Fast and Flexible Bayesian Inference in Time-varying Parameter Regression Models*

NIKO HAUZENBERGER[1†], FLORIAN HUBER[1†], GARY KOOP[2]
and LUCA ONORANTE[3]

[1] *University of Salzburg*
[2] *University of Strathclyde*
[3] *European Central Bank*

November 27, 2019

Abstract

In this paper, we write the time-varying parameter regression model involving K explanatory variables and T observations as a constant coefficient regression model with TK explanatory variables. In contrast with much of the existing literature which assumes coefficients to evolve according to a random walk, this specification does not restrict the form that the time-variation in coefficients can take. We develop computationally efficient Bayesian econometric methods based on the singular value decomposition of the TK regressors. In artificial data, we find our methods to be accurate and much faster than standard approaches in terms of computation time. In an empirical exercise involving inflation forecasting using a large number of predictors, we find our methods to forecast better than alternative approaches and document different patterns of parameter change than are found with approaches which assume random walk evolution of parameters.

Keywords:	Time-varying parameter regression, singular value decomposition, clustering, hierarchical priors
JEL Codes:	C11, C30, E3, D31

*Email: niko.hauzenberger@sbg.ac.at, florian.huber@sbg.ac.at, gary.koop@strath.ac.uk and luca.onorante@ecb.int. We would like to thank the participants of the 6^{th} NBP Workshop on Forecasting (Warsaw, 2019) for helpful comments and suggestions.

†The first two authors gratefully acknowledge financial support by the Austrian Science Fund (FWF): ZK 35 and by funds of the Oesterreichische Nationalbank (Austrian Central Bank, Anniversary Fund, project number: 18127).

1 Introduction

Time-varying parameter (TVP) regressions and Vector Autoregressions (VARs) have shown their usefulness in a range of applications in macroeconomics (e.g. D'Agostino et al., 2013; Cogley and Sargent, 2005; Primiceri, 2005). Particularly when the number of explanatory variables is large, Bayesian methods are typically used since prior information can be essential in overcoming over-parameterization concerns. These priors are often hierarchical and ensure parsimony by automatically shrinking coefficients. Examples include Belmonte et al. (2014), Kalli and Griffin (2014), Huber et al. (2019) and Bitto and Frühwirth-Schnatter (2019). Approaches such as these have two characteristics that we highlight so as to motivate the contributions of our paper. First, they use Markov Chain Monte Carlo (MCMC) methods which can be computationally demanding. They are unable to scale up to the truly large data sets that macroeconomists now work with. Second, the regression coefficients in these TVP models are assumed to follow random walk or autoregressive (AR) processes. In this paper, we develop a new approach which is computationally efficient and scaleable. Furthermore, it allows for more flexible patterns of time variation in the regression coefficients.

We achieve the computational gains by writing the TVP regression as a static regression with a particular, high dimensional, set of regressors. Using the singular value decomposition (SVD) of this set of regressors along with conditionally conjugate priors yields a computationally fast algorithm which scales well in high dimensions. One key feature of this approach is that no approximations are involved. This contrasts with other computationally-fast approaches to TVP regression which achieve computational gains by using approximate methods such as variational Bayes (Koop and Korobilis, 2018), message passing (Korobilis, 2019) or expectation maximization (Rockova and McAlinn, 2018).

Our computational approach avoids large-scale matrix operations altogether and exploits the fact that most of the matrices involved are block diagonal and thus (band) sparse. In large dimensional contexts, this allows fast MCMC-based inference and thus enables the researcher to compute highly non-linear functions of the time-varying regression coefficients while taking parameter uncertainty into account. Compared to estimation approaches based on forward-filtering backward-sampling (FFBS, see Carter and Kohn, 1994; Frühwirth-Schnatter, 1994) algorithms, the computational burden is light. In particular, we show that it rises quadratically in the number of covariates and linearly in the number of observations. For quarterly macroeconomic datasets that feature a few hundred observations, this allows us to estimate and forecast, exploiting all available information without using dimension reduction techniques such as principal components.

Computational tractability is one concern in high dimensional TVP regressions. The curse of dimensionality associated with estimating large dimensional TVP regressions is another. To solve over-parameterization issues and achieve a high degree of flexibility in the type of coefficient change, we use a sparse finite mixture representation (see Malsiner-Walli et al., 2016) for the time-varying coefficients. This introduces shrinkage on the amount of time variation by pooling different time periods into a (potentially) small number of clusters. We also use shrinkage priors which allow for the detection of how many clusters are necessary. Shrinkage towards the cluster means is then introduced by specifying appropriate conjugate priors on the regression coefficients. We propose two different choices for this prior. The first of these is based on Zellner's g-prior (Zellner, 1986). The second is based on the Minnesota prior (Doan et al., 1984; Litterman, 1986).

We investigate the performance of our methods using two applications. Based on synthetic data, we first illustrate computational gains if K and T become large. We then proceed to show that our approach effectively recovers key properties of the data generating process. In a real-data application, we model US inflation dynamics. Our approach provides new insights on how the relationship between unemployment and inflation evolves over time. Moreover, in an extensive forecasting exercise we show that our proposed set of models performs well relative to a wide range of competing models. Specifically, we find that our model yields precise point and density forecasts for one-step-ahead and four-step-ahead predictions. Improvements in forecast accuracy are especially pronounced during recessionary episodes.

The remainder of the paper is structured as follows. Section 2 introduces the static representation of the TVP regression model while Section 3 shows how the SVD can be used to speed up computation. Sections 4 and 5 provide an extensive discussion of our prior setup and briefly sketch the MCMC algorithm used. The model is then applied to synthetic data in Section 6 and real data in Section 7. Finally, the last section summarizes and concludes the paper and an appendix provides additional details on computation and further empirical findings.

