparison of forecasting models for inflation that summarizes the results of a long literature on this topic, Faust and Wright (2013) argue that time variation and shrinkage of information are indeed key principles that emerge from their results. Faust and Wright (2013) specifically mention time variation in the intercept, but combined with results in Koop and Korobilis (2012) and Rossi (2013) and others, time variation in predictors is also of paramount importance. The principle of shrinkage in Faust and Wright (2013) refers to the fact that parsimonious models tend to do much better in forecasting inflation compared to information-rich models. Therefore, a major breakthrough of our empirical approach is to show that a TVP forecasting model of inflation that includes 400+ macroeconomic, financial, and global predictors dominates in most instances, especially when forecasting four and eight quarters ahead, while providing very competitive forecasts in the shorter run. These results are robust across four measures of inflation using U.S. data for the period 1960Q1–2021Q4 and a wide range of competing forecasting models.[5] Although we also explore more parsimonious TVP regressions that feature factors, the specification featuring all predictors has more than 100,000 parameters (442 coefficients that vary across $T = 231$ quarters), making it probably the least parsimonious specification ever considered in the literature of forecasting inflation. Consequently, the excellent forecasting performance even of the least parsimonious versions of our model serves as a verification of the excellent ability of the DVS prior to find parsimonious solutions in high-dimensional settings.

The remainder of the article proceeds as follows: Section 2 introduces the basic principles of VB inference for approximating intractable posteriors, and applies these principles to the problem of estimating a simplified TVP regression model. Section 3 introduces the novel modeling assumptions, namely, DVS and stochastic volatility, and derives an estimation algorithm within the VB framework. In Section 4, we apply the new methodology to the problem of forecasting U.S. inflation using TVP regressions with many predictors. Section 5 concludes the article.

2. VARIATIONAL BAYES INFERENCE IN STATE-SPACE MODELS

By virtue of the fact that VB is not an established estimation methodology in econometrics, we first provide a generic discussion of VB as an approximation methodology for intractable posterior distributions. Detailed reviews of VB can be found in Blei et al. (2017) and Ormerod and Wand (2010), among several others. VB estimation of state-space models is described in the monographs of Beal (2003) and Šmídl and Quinn (2006), as well as research papers such as Beal and Ghahramani (2003), Särkkä and Nummenmaa (2009), Tran et al. (2017), and Wang et al. (2016).

2.1. *Basics of Variational Bayes.*

Consider data y, latent variables s, and (latent) parameters θ. Our interest lies in TVP models that admit a state-space form. Hence, s represents unobserved state variables, such as time-varying regression coefficients and time-varying measurement error variances, and θ represents all other parameters, such as the error covariances in the state equation. The joint posterior of interest is $p(s, \theta|y)$ with associated marginal likelihood $p(y)$ and joint density of data and parameters $p(y, s, \theta)$. When the joint posterior is

mension of the data, the longer the Monte Carlo samples that are needed for inference. Doubling the number of samples collected can only reduce the Monte Carlo standard error by a factor of $\sqrt{2}$. As a result, in high dimensions, approximate inference algorithms may be preferred to MCMC-based posterior algorithms as more computationally reasonable alternatives; see the excellent discussion of these issues in Angelino et al. (2016).

[5] The list of competing models includes simple autoregression, various factor models, parsimonious TVP and structural breaks regressions, alternative machine learning and shrinkage estimators for regression and classification, and flexible nonparametric models.

complex and computationally intractable, we want to find an approximating class of densities $q(s, \theta|y)$ that belongs to a family $\mathcal{F}$ of simpler distributions defined over the parameter space spanned by s, θ. The main idea behind VB inference is to make this approximating posterior distribution $q(s, \theta|y)$ as close as possible to $p(s, \theta|y)$, where "distance" is measured with the Kullback–Leibler divergence[6]

$$KL(q||p) = \int q(s, \theta|y) \log \left\{ \frac{q(s, \theta|y)}{p(s, \theta|y)} \right\} \mathrm{d}s\mathrm{d}\theta. \tag{1}$$

That is, the aim is to find the optimal $q^\star(s, \theta|y)$ that solves

$$q^\star(s, \theta|y) = \underset{q(s,\theta|y)\in\mathcal{F}}{\arg\min}\, KL(q||p). \tag{2}$$

Insight for why $KL(q||p)$ is a desirable distance metric arises from a simple rearrangement involving the log of the marginal likelihood (Ormerod and Wand, 2010, p. 142) where it can be shown that

$$\log p(y) = \log p(y) \int p(s, \theta|y)\mathrm{d}s\mathrm{d}\theta = \int p(s, \theta|y) \log p(y)\mathrm{d}s\mathrm{d}\theta \tag{3}$$

$$= \int q(s, \theta|y) \log \left\{ \frac{p(y, s, \theta)/q(s, \theta|y)}{p(s, \theta|y)/q(s, \theta|y)} \right\} \mathrm{d}s\mathrm{d}\theta \tag{4}$$

$$= \int q(s, \theta|y) \log \left\{ \frac{p(y, s, \theta)}{q(s, \theta|y)} \right\} \mathrm{d}s\mathrm{d}\theta + KL(q||p). \tag{5}$$

Because $KL(q||p)$ is nonnegative (it is exactly zero when $q(s, \theta|y) = p(s, \theta|y)$), the quantity

$$\mathcal{G}(q(s, \theta|y)) = \exp\left[\int q(s, \theta|y) \log \left\{ \frac{p(y, s, \theta)}{q(s, \theta|y)} \right\} ds d\theta \right] \equiv \exp\left[\mathbb{E}_{q(s,\theta|y)} (\log (p(y, s, \theta)) - \log (q(s, \theta|y))) \right], \tag{6}$$

becomes a lower bound for the marginal likelihood $p(y)$.[7] The function $\mathcal{G}(q(s, \theta|y))$ is known as the evidence lower bound (ELBO). Therefore, instead of minimizing the objective function $KL(q||p)$ (which cannot be evaluated), we can find an approximating density $q^\star(s, \theta|y)$ that maximizes the marginal data density $p(y)$ by maximizing the ELBO. We emphasize that $\mathcal{G}$ is a functional on the distribution $q(s, \theta|y)$. As a result, the ELBO can be maximized iteratively using calculus of variations.

If we additionally assume the so-called (in Physics) *mean field* factorization of the form $q(s, \theta|y) = q(\theta|y)q(s|y)$, it can be shown[8] that the optimal choices for $q(s|y)$ and $q(\theta|y)$ are

$$q(s|y) \propto \exp\left[\int q(\theta|y) \log p(s|y, \theta)\mathrm{d}\theta \right] \equiv \exp\left[\mathbb{E}_{q(\theta|y)} (\log p(s|y, \theta)) \right], \tag{7}$$

$$q(\theta|y) \propto \exp\left[\int q(s|y) \log p(\theta|y, s)\mathrm{d}s \right] \equiv \exp\left[\mathbb{E}_{q(s|y)} (\log p(\theta|y, s)) \right]. \tag{8}$$

The first expression denotes the expectation over $q(\theta|y)$ of the conditional posterior for s, and the second expression denotes the expectation over $q(s|y)$ of the conditional posterior for θ.

[6] For notational simplicity, we henceforth abbreviate multiple integrals using a single integration symbol.

[7] In the following, we denote as $\mathbb{E}_{q(\bullet)}$ the expectation w.r.t. to a function $q(\bullet)$.

[8] A formal and thorough derivation of these ideas is given in the excellent monograph of Šmídl and Quinn (2006); see Theorem 3.1 and subsequent results.

Because $q(\theta|y)$ is a function of $q(s|y)$, and vice versa, the above quantities can be approximated iteratively instead of relying on more computationally expensive numerical optimization techniques. Given an initial guess regarding the values of (θ, s), VB algorithms iterate over these two quantities until $\mathcal{G}(q(s, \theta|y))$ has reached a maximum. Due to similarities with the expectation-maximization (EM) algorithm of Dempster et al. (1977), this iterative procedure in its general form is sometimes referred to as the *variational Bayesian EM (VB-EM)* algorithm; see Beal and Ghahramani (2003). It is also worth noting the relationship with Gibbs sampling. Similar to Gibbs sampling, Equations (7) and (8) involve the full conditional posterior distributions. However, the VB-EM algorithm does not repeatedly simulate from posterior conditionals using Monte Carlo, which makes it computationally faster than existing Gibbs sampling approaches. Finally, as is the case with all posterior sampling methods, VB inference using the integrals above is simplified if the complete data likelihood belongs to the exponential family of distributions, and the priors on s and θ are conjugate; see Blei et al. (2017) for a detailed discussion.

As an illustration of the ideas above, the next subsection provides a step-by-step derivation of a VB algorithm in a benchmark TVP regression. Note that a key assumption in all our derivations is the mean-field factorization of the posterior distribution. Giordano et al. (2018) show that mean-field VB inference may misestimate posterior variances, even if posterior mean estimates are accurate. Nevertheless, when the main purpose of statistical inference is prediction, such estimation error is not relevant as long as out-of-sample performance is good in a mean square error sense. Even when one is interested in causal inference (e.g., estimating the effect of a treatment, or obtaining an impulse response from time-series data) where parameter uncertainty is important for testing, there exist ways to utilize biased or approximate machine learning estimators in a meaningful way (see discussion in footnote 3).

2.2. *VB Estimation of a Simple TVP Regression Model.* Before collecting all building blocks of our proposed methodology, we outline a VB algorithm for the univariate TVP regression with time-invariant measurement and state error variances. This simplified model is of the form

$$y_t = \mathbf{x}_t \boldsymbol{\beta}_t + \varepsilon_t, \tag{9}$$

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \tag{10}$$

where y_t is the time t scalar value of the dependent variable, $t = 1, .., T$, $\mathbf{x}_t$ is a $1 \times p$ vector of exogenous predictors and lagged dependent variables, $\varepsilon_t \sim N(0, \sigma^2)$, $\boldsymbol{\eta}_t \sim N(0, \boldsymbol{W})$ with $\boldsymbol{W} = diag(w_1, \ldots, w_p)$ a $p \times p$ diagonal matrix,[9] and $\boldsymbol{w} = [w_1, \ldots, w_p]'$ a $p \times 1$ vector. In likelihood-based analysis of state-space models, inference is simplified by assuming that ε_t and $\boldsymbol{\eta}_t$ are independent of one another and we adopt this assumption here. Finally, we use a notational convention where j, t subscripts denote the jth element of a time-varying state variable, or parameter, observed only at time t, whereas $1 : t$ subscripts denote all the observations of a state variable from period 1 up to period t.

The model in Equations (9) and (10) has unknown parameters $(\boldsymbol{\beta}_{1:T}, \sigma^2, \boldsymbol{w})$. The prior for $\boldsymbol{\beta}_{1:T}$ is provided by the state equation (10), with the additional assumption of a $t = 0$ initial condition of the form $p(\boldsymbol{\beta}_0) = N(\boldsymbol{m}_0, \boldsymbol{P}_0)$, whereas for $\sigma^2, w_1, \ldots, w_p$, we specify conjugate inverse gamma priors. Therefore, the joint prior can be represented as

[9] By restricting $\boldsymbol{W}$ not to be a full covariance matrix, coefficients β_{it} and β_{jt} are uncorrelated a posteriori for $i \neq j$, which might not seem like an empirically plausible assumption. However, allowing for cross-correlation in the state vector $\boldsymbol{\beta}_t$ may result in counterproductive increases in estimation uncertainty, with this problem being significantly more pronounced in higher dimensions. A diagonal $\boldsymbol{W}$ allows for a more parsimonious econometric specification, less cumbersome derivations of posterior distributions, and faster and numerically stable computation; see also Belmonte et al. (2014), Bitto and Frühwirth-Schnatter (2019), and Rockova and McAlinn (2021) who adopt a similar assumption.

$$p(\boldsymbol{\beta}_{1:T}, \sigma^2, \boldsymbol{w}) = p(\boldsymbol{\beta}_{1:T}|\boldsymbol{w}) \times p(\sigma^2) \times p(\boldsymbol{w}) \tag{11}$$

$$= \prod_{t=1}^{T} p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, \boldsymbol{w}) \times p(\sigma^2) \times \prod_{j=1}^{p} p(w_j) \tag{12}$$

$$= \prod_{t=1}^{T} N(\boldsymbol{\beta}_{t-1}, \boldsymbol{W})[Gamma(a_0, b_0)]^{-1} \prod_{j=1}^{p} [Gamma(c_{j,0}, d_{j,0})]^{-1}. \tag{13}$$

The Bayesian posterior is proportional to the product of this prior with the likelihood (measurement equation), that is, it has the form

$$p(\boldsymbol{\beta}_{1:T}, \sigma^2, \boldsymbol{w}|\boldsymbol{y}_{1:T}) \propto p(\sigma^2)\left[\prod_{j=1}^{p} p(w_j)\right] p(\boldsymbol{\beta}_0)\left[\prod_{t=1}^{T} p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, \boldsymbol{w}) p(\boldsymbol{y}_t|\boldsymbol{\beta}_t, \sigma^2)\right]. \tag{14}$$

This joint posterior of all model parameters is a complex product of densities that is not analytically tractable. Nevertheless, the conditional posteriors are easier to derive, as in this case, we assume that all parameters (other than the parameter of interest) are known and their prior distributions become normalizing constants. Under certain conditions, simulating sequentially samples from the conditional posteriors is equivalent to samples from the joint posterior. Therefore, it is not surprising that for TVP and general linear state-space models, the Gibbs sampler is fairly straightforward to apply; see Stock and Watson (2007) for an application of this algorithm. However, repeated sampling from high-dimensional posteriors can be both computationally cumbersome and numerically inefficient (e.g., if covariates are highly correlated), and for that reason, the Gibbs sampler is a sensible choice for small to medium-dimensional problems.

