Variational Bayes inference in high-dimensional time-varying parameter models

Koop, Gary and Korobilis, Dimitris

University of Strathclyde, University of Essex

15 July 2018

Variational Bayes inference in high-dimensional time-varying parameter models

Gary Koop[*]
University of Strathclyde

Dimitris Korobilis [†]
University of Essex

Abstract

This paper proposes a mean field variational Bayes algorithm for efficient posterior and predictive inference in time-varying parameter models. Our approach involves: i) computationally trivial Kalman filter updates of regression coefficients, ii) a dynamic variable selection prior that removes irrelevant variables in each time period, and iii) a fast approximate state-space estimator of the regression volatility parameter. In an exercise involving simulated data we evaluate the new algorithm numerically and establish its computational advantages. Using macroeconomic data for the US we find that regression models that combine time-varying parameters with the information in many predictors have the potential to improve forecasts over a number of alternatives.

Keywords: dynamic linear model; approximate posterior inference; dynamic variable selection; forecasting

JEL Classification: C11, C13, C52, C53, C61

[*] Strathclyde Business School, Glasgow, G4 0GE, UK, email: gary.koop@strath.ac.uk

[†] Essex Business School, Colchester, CO4 3SQ, UK, email: d.korobilis@essex.ac.uk

1 Introduction

Regression models that incorporate stochastic variation in parameters have been used by economists at least since the works of Sarris (1973) and Cooley and Prescott (1976). Granger (2008) argued that time-varying parameter models might become the norm in econometric inference since (as he illustrated via White's theorem) generic time-varying parameter (TVP) models can approximate any form of nonlinearity. Much empirical work shows the benefits of TVP models for forecasting. For instance, Stock and Watson (2007) show that their flexible TVP model with no predictors can forecast inflation more accurately than traditional constant parameter regressions based on the Phillips curve augmented with exogenous predictors. Extending such evidence, recent studies have developed novel Bayesian estimation algorithms that are able to combine time-varying parameter regressions with information in exogenous predictors. Relevant papers include Belmonte et al. (2014), Chan et al. (2012), Dangl and Halling (2012), Groen et al. (2013), Kalli and Griffin (2014), Koop and Korobilis (2012), Kowal et al. (2017), Nakajima and West (2013), Ročková and McAlinn (2018), and Uribe and Lopes (2017).

Such algorithms demonstrate various inventive ways of allowing for many predictor variables in a TVP regression setting. In practice, however, empirical application of all these algorithms is restricted to a handful of predictors and short (quarterly instead of monthly) data, because of their high complexity associated with their reliance on Markov Chain Monte Carlo (MCMC) or other computationally intensive estimation methods.[1] In light of evidence that information in many predictors can be beneficial in constant parameter regressions (Stock and Watson, 2002), the inability of existing

[1] The only exception is the recent work by Ročková and McAlinn (2018) that, alongside an MCMC algorithm, also proposes an expectation-maximization (EM) algorithm that searches for the mode of the posterior.

estimation algorithms to be used in high-dimensional TVP settings is a fundamental shortcoming. Therefore, an open question of interest in applied econometric research is whether models that combine large information sets with time-varying parameters could also be beneficial. This question doesn't just hold for the case of many exogenous predictors, but it is also important when the high dimensionality comes from using monthly or even daily data sets: Regressions with higher frequency data will be more likely to exhibit time-varying parameter behavior.[2]

In this paper, we fill this gap in the literature by developing an iterative algorithm that can handle regressions with many time series observations and/or many predictors in the presence of time-varying parameters. We use variational Bayes (VB) methods which allow us to approximate the true high-dimensional posterior distribution in a simple and straightforward manner. The main idea behind VB methods is to approximate the high-dimensional and intractable posterior distribution using a simpler, tractable distribution. VB methods ensure that the approximation is good by minimizing the Kullback-Leibler distance between the true posterior and the proposed approximation. Following a large literature in physics and engineering where the mean field approximation was first developed, our proposed approximation to the posterior is decomposed into a series of simpler, independent densities that make inference scalable in high dimensions. We tackle computation by means of the an optimization algorithm that has as output the first two moments of the posterior density and resembles the expectation-maximization (EM) algorithm, instead of relying on computationally intensive MCMC methods. The result is an algorithm that combines Kalman filter updates for time-varying coefficients and volatilities with trivial posterior updates of all other model parameters and, hence, we call it the Variational Bayes Kalman Filter (VBKF).

[2] See Bauwens et al. (2015) for a comparison of the number of estimated breaks in monthly vs quarterly macroeconomic time series using a variety of structural breaks and time-varying parameter models.

The use of the VBKF surmounts the computational problem associated with TVP regressions with many predictors. However, on its own it does not surmount over-parameterization concerns. Accordingly, we derive a dynamic version of the stochastic search variable selection (SSVS) prior of George and McCulloch (1993) and incorporate it into the VBKF. This prior allows us to implement dynamic variable selection by stochastically searching for probable predictors at each point in time. While this time-varying variable selection problem is typically of tremendous complexity[3], we are able to integrate it to our efficient VBKF setting. Therefore, the proposed dynamic SSVS prior extends existing dynamic model selection and shrinkage algorithms (e.g. Kalli and Griffin, 2014; Koop and Korobilis, 2012) to high-dimensional regression problems. Finally, we add to the VBKF algorithm for the time-varying regression coefficients, an approximate VBKF estimator for stochastic volatility (SV) models. This latter filter is as fast as the exponentially weighted moving average (EWMA) filter used in Koop and Korobilis (2012), but it is less ad-hoc and can also provide a full characterization of the posterior distribution of the volatility process instead of a point volatility estimate.

The purpose of these computationally efficient approximations, as well as the dynamic shrinkage and variable selection prior, is prediction. While approximation-free parameter estimation is equally important, there are several reasons we don't focus on this aspect of statistical inference using the proposed algorithm. First, even though asymptotic properties of general variational Bayes estimators have been derived in various regression settings (Wang and Blei, forthcoming), establishing consistency of our time-varying parameter estimators under a dynamic hierarchical prior is a non-trivial task. Second, for the kind of high-dimensional inference problems we are interested in, estimation error might be large. For example, our empirical exercise uses up to

[3] A traditional static variable selection problem with p predictors involves a model space of $K = 2^p$ possible models containing combinations of these predictors. The dynamic variable selection has to solve the static problem in all T observations associated with a given time series data set.

118 predictors, all featuring parameters that drift at each time period. In this case, the parameter space is so vast that regardless of whether using exact or approximate estimators the sampling error for TVP problems is high.[4] As a result, having a flexible and subjective shrinkage prior in our proposed algorithm is desirable as it leads to posterior mean estimates that might be biased, but provide a huge reduction in estimation variance (with benefits in terms of mean squared error compared to unbiased estimators that might have extremely large variance). This observation is confirmed by the fact that all the recent contributions in this field (see citations above) focus exclusively on forecasting, and not causal analysis using flexible TVP models.

We show, via a Monte Carlo exercise and an empirical application, that our proposed algorithm works well in high-dimensional sparse time-varying parameter settings. In the Monte Carlo exercise we compare the numerical accuracy of our algorithm against an established algorithm in the literature, namely the Dynamic Model Averaging (DMA) algorithm with forgetting factors and EWMA stochastic volatility used in Raftery et al. (2010) and Koop and Korobilis (2012). We note that, of the Bayesian algorithms in this literature, DMA is the main one which does not involve the use of MCMC methods and, thus, suffers less from the computational burdens associated with MCMC. Thus, we treat DMA as the most important competitor to our proposed VBKF methods. We show that dynamic variable selection VBKF estimates of time-varying parameters and stochastic volatilities are on average more accurate than those obtained by DMA. Most importantly, algorithmic complexity is very low compared to DMA when the number of observations and/or number of predictors increases. Our empirical work follows much of

[4] In addition, when using Markov chain Monte Carlo 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 dimension 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 $\sqrt{2}$. Therefore, in high dimensions approximate inference algorithms may be preferred relative to MCMC-based posterior algorithms; see the excellent discussion of these issues in Angelino et al. (2016).

the relevant literature such as Stock and Watson (2007), Chan et al. (2012) and Kalli and Griffin (2014). That is, we forecast US GDP and price inflation. Using TVP regressions with up to 118 predictors, we compare our algorithm with a wide range of competing state-of-the-art algorithms for estimating TVP regressions including DMA and many which involve use of MCMC methods. We do find evidence in favor of combining time-varying parameters with many predictors, although the dynamic shrinkage/selection prior shrinks heavily the full model towards a TVP regression with few important predictors.

The remainder of the paper proceeds as follows. Section 2 briefly describes the basic principles of VB inference for approximating intractable posteriors. Section 3 introduces the our econometric specification and outlines the proposed VBKF algorithm. Section 4 contains our Monte Carlo study where we document the benefits of using this algorithm against an important competitor: DMA. Section 5 contains our forecasting exercise involving US macroeconomic data which compares our methods to a range of TVP alternatives. Section 6 concludes.

2 Bayesian Inference Using Variational Bayes Methods

Before we describe our specific model and how VB can be used with it, we provide a generic discussion of variational Bayes methods in approximating intractable posterior distributions. Variational Bayes methods have grown in popularity as a way of approximating posterior densities which are difficult to analyze using MCMC methods; see Blei, Kucukelbir and McAuliffe (2017), Ormerod and Wand (2010) and Wand (2017) for recent surveys relating to machine learning and statistics and Hajargasht and Wozniak (2018) for a recent econometric application. Consider data y, latent variables s and parameters θ. Our interest lies in time-varying parameter models which are state space models. Hence, s represents the unobserved time-varying regression coefficients and error variances and θ all other parameters such as the error variances in the state equations. The joint posterior of interest is $p\left(s, \theta | y\right)$ with associated marginal likelihood $p\left(y\right)$ and joint density $p\left(y, s, \theta\right)$. When the joint posterior is computationally intractable, we can define an approximating density $q\left(s, \theta\right)$ that belongs to a family of simpler distributions. The main idea behind variational Bayes inference is to make this approximating density $q\left(s, \theta\right)$ as close as possible to $p\left(s, \theta | y\right)$, where distance is measured using the Kullback-Leibler divergence:

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

Note that $KL \geq 0$, and equals zero iff $q\left(s, \theta\right) = p\left(s, \theta | y\right)$.

Insight for why KL is a desirable distance metric arises from a simple re-arrangement involving the log of the marginal likelihood (see also Ormerod and Wand, 2010, page 142) where it can be shown that

$$\begin{aligned} \log p\left(y\right) &= \log \int p\left(y, s, \theta\right) \mathrm{d}s\mathrm{d}\theta = \log \int q(s, \theta) \frac{p\left(y, s, \theta\right)}{q(s, \theta)} \mathrm{d}s\mathrm{d}\theta & (2) \\ &= \int q\left(s, \theta\right) \log\left\{\frac{p\left(y, s, \theta\right)}{q\left(s, \theta\right)}\right\} \mathrm{d}s\mathrm{d}\theta + KL, & (3) \end{aligned}$$

which finally gives

$$p\left(y\right) \geq \exp\left[\int q\left(s, \theta\right) \log\left\{\frac{p\left(y, s, \theta\right)}{q\left(s, \theta\right)}\right\} d\theta\right] \equiv \mathscr{F}\left(q(s, \theta)\right)$$

where we emphasize that $\mathscr{F}$ is a functional on the distribution $q(s, \theta)$. Maximizing

$\mathscr{F}\left(q(s,\theta)\right)$ over $q\left(s,\theta\right)$ thus amounts to finding an approximation which has an estimated marginal likelihood as close as possible to the correct $p\left(y\right)$. This procedure is also equivalent to minimizing the KL distance between the approximating and the true posterior.

The lower bound $\mathscr{F}\left(q(s,\theta)\right)$ can be maximized iteratively by using calculus of variations. If we use a mean field factorization of the form $q\left(s,\theta\right) = q\left(\theta\right)q\left(s\right)$ then it can be shown that the optimal choices for $q\left(s\right)$ and $q\left(\theta\right)$ are

$$q\left(s\right) \propto \exp\left[\int q(\theta)\log p\left(s|y,\theta\right)\mathrm{d}\theta\right], \tag{4}$$

$$q\left(\theta\right) \propto \exp\left[\int q(s)\log p\left(\theta|y,s\right)\mathrm{d}s\right]. \tag{5}$$

VB algorithms iterate over these two densities until convergence is reached. Due to the similarities with the EM algorithm of Dempster, Laird and Rubin (1977), this iterative procedure in its general form is 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. Like Gibbs sampling, (4) and (5) involve the full conditional posterior distributions. But unlike Gibbs sampling, the VB-EM algorithm does not repeatedly simulate from them and thus, typically, is computationally much faster.

Our implementation of VB methods for time varying parameter regressions with a shrinkage prior leads leads to simple forms for (4) and (5). The scheme we use relies on three assumptions. First, the complete-data likelihood for y, θ and s comes from the exponential family. Second, all priors need to be conditionally conjugate to the likelihood. Third, it assumes a factorization $q\left(s,\theta\right) = q\left(s\right)q\left(\theta\right)$. The first two assumptions are not at all restrictive. Most macroeconometric models assume Normal errors, and conjugate Bayesian analysis is desirable in most settings. The third assumption is harmless if θ and s have low posterior correlation and can thus be safely factorized into independent components. As we show in detail in the next section, this assumption can indeed be fully exploited in the TVP regression setting. For example, for parameters such as state equation error variances, we expect the correlation with the states to be typically weak.