2 A Static Representation of the TVP Regression Model

Let $\{y_t\}_{t=1}^T$ denote a scalar response variable[1] that is described by a TVP regression given by

$$y_t = \boldsymbol{x}_t' \boldsymbol{\beta}_t + \sigma \eta_t, \quad \eta_t \sim \mathcal{N}(0, 1), \tag{1}$$

[1] This setup can be easily extended to VAR models. In particular, recent papers (see, e.g., Carriero et al., 2016; Kastner and Huber, 2017; Koop et al., 2019; Huber et al., 2019) work with a structural VAR specification which allows for the equations to be estimated separately. Accordingly, the size of the system does not penalize the estimation time. This extension is part of our current research agenda.

where $\boldsymbol{x}_t$ is a K-dimensional vector of regressors, $\boldsymbol{\beta}_t$ is a set of K time-varying regression coefficients and σ^2 is the error variance. For now, we assume homoskedastic errors, but will relax this assumption later.

The TVP regression can be written as a static regression model as follows:

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{pmatrix} \boldsymbol{x}_1' & \mathbf{0}_{K\times 1}' & \dots & \mathbf{0}_{K\times 1}' \\ \mathbf{0}_{K\times 1}' & \boldsymbol{x}_2' & \dots & \mathbf{0}_{K\times 1}' \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{K\times 1}' & \mathbf{0}_{K\times 1}' & \dots & \boldsymbol{x}_T' \end{pmatrix}}_{\boldsymbol{Z}} \underbrace{\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_T \end{pmatrix}}_{\boldsymbol{\beta}} + \sigma \underbrace{\begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_T \end{pmatrix}}_{\eta}. \tag{2}$$

Equation 2 implies that the dynamic regression model in Equation 1 can be cast in the form of a standard linear regression model with KT predictors stored in a $T \times KT$-dimensional design matrix $\boldsymbol{Z}$. Notice that the rank of $\boldsymbol{Z}$ is equal to T and inverting $\boldsymbol{Z}'\boldsymbol{Z}$ is not possible. We stress that, at this stage, we are agnostic on the evolution of $\boldsymbol{\beta}_t$ over time.

The researcher may want to investigate whether any explanatory variable has a time-varying coefficient or a constant coefficient or a zero coefficient. In such a case, it proves convenient to work with a different parameterization of the model which decomposes $\boldsymbol{\beta}$ into a time-invariant ($\boldsymbol{\gamma}$) and a time-varying part ($\tilde{\boldsymbol{\beta}}$):

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{Z}\tilde{\boldsymbol{\beta}} + \sigma\boldsymbol{\eta}, \tag{3}$$

with $\boldsymbol{X} = (\boldsymbol{x}_1', \dots, \boldsymbol{x}_T')'$ denoting a $T \times K$ matrix of stacked covariates and $\boldsymbol{\beta} = \boldsymbol{\gamma} + \tilde{\boldsymbol{\beta}}$.

The t^{th} equation is given by

$$y_t = \boldsymbol{x}_t'\boldsymbol{\gamma} + \boldsymbol{x}_t'\tilde{\boldsymbol{\beta}}_t + \sigma\eta_t,$$

with $\tilde{\boldsymbol{\beta}}_t$ being the relevant elements of $\tilde{\boldsymbol{\beta}}$. Thus, we have written the TVP regression as a static regression, but with a huge number of explanatory variables. That is, $\tilde{\boldsymbol{\beta}}$ is a KT-dimensional vector with K, T both being potentially large numbers.

This representation is related to a non-centered parameterization (Frühwirth-Schnatter and Wagner, 2010) of a state space model. The main intuition behind Equation 3 is that parameters tend to fluctuate around a time-invariant regression component $\boldsymbol{\gamma}$, with deviations being driven by $\tilde{\boldsymbol{\beta}}_t$. This parameterization, in combination with the static representation of the state space model, allows us to push the model towards a time-invariant specification during certain points in time, if necessary. This behavior closely resembles the behavior of mixture innovation models, e.g. Giordani and Kohn (2008), and allows the model to decide the points in time when it is necessary to allow for parameter change.

In the theoretical discussion which follows, we will focus on the time-varying part of the regression model

$$\hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\gamma} = \boldsymbol{Z}\tilde{\boldsymbol{\beta}} + \sigma\boldsymbol{\eta},$$

since sampling from the conditional posterior of $\boldsymbol{\gamma}$ (under a Gaussian shrinkage prior and conditional on $\boldsymbol{Z}\tilde{\boldsymbol{\beta}}$) is straightforward. In principle, any shrinkage prior can be introduced on $\boldsymbol{\gamma}$. In our empirical work, we use a hierarchical Normal-Gamma prior of the form:

$$\gamma_j|\tau_j \sim \mathcal{N}(0, \tau_j), \quad \tau_j|\psi \sim \mathcal{G}(\vartheta, \vartheta\psi/2), \quad \psi \sim \mathcal{G}(a_0, a_1),$$

where γ_j is the j^{th} element of $\boldsymbol{\gamma}$ for $j = 1, \dots, K$. We set $\vartheta = 0.1$ and $a_0 = a_1 = 0.01$. We use MCMC methods to learn about the posterior for these parameters. The relevant posterior conditionals are given in Griffin and Brown (2010) and Appendix A.

3 Fast Bayesian Inference using SVDs

3.1 The Homoskedastic Case

In static regressions with huge numbers of explanatory variables, there are several methods for ensuring parsimony that involve compressing the data. Traditionally principal components or factor methods have been used, see Stock and Watson (2011). Random compression methods have also been used with TVP models, see Koop et al. (2019). In this paper, we use the SVD to do the data compression and show that this has some attractive properties.

The SVD of our matrix of explanatory variables, $\boldsymbol{Z}$, is

$$\underbrace{\boldsymbol{Z}}_{T\times KT} = \underbrace{\boldsymbol{U}}_{T\times T}\underbrace{\boldsymbol{\Lambda}}_{T\times T}\underbrace{\boldsymbol{V}'}_{T\times KT}$$

whereby $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices and $\boldsymbol{\Lambda}$ denotes a diagonal matrix with the singular values, denoted by $\boldsymbol{\lambda}$, of $\boldsymbol{Z}$ as diagonal elements.