Here, we follow a different strategy and define the class of approximating functions $q(\boldsymbol{\beta}_{1:T}, \sigma^2, \boldsymbol{w}|\boldsymbol{y}_{1:T})$. Among all possible functions $q(\boldsymbol{\beta}_{1:T}, \sigma^2, \boldsymbol{w}|\boldsymbol{y}_{1:T})$, we want to find the one that has hyperparameters that minimize the relative entropy with the true posterior. Following the discussion earlier in this section, this problem is equivalent to maximizing the ELBO of the log-marginal likelihood, that is, it is the solution to

$$q^\star(\boldsymbol{\beta}_{1:T}, \sigma^2, \boldsymbol{w}|\boldsymbol{y}_{1:T}) = \underset{q(\boldsymbol{\beta}_{1:T}, \sigma^2, \boldsymbol{w}|\boldsymbol{y}_{1:T})}{\arg\max} \int q(\boldsymbol{\beta}_{1:T}, \sigma^2, \boldsymbol{w}|\boldsymbol{y}_{1:T}) \log\left(\frac{q(\boldsymbol{\beta}_{1:T}, \sigma^2, \boldsymbol{w}|\boldsymbol{y}_{1:T})}{p(\boldsymbol{\beta}_{1:T}, \sigma^2, \boldsymbol{w}|\boldsymbol{y}_{1:T})}\right). \tag{15}$$

This maximization problem is simplified once we assume the mean field factorization of the form $q(\boldsymbol{\beta}_{1:T}, \sigma^2, \boldsymbol{w}|\boldsymbol{y}_{1:T}) = q(\boldsymbol{\beta}_{1:T}|\boldsymbol{y}_{1:T})q(\sigma^2|\boldsymbol{y}_{1:T}) \prod_j q(w_j|\boldsymbol{y}_{1:T})$ for all variational densities. As a result, using variational calculus (Šmídl and Quinn, 2006), we can show that the ELBO is maximized by iterating through the following recursions:

$$q(\boldsymbol{\beta}_{1:T}|\boldsymbol{y}_{1:T}) \propto \exp\left(\int \log p(\boldsymbol{\beta}_{1:T}, \sigma^2, \boldsymbol{w}|\boldsymbol{y}_{1:T}) q(\sigma^2|\boldsymbol{y}_{1:T}) \prod_j q(w_j|\boldsymbol{y}_{1:T}) d\sigma^2 d\boldsymbol{w}\right), \tag{16}$$

$$q(\sigma^2|\boldsymbol{y}_{1:T}) \propto \exp\left(\int \log p(\boldsymbol{\beta}_{1:T}, \sigma^2, \boldsymbol{w}|\boldsymbol{y}_{1:T}) q(\boldsymbol{\beta}_{1:T}|\boldsymbol{y}_{1:T}) \prod_j q(w_j|\boldsymbol{y}_{1:T}) d\boldsymbol{\beta}_{1:T} d\boldsymbol{w}\right), \tag{17}$$

$$q(w_j|\boldsymbol{y}_{1:T}) \propto \exp\left(\int \log p(\boldsymbol{\beta}_{1:T}, \sigma^2, \boldsymbol{w}|\boldsymbol{y}_{1:T}) q(\boldsymbol{\beta}_{1:T}|\boldsymbol{y}_{1:T}) q(\sigma^2|\boldsymbol{y}_{1:T}) d\boldsymbol{\beta}_{1:T} d\sigma^2\right), \; j = 1, \ldots, p. \tag{18}$$

The above formulas become equalities after the addition of a normalizing constant. The integrals in these formulas are expectations of the joint posterior w.r.t. the variational posterior

of all other parameters. Therefore, if we replace the log joint posterior with its expression in Equation (14), we can achieve further simplifications by noting that expectations are taken w.r.t. to uncertainty over all other parameters being integrated out. This fact leads to simplifications similar to the ones met in the Gibbs sampler, such that we can rewrite the variational formulas above in terms of conditional posteriors (see also Equations (7)–(8) of the previous subsection). For example, Equation (16) can be written as

$$q(\boldsymbol{\beta}_{1:T}|\boldsymbol{y}_{1:T}) \propto \exp\left[\mathbb{E}_{q(\sigma^2|\boldsymbol{y}_{1:T})}\left(\log\prod_{t=1}^{T} p(y_t|\boldsymbol{\beta}_t, \sigma^2)\right) + \mathbb{E}_{q(\boldsymbol{w}|\boldsymbol{y}_{1:T})}\left(\log\prod_{t=1}^{T} p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, \boldsymbol{w})\right)\right]. \tag{19}$$

This simplification occurs because, w.r.t. $q(\sigma^2|\boldsymbol{y}_{1:T})$ and $q(\boldsymbol{w}|\boldsymbol{y}_{1:T})$, many of the densities involved in the product in Equations (14) become normalizing constants. Additionally, notice in the expression above that we are adding to Gaussian log-kernels and then taking their exponential, which will result in $q(\boldsymbol{\beta}_{1:T}|\boldsymbol{y}_{1:T})$ being Gaussian. This result shows more clearly why finding the variational posterior is trivial only when likelihoods and priors are of the exponential form, but for more complex densities, calculation becomes nontrivial or impossible. Finally, we can further simplify Equations (17)–(18) that lead in these densities being inverse gamma. Detailed derivations for general state-space models can be found in Beal (2003, Chapter 5) and Särkkä and Nummenmaa (2009), among numerous other data science sources.

Algorithm 1 provides pseudocode for the basic VB estimation problem described in this section, without assuming either a (dynamic) variable selection prior, or stochastic volatility in the measurement equation. The algorithm has to iterate until convergence, typically until the value of the ELBO has reached a maximum and does not update substantially between two consecutive iterations. [10] Following Equation (6), the ELBO for the example model of this subsection can be written as

$$ELBO \propto \mathbb{E}\big(\log\big(p(\boldsymbol{y}_{1:T}, \boldsymbol{\beta}_{1:T}, \boldsymbol{\sigma}^2, \boldsymbol{w})\big) - \log\big(q(\boldsymbol{\beta}_{1:T}, \boldsymbol{\sigma}^2, \boldsymbol{w}|\boldsymbol{y}_{1:T})\big)\big) \tag{20}$$

$$\begin{aligned} &= \mathbb{E}\big(\log p(\boldsymbol{y}_{1:T}|\boldsymbol{\beta}_{1:T}, \sigma^2)\big) + \mathbb{E}\big(\log p(\boldsymbol{\beta}_{1:T}, \sigma^2, \boldsymbol{w})\big) - \mathbb{E}\big(\log q(\boldsymbol{\beta}_{1:T}, \boldsymbol{\sigma}^2, \boldsymbol{w}|\boldsymbol{y}_{1:T})\big) \\ &= \mathbb{E}\big(\log p(\boldsymbol{y}_{1:T}|\boldsymbol{\beta}_{1:T}, \sigma^2)\big) + \mathbb{E}(\log p(\boldsymbol{\beta}_{1:T}|w)) + \mathbb{E}\big(\log p(\sigma^2)\big) + \mathbb{E}(\log p(\boldsymbol{w})) \end{aligned} \tag{21}$$

$$-\mathbb{E}(\log q(\boldsymbol{\beta}_{1:T}|\boldsymbol{y}_{1:T})) - \mathbb{E}\big(\log q(\sigma^2|\boldsymbol{y}_{1:T})\big) - \mathbb{E}(\log q(w_j|\boldsymbol{y}_{1:T})), \tag{22}$$

that is, the ELBO is simply the sum of the expected log likelihood and the expected log prior densities, minus the expected log variational posterior (also known as variational entropy). Exactly because both the prior and posterior can factorize (due to the mean field approximation), we can simplify Equation (21) into Equation (22) whose terms are expectations of much simpler densities. Note that all expectations $\mathbb{E}$ are with respect to the components of $q(\boldsymbol{\beta}_{1:T}, \boldsymbol{\sigma}^2, \boldsymbol{w}|\boldsymbol{y}_{1:T})$ and this conditioning is only omitted for the sake of notational simplicity. Therefore, all we have to do is to evaluate these log densities at the posterior mean of the parameters $\boldsymbol{\beta}_{1:T}, \sigma^2, \boldsymbol{w}$ achieved in each iteration. Since these are simple densities (mainly normal and gamma), lengthy, detailed derivations are omitted.

[10] Of course, as with any EM-type iterative algorithm, there are no guarantees that the criterion function (ELBO) will always improve in each iteration, or that a global maximum will be reached. In such cases, we follow much of the literature in machine learning and set a maximum number of iterations (which in all our computations is equal to 200).

ALGORITHM 1 VARIATIONAL BAYES ALGORITHM FOR TVP REGRESSION WITH CONSTANT MEASUREMENT ERROR VARIANCE

1: Choose values of hyperparameters $\boldsymbol{m}_0, \boldsymbol{P}_0, a_0, b_0, c_{j,0}, d_{j,0}$ for $j = 1, ..., p$. Set $r = 1$ and initialize $\sigma^{2(0)}, \boldsymbol{W}^{(0)}$.

2: **while** $\| \mathcal{G}\left(q(\boldsymbol{\beta}^{(r)}, \sigma^{2(r)}, \boldsymbol{w}^{(r)} | \boldsymbol{y}\right) - \mathcal{G}\left(q(\boldsymbol{\beta}^{(r-1)}, \sigma^{2(r-1)}, \boldsymbol{w}^{(r-1)} | \boldsymbol{y}\right) \| \rightarrow 0$ **do**

3: **Step 1:** Approximate, $\forall\ t = 1, ..., T$, the posterior

$$q^{(r)}\left(\boldsymbol{\beta}_t | \boldsymbol{y}_{1:T}\right) \sim N\left(\boldsymbol{m}_t^{(r)}, \boldsymbol{P}_t^{(r)}\right)$$

where $\boldsymbol{m}_t^{(r)}, \boldsymbol{P}_t^{(r)}\ \forall\ t = 1, ..., T$, are approximated using the forward covariance Kalman filter and backward information smoother.

4: **Step 2:** Approximate the posterior

$$q^{(r)}\left(\sigma^{-2} | \boldsymbol{y}_{1:T}\right) \sim G\left(a^{(r)}, b^{(r)}\right)$$

where $a^{(r)} = a_0 + T/2$, $b^{(r)} = b_0 + R/2$ and $R = \sum_{t=1}^{T}\left[\left(y_t - \boldsymbol{x}_t \boldsymbol{m}_t^{(r)}\right)^2 + \boldsymbol{x}_t \boldsymbol{P}_t^{(r)} \boldsymbol{x}_t'\right]$. Set $\sigma^{2(r)} = b^{(r)}/a^{(r)}$.

5: **Step 3:** Approximate, $\forall\ j = 1, ..., p$, the posterior

$$q^{(r)}\left(w_j^{-1} | \boldsymbol{y}_{1:T}\right) \sim G\left(c_j^{(r)}, d_j^{(r)}\right)$$

where $c_j^{(r)} = c_{j,0} + T/2$, $d_j^{(r)} = d_{j,0} + \boldsymbol{D}_{jj}/2$, and $\boldsymbol{D}_{jj}$ is the j^{th} diagonal element of the matrix $\boldsymbol{D} = \sum_{t=1}^{T}(\boldsymbol{m}_t^{(r)} - \boldsymbol{m}_{t-1}^{(r)})'(\boldsymbol{m}_t^{(r)} - \boldsymbol{m}_{t-1}^{(r)}) + \boldsymbol{P}_t^{(r)} + \boldsymbol{P}_{t-1}^{(r)}$. Set $\boldsymbol{W}^{(r)} = diag\left(d_1^{(r)}/c_1^{(r)}, ..., d_p^{(r)}/c_p^{(r)}\right)$.

6: $r = r + 1$

7: **end while**

8: Upon convergence set $q^{\star}\left(\boldsymbol{\beta}_{1:T}, \sigma^2, \boldsymbol{w} | \boldsymbol{y}_{1:T}\right) = q^{(r)}\left(\boldsymbol{\beta}_{1:T} | \boldsymbol{y}_{1:T}\right) \times q^{(r)}\left(\sigma^2 | \boldsymbol{y}_{1:T}\right) \times \prod_{j=1}^{p} q^{(r)}\left(w_j | \boldsymbol{y}_{1:T}\right)$ using the parameters $(\boldsymbol{m}_{1:T}^{(r)}, \boldsymbol{P}_{1:T}^{(r)}, a^{(r)}, b^{(r)}, \boldsymbol{c}_{1:p}^{(r)}, \boldsymbol{d}_{1:p}^{(r)})$ obtained during the last iteration of the *while* loop.

3. VARIATIONAL BAYES INFERENCE IN HIGH-DIMENSIONAL TVP REGRESSIONS

We rewrite for convenience the univariate TVP model

$$y_t = \mathbf{x}_t \boldsymbol{\beta}_t + \varepsilon_t, \tag{23}$$

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \tag{24}$$

where we instead assume that $\varepsilon_t \sim N(0, \sigma_t^2)$ with σ_t^2 a stochastic (time-varying) variance parameter, and $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \boldsymbol{W}_t)$ where $\boldsymbol{W}_t = diag(\boldsymbol{w}_t) = diag(w_{1,t}, \ldots, w_{p,t})$ is a $p \times p$ diagonal matrix. The initial condition (time $t = 0$ prior) of this model is again of the form

$$p\left(\boldsymbol{\beta}_0, \sigma_0^2, \boldsymbol{w}_0\right) = N(\boldsymbol{m}_0, \boldsymbol{P}_0) Gamma(a_0, b_0)^{-1} \prod_j Gamma(c_{j,0}, b_{j,0})^{-1}. \tag{25}$$

As we explain in detail Subsection 4.2, it is trivial to define default noninformative choices for $\boldsymbol{m}_0, \boldsymbol{P}_0, a_0, b_0$, but more care needs to be exercised in the selection of $c_{j,0}, b_{j,0}$ because $\boldsymbol{W}_t$ is high-dimensional; see also the discussion in Amir-Ahmadi et al. (2020).

Introducing time variation in the measurement error variance is important for macroeconomic forecasting, as this assumption allows to account for all the large shocks observed throughout the sample as well as shocks that can possibly occur out-of-sample. Introducing time variation in the state equation covariance is also important for the type of dynamic prior we aim to establish in this section. The assumption that $w_{j,t}$ varies over time, in turn, allows

for an adaptive pattern of time variation in $\beta_{j,t}$. That is, while $\beta_{j,t}$ is very persistent by being centered around $\beta_{j,t-1}$, in the presence of a large, sudden shock at time t the term $\boldsymbol{\eta}_t$ can become temporarily large, thus allowing $\beta_{j,t}$ to adapt to a completely new state. All other assumptions from Subsection 2.2 are maintained. However, now we are also particularly interested in the case where the number of predictors p is large, possibly $p \gg T$. For that reason, we introduce a DVS prior for $\boldsymbol{\beta}_t$ that provides a regularized posterior and prevents overparameterization.