3 VB Inference in High-Dimensional TVP Regressions

In this paper, we work with the univariate[5] TVP regression model with stochastic volatility of the form

$$\begin{aligned} y_t &= x_t\beta_t + \sigma_t\varepsilon_t && (6) \\ \beta_t &= \beta_{t-1} + \eta_t && (7) \\ \log\left(\sigma_t^2\right) &= \log\left(\sigma_{t-1}^2\right) + \zeta_t && (8) \end{aligned}$$

where y_t is the time t value of the dependent variable, $t = 1, .., T$, x_t is a $1 \times p$ vector of predictors and lagged dependent variables, $\varepsilon_t \sim N(0,1)$, $\eta_t \sim N(0, Q_t)$ with Q_t a $p \times p$ diagonal matrix, and $\zeta_t \sim N(0, r_t)$. In likelihood-based analysis of this model it is standard to assume that ε_t, η_t and ζ_t are independent of one another and we adopt this assumption. The assumption of diagonality of the state covariance matrix Q_t is not a standard assumption in the literature, although it has been used in some cases; see for example Belmonte et al. (2014). As argued in the introduction, our interest lies in prediction and not parameter estimation. The diagonality assumption allows for a more parsimonious econometric specification, less cumbersome derivations of posterior

[5] Our estimation methodology can also be adapted to the multivariate case, e.g. the TVP Vector Autoregressive model, with minor adjustments.

distributions, and faster computation – with these three characteristics being particularly important in Big Data forecasting applications. For future reference, note that we use a notational convention where j,t subscripts denote the j^{th} element of a time varying state or parameter and $1:t$ subscripts denoting all the states/parameters/data up to time t.

Variational Bayes methods can be used with state space models such as the TVP model given in (6), (7) and (8). When there are large numbers of predictors it is important to add prior shrinkage to avoid over-parameterization problems. In this paper, we follow ideas in Wang et al (2016) and add to the state space model in (6) and (7) an additional hierarchical prior which shrinks the states towards zero. While these authors use Student-t shrinkage via a Normal-inverse Gamma mixture prior, we instead use a dynamic version of the variable selection mixture prior of George and McCulloch (1993). This dynamic prior takes the form

$$\begin{aligned} \beta_{j,t}|\gamma_{j,t} &\sim (1-\gamma_{j,t})\, N\left(0, \underline{v}_{j,0}^2\right) + \gamma_{j,t} N\left(0, \underline{v}_{j,1}^2\right), && (9) \\ \gamma_{j,t} &\sim Bernoulli\left(\underline{\pi}_{j,0}\right), \quad j = 1,...,p, && (10) \end{aligned}$$

where $\underline{v}_{j,0}^2, \underline{v}_{j,1}^2$ are fixed prior variances with $\underline{v}_{j,0}^2 \to 0$ and $\underline{v}_{j,1}^2 \to \infty$, and $\underline{\pi}_{j,0}$ is a fixed prior hyperparameter. Under this specification the prior hyperparameter $\gamma_{j,t}$ is a Bernoulli variable which decides which mixture component applies as a prior distribution for the coefficient $\beta_{j,t}$. If $\gamma_{j,t} = 1$ the prior of $\beta_{j,t}$ is diffuse (Normal with a very large variance) and estimation of this parameter using the data is unrestricted. If $\gamma_{j,t} = 1$ the prior of $\beta_{j,t}$ is approximately a point mass at zero[6] and the posterior of this coefficient

[6] Notice that $\underline{v}_{j,0}^2$ is set to be small, but not exactly zero. In Bayesian analysis there exist specifications where $\underline{v}_{j,0}^2 = 0$, and then the SSVS prior is simply called a spike and slab prior, where the spike is exactly a point mass at zero. However, as George and McCullogh (1997) argue, posterior inference in the spike and slab case is more cumbersome as it requires several computationally expensive evaluations involving the likelihood function.

will also be restricted to be very close to zero, and the effect of the j-th predictor is removed from the regression at time t. It becomes apparent that under this variable selection prior setting, $\underline{\pi}_{j,0}$ is the prior probability of inclusion of predictor j in the TVP regression. We also adopt conditionally conjugate priors for the state variance parameters:

$$\begin{aligned} q_{j,t}^{-1} &\sim Gamma\left(\underline{c}_0, \underline{d}_0\right), j = 1, ..., p, && (11) \\ r_t^{-1} &\sim Gamma\left(\underline{f}_0, \underline{g}_0\right), && (12) \end{aligned}$$

where $\underline{c}_0, \underline{d}_0, \underline{f}_0, \underline{g}$ are fixed prior hyperparameters. The model is completed by defining the initial condition of the two state variables, namely

$$\begin{aligned} \beta_0 &\sim N\left(\underline{\beta}_0, \underline{P}_0\right), && (13) \\ \log \sigma_0^2 &\sim N\left(\log \underline{\sigma}_0^2, \underline{R}_0\right). && (14) \end{aligned}$$

Up to this point the definitions of likelihood and priors are mainly standard and similar specifications are commonly used with TVP regressions. The novel feature in our specification is the dynamic variable selection prior of equations (9) and (10), so the question arises as to how to incorporate this prior into our methods of posterior computation. First we note that, while equation (7) is the second layer of a hierarchical regression, for the Bayesian it can be viewed as a hierarchical prior for the regression coefficients β_t of the form $\beta_t|\beta_{t-1}, Q_t \sim N\left(\beta_{t-1}, Q_t\right)$. Second, we follow Wang et al. (2016) and write the dynamic SSVS prior as a prior for latent data (pseudo-observations) $z_{j,t} = 0$ which is of the form

$$z_{j,t} \sim N\left(\beta_{j,t}, v_{j,t}\right), \qquad (15)$$

where we define $v_{j,t} = \left(1 - \gamma_{j,t}\right)^2 \underline{v}_{j,0}^2 + \gamma_{j,t}^2 \underline{v}_{j,1}^2$ and V_t is the $p \times p$ diagonal matrix comprising the elements $v_{j,t}$. We show in the Technical Appendix that by combining these two priors for β_t, we obtain the following state equation:

$$\beta_t = \widetilde{F}_t \beta_{t-1} + \widetilde{\eta}_t, \tag{16}$$

where $\widetilde{\eta}_t \sim N\left(0, \widetilde{Q}_t\right)$, with parameter matrices $\widetilde{Q}_t = \left(Q_t^{-1} + V_t^{-1}\right)^{-1}$ and $\widetilde{F}_t = \widetilde{Q}_t Q_t^{-1}$.

The vector of states is $s = (\beta_{1:T}, \log \sigma^2_{1:T})$ and the vector of other parameters is $\theta = (q_{1:T}, \gamma_{1:T}, r_{1:T})$. Consequently, the posterior distribution for the joint vector of states and parameters is of the form

$$\begin{aligned} p\left(s, \theta | y_{1:T}, z_{1:T}\right) \;\; \propto \;\; & \prod_{t=1}^{T} p\left(\beta_t | \beta_{t-1}, Q_t\right) p\left(\log \sigma_t^2 | \log \sigma_{t-1}^2, r_t\right) p\left(y_t | \beta_t, \log \sigma_t^2\right) & (17) \\ & p\left(z_t | \beta_t, V_t\right) p\left(\gamma_t\right) p\left(Q_t\right) p\left(r_t\right). & (18) \end{aligned}$$

While this joint posterior is analytically intractable, the conditional posteriors are tractable and thus MCMC methods can be used. But, when the number of predcitors is large, this would be computationally burdensome. In order to deal with these challenges, in this paper we apply the following mean field VB approximation

$$q\left(s, \theta\right) \equiv q\left(\beta_{1:T}\right) q\left(\log \sigma_{1:T}^2\right) \prod_{t=1}^{T} \left(q\left(r_t\right) \times \prod_{j=1}^{p} q\left(v_{j,t}\right) q\left(\gamma_{j,t}\right) q\left(q_{j,t}\right) \right). \tag{19}$$

Notice that we want to decompose the parameters q_t, v_t, γ_t into components that are independent over t and over j, in order to facilitate computation. However, we don't want to factorize $\beta_{1:T}$ and $\log \sigma_{1:T}^2$ over time, because this means that posterior estimates would be independent at each time period, which is surely not a realistic assumption for TVP regression models that specifically assume that time-varying parameters evolve dynamically as random walks.

Using this mean field approximation we can derive a VBKF that is simple and resembles the popular EM algorithm for maximum likelihood estimation of state-space models that was proposed by Shumway and Stoffer (1982) but includes the SSVS shrinkage prior which is crucial in avoiding over-parameterization concerns. As discussed in the preceding section, the optimal choices for the components that make up $q\left(s,\theta\right)$ are the conditional posterior distributions. These are given in the Technical Appendix. Further details, derivations and theoretical justifications of such VB algorithms are given in Beal (2003).

In the previous section we highlighted the fact that in order to derive the algorithm, two necessary conditions are that the likelihood belongs to the exponential family and that the priors that are conditionally conjugate. With one exception, all of the posterior conditionals of the TVP regression with shrinkage prior meet these conditions. The one exception is for the volatility process. This arises from the fact that the stochastic volatility model is not a linear Normal state space model. Hence, we need to use an alternative approximation for $q\left(\log\sigma^2_{1:T}\right)$.

Note that the state space model with states $\log\sigma^2_t$ can be transformed so as to be a linear state space model with measurement error that is distributed as $\log -\chi^2$ with one degree of freedom. To show this, consider equations (6) and (8) and, assuming β_t known, bring the term $x_t\beta_t$ on the left hand side, take squares and then logarithms. This produces the following state-space model

$$\begin{aligned} \widetilde{y}_t &= \log\sigma^2_t + w_t, & (20) \\ \log\sigma^2_t &= \log\sigma^2_{t-1} + \zeta_t, & (21) \end{aligned}$$

where $\widetilde{y}_t = \log\left(\left(y_t - x_t\beta_t\right)^2\right)$ and $w_t = \log\varepsilon^2_t$.[7]

[7] It is common to add a very small offset constant to the transformed dependent variable to avoid

Kim, Shephard and Chib (1998) apply a mixture of Normals approximation to this $\log - \chi^2$ distributed error w_t. Our VBKF approximation cannot handle the mixture of Normals, so instead we approximate the $\log - \chi^2$ distribution with a single Normal distribution with mean and variance matching those of the $\log - \chi^2$. As we show in the Technical Appendix, by doing such an approximation we lose information in the left tail of the $\log - \chi^2$ which corresponds to large negative values of the log-volatility parameter $\log \sigma_t^2$. In our empirical work, we standardize our data prior to analysis to have unconditional sample variance equal to one so large negative values are unlikely to arise. Additionally, we argue that we are not immediately interested in forecasts of volatility, rather we want forecasts of y_t and these are likely to be only slightly affected by using an approximation which becomes poor only in the tails of the distribution. Finally, having an approximate stochastic volatility estimator should still work much better than the case of having a constant volatility, since it is established that stochastic volatility is extremely important for macro forecasting (see, among many others, Clark and Ravazzolo, 2015). The next two sections establish that this is the case and our volatility estimator works very well – much better than the approximate EWMA volatility estimator used in Koop and Korobilis (2012).

Algorithm 1 below outlines our VBKF algorithm. All the detailed algorithmic steps are provided in the Technical Appendix, and here we only demonstrate the general form of the new algorithm. We have found that that this algorithm will normally iterate only a few times. This takes much less computational resources compared to obtaining tens of thousands of MCMC draws. Convergence is typically achieved by assessing whether the values of the parameters have changed substantially from one iteration to the next. Hence, we define the stopping rule $\| s_t^{(r)} - s_t^{(r-1)} \| \rightarrow 0$, where $s_t = (\beta_t, \log \sigma_t^2)$, the symbol $\| \bullet \|$ denotes the Euclidean norm, r denotes the replication number and $t|t$ subscripts

numerical instabilities.

denote Kalman filter estimates of time t quantities given data through period t.

Algorithm 1 *Variational Bayes Kalman Filter (VBKF) pseudo-algorithm in a TVP regression with stochastic volatility*

```
Initialize $\underline{\beta}_0, \log \underline{\sigma}_0, \underline{P}_0, \underline{R}_0, \underline{c}_0, \underline{d}_0, \underline{f}_0, \underline{g}_0, \underline{v}_{j,0}, \underline{v}_{j,1}, \underline{\pi}_{j,0}$
for $t = 1$ to $T$ do
    r=1;
    while $\|\beta_{t|t}^{(r)} - \beta_{t|t}^{(r-1)}\| \rightarrow 0$ and $\| \left(\log \sigma_{t|t}^{2}\right)^{(r)} - \left(\log \sigma_{t|t}^{2}\right)^{(r-1)} \| \rightarrow 0$ do
        1. Perform Kalman filter updating of $\beta_{t|t}$ based on the state-space model consisting of equations (6) and (16)
        2. Update $\gamma_{j,t}$ and $q_{j,t}$ $\forall$ $j \in 1, p$ from their analytical conditional posteriors (see details in Technical Appendix)
        3. Based on step 2, construct matrices $Q_t$ and $V_t$ (see equation (15)), and subsequently $\widetilde{F}_t$ and $\widetilde{Q}_t$ (see equation (16)), to be used in the next iteration
        4. Perform Kalman filter updating of $\log \sigma_{t|t}^{2}$ based on the state-space model consisting of equations (20) and (21)
        5. Update $r_t$ from its analytical conditional posterior (see details in Technical Appendix)
        $r = r + 1$
    end while
    Upon convergence, set $\beta_t = \beta_t^{(r)}$ and $\log \sigma_t^2 = \left(\log \sigma_{t|t}^{2}\right)^{(r)}$, and do forecasting using standard formulas for dynamic regression models
end for
```

4 Simulation study

In this section we evaluate the performance of the new estimator using artificial data. Although we view the algorithm as primarily a forecasting algorithm, it is also important to investigate its estimation properties in an environment where we know the true data generating process (DGP). Thus, we wish to to establish that the VBKF is able to track time-varying parameters satisfactorily and establish that the dynamic variable selection prior is able to perform shrinkage and selection with high accuracy (at least in cases where we know that the DGP is that of a sparse TVP regression model). We also wish to investigate the computational gains that can be achieved by using our algorithm compared to the dynamic model averaging (DMA) approach of Koop and Korobilis (2012) which is based on a computationally efficient dynamic shrinkage algorithm that does not use MCMC methods.

We do not consider MCMC methods as benchmarks when assessing the numerical precision of VBKF, even though we have several MCMC-based algorithms in the next section when doing a full-fledged forecast comparison using real data. We do know that MCMC methods will converge to the exact posterior whereas VB methods are approximate. On top of that, MCMC estimates of time-varying parameters are less noisy because they are smoothed estimates, while VBKF estimates are filtered. Having smoothed estimates is important for reliable parameter estimation in-sample, but when forecasting smoothing does not play a role. Therefore, there is no practical need to establish numerical precision of MCMC relative to VBKF in-sample, however, it is extremely important to establish their relative performance when forecasting out-of-sample (something we do in the next section). Variational Bayes methods are scalable to very large dimensions where MCMC methods are not and, thus, they can be used for forecasting even when the number of predictors in a TVP regression becomes very large. Accordingly, the main aim of this section is to establish that VBKF methods, although approximate, yield reasonable results and that they are comparable to established approximate algorithms such as DMA.