The usefulness and theoretical soundness of the SVD to compress regressions is demonstrated in Trippe et al. (2019). They use it as an approximate method in the sense that, in a case with K regressors, they only use the part of the SVD corresponding to the largest M singular values, where $M < K$. In such a case, their methods become approximate.[2]

In our case, we can exploit the fact that $\text{rank}(\boldsymbol{Z}) = T$ $(T \ll KT)$ and utilize the SVD of $\boldsymbol{Z}$ as in Trippe et al. (2019). But we do not truncate the SVD using only the M largest singular values, but use all T of them.[3] But since the rank of $\boldsymbol{Z}$ is $T(\ll KT)$, our approach translates into an exactly low rank structure. In this particular setting, with $\text{rank}(\boldsymbol{\Lambda}) = \text{rank}(\boldsymbol{Z})$, applying the SVD on the full matrix $\boldsymbol{Z}$ implies that on the t^{th} position of $\boldsymbol{\lambda}$ we have the eigenvalue of $\boldsymbol{x}_t'$.

Thus, using the SVD we can exactly recover the big matrix $\boldsymbol{Z}$. The reason for using the SVD instead of $\boldsymbol{Z}$ is that we can exploit several convenient properties of the SVD that speed up computation. To be specific, if we use a Gaussian prior, this leads to a computationally particularly convenient expression of the posterior distribution of $\tilde{\boldsymbol{\beta}}$ which avoids complicated matrix manipulations such as inversion and the Cholesky decomposition of high-dimensional matrices. Thus, computation is fast.

We assume a conjugate prior of the form

$$\tilde{\boldsymbol{\beta}}|\sigma^2 \sim \mathcal{N}(\boldsymbol{b}_0, \sigma^2 \boldsymbol{D}_0),$$

with $\boldsymbol{D}_0 = \boldsymbol{I}_T \otimes \boldsymbol{\Psi}$ being a KT-dimensional diagonal prior variance-covariance matrix, where $\boldsymbol{I}_T$ denotes a T-dimensional identity matrix and $\boldsymbol{\Psi}$ a $K \times K$-dimensional diagonal matrix that contains covariate-specific shrinkage parameters on its main diagonal. Our prior will be hierarchical so that $\boldsymbol{\Psi}$ will depend on other prior hyperparameters $\boldsymbol{\theta}$ to be defined later.

Using textbook results for the Gaussian linear regression model with a conjugate prior (conditional on the time-invariant coefficients $\boldsymbol{\gamma}$), the posterior is

$$\tilde{\boldsymbol{\beta}}|Data, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_{\tilde{\beta}}, \sigma^2 \boldsymbol{V}_{\tilde{\beta}}). \tag{4}$$

[2] They derive errors bounds for the approximation and other theoretical properties. One of these they call conservativeness by which they mean the approximate posterior never exhibits less uncertainty than the exact posterior. This contrasts with other approximate approaches such as variational Bayes which can underestimate posterior variances.

[3] In cases with very large T we could follow Trippe et al. (2019) and choose a value $M < T$ to speed up computation. But for macroeconomic data sets T is small enough that we have not found it necessary to resort to such an approximation.

In conventional regression contexts, the computational bottleneck is typically the $KT \times KT$ matrix $\boldsymbol{V}_{\tilde{\beta}}$. However, with our SVD regression, Trippe et al. (2019), show this to take the form:[4]

$$\begin{aligned}
\boldsymbol{V}_{\tilde{\beta}} &= \left(\boldsymbol{D}_0^{-1} + \boldsymbol{V} \text{ diag } (\boldsymbol{\lambda} \odot \boldsymbol{\lambda}) \boldsymbol{V}'\right)^{-1} \\
&= \boldsymbol{D}_0 - \boldsymbol{D}_0 \boldsymbol{V} \ \left(\text{diag } (\boldsymbol{\lambda} \odot \boldsymbol{\lambda})^{-1} + \boldsymbol{V}' \boldsymbol{D}_0 \boldsymbol{V}\right)^{-1} \boldsymbol{V}' \boldsymbol{D}_0 \\
\boldsymbol{\mu}_{\tilde{\beta}} &= \boldsymbol{V}_{\tilde{\beta}} (\boldsymbol{Z}' \hat{\boldsymbol{y}} + \boldsymbol{D}_0^{-1} \boldsymbol{b}_0). \qquad (5)
\end{aligned}$$

Crucially, the matrix $\boldsymbol{\Xi} = \left(\text{diag } (\boldsymbol{\lambda} \odot \boldsymbol{\lambda})^{-1} + \boldsymbol{V}' \boldsymbol{D}_0 \boldsymbol{V}\right)^{-1}$ is a diagonal matrix and thus trivial to compute. The main computational hurdle boils down to computing $\boldsymbol{V} \boldsymbol{\Xi} \boldsymbol{V}'$, but it is a block-diagonal matrix. The resulting computation time, conditional on a fixed T, rises linearly in K because most of the matrices involved are (block) diagonal. The key feature of our algorithm is that we entirely avoid inverting a full matrix. The only inversion involved is the inversion of $\boldsymbol{\Xi}$ which can be carried out in $\mathrm{O}(T)$ steps. This implies that efficient algorithms can be employed to further speed up computation and thus sampling from $\mathcal{N}(\boldsymbol{\mu}_{\tilde{\beta}}, \sigma^2 \boldsymbol{V}_{\tilde{\beta}})$ is computationally simple.[5]

In this sub-section, we have described computationally efficient methods for doing Bayesian estimation in the homoskedastic Gaussian linear regression model when the number of explanatory variables is large. They can be used in any Big Data regression model, but here we are using them in the context of our TVP regression model written in static form as in (3). These methods involve transforming the matrix of explanatory variables using the SVD. If the matrices of prior hyperparameters, $\boldsymbol{b}_0$ and $\boldsymbol{D}_0$, were known and if homoskedasticity were a reasonable assumption, then textbook, conjugate prior, results for Bayesian inference in the Gaussian linear regression model are all that are required. Analytical results are available for this case and there would be no need for MCMC methods. This is the case covered by Trippe et al. (2019). However, in macroeconomic data sets homoskedasticity is often not a reasonable assumption. And it is unlikely that the researcher would be able to make sensible choices for $\boldsymbol{b}_0$ and $\boldsymbol{D}_0$ in this high-dimensional context. Accordingly, the next two sub-sections of this paper will develop methods for adding stochastic volatility and propose a hierarchical prior for the regression coefficients.

[4] More precisely, Trippe et al. (2019) derive the posterior moments under the prior $\sigma^2 \boldsymbol{I}$. In this case, the resulting quantities take an even simpler form.