3.1. *Dynamic Variable Selection.* The core ingredient of our modeling approach is a dynamic variable/model selection strategy. We specify a DVS prior that extends the "static" variable selection prior of George and McCulloch (1993) and is of the form

$$\beta_{j,t}|\gamma_{j,t}, \tau_{j,t}^2 \sim (1-\gamma_{j,t})N\left(0, \underline{c} \times \tau_{j,t}^2\right) + \gamma_{j,t}N\left(0, \tau_{j,t}^2\right), \tag{26}$$

$$\gamma_{j,t}|\pi_{0,t} \sim Bernoulli\left(\pi_{0,t}\right), \tag{27}$$

$$\frac{1}{\tau_{j,t}^2} \sim Gamma(g_0, h_0), \tag{28}$$

$$\pi_{0,t} \sim Beta(1,1), \tag{29}$$

for $j = 1, \ldots, p$, where $\underline{c}$, g_0 and h_0 are fixed prior hyperparameters. Variable selection principles require us to set $\underline{c} \to 0$, such that the first component in the prior for $\beta_{j,t}$ shrinks the posterior toward zero, whereas the second component has variance $\tau_{j,t}^2$ which is "large enough" to allow for unrestricted estimation. The choice between the two components in the prior for $\beta_{j,t}$ is governed by the random variable $\gamma_{j,t}$ that is distributed Bernoulli and takes values either zero or one. If $\gamma_{j,t} = 1$ the prior for $\beta_{j,t}$ has a normal prior with zero mean and variance $\tau_{j,t}^2$, whereas if $\gamma_{j,t} = 0$ the prior variance becomes $\underline{c}\tau_{j,t}^2$.

Early papers such as George and McCulloch (1993) give very broad guidelines on choosing values for $\underline{c}$ and $\tau_{j,t}^2$ such that the first component in Equation (26) has small enough variance (to force shrinkage) and the second component has large enough variance (to allow unrestricted estimation). More recently, Narisetty and He (2014) show that selecting and fixing the prior variances of such mixture priors could, as T and p grow, lead to model selection inconsistency. The authors suggest to specify these parameters to be certain deterministic functions of the data dimensions T and p. In our case, we do fix $\underline{c} = 10^{-4}$ such that the first component has always smaller variance, but we assume $(\tau_{j,t}^2)^{-1}$ is a random variable that has a gamma prior. That way this parameter is always updated by the information in the data likelihood. The choice of a gamma prior for $(\tau_{j,t}^2)^{-1}$ implies that the marginal prior for $\beta_{j,t}$ is a mixture of leptokurtic Student's T distributions whose components could tend to shrink $\beta_{j,t}$ toward zero, regardless of whether $\gamma_{j,t}$ is zero or one. Therefore, the proposed prior is able to find patterns of dynamic sparsity as well as impose dynamic shrinkage in TVPs, a property that is very desirable in high-dimensional settings.[11]

[11] In signal processing a signal (regression coefficient vector) is typically sparse by default, that is, the researcher knows a-priori to expect that estimates of several coefficients will tend to be exactly zero. In economics, the sparsity assumption might not be empirically founded in certain settings; see the discussion in Giannone et al. (2017). In such cases, a dense model may be preferred, that is, a model where all predictors are relevant with varying weights. Although factor models and principal components have been used widely to model dense models in macroeconomics, shrinkage methods are also quite reliable for this task. In particular, we note the result in De Mol et al. (2008) that forecasts from Bayesian shrinkage are highly correlated to forecasts from principal components.

Finally, it becomes apparent that under this variable selection prior setting, $\widehat{\pi}_{0,t} = \mathbb{E}(p(\pi_{0,t})) = \frac{1}{2}$ is the time t prior mean probability of inclusion of all predictors in the TVP regression, whereas the quantity $\widetilde{\pi}_{j,t} = \mathbb{E}(p(\gamma_{j,t}|\boldsymbol{y}_{1:T}))$ is the posterior mean probability of inclusion in the regression of predictor j at time period t, simply referred to as the *posterior inclusion probability (PIP)*. Due to the fact that all of the hyperparameters $\boldsymbol{\gamma}, \pi$ and $\boldsymbol{\tau}^2$ are time-varying, our prior allows to obtain time-varying PIPs whose interpretation extends this of PIPs in constant parameter settings, such as the one in George and McCulloch (1993), in a straightforward way.

The full prior of this model is of the form

$$p(\boldsymbol{\beta}_{1:T}, \boldsymbol{\sigma}^2_{1:T}, \boldsymbol{w}_{1:T}, \boldsymbol{\gamma}_{1:T}, \boldsymbol{\tau}^2_{1:T}, \boldsymbol{\pi}_{0,1:T}) = \prod_{t=1}^{T} \left[p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, \boldsymbol{w}_t)p(\boldsymbol{w}_t)p(\sigma_t^2)p(\boldsymbol{\beta}_t|\boldsymbol{\gamma}_t, \boldsymbol{\tau}_t^2)p(\boldsymbol{\tau}_t^2)p(\boldsymbol{\gamma}_t|\pi_{0,t})p(\pi_{0,t})\right],$$

where in the long product on the right-hand side, the first density is the Gaussian TVP state equation, the second density is the inverse gamma prior implied for the state variance $\boldsymbol{w}_t$, the third density is the prior of the stochastic volatility parameter σ_t^2 (will deal with this later in this section), and the remaining densities represent the DVS prior in (26)–(29). Due to the fact that $p(\boldsymbol{\beta}_t|\boldsymbol{\gamma}_t, \boldsymbol{\tau}_t^2)$ has the mixture representation defined in Equation (26), it would be helpful if we replace it with its alternative single-component representation

$$p(\boldsymbol{\beta}_t|\boldsymbol{V}_t) \sim \prod_{j=1}^{p} N(0, v_{j,t}), \tag{30}$$

where we define $v_{j,t} = (1-\gamma_{j,t})^2\underline{c} \times \tau_{j,t}^2 + \gamma_{j,t}^2\tau_{j,t}^2$ and the $p \times p$ diagonal matrix $\boldsymbol{V}_t = diag(v_{1,t}, \ldots, v_{p,t})$. Multiplying the joint prior with $p(y_t|\boldsymbol{\beta}_t, \sigma^2)$ for each t provides the kernel of the joint posterior, which is a complex product of densities. Applying the formulas in the previous section, and given the hierarchical structure of the DVS prior, the variational posterior of $\beta_{1:T}$ is of the form

$$\begin{aligned} q(\boldsymbol{\beta}_{1:T}|\boldsymbol{y}_{1:T}) \propto \exp\Bigg[& \mathbb{E}_{q(\boldsymbol{\sigma}^2_{1:T}|\boldsymbol{y}_{1:T})}\left(\sum_{t=1}^{T} \log p\left(y_t|\boldsymbol{\beta}_t, \sigma_t^2\right)\right) + \\ & \mathbb{E}_{q(\boldsymbol{w}^2_{1:T}|\boldsymbol{y}_{1:T})}\left(\sum_{t=1}^{T} \log p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, \boldsymbol{W}_t)\right) + \\ & \mathbb{E}_{q(\boldsymbol{v}_{1:T}|\boldsymbol{y}_{1:T})}\left(\sum_{t=1}^{T} \log p(\boldsymbol{\beta}_t|\boldsymbol{V}_t)\right)\Bigg]. \end{aligned} \tag{31}$$

Following ideas in Wang et al. (2016) the expression above can be regarded as the joint distribution of a Gaussian linear state-space model with measurement equation given by Equation (23) and state equation of the form[12]

$$\boldsymbol{\beta}_t = \widetilde{\boldsymbol{F}}_t\boldsymbol{\beta}_{t-1} + \widetilde{\boldsymbol{\eta}}_t, \tag{32}$$

[12] Take Equation (31) for a given period t, that is, ignore the summations, and also ingore the term $p(y_t|\boldsymbol{\beta}_t, \sigma_t^2)$, then we have

$$\begin{aligned} q(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, \boldsymbol{y}_{1:t}) &\propto \exp\left\{\mathbb{E}(\log p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, \boldsymbol{W}_t)) + \mathbb{E}(\log p(\boldsymbol{\beta}_t|\boldsymbol{V}_t))\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})'\boldsymbol{W}_t^{-1}(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}) - \frac{1}{2}\boldsymbol{\beta}_t'\boldsymbol{V}_t^{-1}\boldsymbol{\beta}_t\right\} \end{aligned}$$

where $\widetilde{\boldsymbol{\eta}}_t \sim N(\mathbf{0}, \widetilde{\boldsymbol{W}}_t)$, with parameter matrices $\widetilde{\boldsymbol{W}}_t = ((\boldsymbol{W}_t)^{-1} + (\boldsymbol{V}_t)^{-1})^{-1}$ and $\widetilde{\boldsymbol{F}}_t = \widetilde{\boldsymbol{W}}_t \times (\boldsymbol{W}_t)^{-1}$, where $\boldsymbol{W}_t = diag(w_{1,t}, \ldots, w_{p,t})$ and $\boldsymbol{V}_t = diag(v_{1,t}, \ldots, v_{p,t})$. Under this formulation, we can observe that the joint prior variance for $\beta_{j,t}$ is a function of both $w_{j,t}$ and $v_{j,t}$, $\forall j = 1, \ldots, p$.

Application of VB updates to the transformed state-space model consisting of Equations (23) and (32) provides as output estimates $\boldsymbol{m}_{t|T}$ $\forall t$, that is, the smoothed posterior mean of $q(\boldsymbol{\beta}_t|y_{1:T})$. Conditional on these estimates, derivation of the update steps for $\gamma_{j,t}$, $\tau^2_{j,t}$, and $\pi_{0,t}$ relies also on deriving the expectations of these variables with respect to $q(\boldsymbol{\beta}_t|y_{1:T})$. Therefore, extending the analysis of the previous section to accommodate these new parameters, and similar to derivations found in Gibbs sampling approaches to variable selection (see, e.g., the formulas of the conditional posteriors in George and McCulloch, 1993), the updating steps for the parameters in the DVS prior are the following:

$$\widehat{\tau}^2_{j,t} = \mathbb{E}\Big[q\Big(\tau^2_{j,t}|y_t\Big)\Big] = \Big[h_0 + \Big(m^2_{j,t|T} + P_{jj,t|T}\Big)/2\Big]/[g_0 + 1/2], \tag{33}$$

$$\widehat{\gamma}_{j,t} = \mathbb{E}[q(\gamma_{j,t}|y_t)] = \frac{N\Big(m_{j,t|T}|0, \widehat{\tau}^2_{j,t}\Big)\widehat{\pi}_{0,t}}{N\Big(m_{j,t|T}|0, \widehat{\tau}^2_{j,t}\Big)\widehat{\pi}_{0,t} + N\Big(m_{j,t|T|t}|0, \underline{c} \times \widehat{\tau}^2_{j,t}\Big)(1 - \widehat{\pi}_{0,t})}, \tag{34}$$

$$\widehat{v}_{j,t} = \mathbb{E}[q(v_{j,t}|y_t)] = (1 - \widehat{\gamma}_{j,t})^2 \underline{c}\widehat{\tau}^2_{j,t} + \widehat{\gamma}^2_{j,t}\widehat{\tau}^2_{j,t}, \tag{35}$$

$$\widehat{\pi}_{0,t} = \mathbb{E}[q(\pi_{0,t}|y_t)] = \left(1 + \sum_{j=1}^{p} \widehat{\gamma}_{j,t}\right)/(2 + p), \tag{36}$$

for each $t = 1, .., T$ and $j = 1, \ldots, p$, where, again, expectations $\mathbb{E}$ are with respect to the VB posteriors of each of the parameters showing up on the right-hand side of the equations above.

3.2. *Understanding the Proposed Prior Structure.* Before progressing into enhancing our approach with stochastic volatility and outlining the full estimation steps, a justification of our choice of prior is in order. A natural question to ask is why not insert the variable selection prior of Equations (26)–(29) directly into the state equation (24)? To see why this question is relevant, notice that the TVP regression model in stacked form can be written as the following

$$\begin{aligned} &\propto \exp\left\{-\frac{1}{2}\boldsymbol{\beta}_t'\boldsymbol{W}_t^{-1}\boldsymbol{\beta}_t + \boldsymbol{\beta}_t'\boldsymbol{W}_t^{-1}\boldsymbol{\beta}_{t-1} - \frac{1}{2}\boldsymbol{\beta}_t'\boldsymbol{V}_t^{-1}\boldsymbol{\beta}_t\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_t - \widetilde{\boldsymbol{F}}_t\boldsymbol{\beta}_{t-1})'\widetilde{\boldsymbol{W}}_t^{-1}(\boldsymbol{\beta}_t - \widetilde{\boldsymbol{F}}_t\boldsymbol{\beta}_{t-1})\right\}, \end{aligned}$$

where the simplification occurs due to the fact that the function $q()$ is derived conditional on $\boldsymbol{\beta}_{t-1}$ being known and fixed (i.e., not a random variable). Therefore, the equation above describes the transformed state equation used in filtering and smoothing the VB solution of $\boldsymbol{\beta}_t$, under the impact of both the original TVP state equation $p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}, \boldsymbol{W}_t)$ and the DVS prior $p(\boldsymbol{\beta}_t|\boldsymbol{V}_t)$.

pseudolinear regression model:

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_{T-1} \\ y_T \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{x}_2 & \mathbf{x}_2 & \ddots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{x}_{T-1} & \dots & \mathbf{x}_{T-1} & \mathbf{0} \\ \mathbf{x}_T & \dots & \mathbf{x}_T & \mathbf{x}_T \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{\beta}_1 \\ \Delta\boldsymbol{\beta}_2 \\ \dots \\ \Delta\boldsymbol{\beta}_{T-1} \\ \Delta\boldsymbol{\beta}_T \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{T-1} \\ \varepsilon_T \end{bmatrix}, \tag{37}$$

where Δ is the first difference operator and $\Delta\boldsymbol{\beta}_t = [\Delta\beta_{1,t}, \dots, \Delta\beta_{p,t}]'$.[13] Under this stacked formulation, the TVP regression resembles a regular regression model with a right-hand side matrix that has T observations and Tp predictors. That is, this formulation shows that the TVP regression is a high-dimensional, heavily-parameterized model and one would want to apply a shrinkage prior directly on its coefficients, that is, the prior

$$\Delta\beta_{j,t}|\gamma_{j,t}, \tau_{j,t}^2 \sim N(0, v_{j,t}), \tag{38}$$

for $j = 1, \dots, p, t = 1, \dots, T$, where $v_{j,t} = (1-\gamma_{j,t})^2\underline{c} \times \tau_{j,t}^2 + \gamma_{j,t}^2\tau_{j,t}^2$ and $\gamma_{j,t}$, $\tau_{j,t}$ also have hyperprior distributions as in Equations (27)–(29). In practice, this prior is equivalent to applying the DVS prior to the state variance of the TVP regression,[14] that is, the prior above implies

$$\beta_{j,t}|\beta_{j,t-1}, \gamma_{j,t}, \tau_{j,t}^2 \sim N(\beta_{j,t-1}, v_{j,t}), \tag{39}$$

which is equivalent to the state variance $w_{j,t}$ being replaced with the mixture variance parameter $v_{j,t}$. Although the prior in (39) seems to be a natural extension of George and McCulloch (1993) to the TVP model, it is not appropriate for achieving our modeling aims for high-dimensional inflation forecasting. To see this, consider the static form of the TVP regression in (37) and the shrinkage prior for the first differences of coefficients in Equation (38). A value of $\gamma_{j,t} = 0$ implies that $v_{j,t} \to 0$ and $\Delta\beta_{j,t} \to 0$, which, in turn, implies that $\beta_{j,t} = \beta_{j,t-1}$. Therefore, under sufficient amount of shrinkage over different periods t, the TVP regression model shrinks toward a structural breaks specification. This prior is desirable if the researcher suspects that the correct specification is a structural breaks model.