As a consequence, our Monte Carlo study involves generating data from sparse time-varying parameter DGPs and comparing VBKF against DMA. This latter algorithm is dynamically averaging over many state space models, where the states in each model are estimated using exponential discounting. In particular, the time-varying regression coefficients are estimated using a so-called forgetting factor Kalman filter (FFKF) and the time varying error variance is estimated using an exponentially weighted moving average (EWMA) filter. Given the recursive nature of these filters, time t estimates are readily available given past information. The exponential weighting scheme implies that recent observations take more weight than older observations, that is, it is a rolling estimation scheme with an adaptively changing window of observations that allows faster or slower changes in parameters over different periods. This algorithm being fast allows to enumerate all possible models using p predictors, estimate them all efficiently using a single pass of the Kalman filter algorithm, and then average using some measure of fit. For p predictors DMA requires estimation of all 2^p models, which can be cumbersome for $p >> 20$, even after accounting for the fact that all these models can be estimated easily in parallel using modern multi-core processors. Therefore, DMA can be thought of as "deterministic variable selection" because all 2^p models need to be enumerated and estimated. Our use of the SSVS prior in the VBKF algorithm allows for a more efficient "stochastic variable selection" by visiting probabilistically only the best (according to marginal likelihoods) specifications among all possible models.

For the DMA procedure we set the forgetting and decay factors as in Koop and Korobilis (2012), and the reader is referred to that paper for more information about the effect of such choices and their justification. The forgetting factor is set to 0.96 and the decay factor, which controls the amount of time-variation in the error variance, is set to 0.94. These choices allow for substantial time variation in both regression coefficients and variances, and they are calibrated so as to comply with the amount of time variation we allow in the DGP (which is described next). DMA also involves a model averaging forgetting factor which controls how fast model switching occurs and we set this to 0.99.[8] Additional details and references about the method are provided in the Technical Appendix.

[8] These factors could be estimated from the data by specifying a grid of values for each and optimizing over them. However, this substantially adds to the computational burden.

We use DGPs of the following form:

$$y_t = \beta_{1t}x_{1t} + \beta_{2t}x_{2t} + ... + \beta_{pt}x_{pt} + \sigma_t\varepsilon_t \tag{22}$$

$$\beta_{it} = d_i \times \theta_{it} \tag{23}$$

$$d_i = \begin{cases} 0 & \text{with probability } \ \underline{\pi} \\ 1 & \text{with probability } \ 1-\underline{\pi} \end{cases} \tag{24}$$

$$\theta_t = \underline{c} + \underline{\gamma}\left(\theta_{t-1} - \underline{c}\right) + \underline{\delta}\eta_t \tag{25}$$

$$\log\left(\sigma_t^2\right) = \underline{\mu} + \underline{\phi}\left(\log\left(\sigma_{t-1}^2\right) - \underline{\mu}\right) + \underline{\xi}\zeta_t \tag{26}$$

$$\theta_0 \sim \underline{\theta}, \qquad \log\left(\sigma_t^2\right) = \underline{\sigma}, \tag{27}$$

where $\beta_t = (\beta_{1t}, \beta_{2t}, ..., \beta_{pt})$ is a vector of p regression coefficients at time t, d_i for $i = 1, .., p$ is a Bernoulli random variable that determines whether the coefficients, β_{it}, are zero or not, and $\theta_t = (\theta_{1t}, \theta_{2t}, ..., \theta_{pt})$. The errors in all equations, $\varepsilon_t, \eta_t, \zeta_t$, are standard Normal and independent of one another and over time. All variables with an underscore are fixed so as to define the DGP. We set $\underline{\pi} = 0.5$, $\underline{\gamma} = 0.99$, $\underline{\phi} = 0.98$, $\underline{\delta} = T^{-3/4}$, $\underline{\xi} = T^{-1/2}$, $\underline{\theta} \sim U(-2, 2)$, $\underline{\sigma} = 0.2$, $\underline{c} = \underline{\theta}$, $\underline{\mu} = \underline{\sigma}$. The chosen value of $\underline{\pi}$ implies that, on average, only half of the predictors are included in the TVP regression. Note that all methods estimate time-varying coefficients and variances which evolve as random walks, but the parameters in equations (25) and (26) of the DGP are generated from mean-reverting AR processes. We set $\underline{\gamma}$ and $\underline{\phi}$ to values slightly smaller than one in order to make sure we don't generate explosive values for y_t. Finally, we generate predictor variables from $x_t \sim N(0, S)$, where S is a $p \times p$ matrix of correlations with i, j element generated as $S_{ij} = \rho^{|i-j|}$.

We generate models with different number of predictors p, number of observations T, and correlation coefficient for the predictors ρ. In particular, we generate models with $p = 4, 8, 12$ predictors, $T = 100, 200$ observations and $\rho = 0, 0.9$ correlation intensity for the predictor variables. This gives a total of 12 possible DGPs to compare. Note that the VBKF methodology works with many predictors, but DMA cannot handle very large number of predictors which is why $p = 12$ is the maximum number of predictors we consider in this section. From each DGP, we generate 500 data sets.

For the VBKF we use the following default priors:

$$\begin{aligned}
\beta_{j,t}|\gamma_{j,t} &\sim (1-\gamma_{j,t})\, N\left(0, 0.0001^2\right) + \gamma_{j,t} N\left(0, 2^2\right), && (28)\\
\gamma_{j,t} &\sim \mathit{Bernoulli}\,(0.5)\,, && (29)\\
q_{j,t}^{-1} &\sim \mathit{Gamma}\,(1000, 1)\,, && (30)\\
r_t^{-1} &\sim \mathit{Gamma}\,(100, 1)\,, && (31)\\
\beta_0 &\sim N\,(0, 2 \times I)\,, && (32)\\
\log \sigma_0^2 &\sim N\,(0, 0.1)\,. && (33)
\end{aligned}$$

Before discussing numerical results based on all 500 of the data sets generated from each DGP, we present parameter estimates using a single, randomly generated data set for $T = 200$, $p = 8$ and $\rho = 0.9$. Figure 1 plots the true values of the eight time-varying parameters and the respective VBKF and DMA estimates using this data set. This data set has randomly chosen four of the regression coefficients to be non-zero and time-varying. For these, we can see that parameter tracking in real time for coefficients $\beta_{1t}, \beta_{4t}, \beta_{5t}, \beta_{6t}$ is quite accurate using both methods. This accuracy is particularly noteworthy since both VBKF and FFKF are filtering methods and thus the estimates are not smoothed. VBKF and DMA estimates of coefficient β_{6t} lie slightly below the true value, but some bias is to be expected as both these methods can be thought of as time-varying, Bayesian versions of classical penalized estimators which are known to be biased. As long as interest lies in forecasting, such biases are welcome in high dimensions because they are typically accompanied by much lower variances of estimates and a reduction in mean square error. For the remaining four coefficients that were set to zero in the DGP, both methods accurately indicate that their values are zero. The partial exceptions are for β_{2t} and β_{8t}, where for an initial period the VBKF estimate is slightly different than zero, before eventually being shrunk to zero. Similarly, Figure 2 plots the time-varying volatilities from VBKF and DMA against the true values. It can be seen that both estimates track satisfactorily the true values in real time. Thus, overall we are finding both approaches to estimate time varying coefficients and volatilities quite well.

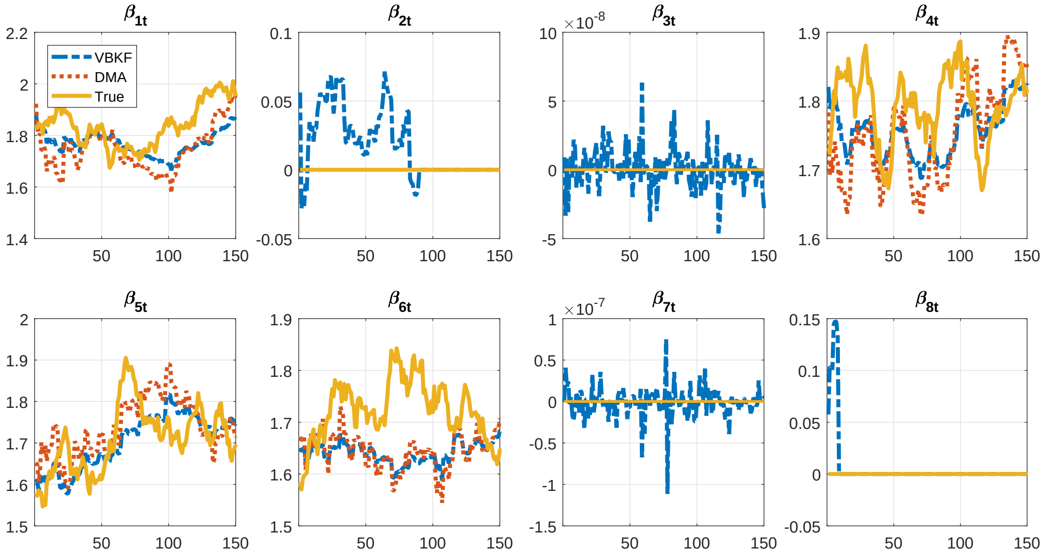

Figure 1: *True values of generated coefficients in the sparse time-varying parameter regression DGP with $T = 200$, $p = 8$, and $\rho = 0.9$, plotted against the VBKF and FFKF estimates. The VBKF uses a dynamic variable selection prior, while the FFKF is combined with a dynamic model averaging (DMA) procedure that enumerates all possible model combinations using the $p = 8$ predictors. The first 50 observations are not plotted in order to remove the effect of initial conditions on both filtering methods.*

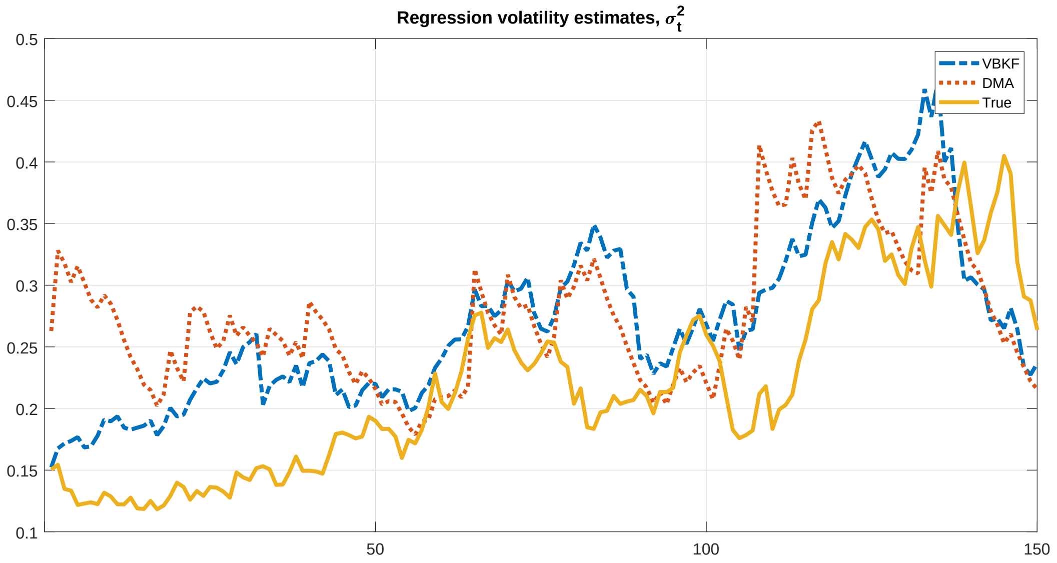

Figure 2: *True generated volatility in the sparse DGP with $T = 200$, $p = 8$, plotted against estimates from VB and EWMA filters. The first 50 observations are not plotted in order to remove the effect of initial conditions on both filtering methods.*

Results averaged over 500 artificially generated data sets for each of our DGPs are presented in Table 1. Entries in this table are mean squared deviations (MSD) averaged over the 500 data sets and T time periods. Results relating to the p-dimensional vector of regression coefficients β_t further average over p. To be precise, if we write the true artificially generated coefficients as $(\beta_t^{true}, \sigma_t^{true})$ and the estimates from VBKF and DMA as $\left(\beta_t^j, \sigma_t^j\right)$, for $j = VBKF, DMA$, we calculate MSD as

$$MSD_\beta^j = \frac{1}{500} \sum_{r=1}^{500} \sum_{t=1}^{T} \sum_{i=1}^{p} \left(\beta_{it}^{true,(r)} - \beta_{it}^{j,(r)}\right)^2, \tag{34}$$

$$MSD_\sigma^j = \frac{1}{500} \sum_{r=1}^{500} \sum_{t=1}^{T} \left(\sigma_t^{true,(r)} - \sigma_t^{j,(r)}\right)^2 \tag{35}$$

where $r = 1, ..., 500$ denotes the number of Monte Carlo iterations. The table also presents CPU times measured in seconds per Monte Carlo draw.

Regarding MSD results for the time-varying coefficients, in most cases VBKF with the dynamic shrinkage prior has lower estimation error than DMA. Note here that, using the algorithm of Koop and Korobilis (2012), we could have presented results for dynamic model selection (DMS) where a single best model is selected at each point in time. For brevity, we do not present such results since we found DMS to be substantially inferior in this Monte Carlo study. Similarly, we do not present results from a simple benchmark such as rolling OLS. Rolling OLS produced MSDs which are several times higher than VBKF and DMA. The poor performance of rolling OLS is due to the fact that the true time-varying parameter vector is sparse, in which case procedures such as (unrestricted) rolling OLS are condemned to be over-parameterized and not track coefficients well. Regarding volatility estimates, the picture is similar. Our approximate VBKF filter performs better in most cases than the EWMA filter used in the DMA algorithm of Koop and Korobilis (2012). Overall, we are finding the VBKF to work well in an absolute sense, but also relative to DMA.