[5] One could imagine an apparently similar strategy using all of the principal components of $\boldsymbol{X}$. But in our TVP context the posterior covariance matrix for $\tilde{\boldsymbol{\beta}}$ would not simplify as it does here and it would be difficult if not impossible to invert it. On the other hand, if the TVP aspect were added through a state equation (e.g. specifying random walk behavior) then the problems caused by the large number of state equations would be enormous.

3.2 Adding Stochastic Volatility

Stochastic volatility typically is an important feature of successful macroeconomic forecasting models (e.g. Clark, 2011). We incorporate this by replacing σ^2 in (3) with $\boldsymbol{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_T^2) \otimes \boldsymbol{I}_K$. This implies that the prior on $\tilde{\boldsymbol{\beta}}$ is

$$\tilde{\boldsymbol{\beta}} | \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{b}_0, \boldsymbol{\Sigma}\boldsymbol{D}_0).$$

Note that the prior in a specific time period is given by

$$\tilde{\boldsymbol{\beta}}_t | \sigma_t^2 \sim \mathcal{N}(\boldsymbol{b}_{0t}, \sigma_t^2 \boldsymbol{\Psi}),$$

with $\boldsymbol{b}_{0t}$ being the relevant block associated with the t^{th} period. Thus, it can be seen that the degree of shrinkage changes with σ_t^2, implying less shrinkage in more volatile times.

We assume that $h_t = log(\sigma_t^2)$ follows an AR(1) process:

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

In our empirical work, we follow Kastner and Frühwirth-Schnatter (2014) and specify a Gaussian prior on the unconditional mean $\mu_h \sim \mathcal{N}(0, 10)$, a Beta prior on the (transformed) persistence parameter $\frac{\rho_h+1}{2} \sim \mathcal{B}(25, 5)$ and a non-conjugate Gamma prior on the process innovation variance $\sigma_h^2 \sim \mathcal{G}(1/2, 1/2)$. Bayesian estimation of the volatilities proceeds using MCMC methods and, in particular, the algorithm of Kastner and Frühwirth-Schnatter (2014).

4 A Hierarchical Prior for the Regression Coefficients

4.1 General Considerations

With hierarchical priors, where $\boldsymbol{b}_0$ and/or $\boldsymbol{D}_0$ depend on unknown parameters, MCMC methods based on the full conditional posterior distributions are typically used. In our case, we would need to recompute the enormous matrix $\boldsymbol{V}_{\tilde{\beta}}$ and its Cholesky factor at every MCMC draw. This contrasts with the non-hierarchical case with fixed $\boldsymbol{b}_0$ and $\boldsymbol{D}_0$ where $\boldsymbol{V}_{\tilde{\beta}}$ is calculated once. Due to this consideration, we wish to avoid using MCMC methods based on the full posterior conditionals.

Many priors, including the two introduced here, have $\boldsymbol{D}_0$ depending on a small number of prior hyperparameters. These can be simulated using a Metropolis Hastings (MH) algorithm. With such an algorithm, updating of $\boldsymbol{V}_{\tilde{\beta}}$ (and computing its Cholesky decomposition) only takes place for accepted draws (in our forecasting exercise roughly 30% of draws are accepted). We find that this leads to vastly reduced computation time and adopt this strategy in our empirical work.

4.2 The Prior Covariance Matrix for the Regression Coefficients

In this paper, we consider two different hierarchical priors for $\tilde{\boldsymbol{\beta}}$. Since our empirical application centers on forecasting inflation, the design matrix $\boldsymbol{x}_t$ will be structured as follows $\boldsymbol{x}_t = (y_{t-1}, \dots, y_{t-p}, \boldsymbol{d}'_{t-1}, \dots, \boldsymbol{d}'_{t-p}, 1)'$, with $\boldsymbol{d}_t$ denoting a set of N exogenous regressors and p is the maximum number of lags to include.

The first prior is inspired by the Minnesota prior (see Litterman, 1986). It captures the idea that own lags are typically more important than other lags and, thus, require separate shrinkage. It also captures the idea that more distant lags are likely to be less important than more recent ones. Our variant of the Minnesota prior translates these ideas to control the amount of time-variation, implying that coefficients on own lags might feature more time-variation while parameters associated with other lags feature less time-variation. The same notion carries over to coefficients related to more distant lags which should feature less time-variation a priori.

This prior involves two hyperparameters to be estimated: $\boldsymbol{\theta} = (\zeta_1, \zeta_2)'$. These prior hyperparameters are used to parameterize $\boldsymbol{\Psi}$ to match the Minnesota prior variances:

$$[\boldsymbol{\Psi}]_{ii} = \begin{cases} \frac{\zeta_1^2}{l^2} \quad \text{on the coefficients associated with } y_{t-l} \ (l = 1, \dots, p) \\ \frac{\zeta_2^2}{l^2} \frac{\hat{\sigma}_y^2}{\hat{\sigma}_j^2} \text{ on the coefficients related to } d_{jt-l} \\ \zeta_2^2 \text{ on the intercept term.} \end{cases}$$

Here, we let $[\boldsymbol{\Psi}]_{ii}$ denote the $(i,i)^{th}$ element of $\boldsymbol{\Psi}$, d_{jt} refers to the j^{th} element of $\boldsymbol{d}_t$, $\hat{\sigma}_y^2$, $\hat{\sigma}_j^2$ denotes the OLS variance obtained by estimating an AR(p) model in y_t and d_{jt}, respectively. The hyperpriors on ζ_1 and ζ_2 follow a Uniform distribution:

$$\zeta_j \sim \mathcal{U}(\mathfrak{s}_{0,j}, \mathfrak{s}_{1,j}) \quad \text{for} \quad j = 1, 2.$$