In contrast, our proposed prior is different as it favors the random walk TVP dynamics on the regression coefficients, which are long established to provide better performance in forecasting inflation over structural break alternatives; see Bauwens et al. (2015) and Pettenuzzo and Timmermann (2017) among numerous others. At the same time, it imposes DVS in the coefficient of interest $\beta_{j,t}$ and not its first difference $\Delta\beta_{j,t}$. In simulations, when the true model is a sparse TVP regression—as is the case in the simulations we undertake in the Supporting Information (see also Subsection 3.5)—our prior performs much better than the alternative prior in (39).[15] Similarly, in the empirical exercise, our proposed prior performs consistently better in forecasting inflation with many predictors than the prior in (39). This observation is

[13] See Korobilis and Shimizu (2021, Section 4.3) for a derivation of this stacked form of the TVP regression model.

[14] Within the context of MCMC inference, Amir-Ahmadi et al. (2020) propose a similar approach of specifying hierarchical distributions that are directly applied to the state variance parameter. However, the focus of Amir-Ahmadi et al. (2020) is not primarily the shrinkage ability of hierarchical priors, rather they use the hierarchical Bayes approach as a means of updating the prior from the data instead of tuning it subjectively. In contrast, our aim is to develop DVS in a forecasting TVP setting, as motivated in Koop and Korobilis (2012).

[15] These results are available from the authors.

consistent with the large body of recent evidence cited in the Introduction, that supports nonlinearities in the inflation process that is of the TVP form (as opposed to other nonlinear formulations proposed in the past).

3.3. *Adding Stochastic Volatility.* A vast recent literature highlights the importance of time-varying volatility in improving point and density forecasts (Clark and Ravazzolo, 2015), and the purpose of this subsection is to accommodate estimation of the parameter $\text{var}(\varepsilon_t) = \sigma_t^2$ in the VB setting. Several elegant algorithms for VB inference in stochastic volatility models exist in the literature. For example, Naesseth et al. (2017) introduce a VB sequential Monte Carlo (SMC) algorithm for stochastic volatility models. Tran et al. (2017) propose a variational Bayes method for intractable likelihoods that does not rely on the mean field approximation, and apply their algorithm to the estimation of a stochastic volatility model.

Nevertheless, such algorithms assume an explicit time-series model for the stochastic volatility parameter, an assumption that is only useful in a setting where one is interested in forecasting the second moment of a series. In a macroeconomic setting, we are interested in forecasting $E(y_{t+h}|y_t, x_t)$ and not its volatility (as it would be the case in empirical asset pricing). At the same time, previous empirical work shows that there are no statistically important differences when forecasting with alternative specifications of macroeconomic volatility.[16] For that reason, our aim here is not only to render estimation of stochastic volatility precise, but also numerically reliable and computationally efficient. To achieve this triple aim, we build on variance discounting ideas for dynamic linear methods as described in West and Harrison (1997); see also Rockova and McAlinn (2021).

Define $\phi_t = \frac{1}{\sigma_t^2}$ to be the precision (inverse variance). Following West and Harrison (1997), we assume that the time $t-1$ posterior of ϕ has the following conjugate form:

$$\phi_{t-1}|y_{1:t-1} \sim Gamma(a_{t-1}, b_{t-1}). \tag{40}$$

We do not specify an explicit time-series model for the dynamics of ϕ (e.g., stochastic volatility or GARCH) because the posterior for ϕ_t would not be conjugate to the likelihood and we would fail to obtain fast updates. To maintain this conjugacy, we specify instead the time t prior of the form

$$\phi_t|y_{1:t-1} \sim Gamma(\delta a_{t-1}, \delta b_{t-1}), \tag{41}$$

for a variance discounting factor $0 < \delta < 1$, subject to a choice of hyperparameters a_0 and b_0. By doing so, we assume that ϕ_t is centered around ϕ_{t-1} as if this parameter had random walk dynamics,[17] since it holds that $\mathbb{E}(\phi_t|y_{1:t-1}) = \mathbb{E}(\phi_{t-1}|y_{1:t-1})$. However, based on the properties of the gamma distribution, the dispersion of ϕ_t is larger to that of ϕ_{t-1}.

Under this scheme, the VB update of ϕ_t, that is, its time t posterior mean has the form

$$\widehat{\phi}_t = \mathbb{E}_{q(\beta_t|y_{1:T})}(\phi_t|y_{1:t}) = a_t/b_t, \tag{43}$$

[16] For example, Clark and Ravazzolo (2015) compare a range of specifications for time-varying variance parameters in univariate and multivariate autoregressive models, and any differences among such specifications are not statistically important (whereas all volatility specifications are always better relative to constant variance specifications).

[17] Even though we have not specified an explicit time-series evolution for ϕ_t, by using results in Uhlig (1994), we can show that the proposed variance discounting methodology is equivalent to assuming the following specification:

$$\phi_t = \gamma_t \phi_{t-1}/\delta, \tag{42}$$

for a parameter $\gamma_t|y_{1:t-1} \sim Beta(\delta a_{t-1}/2, (1-\delta)a_{t-1}/2)$.

where $a_t = 1/2 + \delta a_{t-1}$ and $d_t = \frac{1}{2}[(y_t - \boldsymbol{x}_t \boldsymbol{m}_{t|T}))^2 + \boldsymbol{x}_t \boldsymbol{P}_{t|T} \boldsymbol{x}_t'] + \delta b_{t-1}$, where $\boldsymbol{m}_{t|T}, \boldsymbol{P}_{t|T}$ are the smoothed mean and variance of $\boldsymbol{\beta}_t$. Using this scheme, past information in the data is discounted exponentially by the factor δ. The scalar δ can be seen as a prior hyperparameter whose choice determines how much relative weight we give to recent versus older observations, that is, it determines how fast we expect the precision parameter to change over time. For $\delta = 1$, we obtain the posterior under a standard recursive update scheme (similar to recursive least squares), whereas typical values that would allow for faster time variation in the precision/variance would be between 0.8 and 0.99. Values lower than 0.8 are not empirically advised, since they allow for a large amount of time variation and stochastic variance estimates become very noisy. In the empirical exercise, we set $\delta = 0.8$, a choice that reflects our prior expectation that macroeconomic data have many abrupt breaks in their second moments and excess kurtosis during recessions (implying variances that can move very fast over time).

The previous formulas pertain to the iterative updating of ϕ_t given ϕ_{t-1}. Estimates of ϕ_t can be smoothed using subsequent observations $t+1, \ldots, T$. Following West and Harrison (1997), we can approximate smoothed estimates by running a backward recursive filter of the form

$$\widetilde{\phi}_t = (1-\delta)\widehat{\phi}_t + \delta \widetilde{\phi}_{t+1}, \tag{44}$$

for $t = T-1, \ldots, 1$, where $\widetilde{\phi}_t = \mathbb{E}_{q(\beta_t|y_{1:T})}(\phi_t|y_{t+1})$ and $\widetilde{\phi}_T = \widehat{\phi}_T$. Once we obtain this update for the precision ϕ_t, a posterior mean estimate of the volatility σ_t^2 can be obtained simply as the inverse of $\widetilde{\phi}_t$.

3.4. *The Variational Bayes Dynamic Variable Selection Algorithm.* Here, we provide details of the exact parameter updates that result from VB inference in our proposed specification. Algorithm 2 outlines our proposed *variational Bayes dynamic variable selection (henceforth, VBDVS)* algorithm. This algorithm shows an accurate picture of how this would look like when programmed using a language like MATLAB or R: whereas there are many parameters involved in our specification, the code is short and it mostly involves simple scalar operations within loops over observations T and predictors p (meaning that the worst-case algorithmic complexity of these operations is $\mathcal{O}(Tp)$ per each iteration of the external "while" loop). The only cumbersome operation is the inversion of the $p \times p$ matrix $\boldsymbol{P}_{t+1|t}$ in line 14 that has complexity $\mathcal{O}(p^3)$ for each t. There are four main blocks in this algorithm. Lines 4–12 are a result of straightforward application of the Kalman filter recursions to the state-space model of Equations (23) and (32), and lines 13–17 show the approximate backward (smoothing) recursions. Lines 18–27 update the prior hyperparameters of the DVS prior for $\boldsymbol{\beta}_t$. Finally, lines 28–33 provide updates for the stochastic volatility parameter, as discussed in the previous subsection.

In all subsequent numerical evaluations, we iterate Algorithm 2 until the ELBO criterion has converged or a maximum of 200 iterations have been reached (whatever comes first). The ELBO for this complex model is a generalization of Equation (22), where the parameter space now also includes all the hyperparameters of the variable selection prior $(\boldsymbol{\tau}_{1:T}, \boldsymbol{\gamma}_{1:T}, \boldsymbol{\pi}_{0,1:T})$ and σ^2 is now replaced by $\boldsymbol{\sigma}_{1:T}^2 = \widetilde{\boldsymbol{\phi}}_{1:T}$. To avoid numerical problems that can show up when doing extensive use of the logarithmic and exponential functions, in practice, we evaluate a simpler version of the ELBO that only considers the parameters $\boldsymbol{\beta}_{1:T}, \boldsymbol{w}$, and in each iteration, it ignores the values of $\boldsymbol{\sigma}_{1:T}^2, \boldsymbol{\tau}_{1:T}, \boldsymbol{\gamma}_{1:T}, \boldsymbol{\pi}_{0,1:T}$ as if these were known (and their respective terms in the ELBO become normalizing constants).

3.5. *Numerical Evaluation of VBDVS.* We undertake a large-scale simulation study using synthetic data to find out how fast and accurate the new algorithm is. For the sake of space, results of this study are available in the Supporting Information. Here we only summarize that

ALGORITHM 2 VARIATIONAL BAYES ALGORITHM FOR TVP REGRESSION MODEL WITH DYNAMIC VARIABLE SELECTION AND STOCHASTIC VARIANCE (VBDVS ALGORITHM)

Choose values of $\boldsymbol{m}_0, \boldsymbol{P}_0, a_0, b_0, c_{j,0}, d_{j,0}, g_0, h_0, \underline{c}$, and δ; initialize all vectors/matrices.

$r = 1$

while $\| \mathcal{G}\left(q(\boldsymbol{\beta}^{(r)}, \boldsymbol{w}^{(r)}|\boldsymbol{y}\right) - \mathcal{G}\left(q(\boldsymbol{\beta}^{(r-1)}, \boldsymbol{w}^{(r-1)}|\boldsymbol{y}\right) \| \to 0$ **do**

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

    $\widetilde{\boldsymbol{W}}_t^{(r)} = \left(\left(\boldsymbol{W}_t^{(r-1)}\right)^{-1} + \left(\boldsymbol{V}_t^{(r-1)}\right)^{-1}\right)^{-1}$

    $\widetilde{\boldsymbol{F}}_t^{(r)} = \widetilde{\boldsymbol{W}}_t^{(r)} \left(\boldsymbol{W}_t^{(r-1)}\right)^{-1}$

    $\boldsymbol{m}_{t|t-1}^{(r)} = \widetilde{\boldsymbol{F}}_t^{(r)} \boldsymbol{m}_{t-1|t-1}^{(r)}$ Predicted mean

    $\boldsymbol{P}_{t|t-1}^{(r)} = \widetilde{\boldsymbol{F}}_t^{(r)} \boldsymbol{P}_{t-1|t-1} \widetilde{\boldsymbol{F}}_t^{(r)'} + \widetilde{\boldsymbol{W}}_t^{(r)}$ Predicted variance

    $\boldsymbol{K}_t^{(r)} = \boldsymbol{P}_{t|t-1}^{(r)} \boldsymbol{x}_t' \left(\boldsymbol{x}_t \boldsymbol{P}_{t|t-1}^{(r)} \boldsymbol{x}_t' + \widehat{\sigma}_t^{2\ (r-1)}\right)^{-1}$ Kalman gain

    $\boldsymbol{m}_{t|t}^{(r)} = \boldsymbol{m}_{t|t-1}^{(r)} + \boldsymbol{K}_t^{(r)} \left(y_t - \boldsymbol{x}_t \boldsymbol{m}_{t|t-1}^{(r)}\right)$ Filtered mean of $\boldsymbol{\beta}_t$

    $\boldsymbol{P}_{t|t}^{(r)} = \left(\boldsymbol{I}_p - \boldsymbol{K}_t^{(r)} \boldsymbol{x}_t\right) \boldsymbol{P}_{t|t-1}^{(r)}$ Filtered variance of $\boldsymbol{\beta}_t$

  end for

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

    $\boldsymbol{C} = \boldsymbol{P}_{t|t}^{(r)} \widetilde{\boldsymbol{F}}_t^{(r)} \left(\boldsymbol{P}_{t+1|t}^{(r)}\right)^{-1}$

    $\boldsymbol{m}_{t|T}^{(r)} = \boldsymbol{m}_{t|t}^{(r)} + \boldsymbol{C}\left(\boldsymbol{m}_{t+1|T}^{(r)} - \boldsymbol{m}_{t+1|t}^{(r)}\right)$ Smoothed mean of $\boldsymbol{\beta}_t$

    $\boldsymbol{P}_{t|T}^{(r)} = \boldsymbol{P}_{t|t}^{(r)} + \boldsymbol{C}\left(\boldsymbol{P}_{t+1|T}^{(r)} - \boldsymbol{P}_{t+1|t}^{(r)}\right) \boldsymbol{C}'$ Smoothed variance of $\boldsymbol{\beta}_t$

  end for

  $\boldsymbol{D}_t = \boldsymbol{P}_{t|T}^{(r)} + \boldsymbol{m}_{t|T}^{(r)} \boldsymbol{m}_{t|T}^{(r)'} + \left(\boldsymbol{P}_{t-1|T}^{(r)} + \boldsymbol{m}_{t-1|T}^{(r)} \boldsymbol{m}_{t-1|T}^{(r)'}\right)\left(I_p - 2\widetilde{F}_t^{(r)}\right)'$ Squared error in state eq.