In terms of computation times, DMA is faster when using four variables. This is because with DMA one needs to estimate 2^4 models but in each only one run of the Kalman filter is required. By constrast, VBKF involves running the Kalman filter until a convergence criterion is met. In practice in this Monte Carlo study, this amounts to running the equivalent of five to 10 Kalman filter iterations. However, as the number of predictors increases, DMA clearly reveals its computational disadvantage. The number of models DMA estimates is 2^p and, thus, computation increases commensurately. Computation time for VKBF, in contrast, increases at an approximately linear rate. Thus, VKBF is a computationally feasible algorithm, even with hundreds or more predictors, whereas the computational burden of DMA becomes enormous even when $p = 20$. Clearly, VBKF is a scalable algorithm whereas DMA is not.

Table 1: *Mean squared deviations and average CPU time per Monte Carlo iteration*

	$p = 4$ predictors				$p = 8$ predictors				$p = 12$ predictors			
	$T = 100$		$T = 200$		$T = 100$		$T = 200$		$T = 100$		$T = 200$	
	$\rho = 0$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.9$
					MSD (PREDICTOR COEFFICIENTS)							
VBKF	**0.0519**	**0.1002**	**0.0233**	0.0701	**0.0850**	0.1636	**0.0398**	0.0957	0.1480	**0.2230**	**0.0529**	**0.0849**
DMA	0.0531	0.1256	0.0244	**0.0527**	0.0896	**0.1583**	0.0486	**0.0905**	**0.1222**	0.2504	0.0584	0.1005
					MSD (STOCHASTIC VOLATILITY)							
VBKF	**0.0991**	**0.1526**	0.0467	0.0560	**0.1280**	**0.3809**	**0.1724**	**0.2959**	**0.5122**	**1.3126**	**0.4942**	**0.1651**
DMA	0.1152	0.1604	**0.0435**	**0.0405**	0.9036	1.6338	0.3090	0.6047	1.6993	7.1928	1.2160	3.2218
					CPU TIME (SEC)							
VBKF	0.20	0.34	0.51	0.36	**0.59**	**0.71**	**1.12**	**1.29**	**0.83**	**0.79**	**1.07**	**1.80**
DMA	**0.04**	**0.04**	**0.08**	**0.08**	1.99	1.98	4.01	4.06	81.83	86.84	159.11	164.41

Notes: CPU times are based on MATLAB 2017b 64-bit on a Windows 7 machine with Core i7-8700K processor running in stock clock speed. All calculations rely on MATLAB's built-in capabilities as well as the statistics toolbox, but without utilizing the parallel computing toolbox. Both VBKF and DMA can be trivially parallelized but in completely different ways, meaning that parallel processing times can differ substantially from the times we report in this table.

5 Macroeconomic Forecasting with Many Predictors

5.1 Data and forecasting models

In this section we investigate the performance of the new VBKF algorithm in an application that involves forecasting two important macroeconomic variables, GDP growth and inflation, using many predictors. Our data set includes these two variables and 116 other quarterly US time series variables for the period 1959Q1 - 2015Q4. A detailed description of our data set and the transformations done to each variable are provided in the Data Appendix.

Our largest TVP regressions thus involve a dependent variable (inflation or GDP growth) along with 117 exogenous predictors (the 116 other variables plus either inflation or GDP growth) along with an intercept and two lags of the dependent variable. Thus, they contain 120 right-hand side variables. We also select two subsets of the exogenous predictors involving five and 16 potentially important predictors which have been commonly used for macroeconomic forecasting in other studies.[9] These smaller data sets, as noted below, are used with some of the comparative methods which are too computationally burdensome to use with the full data set. The predictors and dependent variables are standardized, and then forecasts are transformed back to the original scale. We use the direct method of forecasting. Forecasts are evaluated over the last 50% of the sample, for horizons $h = 1, 2, 3, 4$ quarters ahead.

We forecast with four different variants of the VBKF which involve different numbers of predictors as well as a range of popular competitors which involve either time variation in parameters or structural breaks in regression or AR models. We include a variety of specifications for parameter change and a variety of data configurations:

- **VBKF1**: TVP regression with only the intercept and the two lags of the dependent variable. These parameters are always included in each specification, so they have an unrestricted Normal prior. That is, their unrestricted prior is a special case of the the dynamic SSVS prior where $\gamma_{j,t} = 1$ for all t and for all j corresponding to intercept and lags of the endogenous variable.
- **VBKF2**: Extends VBKF1 by adding the set of five important predictors.
- **VBKF3**: Extends VBKF1 by adding the set of 16 important predictors.
- **VBKF4**: Extends VBKF1 by adding all available 117 predictors.
- **KP-AR**: Structural break AR(2) model based on Koop and Potter (2007).
- **GK-AR**: Structural break AR(2) model based on Giordani and Kohn (2008).

[9] The 16 variables have mnemonics 'EXUSUK' 'OILPRICEx' 'HOUST' 'S&P 500' 'T10YFFM' 'CUMFNS' 'HWI' 'AWHMAN' 'AWOTMAN' 'AMDMNOx' 'AMDMUOx' 'TB3MS' 'AAAFFM' 'BAAFFM' 'PPICMM' 'CES3000000008'. The five-variable data set uses the first five of these 16 variables. See the Data Appendix for exact definitions.

- **TVP-AR**: TVP-AR(2) model with stochastic volatility similar to Pettenuzzo and Timmerman (2017).

- **UCSV**: The unobserved components stochastic volatility model of Stock and Watson (2007) is a special case of a TVP regression with no predictors - it is a local level state-space model featuring stochastic volatility in the state equation.

- **TVD**: The time-varying dimension (TVD) model of Chan et al. (2012) using five predictors. This is the first of three alternative TVD specifications proposed by the authors. To ease the computational burden (and following Chan et al.) we do dynamic model selection over a model space containing models with a single predictor or all five predictors (but not 2, 3 or 4 predictors).

- **TVS**: The time-varying shrinkage (TVS) algorithm of Kalli and Griffin (2014) using five predictors.

- **TVP-BMA**: Groen et al (2013) develop methods for doing Bayesian model averaging with TVP regressions. We use their algorithm with 16 predictors.

- **TVP-LASSO**: Belmonte et al. (2014) show how to incorporate the Bayesian lasso prior in TVP regressions, in order to shrink coefficients either towards zero or towards a constant parameter specification. We use this approach with 16 predictors.

- **DMA**: The DMA algorithm as implemented in Koop and Korobilis (2012) with 16 predictors.

- **SSVS**: The constant parameter regression version of the SSVS prior was first developed in George and McCulloch (1993). We use this algorithm with the full set of 117 predictors.

We stress that, with the exception of VBKF and the static SSVS algorithm, the computational demands of the other approaches become overwhelming with the full data set, which is why the other approaches are limited to 16 or fewer exogenous predictors. In addition, we have one constant coefficient regression with shrinkage of a similar sort to that used in our VBKF so as to investigate the importance of time-variation in parameters. All models, except for UCSV, include at least an intercept and two lags of the dependent variable. Prior shrinkage is only done on the exogenous predictors and not on the intercept or AR lags. The prior for VBKF methods is the one specified in the Monte Carlo study. The following is a list which summarizes and offers a brief description of all the forecasting methods. Appendix C provides details (including prior hyperparameter choices) of all the competing methods.

5.2 Estimation Results

Before presenting the results of the forecasting comparison, we demonstrate some evidence on what VBKF is estimating in the TVP regression model involving all 117 predictors. We focus on the $h = 1$ case.

Figures 3 and 4 plot the time-varying posterior inclusion probabilities for the most important predictors of GDP growth and inflation. The first point to note about both of these figures is that our dynamic shrinkage prior is indeed shrinking a large number of coefficients to zero. Out of 117 possible predictors, only a small number (21 for GDP growth and 18 for inflation) have high posterior inclusion probabilities for appreciable periods of time. In both cases, approximately 100 predictors are being shrunk to zero in all periods. A second point to stress is that there is a great deal of time variation in these inclusion probabilities. If a predictor were always important, then the posterior inclusion probability would be near one for the entire sample. No variable exhibits this characteristic.

For both inflation and GDP growth, there is a tendency (with several exceptions) for posterior inclusion probabilities to be highest in the late 1970 through the 1980s and be lowest at the beginning and end of the sample. Interesting exceptions to this occur for the inflation forecasts where two variables (wage inflation in manufacturing and the growth in real personal income) become important predictors only around the time of the financial crisis.

These estimation results establish that our VKBF methods with hierarchical shrinkage can effectively ensure parsimony in a time-varying manner in a TVP regression.

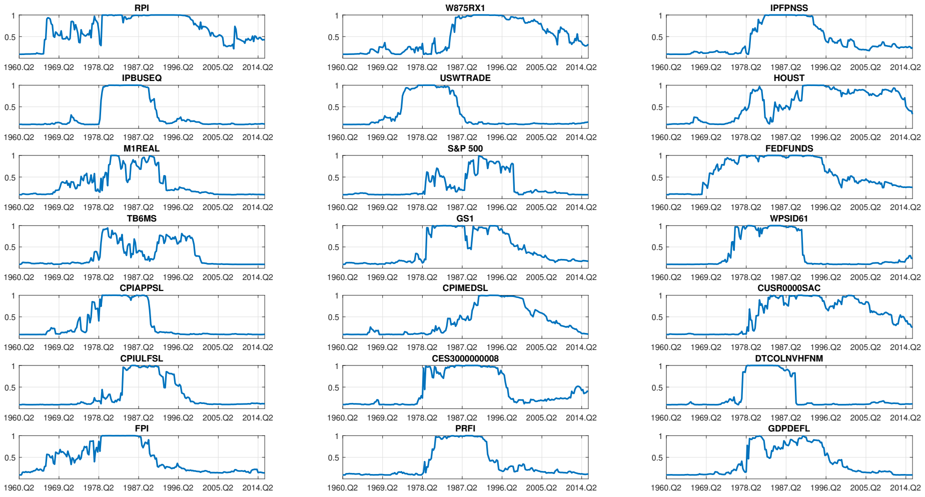

Figure 3: *Posterior inclusion probabilities for the most important predictors of GDP growth ($h = 1$). Only predictors which have probability higher than* 0.5 *for at least 10 quarters are plotted.*

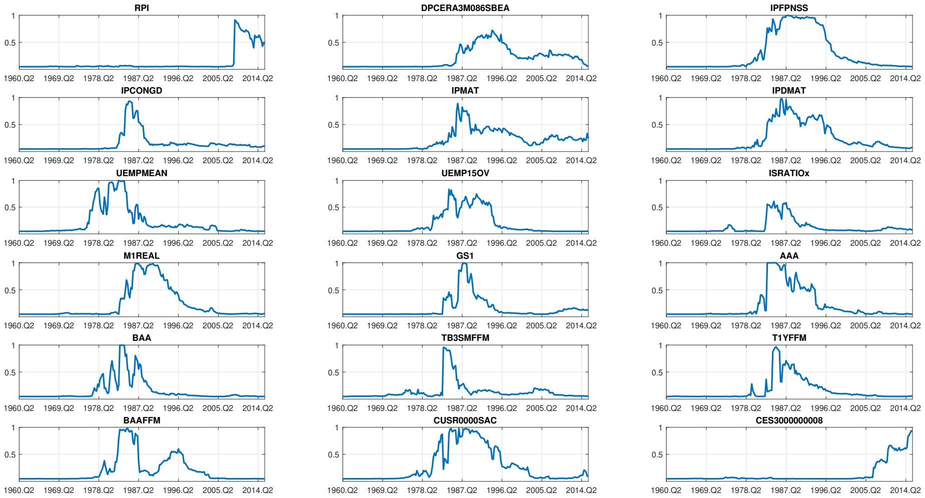

Figure 4: *Posterior inclusion probabilities for the most important predictors of inflation ($h = 1$). Only predictors which have probability higher than* 0.5 *for at least 10 quarters are plotted.*

Figure 5 presents the volatility estimates from the VBKF4 model compared to those produced by DMA. Note that for DMA we use the smaller data set of 16 predictors. Although broadly similar, there are differences between the VKBF and DMA volatility estimates with the former being more stable and less erratic than the latter. Note that for GDP growth DMA is producing very high and erratic volatilities both at the beginning of the sample and around the time of the financial crisis. These features are greatly muted by VBKF. For inflation, VBKF and DMA volatility estimates are mostly similar, but at the time of the financial crisis DMA is producing a large, "noisy" spike in volatility which is absent for VBKF.

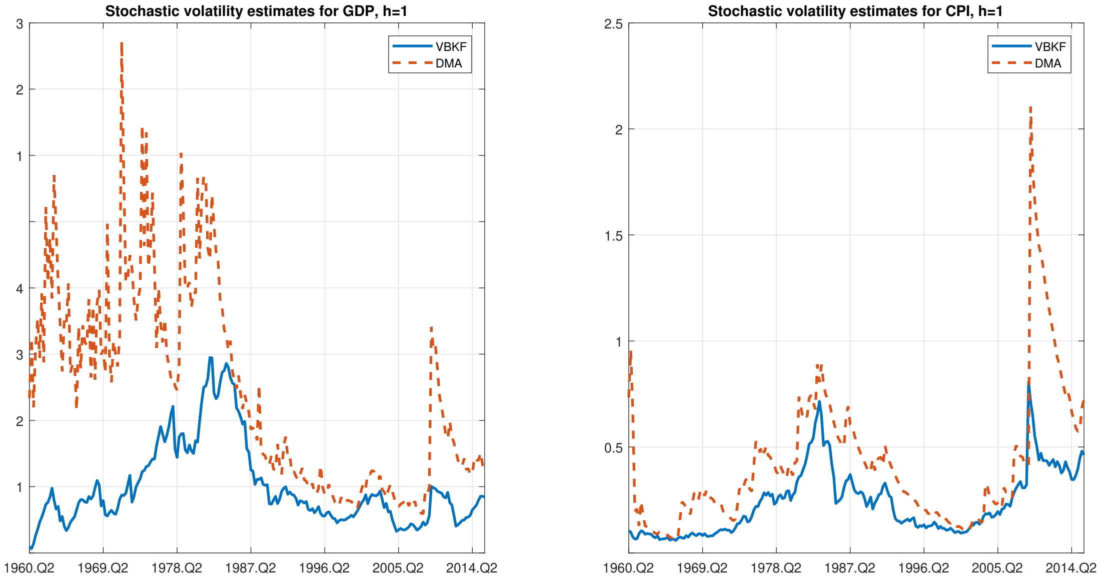

Figure 5: *Stochastic volatility estimates for GDP growth (left panel) and inflation (right panel). The blue solid line is for VKBF, the red dashed line is for DMA.*

5.3 Forecasting results

In this subsection we report the results of our forecast comparison using Mean Squared Forecast Errors (MSFEs) and averages of log predictive likelihoods (APLs) as measures of point and density forecast performance, respectively. Both are benchmarked against the AR(2). For MSFEs we present ratios of the MSFE of a given model relative to that of the AR(2), such that values lower than one signify better performance of the model relative to the benchmark. For APLs we subtract off the AR(2) APL and, thus, positive numbers indicate a forecasting method is beating the benchmark.[10]

We, thus, have 2 forecast metrics, 4 forecast horizons and 2 variables which makes 16 comparisons possible. Different forecasting approaches do well in some cases and less well in others. But a general story we are finding is that VBKF often forecasts best,

[10] To aid in interpretation, note that sums of log predictive likelihoods, which can be interpreted in a similar fashion as marginal likelihoods or information criteria, can be obtained by multiplying APLs by the number of observations in the forecast evaluation period. The latter is $112 - h$.

particularly when we use APLs as our measure of forecast performance. And it never forecasts poorly in the sense that is always easily beats the AR(2) benchmark. Other approaches do not have these properties. We provide evidence on these points in the remainder of this subsection.