The second prior we use is a variant of the g-prior involving a single prior hyperparameter: $\theta = \xi$. This specification amounts to setting $\boldsymbol{\Psi} = \xi \times \boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ is a diagonal matrix with the $(i,i)^{th}$ element being defined as $[\boldsymbol{\Omega}]_{ii} = \hat{\sigma}_y^2/\hat{\sigma}_j^2$. For reasons outlined in Doan et al. (1984), we depart from using the diagonal elements of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$

to scale our prior and rely on the OLS variances of an AR(p) model as in the case of the Minnesota-type prior. Similar to the Minnesota prior we use again a Uniform prior:

$$\xi \sim \mathscr{U}(\mathfrak{s}_0, \mathfrak{s}_1).$$

Since we aim to infer ξ, ζ_1 and ζ_2 from the data we set $\mathfrak{s}_0 = \mathfrak{s}_{0,1} = \mathfrak{s}_{0,2} = 10^{-10}$ close to zero and $\{\mathfrak{s}_1, \mathfrak{s}_{1,1}, \mathfrak{s}_{1,2}\}$ is specified as follows:

$$\mathfrak{s}_1 = \mathfrak{s}_{1,j} = \kappa \frac{T}{K^2} \quad \text{for} \quad j = 1, 2. \tag{6}$$

Here, κ is a constant being less or equal than unity to avoid excessive overfitting in light of large K and T. Since large values of κ translate into excessive time variation in $\tilde{\boldsymbol{\beta}}_t$, we need to select κ carefully. In the empirical application, we infer κ over a grid of values and select the κ that yields the best forecasting performance in terms of log predictive scores. Further discussion of and empirical evidence relating to κ is given in Appendix C.

The methods developed in this paper will hold for any choice of prior covariance matrix, $\boldsymbol{D}_0$, although assuming it to be diagonal greatly speeds up computation. In this sub-section, we have proposed two forms for it which we shall (with some abuse of terminology) refer to as the Minnesota and g-prior forms, respectively, in the following material.

4.3 The Prior Mean of the Regression Coefficients

As for the prior mean, $\boldsymbol{b}_0$, it can take a range of possible forms. The simplest thing is to set it to zero. After all, from (3) it can be seen that $\tilde{\boldsymbol{\beta}}_t$ measures the deviation from the constant coefficient case which, on average, is zero. This is what we do with the Minnesota prior.[6] However, it is possible that we can gain estimation accuracy through pooling information across coefficients by adding extra layers to the prior hierarchy. In this paper, we do so using a sparse finite location mixture of Gaussians and adapt the methods of Malsiner-Walli et al. (2016) to the TVP regression context. We refer to these two treatments of the prior mean as non-clustered and clustered, respectively, below. With the g-prior, we consider both clustered and non-clustered approaches.

We emphasize that both of these specifications for the prior mean are very flexible and let the data decide the form that the change in parameters takes. This contrasts with TVP regression models, where it is common to assume that the states evolve according to random walks. This implies that the prior mean of $\boldsymbol{\beta}_t$ is $\boldsymbol{\beta}_{t-1}$.

[6] Using the Minnesota prior in combination with the clustering specification introduced in this sub-section is less sensible. That is, its form, involving different treatments of coefficients on lagged dependent variables and exogenous variables and smaller prior variances for longer lag length already, in a sense, clusters the coefficients into groups.

With the clustered approach, we assume that each $\tilde{\boldsymbol{\beta}}_t$ has a prior of the following form:

$$f_{\mathcal{N}}(\tilde{\boldsymbol{\beta}}_t|\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_G,\boldsymbol{w},\boldsymbol{V}_{\tilde{\beta},t}) = \sum_{g=1}^{G} w_g f_{\mathcal{N}}(\tilde{\boldsymbol{\beta}}_t|\boldsymbol{\mu}_g,\sigma_t^2\boldsymbol{\Psi}),$$

where $f_{\mathcal{N}}$ denotes the density of a Gaussian distribution and $\boldsymbol{w}$ are component weights with $\sum_{g=1}^{G} w_g = 1$ and $w_g \geq 0$ for all g. $\boldsymbol{\mu}_g$ $(g = 1,\dots,G)$ denotes G component-specific means with G being a potentially large integer that is much smaller than T (i.e. $G \ll T$).

An equivalent representation, based on auxiliary variables δ_t, is

$$\tilde{\boldsymbol{\beta}}_t|\delta_t = g \sim \mathcal{N}(\boldsymbol{\mu}_g,\sigma_t^2\boldsymbol{\Psi}),$$

with $Prob(\delta_t = g) = w_g$ being the probability that $\tilde{\boldsymbol{\beta}}_t$ is assigned to group g. Before proceeding to the exact prior setup, it is worth noting that the mixture model is not identified with respect to relabeling the latent indicators. In the forecasting application, we consider functions of the states which are not affected by label switching. Thus, we apply the random permutation sampler of Frühwirth-Schnatter (2001) to randomly relabel the states in order to make sure that our algorithm visits the different modes of the posterior. In what follows, we define $\boldsymbol{m}_t = \boldsymbol{\mu}_g$ if $\delta_t = g$. Using this notation, the prior mean is given by $\boldsymbol{b}_0 = (\boldsymbol{m}_1',\dots,\boldsymbol{m}_T')'$.

For the weights $\boldsymbol{w} = (w_1,\dots,w_G)'$, we use a Dirichlet prior:

$$\boldsymbol{w}|\pi \sim \mathrm{Dir}(\pi,\dots,\pi).$$

Here, π denotes the intensity parameter that determines how the model behaves in treating superfluous components. If $\pi \leq K/2$, irrelevant components are emptied out while if $\pi > K/2$, the model tends to duplicate component densities to handle overfitting issues. This implies that careful selection of π is crucial since it influences the number of breaks in $\tilde{\boldsymbol{\beta}}_t$. The literature suggests different strategies based on using traditional model selection criteria or reversible jump MCMC algorithms to infer π from the data. Our approach closely follows Malsiner-Walli et al. (2016) and uses a shrinkage prior on π. The prior we adopt follows a Gamma distribution with

$$\pi \sim \mathcal{G}(a, aG),$$

with $a = 10$ being a hyperparameter that determines the tightness of the prior. This prior choice is based on simulation evidence discussed in Malsiner-Walli et al. (2016). Notice that this prior choice implies that $\mathbb{E}(\pi) = 1/G$ and $\mathbb{V}(\pi) = 1/(aG^2)$.