  $R_t = \left[\left(y_t - \boldsymbol{x}_t \boldsymbol{m}_{t|T}^{(r)}\right)^2 + \boldsymbol{x}_t \boldsymbol{P}_{t|T} \boldsymbol{x}_t'\right]$ Squared error in measurement eq.

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

    for $j = 1$ **to** p **do**

      $\widehat{\tau}_{j,t}^{-2\ (r)} = (g_0 + 0.5) / \left(h_0 + 0.5\left[\left(m_{j,t|T}^{(r)}\right)^2 + P_{jj,t|T}^{(r)}\right]\right)$ Posterior mean of $\frac{1}{\tau_{j,t}^2}$

      $\widehat{\gamma}_{j,t}^{(r)} = \frac{N\left(m_{j,t|T}^{(r)}|0, \widehat{\tau}_{j,t}^{2\ (r)}\right)\widehat{\pi}_{0,t}^{(r-1)}}{N\left(m_{j,t|T}^{(r)}|0, \widehat{\tau}_{j,t}^{2\ (r)}\right)\widehat{\pi}_{0,t}^{(r-1)} + N\left(m_{j,t|T}^{(r)}|0, \underline{c} \times \widehat{\tau}_{j,t}^{2\ (r)}\right)\left(1 - \widehat{\pi}_{0,t}^{(r-1)}\right)}$ Posterior mean of $\gamma_{j,t}$

      $\widehat{v}_{j,t}^{(r)} = \left(1 - \widehat{\gamma}_{j,t}^{(r)}\right)^2 \underline{c}\widehat{\tau}_{j,t}^{2\ (r)} + \left(\widehat{\gamma}_{j,t}^{(r)}\right)^2 \widehat{\tau}_{j,t}^{2\ (r)}$ Posterior mean of $v_{j,t}$

      $\widehat{w}_{j,t}^{-1(r)} = (c_0 + 0.5) / (d_0 + 0.5\boldsymbol{D}_{jj,t})$ Posterior mean of $\frac{1}{w_{j,t}}$

    end for

    $\boldsymbol{W}^{(r)} = diag\left(\widehat{w}_{1,t}^{(r)}, ..., \widehat{w}_{p,t}^{(r)}\right)$ State equation cov. matrix

    $\boldsymbol{V}^{(r)} = diag\left(\widehat{v}_{1,t}^{(r)}, ..., \widehat{v}_{p,t}^{(r)}\right)$ Variable selection prior cov. matrix

    $\widehat{\pi}_{0,t}^{(r)} = \left(1 + \sum_{j=1}^{p} \widehat{\gamma}_{j,t}^{(r)}\right) / (2 + p)$ Posterior mean of $\pi_{0,t}$

    $\widehat{\phi}_t^{(r)} = (\delta a_{t-1} + 0.5) / (\delta b_{t-1} + 0.5 R_t)$ Filtered mean of $\frac{1}{\sigma_t^2}$

  end for

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

    $\widetilde{\phi}_t^{(r)} = (1 - \delta)\widehat{\phi}_t^{(r)} + \delta\widetilde{\phi}_{t+1}^{(r)}$ Smoothed mean of $\frac{1}{\sigma_t^2}$

  end for

  $r = r + 1$

end while

when the true data-generating process (DGP) is that of a TVP model, VBDVS is always able to recover with accuracy parameters and DVS probabilities, especially in high-dimensional cases. Such results verify that the proposed algorithm is numerically correct and converges to the desired posterior moments.

In practical situations, the true underlying DGP is not known, and estimation of various algorithms can be erratic. This is the topic of the next section, where VBDVS is evaluated using a real data set comprising a large panel of macroeconomic and financial variables. Finally, the Monte Carlo experiments verify that, in high-dimensions, the VBDVS algorithm is orders of magnitude faster than the fast MCMC sampler of Chan and Jeliazkov (2009). This means that the algorithm will be even faster compared to several recent MCMC algorithms that allow for dynamic shrinkage and variable selection (Giordani and Kohn, 2008; Chan et al., 2012; Nakajima and West, 2013; Kalli and Griffin, 2014, to name a few).

4. MACROECONOMIC FORECASTING WITH MANY PREDICTORS

4.1. *A New Large Data Set for Forecasting Inflation.* Following a large literature on TVP models in macroeconomics, our primary target is to forecast quarterly U.S. inflation. There exists mixed empirical evidence about the potential of very large data sets to improve forecasts of inflation, and in many cases, small, simple models are extremely hard to beat. Therefore, our aim is to demonstrate that the new DVS methodology can extract meaningful predictive information from a large number of predictors, even if it is not always and everywhere the best forecasting methodology. For that reason, we build a novel, high-dimensional data set that merges predictors from several mainstream aggregate macroeconomic and financial data sets.[18] Our building block is the FRED-QD data set of McCracken and Ng (2020), which we augment with portfolio data used in Jurado et al. (2015), stock market predictors from Welch and Goyal (2007), survey data from University of Michigan consumer surveys, commodity prices from the World Bank's Pink Sheet database, and key macroeconomic indicators from the Federal Reserve Economic Data for four economies (Canada, Germany, Japan, United Kingdom). All data are quarterly, and span the period 1960Q1–2021Q4. When variables are originally measured at monthly level, quarterly values are constructed by taking the average over the quarter. All variables are preadjusted from their respective sources for seasonality (where relevant), and we additionally remove extreme outliers.[19]

The data set has 440 variables in total. Out of these, we forecast the series (FRED-QD mnemonics in parentheses): GDP deflator (GDPCTPI), total CPI (CPIAUCSL), core CPI (CPILFESL), and PCE deflator (PCECTPI). When each of these price series, P_t, is used as the dependent variable to be forecast h-quarters ahead, we transform it according to the formula $y_{t+h} = (400/h)\ln(P_{t+h} - P_t)$. We forecast these transformed series one at a time, and the remaining three price series are included in the list of exogenous predictor variables (439 in total). The predictor variables are transformed using standard norms in the literature (see, e.g., McCracken and Ng, 2020): (i) levels for variables that are already expressed in rates (e.g., unemployment, interest); (ii) first differences of logarithm for variables measuring population (e.g., employment), variables expressed in dollars (e.g., GDP), commodity prices, and some indexes (e.g., industrial production); and (iii) second differences of logarithm for price and consumption indexes, as well as deflator series. The Supporting Information describes in detail all variables and transformations, and provides links to all sources.

[18] Although one could also think of potential predictors in disaggregated panels obtained in surveys, Internet, or documents (text data), such novel sources are typically proprietary or require subjective data processing (in the case of text data), such that forecasting results would be hard to replicate.

[19] Following Stock and Watson (2016), we replace outliers using the median of the preceding five observations. An outlier is defined to be any observation that satisfies $|y_t - m|/iqr > \kappa$, where m is the median of y, iqr is the interquantile range, and $\kappa = 4.5$.

TABLE 1
HYPERPARAMETER CHOICES FOR SENSITIVITY ANALYSIS

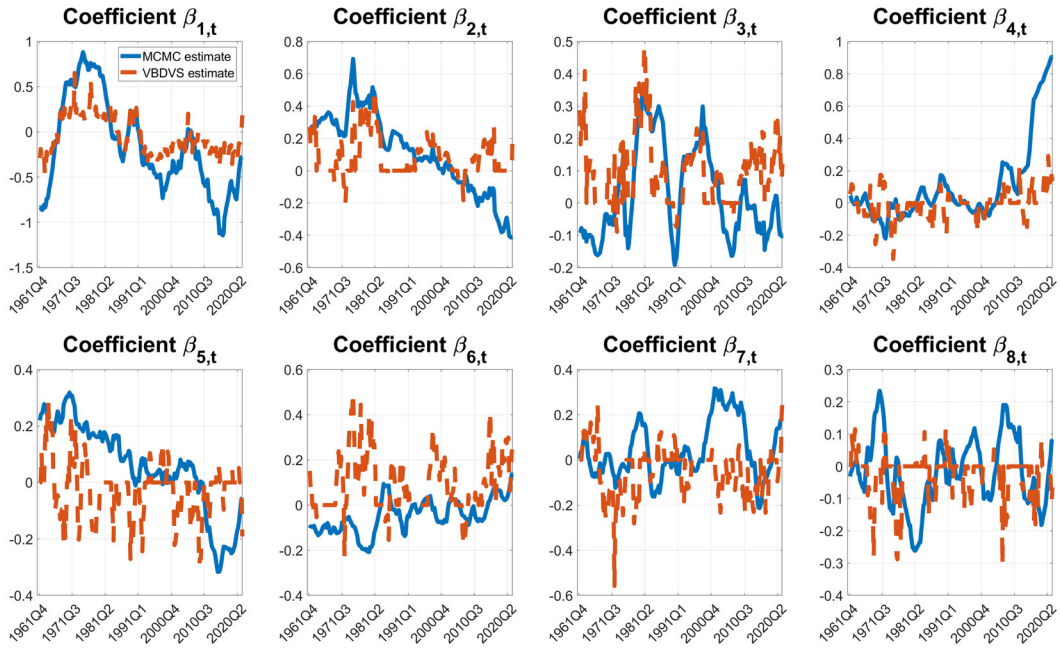
	Case 1	Case 2	Case 3	Notes
IMPORTANT HYPERPARAMETERS				
h_0	1	1	12	See Equation (28)
$c_{j,0}$	1	100	100	See Equation (25)
FIXED HYPERPARAMETERS				
g_0	1	1	1	See Equation (28)
$d_{j,0}$	1	1	1	See Equation (25)
$\underline{c}$	10^{-4}	10^{-4}	10^{-4}	See Equation (26)
a_0	0.01	0.01	0.01	See Equation (25)
b_0	0.01	0.01	0.01	See Equation (25)
δ	0.8	0.8	0.8	See Equation (41)
$m_{j,0}$	0	0	0	See Equation (25)
$P_{j,0}$	4	4	4	See Equation (25)

4.2. *Dynamic Variable Selection at Work: An In-Sample Assessment.* Before we set up a comprehensive out-of-sample forecasting exercise, we first assess in-sample estimates from the VBDVS by performing a sensitivity analysis to some benchmark prior choices. This exercise is intended to demonstrate that the new algorithm provides reasonable estimates of trends, volatilities, and other parameters. Most importantly, it serves as a way to clarify that—despite the fact that our prior is heavily parameterized—prior elicitation in the VBDVS algorithm becomes a reasonably straightforward task. As it is impossible to present estimates of the TVP model using all variables in our data set as predictors, we focus on the results of a small model. Using the real macro data and the full sample 1960Q1–2021Q4, we illustrate the effect of different prior choices to the estimation of TVPs in the following regression:

$$\pi_t = \beta_{1,t} + \beta_{2,t-1}\pi_t + \beta_{3,t}\pi_{t-2} + \sum_{j=1}^{5} \beta_{(3+j),t} f_{j,t-1}^{pca} + \varepsilon_t, \tag{45}$$

where π_{t+1} in this example is GDP deflator growth, and $f_{j,t}^{pca}$ is the jth principal component extracted from the remaining 439 exogenous predictors.

Out of all parameters and hyperparameters defined in our algorithm, it is only a handful that are crucial for inference and forecasting, whereas others can be fixed to reasonable or uninformative values. Table 1 lists all hyperparameters that need to be chosen in the VBDVS algorithm, separating them into "Important" and "Fixed" hyperparameters. Starting from the latter group, a_0 and b_0 are the initial scale and rate parameters of the initial condition of the precision parameter in Equation (41). Setting $a_0 = b_0 = 0.01$ implies that the precision has prior mean one and variance 100, which is a reasonable uninformative choice. Next, we want the evolution of inflation stochastic volatility to be fairly fast-moving, so we set $\delta = 0.8$ for reasons explained in Subsection 3.3. Given that p is very large to allow us to obtain meaningful prior information about the regression coefficients $\boldsymbol{\beta}_t$ (e.g., using a training sample), we allow their initial condition $\boldsymbol{\beta}_0$ to be fairly uninformative by setting $\boldsymbol{m}_0 = \mathbf{0}$ and $\boldsymbol{P}_0 = 4\boldsymbol{I}_p$. The parameter $\underline{c}$ in the DVS prior has to be small (see discussion in Subsection 3.1) and its exact value affects the way the algorithm selects each of the two normal components in the spike and slab prior—that is, it affects the choice between a certain $\beta_{j,t}$ being restricted or not. We prefer to fix this parameter to $\underline{c} = 0.0001$ and allow only $\tau_{j,t}^2$ and its prior to determine the ratio of the prior variances of the two normal components in the mixture prior. The parameters that are important in our high-dimensional setting are the ones affecting the two prior variances of the time-varying coefficients $\boldsymbol{\beta}_t$, namely, the hyperparameters of $\tau_{j,t}^2$ and $w_{j,t}$. Both prior parameters are inverse gamma distributed, so they depend on two tuning hyperparameters each. To make tuning easy, we follow a standard norm in Bayesian inference



NOTES: Solid lines are posterior means from a TVP model with the same predictors estimated using MCMC without variable selection prior.