Probably the best overall approach, other than VBKF, is the simple UCSV model. When using MSFE as a forecast metric, UCSV beats VBKF for GDP growth forecasting and for one quarter ahead inflation forecasting. But this ranking is overturned when using APLs where VBKF approaches beat UCSV. Furthermore, there is a case ($h = 2$ inflation forecasts) where UCSV forecasts very poorly, losing out to the AR(2) benchmark. VBKF methods never are beaten by the AR(2). The TVP-AR(2) model exhibits similar patterns, but with a slightly worse forecast performance overall.

DMA is found to be a robust method, never losing out to the AR(2) benchmark. But (with only a couple of exceptions involving MSFE performance of long run forecasts) VBKF forecasts better.

Of the remaining, MCMC-based, methods (regardless of whether they are structural break or TVP models), none of them provides a consistently better forecast performance than VBKF. Indeed TVP-BMA and TVP-LASSO tend to forecast quite poorly, often being beaten by the AR(2) benchmark and never being selected as the best forecasting method for either variable for any forecast horizon. TVD and TVS tend to forecast better and sometimes beat VBKF (e.g. TVS forecasts very well at short horizons).

Another issue worth discussing is whether including a large number of predictors can improve forecast performance. Here the evidence is more mixed. The relatively good performance of methods with no predictors such as UCSV and TVP-AR(2) lends some support to the idea that simple parsimonious methods are adequate (although we do stress that these methods are typically beaten by VBKF with large numbers of predictors). Of course, even if a small number of predictors is enough to forecast US inflation and GDP growth, that does not undermine the contribution of the present paper. Developing econometric methods which will work even with a huge number of predictors is useful, even if forecast improvements in one particular empirical application are not large. But if we compare VBKF1, VBKF2, VBKF3 and VBKF4 (which differ only in the number of exogenous predictors included), we do (with some exceptions) tend to see clear improvements in forecast performance as more predictors are included. Particularly at longer forecast horizons, these improvements are appreciable. See, for instance, the large improvements in APLs and MSFEs for $h = 3$ and $h = 4$ for both variables obtained by VBKF4 relative to VBKF1. With some exceptions, a similar pattern is found with shorter forecast horizons as well.

The discussion of the previous paragraph raises the issue as to whether VBKF is forecasting well simply because it can handle more variables. If this were true, this would only strengthen our argument that developing econometric methods capable of handling more variables is useful. But even when we compare approaches with the same number of predictors (e.g. comparing VBKF3 to TVP-LASSO which both involve 16 predictors), we find VBKF to be forecasting as well or better than other approaches. Of the methods which use 5 predictors, TVS forecasts very well and (with some exceptions) forecasts GDP growth better than VBKF2 (which also has 5 predictors). However, for inflation VBKF2 tends to forecast slightly better than TVS. These two methods will only differ in the way prior shrinkage is done and in the way computation is done. Hence, it is reassuring to see the approximate VBKF method is forecasting as well as a state-of-the-art dynamic shrinkage prior in a case where such a comparison is possible. Of course, we only estimate TVS with the 5 variable data set since TVS will be much too computationally burdensome with larger data sets.

Finally, forecasts produced using the SSVS prior in the constant coefficient model are often very good. But there are exceptions where forecasts are very poor, failing to beat the AR(2) benchmark. See, for instance, the very poor MSFEs produced for long run inflation forecasts.

Overall, we are finding VBKF methods with an SSVS-based dynamic shrinkage prior to forecast well. They are comparable with the best alternatives where such a comparison is possible. But the key benefit of VBKF is that it can handle much larger number of predictors than other approaches.

Table 2: *MSFEs relative to AR(2) benchmark*

	GDP				CPI			
	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 1$	$h = 2$	$h = 3$	$h = 4$
Structural Breaks & TVP models - no predictors, MCMC-based								
KP-AR(2)	0.868	0.764	1.351	1.247	0.920	0.783	0.702	0.589
GK-AR(2)	1.081	1.087	1.075	1.069	0.945	0.934	0.920	0.936
TVP-AR(2)	0.872	0.916	1.029	0.995	1.034	1.052	0.805	0.835
UCSV	0.813	0.823	0.769	0.744	0.854	1.108	0.947	0.776
TVP models with predictors, MCMC-based								
TVD (2 lags, 5 predictors)	0.899	0.836	0.824	0.813	0.932	1.160	1.318	1.228
TVS (2 lags, 5 predictors)	0.957	0.863	0.792	0.757	0.943	0.833	0.827	0.757
TVP-BMA (2 lags, 16 predictors)	1.472	1.601	1.753	1.569	1.790	1.247	1.412	0.859
TVP-LASSO (2 lags, 16 predictors)	1.137	1.603	1.076	0.831	0.952	1.099	1.121	1.270
TVP models, not based on MCMC								
DMA (2 lags, 5 predictors)	0.902	0.740	0.703	0.685	0.961	0.806	0.729	0.668
VBKF1 (2 lags)	0.930	0.903	0.872	0.864	0.964	0.941	0.961	0.971
VBKF2 (2 lags, 5 predictors)	0.934	0.902	0.875	0.869	0.950	0.935	0.954	0.979
VBKF3 (2 lags, 16 predictors)	0.930	0.906	0.869	0.868	0.950	0.816	0.747	0.708
VBKF4 (2 lags, 118 predictors)	0.924	0.875	0.866	0.754	0.927	0.792	0.725	0.691
Constant parameter models with predictors, MCMC-based								
SSVS (2 lags, 118 predictors)	0.848	0.929	0.919	0.924	0.877	1.128	1.202	1.215

Table 3: *APLs relative to AR(2) benchmark*

	GDP				CPI			
	$h=1$	$h=2$	$h=3$	$h=4$	$h=1$	$h=2$	$h=3$	$h=4$
Structural Breaks & TVP models - no predictors, MCMC-based								
KP-AR(2)	0.003	-0.179	-0.004	0.076	0.127	0.058	0.235	0.655
GK-AR(2)	0.043	0.018	-0.003	0.014	0.073	0.054	0.181	0.022
TVP-AR(2)	0.082	0.215	-0.001	-0.048	0.273	0.167	0.128	0.037
UCSV	0.121	0.413	0.001	-0.110	0.474	0.279	0.075	0.517
TVP models with predictors, MCMC-based								
TVD (2 lags, 5 predictors)	0.161	0.610	0.276	0.172	0.674	0.392	0.223	0.067
TVS (2 lags, 5 predictors)	0.200	0.807	0.456	0.234	0.874	0.504	0.308	0.082
TVP-BMA (2 lags, 16 predictors)	-0.157	-0.202	-0.151	-0.235	0.207	0.018	0.159	0.093
TVP-LASSO (2 lags, 16 predictors)	-0.069	-0.558	0.234	0.446	0.189	0.661	0.076	0.823
TVP models, not based on MCMC								
DMA (2 lags, 5 predictors)	0.015	0.137	0.206	0.286	0.051	0.445	0.222	0.916
VBKF1 (2 lags)	0.126	0.371	0.496	0.537	0.386	0.912	0.360	0.637
VBKF2 (2 lags, 5 predictors)	0.108	0.379	0.538	0.344	0.387	0.797	0.359	0.737
VBKF3 (2 lags, 16 predictors)	0.092	0.337	0.476	0.647	0.273	0.706	0.633	0.706
VBKF4 (2 lags, 118 predictors)	0.391	0.947	0.598	0.639	0.236	0.599	0.724	0.763
Constant parameter models with predictors, MCMC-based								
SSVS (2 lags, 118 predictors)	-0.156	0.166	0.282	0.738	0.033	0.039	0.145	0.033

6 Conclusions: Feasible and Reasonable

In this paper, we have developed a method for doing Variational Bayesian inference in TVP regressions with stochastic volatility with a large number of predictors. Our findings may be summarized as: VKBF is feasible and reasonable. That is, it is computationally feasible even with over 100 predictors and can be scaled up to huge dimensions in a way other approaches cannot. And the empirical results (both in terms of estimation and forecasting) are reasonable. That is, VKBF forecasting results are typically among the best regardless of variable choice, forecast horizon and forecast metric despite the fact that VBKF is only approximating the posterior and predictive densities. In some cases, they are beaten by other approaches, but those other approaches cannot handle the large number of predictors that are increasingly being used in empirical macroeconomics and other fields. Furthermore, our approach (unlike all the others) never goes too far wrong. Thus, we have shown that VBKF is doing as well or better than existing approaches in models of dimension where such a comparison is possible and is computationally feasible in models of dimension where such a comparison is impossible.

References

[1] Angelino, E., Johnson, M. J. and R. P. Adams (2016). Patterns of scalable Bayesian inference. *Foundations and Trends$^{®}$ in Machine Learning* 9(2-3), 119-247.

[2] Bauwens, L., Koop, G., Korobilis, D. and J. V. K. Rombouts (2015). The contribution of structural break models to forecasting macroeconomic series. *Journal of Applied Econometrics* 30(4), 596-620.

[3] Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference.* PhD Thesis, Gatsby Computational Neuroscience Unit, University College London.

[4] Beal, M. J. and Z. Ghahramani (2003). The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures. In Bernardo, J.M., Dawid, A.P., Berger, J.O., West, M., Heckerman, D and, Bayarri, M.J., (Eds.) *Bayesian Statistics 7* 453-464, Oxford University Press.

[5] Belmonte, M., Koop, G. and D. Korobilis (2014). Hierarchical shrinkage in time-varying coefficients models. *Journal of Forecasting* 33, 80-94.

[6] Blei, D., Kucukelbir, A. and J. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* 112, 859-877.

[7] Chan, J., Koop, G., Leon-Gonzalez, R. and R. Strachan (2012). Time varying dimension models. *Journal of Business and Economic Statistics* 30(3), 358-367.

[8] Clark, T. E. and F. Ravazzolo (2015). Macroeconomic forecasting performance under alternative specifications of timevarying volatility. *Journal of Applied Econometrics* 30(4), 551-575.

[9] Cooley, T. F. and E. C. Prescott (1976). Estimation in the presence of stochastic parameter variation. *Econometrica* 44(1), 167-184.

[10] Dangl, T. and M. Halling (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics* 106, 157-181.

[11] Dempster, A. P., Laird, N. M. and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)* 39, 1-38.

[12] Frühwirth-Schnatter, S. and H. Wagner (2010). Stochastic model specification search for Gaussian and partial non-Gaussian state space models. *Journal of Econometrics* 154(1), 85-100.

[13] George, E. I. and R. E. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88, 881-889.

[14] George, E. I. and R. E. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica Sinica* 7, 339-373.

[15] Giordani, P. and R. Kohn (2008). Efficient Bayesian inference for multiple change-point and mixture innovation models. *Journal of Business and Economic Statistics*, 26(1) 66-77.

[16] Groen, J. J. J., Paap, R. and F. Ravazzollo (2013). Real time inflation forecasting in a changing world. *Journal of Business and Economic Statistics*, 31(1) 29-44.

[17] Hajargasht, G. and T. Wozniak (2018). Variational Bayes inference for large vector autoregressions. Manuscript.

[18] Kalli, M. and J. E. Griffin (2014). Time-varying sparsity in dynamic regression models. *Journal of Econometrics* 178, 779-793.

[19] Koop, G. and D. Korobilis (2012). Forecasting inflation using dynamic model averaging. *International Economic Review* 53, 867-886.

[20] Koop, G. and S. Potter (2007). Estimation and forecasting in models with multiple breaks. *Review of Economic Studies* 74(3), 763-789.

[21] Kowal, D. R., Matteson, D. S. and D. Ruppert (2017). Dynamic shrinkage processes. arXiv:1707.00763.

[22] Kulhavý, R. and F. Kraus (1996). On duality of regularized exponential and linear forgetting. *Automatica* 32,1403-1416.

[23] Kuo, L., and B. Mallick (1997). Variable selection for regression models. *Shankya: The Indian Journal of Statistics* 60 (Series B), 65-81.

[24] Nakajima, J. and M. West (2013). Bayesian analysis of latent threshold dynamic models. *Journal of Business and Economic Statistics* 31, 151164.

[25] Ormerod, J. and M. Wand (2010). Explaining variational approximations. *American Statistician* 64, 140-153.

[26] Pettenuzzo, D. and A. Timmermann (2017). Forecasting macroeconomic variables under model instability. *Journal of Business and Economic Statistics* 35(2), 183-201.

[27] Raftery, A., Karny, M., and P. Ettler (2010), Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics* 52, 52-66.

[28] Ročková, V. and K. McAlinn (2018). Dynamic variable selection with spike-and-slab process priors. Technical report, Booth School of Business, University of Chicago.

[29] Sarris, A. H. (1973). Kalman filter models: A Bayesian approach to estimation of time-varying regression coefficients. In: Berg, V. G. (Ed.) *Annals of Economic and Social Measurement* 2(4), 501-523. NBER: Cambridge, MA.