To assess which elements in $\boldsymbol{\mu}_g$ determine the group membership, we use yet another shrinkage prior on the component means:

$$\boldsymbol{\mu}_g|\mathbf{\Pi}, \tilde{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{\Pi}),$$

whereby $\mathbf{\Pi} = \boldsymbol{\Upsilon} \boldsymbol{R} \boldsymbol{\Upsilon}$ with $\boldsymbol{\Upsilon} = \text{diag}(\sqrt{\upsilon_1}, \dots, \sqrt{\upsilon_K})$ and $\boldsymbol{R} = \text{diag}(R_1^2, \dots, R_K^2)$. We let R_j denote the range of $\tilde{\boldsymbol{\beta}}_j = (\tilde{\beta}_{j1}, \dots, \tilde{\beta}_{jT})$. The prior on υ_j $(j = 1, \dots, K)$ follows a Gamma distribution,

$$\upsilon_j \sim \mathcal{G}(c_0, c_1),$$

translating into the Normal-Gamma prior of Griffin and Brown (2010). In the empirical application we set $c_0 = c_1 = 0.6$, with $c_0 < 1$ being crucial for pushing the idiosyncratic group means $\boldsymbol{\mu}_g$ strongly towards the common mean $\boldsymbol{\mu}_0$ (Malsiner-Walli et al., 2016). For $\boldsymbol{\mu}_0$, we use an improper Gaussian prior with mean set equal zero and infinite variance.

This location mixture model is extremely flexible in the types of parameter change that are possible. It allows us to capture situations where the breaks in parameters are large or small and frequent or infrequent. It can effectively mimic the behavior of break point models, standard time-varying parameter models, mixture innovation models and many more. In addition, notice that the presence of stochastic volatility implies that the mixture components share a common variance factor that implicitly affects the tightness of the prior.

5 Posterior Computation

We carry out posterior inference using a relatively straightforward MCMC algorithm. Most steps involved are standard and the exact forms for each conditional posterior are given in Appendix A. In this section, we briefly summarize the main steps of the algorithm. After specifying appropriate starting values, we repeat the following steps $30,000$ times and discard the first $10,000$ draws as burn-in.

1. Draw $\boldsymbol{\gamma}$ from a multivariate Gaussian distribution (see Equation A.1).
2. Draw the local shrinkage parameters τ_j $(j = 1, \dots, K)$ from a generalized inverse Gaussian (GIG) distribution (see Equation A.2).

3. Draw the global shrinkage parameter ψ from a Gamma distribution (see Equation A.3).

4. Draw $\tilde{\boldsymbol{\beta}}$ from a TK-dimensional Gaussian distribution (see Equation 4).

5. Draw the volatilities $\sigma_1, \dots, \sigma_T$ as well as the parameters of the state equation of h_t using the algorithm of Kastner and Frühwirth-Schnatter (2014).[7]

6. Draw $\boldsymbol{\theta}$ using a random walk Metropolis-Hastings (RWMH) step.

7. Draw the weights $\boldsymbol{w}$ from a Dirichlet distribution (see Equation A.5).

8. Draw δ_t for each $\tilde{\boldsymbol{\beta}}_t$ from a Multinomial distribution (see Equation A.6).

9. Draw $\boldsymbol{\mu}_0$ from a multivariate Gaussian distribution (see Equation A.8).

10. Draw υ_j $(j = 1, \dots, K)$ from a GIG distribution (see Equation A.9).

For the non-clustered approaches, the final four steps are not required.

6 Illustration Using Artificial Data

In this section we illustrate our modeling approach that utilizes the g-prior and clustering by means of synthetic data simulated from a simple data generating process (DGP).

We begin by illustrating the computational advantages arising from using the SVD, relative to a standard Bayesian approach to TVP regression which involves random walk evolution of the coefficients and the use of FFBS. Figure 1(a) shows a comparison of the time necessary to generate a draw from $p(\tilde{\boldsymbol{\beta}}|Data, \boldsymbol{\gamma}, \sigma^2)$ using our algorithm based on the SVD and the FFBS algorithm as a function of $K \in \{1, 2, \dots, 400\}$ and for $T = 50$.[8]

To illustrate how computation times change with T, Figure 1(b) shows computation times as a function of $T \in \{50, \dots, 250\}$ for $K = 200$. The solid black line denotes the actual time (based on a MacBook Pro late 2016 with a 2.9 GHz Intel Core i5) to simulate from the full conditional of the latent states while the red dotted line denotes a non-linear trend.

In panel (a), we fit a quadratic (cubic) trend on the empirical estimation times of the SVD (FFBS) approach. This implies that while the computational burden is cubic in the number of covariates K for the FFBS approach,

[7] This is implemented in the R package `stochvol`.

[8] An alternative approach to estimating TVP regressions is based on simulating the latent states *all without a loop* (AWOL, see Chan and Jeliazkov, 2009; McCausland et al., 2011; Kastner and Frühwirth-Schnatter, 2014) with an efficient implementation provided in the R package `shrinkTVP` (Bitto-Nemling et al., 2019). These techniques, however, are comparable to using FFBS-based approaches (if implemented efficiently in a low-level computing environment) and our SVD algorithm strongly improves upon them for large K, T.

our technique based on using the SVD suggests that runtimes increase quadratically in K. Notice that the figure clearly suggests that traditional algorithms based on FFBS quickly become infeasible in high dimensions. Up to $K \approx 50$, our algorithm is slightly slower while the computational advantage increases remarkably with K, being more than twice as fast for $K = 100$ and almost six times as fast for $K = 400$.

Panel (b) of the figure shows that, for fixed K, computation times increase linearly in T for both approaches used. It is noteworthy, however, that the slope of the line referring to FFBS is much steeper. This reflects the fact that one needs to perform a filtering (that scales linearly in T) and smoothing step (that is also linear in T). This brief discussion shows that the SVD algorithm scales well and renders estimation of huge dimensional models feasible.