FIGURE 1

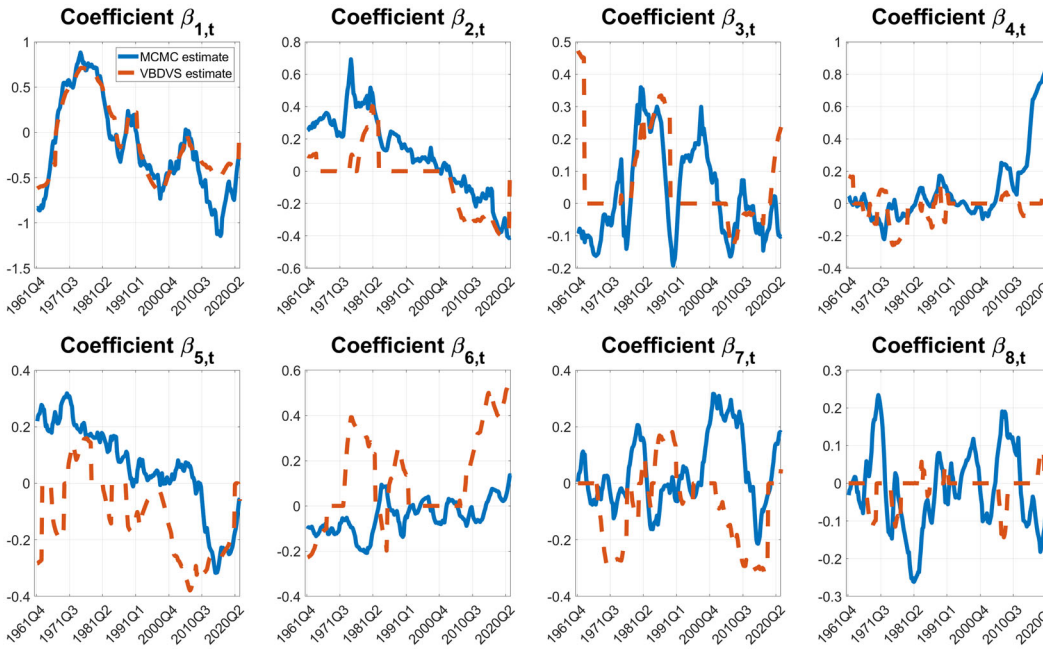
POSTERIOR MEANS OF TIME-VARYING COEFFICIENT ESTIMATES FROM VBDVS (RED DASHED LINES) USING THE PRIOR VALUES OF CASE 1 OF TABLE 1

and fix the one of the two parameters of the inverse gamma priors to one. In the case of $\tau_{j,t}^{-2} \sim Gamma(g_0, h_0)$, we set $g_0 = 1$, whereas for $w_{j,t}^{-1} \sim Gamma(c_{j,0}, d_{j,0})$, we set $d_{j,0} = 1$ for all j. Doing so means that h_0 can be used to tune the amount of time variation in $\beta_{j,t}$, whereas $c_{j,0}$ can be used to tune the amount of shrinkage toward zero.

We test, in turn, the effect of using the values in Cases 1–3 of Table 1 to the outcomes of coefficients $\beta_{1,t}$ to $\beta_{8,t}$. Our first prior choice, denoted as "Case 1" in Table 1, selects the fairly uninformative values $c_{j,0} = 1, d_{j,0} = 1$ resulting in the two prior variances of $\beta_{j,t}$ (namely, $\tau_{j,t}^2$ and $w_{j,t}$) to be affected more by the likelihood. As is expected in high-dimensional state-space models, uninformative choices will be numerically unstable and result to noisy estimates. The former is not the case as VBDVS (unlike MCMC) does not involve sampling, and the TVPs can be estimated without numerical problems.[20] However, as Figure 1 reveals, VBDVS does give quite noisy estimates when adopting a more uninformative prior on the variances of $\beta_{j,t}$. Figure 1 shows posterior mean estimates from VBDVS (dashed red lines) and it contrasts these with posterior mean estimates from the same model estimated using MCMC and no variable selection prior (solid blue line).[21] All eight coefficients estimated with VBDVS provide quite noisy outcomes. The time-varying posterior inclusion probabilities associated with these TVPs (not shown in this figure) switch almost every period between zero and one in a noisy fashion. The VBDVS estimates are more regularized toward zero relative to their MCMC counterparts; however, this first diffuse prior does not identify any patterns of sparsity, for example, consecutive periods where TVPs are zero followed by consecutive periods

[20] The stability of VBDVS also holds even in the presence of many observations T that would make $\boldsymbol{\beta}_t$ explosive with high probability, as the TVPs evolve as a nonstationary random walk. This point of numerical stability is illustrated in the Supporting Information where TVP models for weekly data are estimated without any numerical issues showing up.

[21] Discussion of the prior settings of the MCMC algorithm is provided in the Supporting Information.



NOTES: Solid lines are posterior means from a TVP model with the same predictors estimated using MCMC without variable selection prior.

FIGURE 2

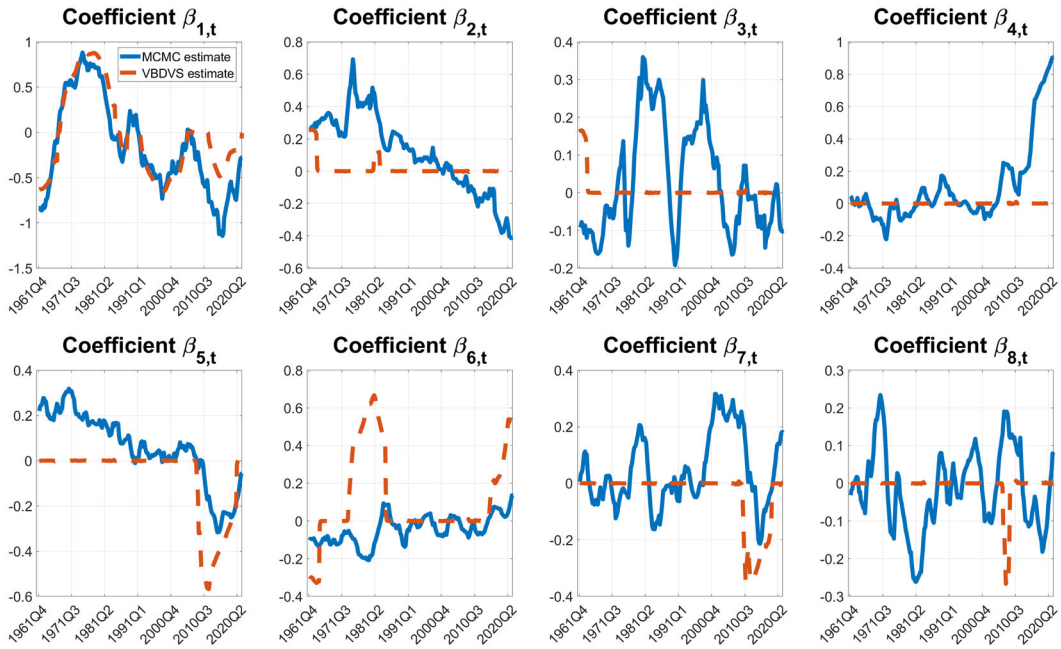
POSTERIOR MEANS OF TIME-VARYING COEFFICIENT ESTIMATES FROM VBDVS (RED DASHED LINES) USING THE PRIOR VALUES OF CASE 2 OF TABLE 1

that they are not zero. Sparsity patterns using these prior values are quite random and noisy that, in general, is not a desirable feature for economic data.

To prevent such noisy estimates of the TVPs, the remaining two priors in Table 1 (Cases 2–3) illustrate the effects of the value $c_{j,0} = 100$, a choice that imposes a conservative evolution for each $\beta_{j,t}$. The assumption that variation in TVPs should be smooth and contained is used widely in empirical macroeconomic research.[22] Fixing $c_{j,0}$ only leaves choice of h_0 to be important for estimation. Cases 2 and 3 explore two values of h_0. The value $h_0 = 1$ in Case 2 favors some shrinkage toward zero but also allows the parameters to vary substantially in some periods. The TVP estimates under this prior are shown in Figure 2, where, again, the time-varying intercept and first lag are not shrunk, but coefficients of the second own lag and the five principal components can be zeroed out in some periods and vary smoothly in others. Figure 3 shows estimates under the choice $h_0 = 12$ that favors more shrinkage toward zero. In general, larger values shrink more, for example, the choice $h_0 = 100$ (not shown here) effectively shrinks the TVP regression toward the unobserved components stochastic volatility (UCSV) specification of Stock and Watson (2007), which is a model with only a time-varying trend.

This graphical analysis establishes that the priors in Cases 2 and 3 are reasonable default choices, and as such we built on these choices in the next subsection when forecasting inflation. To have a visual assessment of the time pattern of DVS and shrinkage, Panel (a) of Figure 4 plots the posterior inclusion probabilities of each regressor associated with the time-varying coefficient estimates presented in Figure 2 (Case 2 prior). These seem to show the exact periods where each coefficient moves from a state of being restricted to zero to a state

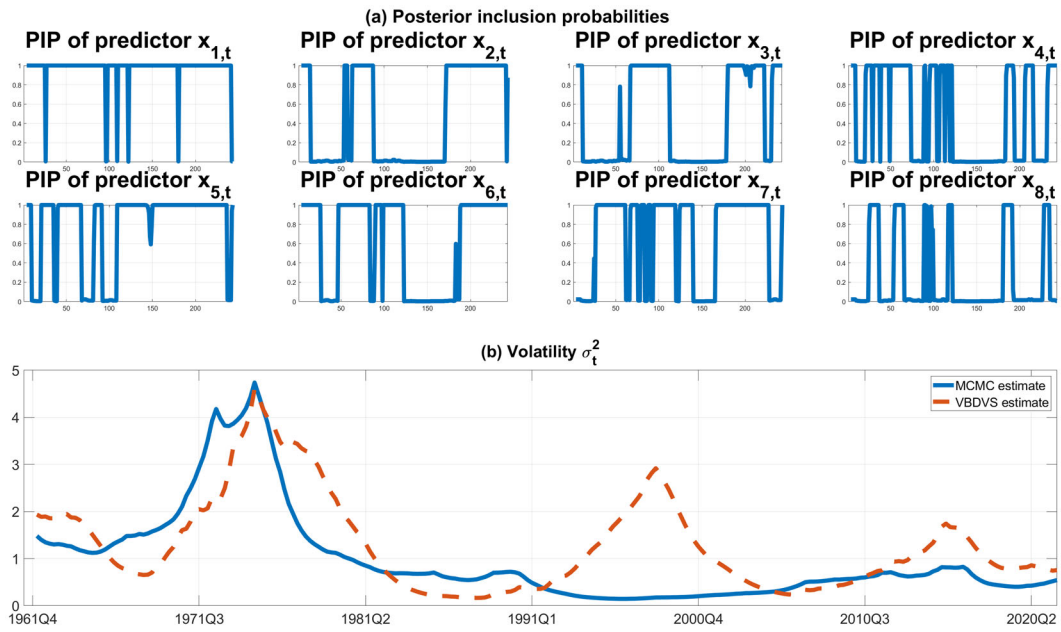
[22] See, for example, the "business as usual" prior motivated in Cogley and Sargent (2005) for the case of a vector autoregression with TVPs.



Notes: Solid lines are posterior means from a TVP model with the same predictors estimated using MCMC without variable selection prior.

Figure 3

posterior means of time-varying coefficient estimates from vbdvs (red dashed lines) using the prior values of case 3 of table 1



Notes: Panel (b) shows posterior means of time-varying volatility estimates from VBDVS using the same prior case (red dashed line) versus MCMC (solid blue line).

Figure 4

panel (a) shows time-varying posterior inclusion probabilities (pips) from vbdvs algorithm, using the prior values of case 2 of table 1

where it is not zero. Panel (b) of the same figure shows the posterior mean of the stochastic volatility estimate from VBDVS versus the estimate from MCMC. Overall, these two estimates are comparable throughout the sample. Differences in volatility estimates observed in the second half of the sample reflect the fact that the two algorithms assume different specification of σ_t^2, and they also rely on different priors in the estimation of $\boldsymbol{\beta}_t$. In particular, the fact that the volatility of VBDVS is elevated in the second half of the sample has to do with the shrinkage of the TVPs: if more TVPs are shrunk toward zero, then the volatility estimates will compensate by trying to capture the excess time variation. Note that higher variance for the VBDVS does not translate into worse fit both in-sample and out-of-sample. Total fit depends jointly on the error variance and the number of parameters. VBDVS provides a sparse solution with more degrees of freedom than MCMC estimates that seem to overfit the data (in MCMC estimates of the previous figures, all eight coefficients have large values and swing substantially over time).

4.3. *Forecasting Inflation.* We forecast inflation using models of the form

$$y_{t+h} = \alpha_t + \phi_{1,t} y_t + \phi_{2,t} y_{t-1} + \boldsymbol{x}_t \boldsymbol{\beta}_t + \varepsilon_{t+h}, \tag{46}$$

where y_{t+h} is h-step ahead inflation (see Subsection 4.1 for a definition) regressed on an intercept, two own lags, and exogenous predictors. We use a variety of forecasting models. Some benchmark models are based on Equation (46) but assume constant coefficients (i.e., $\alpha_t = \alpha$, $\phi_{1,t} = \phi_1$, and so on), whereas others assume different sets of exogenous predictors. However, what all models have in common is that they always include an intercept and two own lags of inflation. Given that our data set is much larger than data sets used before for forecasting inflation, and to avoid confusion by specifying different combinations or subsets of predictors, we only distinguish four simple categories of models: (i) models with no predictors (i.e., only intercept and autoregressive terms); (ii) models with first five principal components as predictors; (iii) models with 60 principal components as predictors; and (iv) models with all 439 exogenous predictors. Our list of models representing each category is the following:

- **AR**: benchmark AR(2) with intercept, estimated with ordinary least squares ordinary least squares (OLS)
- **SBAR**: Bayesian structural breaks AR(2) that allows to simulate break-points out-of-sample following Koop and Potter (2007) and Bauwens et al. (2015)
- **UCSV**: Unobserved components stochastic volatility model proposed by Stock and Watson (2007)
- **TVPAR**: TVP version of the AR model, with stochastic volatility, estimated with MCMC
- **FAC5**: Builds on benchmark AR specification by augmenting it with first five principal components estimated with OLS
- **BAG/FAC5**: Same predictors as FAC5, estimated as constant parameter regression using the Bagging algorithm of Breiman (1996)
- **DMA/FAC5**: Same predictors as FAC5, estimated as TVP regression using the dynamic model averaging (DMA) algorithm of Koop and Korobilis (2012)
- **TVD/FAC5**: Same predictors as FAC5, estimated as TVP regression using the time-varying dimension (TVD) algorithm of Chan et al. (2012)
- **GPR/FAC5**: Same predictors as FAC5, estimated as a Gaussian process regression
- **VBDVS/FAC5**: Same predictors as FAC5, estimated as TVP regression using our DVS prior with VB
- **SSVS/FAC60**: Builds on benchmark AR specification by augmenting it with first 60 principal components, estimated using the SSVS prior with MCMC of George and McCulloch (1993)

- **ELN/FAC60**: Same predictors as SSVS/FAC60, estimated as a constant parameter regression using the Elastic Net algorithm of Zou and Hastie (2005)
- **VBDVS/FAC60**: Same predictors as SSVS/FAC60, estimated as a TVP regression using our DVS prior with VB
- **ELN/X**: Builds on benchmark AR specification by augmenting it with all 439 predictors, estimated using the Elastic Net algorithm of Zou and Hastie (2005)
- **PLS/X**: Same predictors as in ELN/X, estimated as a constant parameter Partial Least Squares (PLS) regression
- **VBDVS/X**: Same predictors as ELN/X, estimated as a TVP regression using our DVS prior with VB

The choice of models is motivated by their simplicity and replicability. In particular, the Gaussian process regression, PLS, and Elastic Net algorithms are based on built-in functions in MATLAB's Statistics and Machine Learning Toolbox (MATLAB, 2020), and are fairly easy to set up. Estimation of these models is done using default settings in MATLAB or default choices proposed by their respective creators.[23] Exact details of these algorithms and their default settings are provided in the Supporting Information.