[30] Shumway, R. H. and D. S. Stoffer (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis* 3(4), 253-264.

[31] Stock, J. H. and M. W. Watson (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* 20(2), 147-162.

[32] Stock, J. H. and M. W. Watson (2007). Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking* 39(1), 3-33.

[33] Uribe, P. V. and H. F. Lopes (2017). Dynamic sparsity on dynamic regression models. Manuscript, available at http://hedibert.org/wp-content/uploads/2018/06/uribe-lopes-Sep2017.pdf.

[34] Wand, M. P. (2017). Fast approximate inference for arbitrarily large semiparametric regression models via message passing. *Journal of the American Statistical Association* 112, 137-168.

[35] Wang, H., Yu, H., Hoy, M., Dauwels, J. , and H. Wang (2016). Variational Bayesian dynamic compressive sensing, 2016 IEEE International Symposium on Information Theory.

[36] Wang, Y. and D. M. Blei (forthcoming). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, https://doi.org/10.1080/01621459.2018.1473776.

A Data Appendix

All data are obtained from St Louis Federal Reserve Bank economic database (FRED - https://fred.stlouisfed.org/). Series which are originally observed at monthly or higher frequencies are converted into quarterly values by taking averages over the quarter. Table A1 below gives the FRED mnemonics of each variable along with its description. The column Tcode denotes the transformations applied in order to convert variables to stationarity.

In particular, if $w_{i,t}$ is the original untransformed series in levels, when the series is used as a predictor in our models it is transformed according to the codes: 1 - no transformation (levels), $x_{i,t} = w_{i,t}$; 2 - first difference, $x_{i,t} = w_{i,t} - w_{i,t-1}$; 3- second difference, $x_{i,t} = \Delta w_{i,t} - \Delta w_{i,t-1}$ 4 - logarithm, $x_{i,t} = \log w_{i,t}$; 5 - first difference of logarithm, $x_{i,t} = \log w_{i,t} - \log w_{i,t-1}$; 6 - second difference of logarithm, $x_{i,t} = \Delta \log w_{i,t} - \Delta \log w_{i,t-1}$.

When the series is used as the variable to be predicted (i.e. as the dependent variable in the regression) the transformation codes are: 1 - no transformation (levels), $y_{i,t+h} = w_{i,t+h}$; 2 - first difference, $y_{i,t+h} = w_{i,t+h} - w_{i,t}$; 3- second difference, $y_{i,t+h} = \frac{1}{h}\Delta^h w_{i,t+h} - \Delta w_{i,t}$ 4 - logarithm, $y_{i,t+h} = \log w_{i,t+h}$; 5 - first difference of logarithm, $y_{i,t+h} = \log w_{i,t+h} - \log w_{i,t}$; 6 - second difference of logarithm, $y_{i,t+h} = \frac{1}{h}\Delta^h \log w_{i,t+h} - \Delta \log w_{i,t}$. In these transformations, $\Delta^h w_{t+h} = w_{t+h} - w_t$.

Table A1: *Quarterly US macro data set*

No	Mnemonic	Tcode	Description
1	RPI	5	Real Personal Income
2	W875RX1	5	RPI ex. Transfers

Table A1 (continued)

3	DPCERA3M086SBEA	5	Real PCE
4	CMRMTSPLx	5	Real M&T Sales
5	RETAILx	5	Retail and Food Services Sales
6	INDPRO	5	IP Index
7	IPFPNSS	5	IP: Final Products and Supplies
8	IPFINAL	5	IP: Final Products
9	IPCONGD	5	IP: Consumer Goods
10	IPDCONGD	5	IP: Durable Consumer Goods
11	IPNCONGD	5	IP: Nondurable Consumer Goods
12	IPBUSEQ	5	IP: Business Equipment
13	IPMAT	5	IP: Materials
14	IPDMAT	5	IP: Durable Materials
15	IPNMAT	5	IP: Nondurable Materials
16	IPMANSICS	5	IP: Manufacturing
17	IPB51222S	5	IP: Residential Utilities
18	IPFUELS	5	IP: Fuels
19	CUMFNS	2	Capacity Utilization: Manufacturing
20	HWI	2	Help-Wanted Index for US
21	HWIURATIO	2	Help Wanted to Unemployed ratio
22	CLF16OV	5	Civilian Labor Force
23	CE16OV	5	Civilian Employment
24	UNRATE	2	Civilian Unemployment Rate
25	UEMPMEAN	2	Average Duration of Unemployment
26	UEMPLT5	5	Civilians Unemployed $\leq$ 5 Weeks
27	UEMP5TO14	5	Civilians Unemployed 5-14 Weeks

Table A1 (continued)

28	UEMP15OV	5	Civilians Unemployed > 15 Weeks
29	UEMP15T26	5	Civilians Unemployed 15-26 Weeks
30	UEMP27OV	5	Civilians Unemployed > 27 Weeks
31	CLAIMSx	5	Initial Claims
32	PAYEMS	5	All Employees: Total nonfarm
33	USGOOD	5	All Employees: Goods-Producing
34	CES1021000001	5	All Employees: Mining and Logging
35	USCONS	5	All Employees: Construction
36	MANEMP	5	All Employees: Manufacturing
37	DMANEMP	5	All Employees: Durable goods
38	NDMANEMP	5	All Employees: Nondurable goods
39	SRVPRD	5	All Employees: Service Industries
40	USTPU	5	All Employees: TT&U
41	USWTRADE	5	All Employees: Wholesale Trade
42	USTRADE	5	All Employees: Retail Trade
43	USFIRE	5	All Employees: Financial Activities
44	USGOVT	5	All Employees: Government
45	CES0600000007	5	Hours: Goods-Producing
46	AWOTMAN	2	Overtime Hours: Manufacturing
47	AWHMAN	5	Hours: Manufacturing
48	HOUST	5	Starts: Total
49	HOUSTNE	5	Starts: Northeast
50	HOUSTMW	5	Starts: Midwest
51	HOUSTS	5	Starts: South
52	HOUSTW	5	Starts: West

Table A1 (continued)

53	AMDMNOx	5	Orders: Durable Goods
54	AMDMUOx	5	Unfilled Orders: Durable Goods
55	BUSINVx	5	Total Business Inventories
56	ISRATIOx	2	Inventories to Sales Ratio
57	M2REAL	5	Real M2 Money Stock
58	S&P 500	5	S&P 500
59	S&P: indust	5	S&P Industrial
60	S&P div yield	2	S&P Divident yield
61	S&P PE ratio	5	S&P Price/Earnings ratio
62	FEDFUNDS	2	Effective Federal Funds Rate
63	CP3M	2	3-Month AA Comm. Paper Rate
64	TB3MS	2	3-Month T-bill
65	TB6MS	2	6-Month T-bill
66	GS1	2	1-Year T-bond
67	GS5	2	5-Year T-bond
68	GS10	2	10-Year T-bond
69	AAA	2	Aaa Corporate Bond Yield
70	BAA	2	Baa Corporate Bond Yield
71	COMPAPFF	1	CP - FFR spread
72	TB3SMFFM	1	3 Mo. - FFR spread
73	TB6SMFFM	1	6 Mo. - FFR spread
74	T1YFFM	1	1 yr. - FFR spread
75	T5YFFM	1	5 yr. - FFR spread
76	T10YFFM	1	10 yr. - FFR spread
77	AAAFFM	1	Aaa - FFR spread

Table A1 (continued)

78	BAAFFM	1	Baa - FFR spread
79	EXSZUS	5	Switzerland / U.S. FX Rate
80	EXJPUS	5	Japan / U.S. FX Rate
81	EXUSUK	5	U.S. / U.K. FX Rate
82	EXCAUS	5	Canada / U.S. FX Rate
83	WPSFD49107	5	PPI: Final demand less energy
84	WPSFD49501	5	PPI: Personal cons
85	WPSID61	5	PPI: Processed goods
86	WPSID62	5	PPI: Unprocessed goods
87	OILPRICEx	5	Crude Oil Prices: WTI
88	PPICMM	5	PPI: Commodities
89	CPIAUCSL	5	CPI: All Items
90	CPIAPPSL	5	CPI: Apparel
91	CPITRNSL	5	CPI: Transportation
92	CPIMEDSL	5	CPI: Medical Care
93	CUSR0000SAC	5	CPI: Commodities
94	CUUR0000SAD	5	CPI: Durables
95	CUSR0000SAS	5	CPI: Services
96	CPIULFSL	5	CPI: All Items Less Food
97	CUUR0000SA0L2	5	CPI: All items less shelter
98	CUSR0000SA0L5	5	CPI: All items less medical care
99	PCEPI	5	PCE: Chain-type Price Index
100	DDURRG3M086SBEA	5	PCE: Durable goods
101	DNDGRG3M086SBEA	5	PCE: Nondurable goods
102	DSERRG3M086SBEA	5	PCE: Services

Table A1 (continued)

103	CES0600000008	5	Ave. Hourly Earnings: Goods
104	CES2000000008	5	Ave. Hourly Earnings: Construction
105	CES3000000008	5	Ave. Hourly Earnings: Manufacturing
106	MZMSL	5	MZM Money Stock
107	DTCOLNVHFNM	5	Consumer Motor Vehicle Loans
108	DTCTHFNM	5	Total Consumer Loans and Leases
109	INVEST	5	Securities in Bank Credit
110	GDP	5	Real Gross Domestic Product
111	PCDG	5	PCE: Durable Goods
112	PCESV	5	PCE: Services
113	PCND	5	PCE: Nondurable Goods
114	FPI	5	Fixed Private Investment
115	PRFI	5	Private Residential Fixed Investment
116	GCEC1	5	Government Cons Expenditures & Gross Inv
117	GDPDEFL	6	GDP deflator
118	PCEDEFL	5	PCE deflator

B Technical Appendix

In this appendix, we provide derivations and details of our VBKF algorithm for TVP regression with hierarchical prior shrinkage. We begin with the homoskedastic case. Subsequently we derive an approximate variational Bayes algorithm for estimation of stochastic volatility.

B.1 Variational Bayes inference in the homoskedastic TVP regression with variable selection prior

In this subsection, we use a regression model with time-varying coefficients and constant error variance of the form

$$\begin{aligned} y_{t+h} &= x_t \beta_t + \varepsilon_{t+h}, && \text{(B.1)} \\ \beta_t &= \beta_{t-1} + \eta_t, && \text{(B.2)} \end{aligned}$$

where β_t is a $p \times 1$ vector of time-varying parameters, $\varepsilon_{t+h} \sim N\left(0, \underline{\sigma}^2\right)$ with $\underline{\sigma}^2$ underlined to denote that in this subsection is considered a fixed/known parameter and $\eta_t \sim N\left(0, Q_t\right)$ with Q_t a $p \times p$ the state equation error covariance matrix. Notice that the state equation (B.2) implies a conditional prior on β_t of the form

$$\beta_t | \beta_{t-1}, Q_t \sim N\left(\beta_{t-1}, Q_t\right), \tag{B.3}$$

subject to the initial condition $\beta_0 \sim N\left(\underline{\beta}_0, \underline{P}_0\right)$. We assume that Q_t is a diagonal matrix with elements $q_{j,t}$, $j = 1, ..., p$, where

$$q_{j,t}^{-1} \sim Gamma\left(\underline{c}_0, \underline{d}_0\right). \tag{B.4}$$

We also impose an SSVS prior on $\beta_{j,t}$ of the form

$$\beta_{j,t}|\gamma_{j,t} \sim (1-\gamma_{j,t}) N\left(0, \underline{v}_{j,0}^2\right) + \gamma_{j,t} N\left(0, \underline{v}_{j,1}^2\right), \tag{B.5}$$

$$\gamma_{j,t} \sim Bernoulli\left(\underline{\pi}_0\right), \quad j = 1, ..., p. \tag{B.6}$$

This is a dynamic version of the SSVS prior of George and McCulloch (1993). With this prior $\underline{v}_{j,0}^2$ is chosen to be small and $\underline{v}_{j,1}^2$ is chosen to be large. If $\gamma_{j,t} = 0$, then $\beta_{j,t}$ has a small prior small variance $\underline{v}_{j,0}^2$ and the coefficient is shrunk to be near zero. Otherwise the coefficient evolves according to a random walk. We highlight the fact that, unlike other approaches to time varying shrinkage such as Chan et al. (2012), our dynamic SSVS prior is independent over time allowing for a high degree of flexibility. Discussion of what constitutes a "small" and "large" prior variance is given in George and McCulloch (1993). Our prior hyperparameter choices are given in Section 4.

In order to derive the posterior, we use a similar strategy to Wang et al. (2016) and write the SSVS prior in terms of pseudo-observations. To be precise, the SSVS prior, $p\left(\beta_{j,t}|\gamma_{j,t}\right)$, can be written as $p\left(z_{j,t}|\beta_{j,t}, v_{j,t}\right) \equiv N\left(\beta_{j,t}, v_{j,t}\right)$ for the pseudo-observations $z_{j,t} = 0$, $\forall j, t$, where we define $v_{j,t} = \left(1-\gamma_{j,t}\right)^2 \underline{v}_{j,0}^2 + \gamma_{j,t}\underline{v}_{j,1}^2$. The resulting posterior is of the form

$$p\left(\beta_{1:T}, Q_{1:T}, V_{1:T}|y_{1:T}, z_{1:T}\right) \propto \prod_{t=1}^{T} p\left(\beta_t|\beta_{t-1}, Q_t\right) p\left(y_t|\beta_t, \underline{\sigma}^2\right) p\left(z_t|\beta_t, V_t\right) p\left(\gamma_t\right) p\left(Q_t\right), \tag{B.7}$$

where we define $V_t = (v_{1,t}, ..., v_{p,t})$.

The objective of variational Bayes inference is to approximate the intractable joint posterior $p\left(\beta_{1:T}, Q_{1:T}, V_{1:T}|y_{1:T}, z_{1:T}\right)$ with a tractable distribution $q\left(\beta_{1:T}, Q_{1:T}, V_{1:T}\right)$.