We now assume that y_t is generated by the following DGP:

$$y_t = \tilde{\beta}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, 0.1^2)$$

for $t = 1, \dots, 160$, $\gamma = 0$, and $\tilde{\beta}_t \sim \mathcal{N}(m_t, 0.1^2)$. The $\tilde{\beta}_t'$s evolve according to the following law of motion:

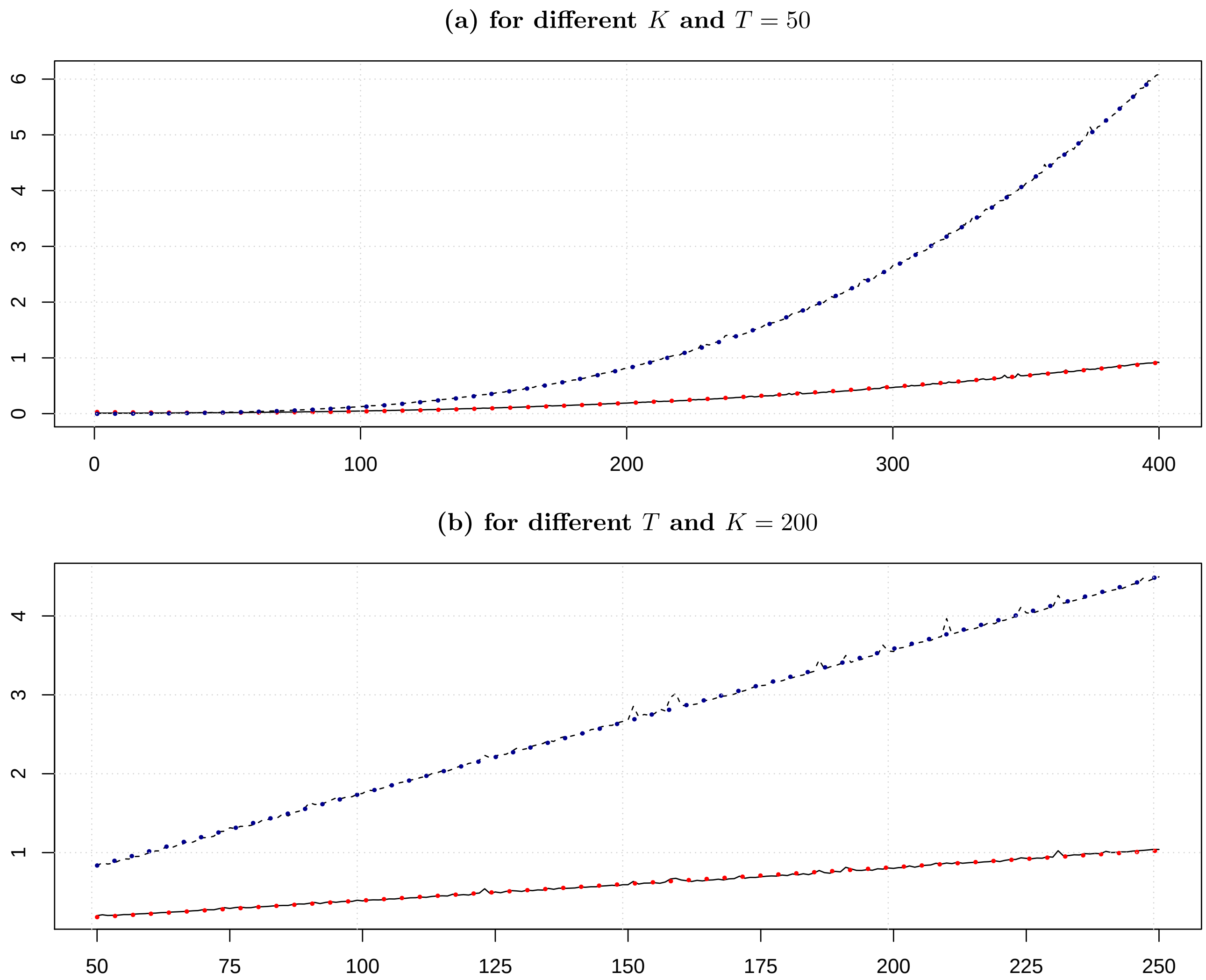
$$m_t = \begin{cases} 3 & \text{if} \quad t \leq 60 \\ 1 & \text{if} \quad 60 < t \leq 85 \\ -3 & \text{if} \quad 86 < t \leq 120 \\ -1 & \text{if} \quad t > 120. \end{cases}$$

Analyzing this stylized DGP allows us to illustrate how our approach can be used to infer the number of latent clusters that determine the dynamics of $\tilde{\beta}_t$. In what follows, we simulate a single path of y_t and use this for estimating our model. We estimate the model using the g-prior with clustering and set $G = 12$. In this application, we show quantities that depend on the labeling of the latent indicators. This calls for appropriate identifying restrictions and we introduce the restriction that $\mu_1 < \cdots < \mu_G$. This is not necessary if interest centers purely on predictive distributions and, thus, we do not impose this restriction in the forecasting section of this paper.

Before discussing how well our model recovers the true state vector $\tilde{\beta}_t$, we show how our modeling approach can be used to infer the number of groups G. Following Malsiner-Walli et al. (2016), the number of groups is estimated during MCMC sampling as follows:

$$G_0^{(j)} = G - \sum_{g=1}^{G} I\left(T_g^{(j)} = 0\right)$$

with $T_g^{(j)}$ denoting the number of observations in cluster g for the j^{th} MCMC draw and $I(\bullet)$ being the indicator function. This yields a posterior distribution for G_0. Its posterior mode can be used as a point estimate of G. In Table 1, we report the posterior probability of a given number of regimes by simply computing the fraction of



(a) for different K and $T = 50$

(b) for different T and $K = 200$

Notes: The figure shows the actual and theoretical time necessary to obtain a draw of $\tilde{\boldsymbol{\beta}}$ using our proposed SVD algorithm and the FFBS algorithm. The solid black line refers to the SVD approach while the dashed black line refers to the FFBS algorithm. The dots refer to theoretical run times.

Figure 1: Runtime comparison: SVD and FFBS

draws that $G_0 = g$ for $g = 1, \ldots, 12$. The table suggests that the probability that $G_0 = 4$ is around 66 percent. This indicates that our algorithm successfully selects the correct number of groups, since the mode of the posterior distribution equals four. It is also worth noting that the posterior mean of π is very small at 0.09, suggesting that our mixture model handles irrelevant components by emptying them instead of replicating them (which would be the case if π becomes large). Notice, however, that $G_0 = 5$ also receives some posterior support. We have a probability of about 26 percent associated with a too large number of regimes. In the present model, this slight overfitting behavior might be caused by additional noise driven by the shocks to the states $\tilde{\beta}_t$, with our mixture model trying to fit the noise.