In terms of their theoretical and empirical properties, all these models cover a wide spectrum of forecasting specifications. The AR(2) is a standard benchmark in macroeconomic time-series forecasting, and typically performs better than a random walk (which is the benchmark for financial data). TVP counterparts of the simple AR(2) (models SB, UCSV, and TVPAR) have been shown to forecast inflation well, see Stock and Watson (2007) and Bauwens et al. (2015). Extracting the first few principal components (factors) is possibly the most popular way of representing parsimoniously the information in a large data set, see Stock and Watson (2016). A naive factor model uses least squares estimation on a model that has the first five principal components as exogenous predictors, whereas a second factor model replaces OLS with the Bagging algorithm of Breiman (1996) that allows to select the "best" factors in a static way. Next, the DMA algorithm described in Koop and Korobilis (2012), the TVD model of Chan et al. (2012), and our VBDVS algorithm allow to implement DVS in a TVP setting using the same first five principal components. The Gaussian process regression is a very flexible kernel-based method that allows us to understand whether inflation can better be forecasted by TVPs, or some more complex form of nonlinearity. Moving on to models with 60 factors, we have to drop many previous specifications for computational reasons.[24] For that reason, we use the SSVS algorithm of George and McCulloch (1993), which can be thought of as the static equivalent of our VBDVS algorithm. The Elastic Net of Zou and Hastie (2005) is a popular penalized likelihood estimator for high-dimensional data. Finally, our VBDVS algorithm is also estimated with a larger number of factors to find out whether its dynamic shrinkage properties are useful relative to the naive selection of the first five factors. Finally, we estimate models using all 439 exogenous predictors. The Elastic Net is again on the list, and we also include PLS regression. PLS is similar to principal component analysis, with the main difference being that factors are extracted with reference to the variable to be predicted. Principal components instead only explain the variability in the exogenous predictors, and it may be the case that their estimates do not carry predictive information for the target variable. Finally, our VBDVS algorithm is also applied to this full model with all 439 predictors.

In terms of the prior choices used when forecasting with our VBDVS algorithm, these follow the arguments in the previous subsection, see Table 1. We only adapt how "aggressively" we shrink parameters based on the total number of predictors in each model, which is a procedure that also has theoretical/asymptotic justification; see the discussion in Narisetty and He (2014). For model VBDVS/FAC5, we set $h_0 = 1$; for VBDVS/FAC60, we set $h_0 = 12$; and for VBDVS/X, we set $h_0 = 100$. These values constitute choices that can be broadly interpreted as

[23] As an example, the penalty parameter in the Elastic Net is estimated using 10-fold cross-validation.

[24] For example, DMA and TVD cannot scale up to these large dimensions, Gaussian process regression becomes overparameterized, and Bagging becomes numerically unstable in some periods of the forecasting exercise.

TABLE 2
RELATIVE MSFES FOR GDP DEFLATOR (GDPCTPI) AND PCE DEFLATOR (PCECTPI)

	GDP deflator				PCE deflator			
	$h = 1$	$h = 4$	$h = 8$	$h = 12$	$h = 1$	$h = 4$	$h = 8$	$h = 12$
MODELS WITH NO PREDICTORS								
AR	*0.0545*	*0.0479*	*0.0485*	*0.0611*	*0.1568*	*0.1142*	*0.0921*	*0.0989*
SBAR	0.97	0.91	0.72	0.57	0.96	0.81	0.60	0.45
UCSV	1.05	0.97	0.77	0.59	0.98	0.84	0.55	**0.41**
TVPAR	1.29	1.19	0.88	0.65	1.23	0.91	0.64	0.46
MODELS WITH FIVE FACTORS								
FAC5	0.94	1.34	1.64	1.34	1.02	1.31	1.49	1.31
BAG/FAC5	0.92	1.29	1.59	1.25	1.05	1.30	1.49	1.31
DMA/FAC5	0.97	0.90	1.02	0.88	1.00	0.90	0.93	1.23
TVD/FAC5	**0.82**	0.96	1.02	0.89	1.05	1.03	0.91	0.83
GPR/FAC5	0.92	0.94	0.80	0.64	0.81	0.83	0.60	0.57
VBDVS/FAC5	0.95	1.04	0.76	0.59	0.81	0.87	0.59	**0.41**
MODELS WITH 60 FACTORS								
SSVS/FAC60	0.88	1.26	1.48	1.16	0.94	1.17	1.43	1.10
ELN/FAC60	0.89	1.03	1.47	1.14	0.84	0.95	1.09	0.96
VBDVS/FAC60	1.06	0.87	0.74	0.62	1.00	0.68	0.49	0.48
MODELS WITH 439 PREDICTORS								
ELN/X	0.84	1.26	1.58	1.22	**0.69**	0.98	1.15	0.76
PLS/X	0.92	1.11	1.50	1.28	0.72	0.92	0.98	0.89
VBDVS/X	0.98	**0.80**	**0.64**	**0.53**	0.92	**0.67**	**0.51**	0.48

[25]NOTES: The AR model serves as a benchmark and its entries (shown in italics) are the regular values of the mean squared forecast error (MSFE) for each forecast horizon $h = 1, 4, 8, 12$ quarters. Entries for each subsequent model are MSFEs relative to the values of the AR benchmark. Entries in boldface indicate the best performing model for each forecast horizon.

"low," "medium," "high" shrinkage values. If a practitioner is concerned about choosing prior hyperparameters in a data rigorous way, it is always possible to follow the suggestions in Koop and Korobilis (2012) and estimate each model size using a large grid of values for the hyperparameter h_0. In this case, the best forecasting specification can be selected out of all possible models with different values of h_0. The VBDVS algorithm is very fast and allows for such grid searches.

We forecast $h = 1, 4, 8$, and 12 quarters ahead. We use 50% of the sample as our initial estimation period that, for example, for $h = 1$ translates to using data for the period 1960Q4–1990Q1 to forecast 1990Q2. We then add one new observation to the estimation sample and forecast h-step ahead, until the full sample is exhausted. Since all models that have exogenous predictors rely on the direct forecasting regression (46), for comparability, we produce direct AR(2) forecasts as a special case of this equation with no predictors.[25] We measure forecast accuracy using the mean squared forecast error (MSFE) that is the square of the forecast error (difference between forecast and real value of y_{t+h}) averaged over the out-of-sample evaluation period.

Tables 2 and 3 present MSFEs for GDP deflator, PCE deflator, CPI, and Core CPI, for all competing models and all considered forecast horizons. To be precise, results for the benchmark AR(2) are the actual values of the MSFE statistic, whereas results for all other models are relative to those for the AR(2). This is obtained as the ratio of MSFE, such that a number lower (higher) than one means that a certain model performs better (worse) than the AR(2). The immediate message from these tables is that the VBDVS/X is the model that performs overall best, especially when looking at horizons of 1 and 2 years ahead ($h = 4, 8$).

[25] The alternative would be to specify an AR(2) model linking y_t with y_{t-1} and y_{t-2} and then iterate the process h periods ahead, a procedure also known as iterated forecasting. By using direct AR(2) forecasts as the benchmark, we can explicitly assess the exact contribution of various models that introduce exogenous predictors.

TABLE 3
RELATIVE MSFES FOR CPI (CPIAUCSL) AND CORE CPI (CPILFESL)

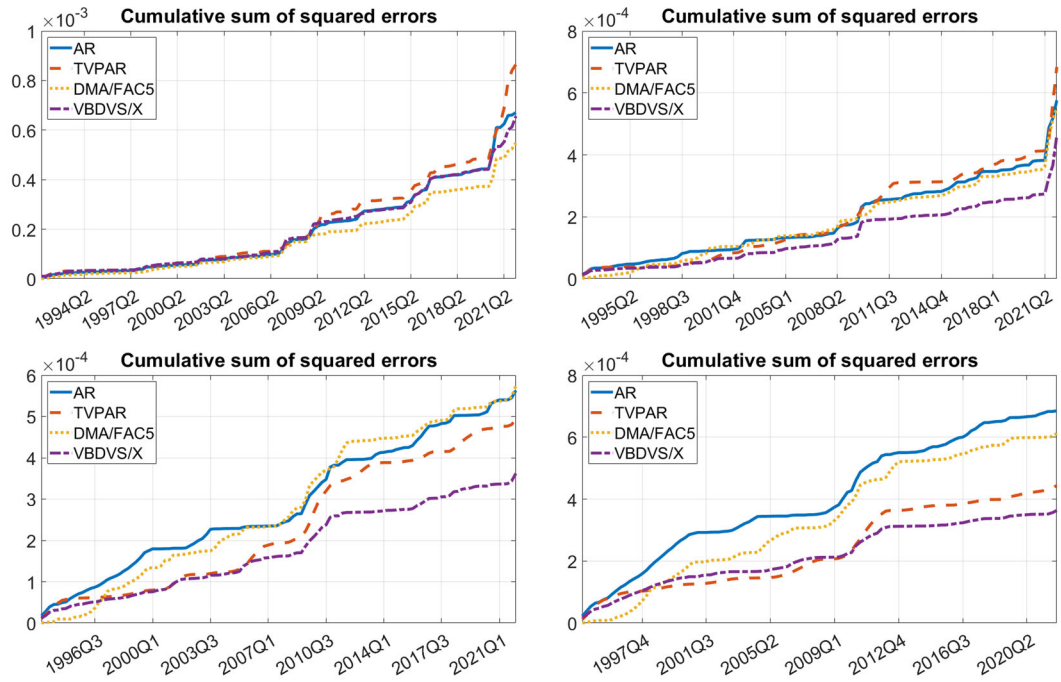
	Total CPI				Core CPI			
	$h = 1$	$h = 4$	$h = 8$	$h = 12$	$h = 1$	$h = 4$	$h = 8$	$h = 12$
MODELS WITH NO PREDICTORS								
AR	*0.2087*	*0.1719*	*0.1271*	*0.1331*	*0.0539*	*0.0371*	*0.0385*	*0.0509*
SBAR	0.94	0.68	0.49	0.49	**0.92**	0.89	0.73	0.52
UCSV	0.92	0.91	0.59	0.44	1.22	**0.85**	0.74	0.42
TVPAR	1.09	0.84	0.62	0.45	1.25	1.19	0.80	0.50
MODELS WITH FIVE FACTORS								
FAC5	0.86	1.08	1.14	0.99	1.15	1.39	1.74	1.47
BAG/FAC5	0.85	1.05	1.07	0.98	1.09	1.32	1.52	1.29
DMA/FAC5	0.83	0.70	0.81	1.00	1.01	1.00	1.11	0.97
TVD/FAC5	0.83	0.91	0.73	0.63	0.94	1.05	0.88	0.60
GPR/FAC5	1.00	0.70	0.60	0.47	0.99	1.03	0.73	0.60
VBDVS/FAC5	0.84	0.84	0.57	0.46	1.00	1.00	0.75	**0.41**
MODELS WITH 60 FACTORS								
SSVS/FAC60	0.76	0.88	0.89	0.66	1.11	1.34	1.53	1.27
ELN/FAC60	0.86	0.90	0.89	0.74	1.18	1.24	1.24	1.17
VBDVS/FAC60	0.89	0.68	**0.50**	**0.42**	1.00	0.96	0.70	0.51
MODELS WITH 439 PREDICTORS								
ELN/X	0.86	0.84	1.01	0.70	1.22	1.19	1.14	1.29
PLS/X	**0.81**	0.86	1.01	0.89	1.32	1.61	1.46	1.33
VBDVS/X	0.90	**0.66**	0.51	0.43	1.00	1.00	**0.68**	0.43

[25]NOTES: The AR model serves as a benchmark and its entries (shown in italics) are the regular values of the mean squared forecast error (MSFE) for each forecast horizon $h = 1, 4, 8, 12$ quarters. Entries for each subsequent model are MSFEs relative to the values of the AR benchmark. Entries in boldface indicate the best performing model for each forecast horizon.

VBDVS/FAC5 and VBDVS/FAC60 are also doing well at horizon $h = 12$, even though they marginally outperform the UCSV model that has TVPs but no exogenous predictors.

There are several interesting stylized facts we can derive from these two tables. First, all TVP models, whether they feature exogenous predictors or not, do a good job in forecasting all measures of inflation especially at longer horizons. Second, in horizon $h = 1$, time-varying parameter models are overall better than the benchmark AR(2); however, models with constant parameters and many exogenous predictors seem to be doing even better. Third, while all three TVP models with no predictors (SBAR-UCSV-TVPAR) clearly improve over the benchmark as the forecast horizon increases, this is not true for all models with exogenous predictors. The usefulness of exogenous predictors seems to be hit-or-miss, greatly depending on the specification (linear or nonlinear) as well as the efficiency of penalization. For example, the GP regression seems to be improving with the horizon, whereas ELN/X forecasts can deteriorate as the forecast horizon increases. The important thing to notice here is that VBDVS/FAC5, VBDVS/FAC60, and VBDVS/X are all improving with the forecast horizon, showing that the algorithm is able to penalize sufficiently many of the irrelevant predictors as well as shrink excess time variation in parameters. For example, for total CPI, the performance of the heavily parameterized VBDVS/X follows quite close that of the parsimonious structural breaks autoregressive (SBAR) model, indicating that the algorithm shrinks VBDVS/X toward significantly smaller dimensions. However, for other dependent variables, the performance of VBDVS/X is completely different (and many times better) than more parsimonious TVP models, indicating that the algorithm is utilizing important information in exogenous predictors (and does so in a superior way to constant parameter regression algorithms such as ELN/X and PLS/X).

To obtain a more detailed picture around patterns of predictability, Figures 5 and 6 plot the cumulative sum of squared forecast errors over the full out-of-sample evaluation period.



NOTES: The *y*-axis shows unnormalized squared forecast errors (that is, not relative to a benchmark) such that lower values signify better forecast performance.