Applying the mean field approximation we obtain the factorization

$$q\left(\beta_{1:T}, Q_{1:T}, V_{1:T}\right) = q\left(\beta_{1:T}\right) \prod_{t=1}^{T} \prod_{j=1}^{p} q\left(v_{j,t}\right) q\left(q_{j,t}\right). \tag{B.8}$$

The optimal form for $q\left(\beta_{1:T}\right)$ is a Normal linear state space model with measurement and state equations

$$\begin{aligned} q\left(y_t|\beta_t\right) &\propto N\left(x_t\beta_t, \underline{\sigma}^2\right) & \text{(B.9)} \\ q\left(\beta_t|\beta_{t-1}\right) &\propto N\left(\widetilde{F}_t\beta_{t-1}, \widetilde{Q}_t^{-1}\right), & \text{(B.10)} \end{aligned}$$

where $\widetilde{Q}_t = \left(Q_t^{-1} + V_t^{-1}\right)^{-1}$ and $\widetilde{F}_t = \widetilde{Q}_t Q_t^{-1}$.[11] Thus, conditional on values of the other parameters in the model, $q\left(\beta_{1:T}\right)$ can be evaluated using the transformed state space model above and standard Kalman filter recursions.

The form for $q\left(v_{j,t}\right)$ can be obtained using standard SSVS prior derivations

$$\begin{aligned} v_{j,t} &= \begin{cases} \underline{v}_0^2, & if \quad \gamma_{j,t} = 0, \\ \underline{v}_1^2, & if \quad \gamma_{j,t} = 1, \end{cases} & \text{(B.15)} \\ q\left(\gamma_{j,t}\right) &\propto Bernoulli\left(\pi_{j,t}\right), & \text{(B.16)} \end{aligned}$$

where $\pi_{j,t} = \frac{N\left(\beta_{j,t}|0,\underline{v}_{j,1}^2\right)\underline{\pi}_0}{N\left(\beta_{j,t}|0,\underline{v}_{j,1}^2\right)\underline{\pi}_0 + N\left(\beta_{j,t}|0,\underline{v}_{j,0}^2\right)\left(1-\underline{\pi}_0\right)}$ and $N\left(x;a,b\right)$ denotes a Normal p.d.f. evaluated at the point x. Thus, conditional on other model parameters, the form for

[11] This uses the form of the state equation given in (16) which can be derived as follows:

$$\begin{aligned} q\left(\beta_t|\beta_{t-1}\right) &\propto \exp\left\{E\left(\log p\left(\beta_t|\beta_{t-1}, Q_t\right)\right) + E\left(\log p\left(z_t|\beta_t, V_t\right)\right)\right\} & \text{(B.11)} \\ &\propto \exp\left\{-\frac{1}{2}\left(\beta_t - \beta_{t-1}\right)' Q_t^{-1}\left(\beta_t - \beta_{t-1}\right) - \frac{1}{2}\beta_t' V_t^{-1}\beta_t\right\} & \text{(B.12)} \\ &\propto \exp\left\{-\frac{1}{2}\beta_t' Q_t^{-1}\beta_t + \beta_t' Q_t^{-1}\beta_{t-1} - \frac{1}{2}\beta_t' V_t^{-1}\beta_t\right\} & \text{(B.13)} \\ &\propto \exp\left\{-\frac{1}{2}\left(\beta_t - \widetilde{F}_t\beta_{t-1}\right)' \widetilde{Q}_t^{-1}\left(\beta_t - \widetilde{F}_t\beta_{t-1}\right)\right\}. & \text{(B.14)} \end{aligned}$$

$q\left(v_{j,t}\right)$ allows for easy updating.

Finally, conditional on the other parameters in the model, the optimal form for $q\left(q_{j,t}\right)$ is of the form

$$q\left(q_{j,t}\right) \propto Gamma\left(c_{j,t}, d_{j,t}\right) \tag{B.17}$$

where $c_{j,t} = \underline{c}_0 + 1/2$ and $d_{j,t} = \underline{d}_0 + D_{j,j}/2$ and $D_{j,j}$ the j-th diagonal element of $D = P_{t|t} + \beta_{t|t}\beta'_{t|t} + \left(P_{t-1|t-1} + \beta_{t-1|t-1}\beta'_{t-1|t-1}\right)\left(I_p - 2\widetilde{F}_t\right)'$, where $\beta_{t|t}$ and $P_{t|t}$ are time t filtered estimates of the posterior mean and variance of β_t.

The VB algorithm using these formulas is presented in Algorithm 2.

Algorithm 2 *Variational Bayes Kalman Filter (VBKF) with variable selection prior and known (fixed) variance*

```
Initialize $\underline{\beta}_0, \underline{P}_0, \underline{a}_0, \underline{b}_0, \underline{c}_0, \underline{d}_0$
for $t = 1$ to $T$ do
    $r = 1$
    while $\|\beta_{t|t}^{(r)} - \beta_{t|t}^{(r-1)}\| \to 0$ do
        UPDATE STATE-SPACE MATRICES:
        $\widetilde{Q}_t^{(r)} = \left[\left(Q_t^{(r-1)}\right)^{-1} + \left(V_t^{(r-1)}\right)^{-1}\right]^{-1}$
        $\widetilde{F}_t^{(r)} = \widetilde{Q}_t^{(r)}\left(Q_t^{(r-1)}\right)^{-1}$

        POSTERIOR OF $\beta_t$:
        $\beta_{t|t-1}^{(r)} = \widetilde{F}_t^{(r)}\beta_{t-1}$                                              Predicted mean
        $P_{t|t-1}^{(r)} = \widetilde{F}_t^{(r)} P_{t-1} \widetilde{F}_t^{(r)'} + \widetilde{Q}_t^{(r)}$                          Predicted variance
        $K_t^{(r)} = P_{t|t-1}^{(r)} x_t' \left(x_t P_{t|t-1}^{(r)} x_t' + \underline{\sigma}^2\right)^{-1}$                     Kalman gain
        $\beta_{t|t}^{(r)} = \beta_{t|t-1}^{(r)} + K_t^{(r)}\left(y_t - x_t\beta_{t|t-1}^{(r)}\right)$                Posterior mean of $\beta_t$
        $P_{t|t}^{(r)} = \left(I_p - K_t^{(r)} x_t\right) P_{t|t-1}^{(r)}$                       Posterior variance of $\beta_t$

        POSTERIORS OF $q_t$ AND $\tau_t$:
        $D^{(r)} = P_{t|t}^{(r)} + \beta_{t|t}^{(r)}\beta_{t|t}^{(r)'} + \left(P_{t-1}^{(r)} + \beta_{t-1}^{(r)}\beta_{t-1}^{(r)'}\right)\left(I_p - 2\widetilde{F}_t^{(r)}\right)'$
        for $j = 1$ to $p$ do
            $\pi_{j,t}^{(r)} = \frac{N\left(\beta_{j,t|t}^{(r)}|0,\underline{v}_{j,1}^2\right)\underline{\pi}_0}{N\left(\beta_{j,t|t}^{(r)}|0,\underline{v}_{j,1}^2\right)\underline{\pi}_0+N\left(\beta_{j,t|t}^{(r)}|0,\underline{v}_{j,0}^2\right)(1-\underline{\pi}_0)}$
            $v_{j,t}^{(r)} = \left(1 - \pi_{j,t}^{(r)}\right)^2 \underline{v}_{j,0}^2 + \pi_{j,t}^{(r)} \underline{v}_{j,1}^2$             Posterior mean of $v_{j,t}$
            $c_{j,t}^{(r)} = \underline{c}_0 + 1/2,$
            $d_{j,t}^{(r)} = \underline{d}_0 + D_{jj}^{(r)}/2$
            $q_{j,t}^{(r)} = d_{j,t}^{(r-1)}/c_{j,t}^{(r-1)}$                                Posterior mean of $q_{j,t}$
        end for
        Set $V_t^{(r)} = diag\left(v_t^{(r)}\right)$ and $Q_t^{(r)} = diag\left(q_t^{(r)}\right)$
        $r = r + 1$
    end while
    Set $\beta_t = \beta_{t|t}^{(r)}$, $P_t = P_{t|t}^{(r)}$, $Q_t = diag\left(q_t^{(r)}\right)$ and $V_t = diag\left(\tau_t^{(r)}\right)$
end for
```

B.2 Incorporating stochastic volatility

We now extend the preceding algorithm to incorporate stochastic volatility and the TVP regression model accordingly becomes:

$$\begin{aligned} y_{t+h} &= x_t\beta_t + \sigma_t\epsilon_{t+h}, &\text{(B.18)}\\ \beta_t &= \beta_{t-1} + \eta_t, &\text{(B.19)}\\ \log\sigma_t^2 &= \log\sigma_t^2 + \zeta_t, &\text{(B.20)} \end{aligned}$$

where $\epsilon_t \sim N(0,1)$, $\zeta_t \sim N(0, r_t)$. We use a prior for r_t of the form

$$r_t^{-1} \sim Gamma\left(\underline{f}_0, \underline{g}_0\right), \tag{B.21}$$

and an initial condition $\log\sigma_0^2 \sim N(\log\underline{\sigma}_0^2, \underline{R}_0)$.

The mean field approximation used in our VB algorithm has the following form:

$$q\left(\beta_{1:T}, Q_{1:T}, V_{1:T}, \log\sigma_{1:T}^2, r_{1:T}\right) = q\left(\beta_{1:T}\right) q\left(\log\sigma_{1:T}^2\right) q\left(r_{1:T}\right) \prod_{t=1}^{T}\prod_{j=1}^{p} q\left(v_{j,t}\right) q\left(q_{j,t}\right). \tag{B.22}$$

The preceding sub-section describes the forms for $q(\beta_{1:T})$, $q(q_{j,t})$ and $q(v_{j,t})$. The presence of stochastic volatility leads to a nonlinear state space model which complicates things. Tran, Nott and Kohn (2017) derive a VB algorithm for models such as this which involve intractable likelihoods. However, their methods are more demanding than the Kalman filter methods used in our paper since they require stochastic optimization and the evaluation of the stochastic volatility likelihood using a particle filter. We use a much simpler (albeit approximate) approach based on the transformed stochastic volatility model given in (20) and (21). This state-space model has measurement variance $\log(\epsilon_t^2)$ which is a log-χ^2 density with one degree of freedom. We approximate this density using a $N(-1.2704, 4.937)$ distribution whose moments match the mean and variance of the log-χ^2 distribution.

The accuracy of this approximation is displayed in Figure 6. It can be seen to be good for relatively high values of log-volatilities (approximately -10 to 3, which correspond to values of the variance parameter between $4.5e-5$ and 20). The approximation is poor in the far left tail of the distribution. This region corresponds to very small values of σ_t^2. As noted in Section 3, we can help avoid this region of the parameter space by standardizing our variables to have sample variance of one. Our Monte Carlo experiment in Section 4 suggests that the approximation is a good one for our purposes.

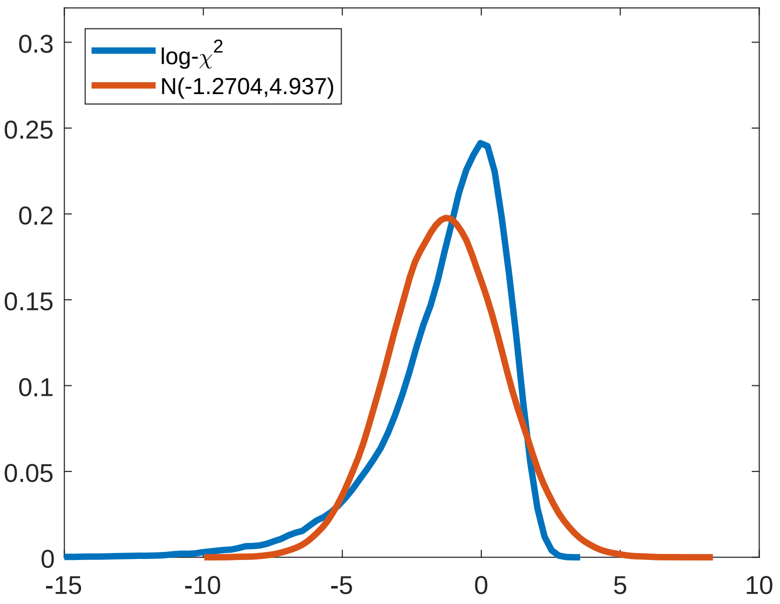

Figure 6: *This figure demonstrates the informational loss by approximating the error variance of a linearized stochastic volatility model (which is log-χ^2 with one degree of freedom), by a $N(-1.2704, 4.937)$ distribution.*

Given this approximation, the state space model for the log-volatilities is linear and Normal and the methods of the preceding subsection can be used. Finally, the form for $q(r_{1:T})$ is standard, involving textbook manipulations involving the Gamma distribution.

Algorithm 3 summarizes the steps needed to implement mean field variational Bayes inference in the stochastic volatility model.

Algorithm 3 *Variational Bayes Kalman Filter (VBKF) for updating log-volatilities*

```
Initialize $\log \underline{\sigma}_0^2, \underline{R}_0, \underline{f}_0, \underline{g}_0$
for $t = 1$ to $T$ do
    $r = 1$
    while $\left\| \log\left(\sigma_{t|t}^2\right)^{(r)} - \log\left(\sigma_{t|t}\right)^{(r-1)} \right\| \to 0$ do
        1. Obtain $\beta_{t|t}^{(r)}$ as in Algorithm 2 above (but whenever $\underline{\sigma}^2$ shows up in this Algorithm, replace
    it with $\exp\left(\log\left(\sigma_t^2\right)^{(r)}\right)$ defined below)
        2. Construct $\widetilde{y}_t = \log\left(\left(y_t - x_t\beta_{t|t}^{(r)}\right)^2 + 10^{-10}\right) - 1.2704$

        POSTERIOR OF $\log \sigma_t^2$:
        $\log\left(\sigma_{t|t-1}^2\right)^{(r)} = \log \sigma_{t-1}^2$                                   Predicted mean
        $R_{t|t-1}^{(r)} = R_{t-1} + r_t^{(r)}$                                                  Predicted variance
        $K_t^{(r)} = R_{t|t-1}^{(r)}\left(R_{t|t-1}^{(r)} + 4.937\right)^{-1}$                                  Kalman gain
        $\log\left(\sigma_{t|t}^2\right)^{(r)} = \log\left(\sigma_{t|t-1}^2\right)^{(r)} + K_t^{(r)}\left(\widetilde{y}_t - \log\left(\sigma_{t|t-1}^2\right)^{(r)}\right)$     Post. mean of $\log \sigma_t^2$
        $R_{t|t}^{(r)} = \left(1 - K_t^{(r)}\right) R_{t|t-1}^{(r)}$                               Posterior variance of $\log \sigma_t^2$

        POSTERIOR OF $r_t$:
        $C^{(r)} = R_{t|t} + \log\left(\sigma_{t|t}^2\right)^{(r)} \times \log\left(\sigma_{t|t}^2\right)^{(r)} - \left(R_{t-1} + \log \sigma_{t-1}^2 \times \log \sigma_{t-1}^2\right)$
        $f_t^{(r)} = \underline{f}_0 + 1/2$
        $g_t^{(r)} = \underline{g}_0 + C^{(r)}/2$
        $r_t^{(r)} = g_t^{(r-1)}/f_t^{(r-1)}$                                         Posterior mean of $r_{j,t}$

        $r = r + 1$
    end while
    Set $\log \sigma_t^2 = \log\left(\sigma_{t|t}^2\right)^{(r)}$ and $R_t = R_{t|t}^{(r)}$.
end for
```

Algorithm 1, given in Section 3, combines Algorithms 2 and 3.