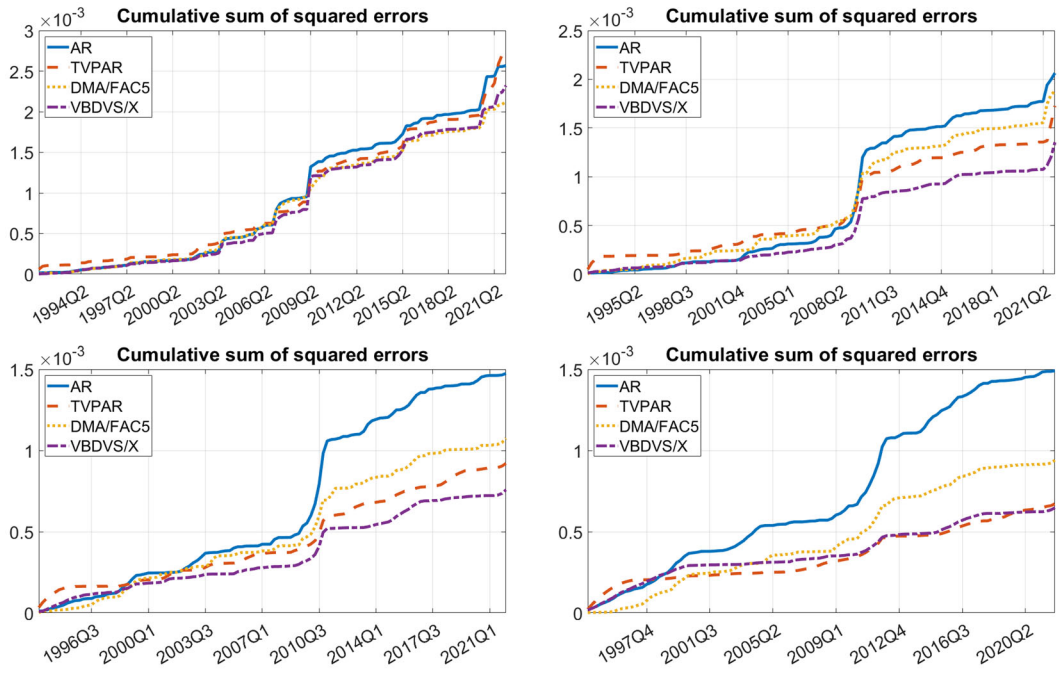
FIGURE 5

CUMULATIVE SQUARED FORECAST ERRORS OF GDP DEFLATOR INFLATION FORECASTS FOR FOUR KEY CLASSES OF MODELS, FOR $h = 1, 4, 8, 12$

These cumulated values are not normalized relative to a benchmark, as in the previous tables, which means that the lower a line in this plot is, the lower the squared forecast error of the model associated with that line. For the sake of clarity, we only present results for the AR(2) benchmark, the TVP-AR(2), the DMA/FAC5, and the full VBDVS/X models. It is clear that for $h = 1$, substantial differences in the forecasting performance of different models occur mainly right after the beginning of the Global Recession (circa 2007–08), whereas in longer horizons, TVP regressions have substantial gains over the AR(2) early in the evaluation sample. When forecasting GDP deflator at horizons $h = 1$, the AR, TVP-AR, and VBDVS models are indistinguishable, whereas the DMA method dominates. However, for one-step ahead forecasts of CPI, all three TVP specifications, whether they have predictors or not, clearly dominate the AR(2) benchmark upon the outburst of the Global Recession. It is interesting to observe in the longest horizon, $h = 8$, that the TVP-AR(2) and VBDVS models provide identical forecasts up around 2005 (GDP deflator) or 2003 (CPI), but after that, VBDVS clearly dominates. Therefore, these patterns comply with previous evidence documented in papers such as Stock and Watson (2007), Faust and Wright (2013), and Koop and Korobilis (2012): TVP and general structural break autoregressions have long been relevant for forecasting inflation, but also sporadically throughout the sample and especially after the Global Recession, there are several, potentially short-lived, predictors that can be relevant for forecasting inflation.

5. CONCLUSIONS

We introduce a comprehensive methodology for forecasting time-series data using TVP dynamic regression models. The algorithm allows to estimate regressions with hundreds of pre-



NOTES: The *y*-axis shows unnormalized squared forecast errors (that is, not relative to a benchmark) such that lower values signify better forecast performance.

FIGURE 6

CUMULATIVE SQUARED FORECAST ERRORS OF TOTAL CPI INFLATION FORECASTS FOR FOUR KEY CLASSES OF MODELS, FOR $h = 1, 4, 8, 12$

dictors, that is, a much larger information set than ever considered in the literature of macroeconomic TVP regressions (a literature that spans half a century; see Cooley and Prescott, 1976). Although it is rarely the case in macroeconomic forecasting for excessively large models to be among the best performing in forecasting, we test our new algorithm in a heavily parameterized specification (VBDVS/X) and find it provides consistently good forecasts, especially in longer horizons. Most importantly, the algorithm is able to work well even when applied to smaller information sets (e.g., when five factors are used as predictors), showing its potential as a fast and flexible forecasting tool in the toolbox of applied economists. Code that replicates the results in the Monte Carlo and empirical exercises is provided by the authors.[26]

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1
Data S1

REFERENCES

AMIR-AHMADI, P., C. MATTHES, and M.-C. WANG, "Choosing Prior Hyperparameters: With Applications to Time-Varying Parameter Models," *Journal of Business & Economic Statistics* 38 (2020), 124–36.

[26] https://sites.google.com/site/dimitriskorobilis/matlab/vbdvs

Angelino, E., M. J. Johnson, and R. P. Adams, "Patterns of Scalable Bayesian Inference," *Foundations and Trends®in Machine Learning* 9 (2016), 119–247.

Bauwens, L., G. Koop, D. Korobilis, and J. V. Rombouts, "The Contribution of Structural Break Models to Forecasting Macroeconomic Series," *Journal of Applied Econometrics* 30 (2015), 596–620.

Beal, M. J., "Variational Algorithms for Approximate Bayesian Inference," Ph.D. Thesis, University College London, 2003.

———, and Z. Ghahramani, "The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures," in J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, eds., *Bayesian Statistics*, Vol. 7 (Oxford: Oxford University Press, 2003), 453–64.

Belloni, A., and V. Chernozhukov, "Least Squares After Model Selection in High-Dimensional Sparse Models," *Bernoulli* 19 (2013), 521–47.

Belmonte, M. A., G. Koop, and D. Korobilis, "Hierarchical Shrinkage in Time-Varying Parameter Models," *Journal of Forecasting* 33 (2014), 80–94.

Bitto, A., and S. Frühwirth-Schnatter, "Achieving Shrinkage in a Time-Varying Parameter Model Framework," *Journal of Econometrics* 210 (2019), 75–97, annals Issue in Honor of John Geweke "Complexity and Big Data in Economics and Finance: Recent Developments from a Bayesian Perspective."

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe, "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association* 112 (2017), 859–77.

Breiman, L., "Bagging Predictors," *Machine Learning* 24 (1996), 123–40.

Byrne, J. P., D. Korobilis, and P. Ribeiro, "On the Sources of Uncertainty in Exchange Rate Predictability," *International Economic Review* 59 (2018), 329–57.

Chan, J., and I. Jeliazkov, "Efficient Simulation and Integrated Likelihood Estimation in State Space Models," *International Journal of Mathematical Modelling and Numerical Optimisation* 1 (2009), 101–20.

Chan, J. C., G. Koop, R. Leon-Gonzalez, and R. W. Strachan, "Time Varying Dimension Models," *Journal of Business & Economic Statistics* 30 (2012), 358–67.

Clark, T. E., and F. Ravazzolo, "Macroeconomic Forecasting Performance under Alternative Specifications of Time-Varying Volatility," *Journal of Applied Econometrics* 30 (2015), 551–75.

Cogley, T., and T. J. Sargent, "Drifts and Volatilities: Monetary Policies and Outcomes in the Post WWII US," *Review of Economic Dynamics* 8 (2005), 262–302.

Cooley, T. F., and E. C. Prescott, "Estimation in the Presence of Stochastic Parameter Variation," *Econometrica* 44 (1976), 167–84.

Dangl, T., and M. Halling, "Predictive Regressions with Time-Varying Coefficients," *Journal of Financial Economics* 106 (2012), 157–81.

De Mol, C., D. Giannone, and L. Reichlin, "Forecasting Using A Large Number of Predictors: Is Bayesian Shrinkage a Valid Alternative to Principal Components?" *Journal of Econometrics* 146 (2008), 318–28, honoring the research contributions of Charles R. Nelson.

Dempster, A. P., N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1977), 1–38.

Faust, J., and J. H. Wright, "Forecasting Inflation," in *Handbook of Forecasting*, Handbook of Economic Forecasting Edited by Graham Elliott, Allan Timmermann Volume 2, Chapter 1 (Elsevier, 2013), 2–56.

Frazier, D. T., R. Loaiza-Maya, and G. M. Martin, "Variational Bayes in State Space Models: Inferential and Predictive Accuracy," Technical Report, arXiv:2106.12262, ArXiV, 2022.

George, E. I., and R. E. McCulloch, "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association* 88 (1993), 881–9.

Gerlach, R., C. Carter, and R. Kohn, "Efficient Bayesian Inference for Dynamic Mixture Models," *Journal of the American Statistical Association* 95 (2000), 819–28.

Giannone, D., M. Lenza, and G. E. Primiceri, "Economic Predictions with Big Data: The Illusion Of Sparsity," CEPR Discussion Papers 12256, C.E.P.R. Discussion Papers, August 2017.

Giordani, P., and R. Kohn, "Efficient Bayesian Inference for Multiple Change-Point and Mixture Innovation Models," *Journal of Business & Economic Statistics* 26 (2008), 66–77.

Giordano, R., T. Broderick, and M. I. Jordan, "Covariances, Robustness, and Variational Bayes," *Journal of Machine Learning Research* 19 (2018), 1–49.

Granger, C., "Non-Linear Models: Where Do We Go Next - Time Varying Parameter Models?," *Studies in Nonlinear Dynamics & Econometrics* 12 (2008), 1–9.

Jurado, K., S. C. Ludvigson, and S. Ng, "Measuring Uncertainty," *American Economic Review* 105 (March 2015), 1177–216.

Kalli, M., and J. E. Griffin, "Time-Varying Sparsity in Dynamic Regression Models," *Journal of Econometrics* 178 (2014), 779–93.

Koop, G., and D. Korobilis, "Forecasting Inflation Using Dynamic Model Averaging," *International Economic Review* 53 (2012), 867–86.
———, and S. M. Potter, "Estimation and Forecasting in Models with Multiple Breaks," *The Review of Economic Studies* 74 (2007), 763–89.
Korobilis, D., "High-Dimensional Macroeconomic Forecasting Using Message Passing Algorithms," *Journal of Business & Economic Statistics* 39 (2021), 493–504.
———, and K. Shimizu, "Bayesian Approaches to Shrinkage and Sparse Estimation," Technical Report, arXiv:2112.11751, ArXiV, 2021.
Kowal, D. R., D. S. Matteson, and D. Ruppert, "Dynamic Shrinkage Processes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81 (2019), 781–804.
MATLAB, "MATLAB Statistics and Machine Learning Toolbox," The MathWorks, Natick, MA, (2020).
McCracken, M., and S. Ng, "FRED-QD: A Quarterly Database for Macroeconomic Research," Working Paper 26872, National Bureau of Economic Research, March 2020.
McCulloch, R. E., and R. S. Tsay, "Bayesian Inference and Prediction for Mean and Variance Shifts in Autoregressive Time Series," *Journal of the American Statistical Association* 88 (1993), 968–78.
Naesseth, C. A., S. W. Linderman, R. Ranganath, and D. M. Blei, "Variational Sequential Monte Carlo," Technical Report, arXiv:1705.11140, ArXiV, 2017.
Nakajima, J., and M. West, "Bayesian Analysis of Latent Threshold Dynamic Models," *Journal of Business & Economic Statistics* 31 (2013), 151–64.
Narisetty, N. N., and X. He, "Bayesian Variable Selection with Shrinking and Diffusing Priors," *The Annals of Statistics* 42 (2014), 789–817.
Ormerod, J. T., and M. P. Wand, "Explaining Variational Approximations," *The American Statistician* 64 (2010), 140–53.
Pettenuzzo, D., and A. Timmermann, "Forecasting Macroeconomic Variables under Model Instability," *Journal of Business & Economic Statistics* 35 (2017), 183–201.
Rockova, V., and K. McAlinn, "Dynamic Variable Selection with Spike-and-Slab Process Priors," *Bayesian Analysis* 16 (2021), 233–69.
Rossi, B., "Chapter 21 - Advances in Forecasting under Instability," in G. Elliott and A. Timmermann, eds., *Handbook of Economic Forecasting*, Volume 2 (Elsevier, Amsterdam, 2013), 1203–324.
Särkkä, S., and A. Nummenmaa, "Recursive Noise Adaptive Kalman Filtering by Variational Bayesian Approximations," *IEEE Transactions on Automatic Control* 54 (March 2009), 596–600.
Shively, T. S., and R. Kohn, "A Bayesian Approach to Model Selection in Stochastic Coefficient Regression Models and Structural Time Series Models," *Journal of Econometrics* 76 (1997), 39–52.
Šmídl, V., and A. Quinn, *The Variational Bayes Method in Signal Processing*, Signals and Communication Technology (Springer, Berlin, 2006).
Stock, J. H., and M. W. Watson, "Why Has U.S. Inflation Become Harder to Forecast?," *Journal of Money, Credit and Banking* 39 (2007), 3–33.
———, and ———, "Chapter 8 - Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics," in J. B. Taylor and H. Uhlig, eds., *Handbook of Macroeconomics*, Volume 2 (Elsevier, Saint Louis, 2016), 415–525.
Tran, M.-N., D. J. Nott, and R. Kohn, "Variational Bayes with Intractable Likelihood," *Journal of Computational and Graphical Statistics* 26 (2017), 873–82.
Uhlig, H., "On Singular Wishart and Singular Multivariate Beta Distributions," *Annals of Statistics* 22 (1994), 395–405.
Uribe, P., and H. Lopes, "Dynamic Sparsity on Dynamic Regression Models," Technical Report, Available at http://hedibert.org/wp-content/uploads/2018/06/uribe-lopes-Sep2017.pdf, 2017.
Wang, H., H. Yu, M. Hoy, J. Dauwels, and H. Wang, "Variational Bayesian Dynamic Compressive Sensing," in *2016 IEEE International Symposium on Information Theory (ISIT)* (2016), 1421–25.
Wang, Y., and D. M. Blei, "Frequentist Consistency of Variational Bayes," *Journal of the American Statistical Association* 114 (2019), 1147–61.
Welch, I., and A. Goyal, "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction," *The Review of Financial Studies* 21 (2007), 1455–508.
West, M., and J. Harrison, *Bayesian Forecasting and Dynamic Models*, 2ndedition (Berlin, Heidelberg: Springer-Verlag, 1997).
Zou, H., and T. Hastie, "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67 (2005), 301-20.