C Competing Forecasting Models

C.1 A Constant Coefficient Model

C.1.1 Stochastic search variable selection

This approach uses a SSVS shrinkage prior in a homoskedastic constant coefficient regression and the notation is as in Section 3 except that t sub-scripts have been removed from all regression coefficients and the variance parameter. The full hierarchical representation of the SSVS prior is

$$\begin{aligned} p\left(\beta_i|\gamma_i\right) &\sim (1-\gamma_i)N\left(0,\tau_0^2\right)+\gamma_i N\left(0,\tau_1^2\right), && \text{(C.1a)} \\ p\left(\gamma_i|\pi\right) &\sim Bernoulli(\pi_i), && \text{(C.1b)} \end{aligned}$$

where we set $\pi = 0.5$, $\tau_0 = 0.001$ and $\tau_1 = 4$, and the regression variance parameter has a diffuse prior. Posterior computation can be done using MCMC methods as described in George and McCulloch (1993).

C.2 Competing specifications: Time-varying parameter algorithms

The h-step ahead direct forecasting regression with time-varying coefficients and stochastic volatility is of the form

$$\begin{aligned} y_{t+h} &= x_t\beta_t+\varepsilon_{t+h}, && \text{(C.2)} \\ \beta_t &= \beta_{t-1}+\eta_t, && \text{(C.3)} \end{aligned}$$

where β_t is a $p \times 1$ vector of time-varying parameters, $\varepsilon_{t+h} \sim N\left(0, \sigma_t^2\right)$ with σ_t^2 a time-varying measurement variance, and $\eta_t \sim N\left(0, Q\right)$ with Q a $p \times p$ state covariance matrix.

C.2.1 KP-AR, Koop and Potter (2007)

The specification of Koop and Potter (2007) is a structural break model. It can be written as a state space model and be viewed as a special case of the time-varying parameter regression. The KP-AR model is of the form

$$y_{t+h} = x_t \beta_{s_t} + \varepsilon_{t+h}, \tag{C.4}$$

$$\beta_{s_t} = \beta_{s_{t-1}} + \eta_{s_t}, \tag{C.5}$$

where x_t includes only an intercept and lags, $s_t \in \{1, 2, ..., K\}$ is a Markov switching process with K states. We follow much of the Bayesian structural breaks literature and assume that the transition probabilities matrix is block diagonal, such that we can move from one regime to the next and never come back (which is the distinguishing feature of structural breaks compared to standard regime-switching specifications). We follow Bauwens et al (2015) and specify a maximum number of $K_{max} = 10$ and allow the Gibbs sampler to determine how many structural breaks are relevant (up to the maximum of K_{max}). Priors and initial conditions are the same as those used in Bauwens et al. (2015), and the reader is referred to that paper and its online Appendix (Section B) for details of posterior computation.

C.2.2 GK-AR, Giordani and Kohn (2008)

The Giordani and Kohn (2008) model is also a structural breaks model which can be written in state space form. It is a dynamic mixture model of the form

$$\begin{aligned} y_{t+h} &= x_t\beta_t + \varepsilon_{t+h}, & \text{(C.6)} \\ \beta_t &= \beta_{t-1} + K_t\eta_t, & \text{(C.7)} \end{aligned}$$

where x_t includes only an intercept and lags, $K_t \in \{0,1\}$. Details of prior hyperparameter choice and the MCMC algorithm used for posterior computation are exactly as described in Section 2.5 of Bauwens et al. (2015).

C.2.3 UCSV, Stock and Watson (2007)

The Stock and Watson (2007) unobserved components stochastic volatility (UCSV) model only allows for a time-varying intercept:

$$\begin{aligned} y_{t+h} &= \tau_t + \varepsilon_{t+h}, & \text{(C.8)} \\ \tau_t &= \tau_{t-1} + \eta_t, & \text{(C.9)} \end{aligned}$$

where not only the measurement error ε_{t+h} features stochastic volatility, but also the variance of state error η_t. This model has been specifically proposed for forecasting inflation, but it is a parsimonious and flexible nonlinear specification that may be able to fit other series as well. Posterior computation is done using standard MCMC methods and prior hyperparameters are identical to the ones described in Section 2.6 of Bauwens et al. (2015).

C.2.4 TVP-AR, Pettenuzzo and Timmermann (2017)

This is a TVP regression model of (C.2) and (C.3) involving only an intercept and lags of the dependent variable. Stochastic volatility is added to the measurement equation, but (unlike UC-SV) the state equation is homoskedastic. Pettenuzzo and Timmermann (2017) is a recent, representative study that uses this model and finds that it beats a large number of alternative models when forecasting inflation. All priors we use for estimation of this model also follow the default values described in Section 2.5 of Bauwens et al. (2015), and the reader is referred to that paper for more details. Posterior computation can be done using MCMC methods for state space models.

C.2.5 TVP-BMA, Groen, Paap and Ravazzolo (2013)

TVP-BMA is a simplified version of a model developed in Groen, Paap and Ravazzolo (2013).[12] It generalizes a variable selection method for the constant coefficient regression developed by Kuo and Mallick (1998) to the TVP case as follows:

$$\begin{aligned} y_{t+h} &= \sum_{j=1}^{p} x_{jt} s_j \beta_{j,t} + \varepsilon_{t+h}, && \text{(C.10)} \\ \beta_t &= \beta_{t-1} + \eta_t, && \text{(C.11)} \end{aligned}$$

where s_j is an indicator variable such that when $s_j = 0$ the j^{th} predictor is removed from the regression in all periods, while when $s_j = 1$ the predictor is included. Details of posterior computation are given in Groen, Paap and Ravazzolo (2013). Prior hyperparameter choices are identical to the TVP-AR model, with the addition of a prior for s_j. In particular, we assume that s_j has a Bernoulli prior with prior probability of inclusion of each variable equal to 0.5.

[12] In particular, their model also features a dynamic mixture as in Giordani and Kohn (2008), but since we also estimate the GK specification separately, we don't add the dynamic mixture part in the Groen, Paap and Ravazzolo (2013) specification.

C.2.6 TVP-LASSO, Belmonte, Koop and Korobilis (2014)

Belmonte, Koop and Korobilis (2014), following Frühwirth-Schnatter and Wagner (2010) use the following non-centered parameterization of the time-varying parameter regression model

$$\begin{aligned} y_{t+h} &= x_t\alpha + x_t\Omega\widetilde{\alpha}_t + \varepsilon_{t+h}, \qquad \text{(C.12)} \\ \widetilde{\alpha}_t &= \widetilde{\alpha}_{t-1} + \widetilde{\eta}_t, \qquad \text{(C.13)} \end{aligned}$$

where Ω is a diagonal matrix of parameters, and now the state equation has disturbance $\widetilde{\eta}_t \sim N\left(0, I_p\right)$ and initial condition $\widetilde{\alpha}_0 = 0$. Written like this, the model consists of a standard constant parameter part (with coefficients α) plus the additional time variation introduced by $\alpha = \Omega \times \widetilde{\alpha}_t$. It can be seen that, compared to the TVP regression specification used in original specification in eqs it holds that $\beta = \alpha + \alpha_t = \alpha + \Omega \times \widetilde{\alpha}_t$, and that $Q = \Omega^2$ where Q in this case is diagonal.

By doing this transformation, Belmonte, Koop and Korobilis (2014) choose to use the Bayesian lasso prior on the parameters α and $\omega = diag(\Omega)$. Given the ability of the lasso to shrink coefficients towards zero, and the fact that α and ω are a-priori independent, the specification above allows predictor j to: i) enter the regression with no restrictions, ii) enter the regression with constant coefficients only, iii) enter the regression with time-varying coefficients only, and iv) not enter the regression at all.

The MCMC algorithm for estimating this model is described in Belmonte, Koop and Korobilis (2014). We use the full model described by these authors (i.e. not any its restricted versions) and we use the default prior hyperparameters described in the empirical section of this paper.

C.2.7 TVD, Chan et al. (2012)

The time-varying dimension (TVD) model of Chan et al. (2012) takes the following form

$$\begin{aligned} y_{t+h} &= \sum_{j=1}^{p} x_{j,t} s_{j,t} \beta_{j,t} + \varepsilon_{t+h}, & \text{(C.14)} \\ \beta_t &= \beta_{t-1} + \eta_t, & \text{(C.15)} \end{aligned}$$

where $s_{j,t}$ is an indicator variable which follows a Markov process such that when $s_{j,t} = 0$ the j^{th} predictor is removed from the regression model in period t only, and when $s_{j,t} = 1$ it is included in the regression. This is a very flexible specification that generalizes the TVP-BMA specification to allow for a predictor to exit the regression only for certain periods. This specification is the first of three alternative time-varying dimension specifications presented in Chan et al. (2012). All other settings follow these authors – see the online Appendix associated with Chan et al. (2012). This pertains to default prior choices (see end of Section 1.1 of that Appendix), as well as other choices the authors make. For example, for computational reasons the authors only consider models with no predictors, one predictor, or all predictors, rather than consider all possible 2^p models with different number of predictors.

C.2.8 TVS, Kalli and Griffin (2014)

The time varying sparsity (TVS) model of Kalli and Griffin is of the form

$$\begin{aligned} y_{t+h} &= \sum_{j=1}^{p} x_{j,t} \beta_{j,t} + \varepsilon_{t+h}, & \text{(C.16)} \\ \beta_{j,t} &= (1 - \alpha_j) \rho_{j,t} \beta_{j,t-1} + \alpha_j \eta_{j,t}, & \text{(C.17)} \end{aligned}$$

where $\rho_{j,t} = \sqrt{\frac{\psi_{j,t}}{\psi_{j,t-1}}}$ and $var\left(\eta_{j,t}\right) = \psi_{j,t}$. In this specification, $\alpha_j \in [0,1]$ is a parameter controlling the temporal correlation, and $\psi_{j,t}$ is an autoregressive gamma process. Thus, the implied prior for $\beta_{j,t}$ is of normal-gamma autoregressive process form, which generalizes the traditional Normal-Gamma priors in linear regression, see Griffin and Brown (2010). Such priors have very good shrinkage properties and the coefficient of each predictor can be shrunk flexibly only in some periods, while be unrestricted in others. Note that these authors specify a Gamma autoregressive process for the error variance, instead of the stochastic volatility process that all previous methods use. We follow Kalli and Griffin (2014) and as these authors do in their Section 5 for forecasting inflation we choose $s^\star = 0.1$ and $b^\star = 0.1$. All other choices and initial conditions are exactly those used also by the authors.

C.2.9 DMA, Koop and Korobilis (2012)

Koop and Korobilis (2012) follow DMA methods introduced in Raftery et al. (2010). DMA involves a model space consisting of many time-varying parameter regressions:

$$y_{t+h} = x_t^{(k)} \beta_t^{(k)} + \varepsilon_{t+h}, \tag{C.18}$$

$$\beta_t^{(k)} = \beta_{t-1}^{(k)} + \eta_t, \tag{C.19}$$

where (k) indexes the model that applies. DMA involves $K = 2^p$ models each of which uses a sub-set of the p potential explanatory variables. It involves estimating and forecasting with each of these models and then averaging over the results in a dynamic fashion. Therefore, in the equations above, $k = 1, ..., K$ indexes each of the various TVP regressions that have different number of predictors. Note that since each of the K models has different predictors and different coefficients $\beta_t^{(k)}$, the associated variances will also be different, that is, $var\left(\varepsilon_{t+h}\right) = \left(\sigma_t^2\right)^{(k)}$ and $var\left(\eta_t\right) = Q^{(k)}$.

In order to be able to enumerate and estimate all possible model combinations, one has to be able to estimate each model very quickly. This motivates the use of the FFKF and EWMA. FFKF is an approximate method that has been popular in engineering; see for example Kulhavý and Kraus (1996) and references therein. Once all models are estimated one can obtain measures of fit for each model at each point in time.[13] DMA generalizes static Bayesian model averaging by allowing different predictors to enter/exit the TVP regression model at each point in time.

Estimation of the model of Koop and Korobilis (2012) relies on crucial selection of forgetting/decay factors $(\alpha, \lambda, \kappa)$. These determine how quickly models, regression coefficients, and volatilities, respectively, evolve over time. We set these to the following default values $\alpha = 0.96$, $\lambda = 0.98$, $\kappa = 0.94$. We also initialize the $\beta_t^{(k)}$ for all models to $\beta_0^{(k)} \sim N\left(0^{(k)}, 4I^{(k)}\right)$, where the vector of zeros $0^{(k)}$ and the identity matrix $I^{(k)}$ comply with the number of elements in $\beta_t^{(k)}$. Finally, the initial value of the volatility parameter is $\sigma_0 = 0.1$ in for all K models.

[13] The Kalman filter allows for the evaluation of the data likelihood as well as the predictive likelihood at each point in time, so one can use various measures to construct model probabilities. We, following most of the DMA literature, use discounted predictive likelihoods to do model averaging at each point in time, but it is worth noting that other metrics such as information criteria (e.g. BIC) or measures of point forecast performance (e.g. MSFE) could be used.