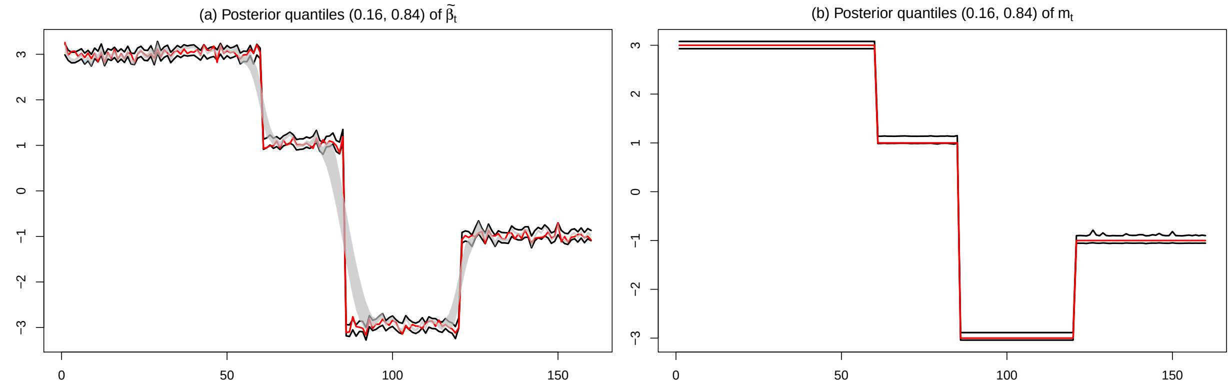
Table 1: Posterior probabilities for a given number of groups G

$G_0 =$	1	2	3	4	5	6	7	8	9	10	11	12
	0.00	0.00	0.00	0.66	0.26	0.07	0.01	0.00	0.00	0.00	0.00	0.00

Next, we assess whether our model is able to recover $\tilde{\beta}_t$ and m_t. Figure 2 shows the 16th and 84th percentiles of the posterior distribution (in solid black) of $\tilde{\beta}_t$ (see panel (a)) and m_t (see panel (b)) over time. The gray shaded areas represent the 16th and 84th percentiles of the posterior of $\tilde{\beta}_t$ obtained from estimating a standard TVP regression model with random walk state equations and stochastic volatility. Apart from the assumption of random walk evolution of the states, all other specification choices are made so as to be as close as possible to our SVD approach. In particular, this model features the hierarchical Normal-Gamma prior (see Griffin and Brown, 2010) on both the time-invariant part of the model and the signed square root of the state innovation variances (Bitto and Frühwirth-Schnatter, 2019). It is estimated using a standard FFBS algorithm. We refer to this model as TVP-RW-FFBS.

In Figure 2 the red lines denotes the true value of $\tilde{\beta}_t$ and m_t, respectively. Panel (a) clearly shows that our model successfully detects major breaks in the underlying states, with the true value of $\tilde{\beta}_t$ almost always being located within the credible intervals. Our modeling approach not only captures low frequency movements but also successfully replicates higher frequency changes. By contrast, the posterior distribution of the TVP-RW-FFBS specification is not capable of capturing abrupt breaks in the latent states. Instead of capturing large and infrequent changes, the TVP-RW-FFBS approach yields a smooth evolution of $\tilde{\beta}_t$ over time, suggesting that our proposed approach performs comparatively better in learning about sudden breaks in the regression coefficients.

Considering panel (b) of Figure 2 yields a similar picture. Our approach yields credible sets that include the actual outcome of m_t for all t. This discussion shows that our model also handles cases with infrequent breaks in



Notes: Panel (a) shows 16th/84th posterior percentiles of $\tilde{\beta}_t$ for our proposed model (solid black lines) and a standard TVP regression with random walk state equation (gray shaded area). The red line denotes the actual outcome. Panel (b) shows the 16th/84th percentiles of the posterior distribution of m_t (in solid black) and the true value of m_t (in solid red).

Figure 2: Posterior distribution of $\tilde{\beta}_t$ and m_t

the regression coefficients rather well. As compared to standard TVP regressions that imply a smooth evolution of the states, using a mixture model to determine the state evolution enables us to capture large and abrupt breaks.

7 An Application to US Inflation

In this section, we investigate the performance of our methods in an inflation forecasting exercise. In the empirical work we use the popular large dimensional quarterly US data set described in McCracken and Ng (2016). Sub-section 7.1 provides details on model specification, the dataset and shows some selected in-sample features. Sub-section 7.2 presents the forecasting results.

7.1 Selected in-sample features

Modeling and forecasting inflation is of great value for economic agents and policymakers. In most central banks, inflation is the main policy objective. Across several major central banks, the workhorse forecasting model is based on some version of the Phillips curve. In general, the relationship of inflation with the real economy constitutes a fundamental building block of modern macroeconomics. The practical forecasting of inflation is, however, difficult (see Stock and Watson, 2007), and the persistence of low inflation in the presence of a closing output gap in recent years has led to a renewed debate about the usefulness of the curve as a policy instrument in the United States (see, e.g., Ball and Mazumder, 2011; Coibion and Gorodnichenko, 2015).

There are three main issues when forecasting inflation. A first problem is that the theoretical literature relating to the Phillips curve and the determination of inflation includes a large battery of very different specifications, emphasizing domestic vs. international variables, forward vs. backward looking expectations or including factors such as labour market developments. The overall number of potential predictors can be quite large (see Stock and Watson, 2008). Second, within each econometric specification there is considerable uncertainty about which indicator should be used to proxy for the economic cycle (see Moretti et al., 2019). Third, there are structural breaks that make different variables and specifications more or less important at different times. The Great Recession, for example, is universally considered as a structural break that requires appropriate econometric techniques. Inflation persistence has also changed over time (see Watson, 2014). Even the link between inflation and the economic cycle seems to have weakened after 2008, making inflation difficult to forecast ex ante and to explain ex post. In general, the relevant predictors for inflation have been found to change over time (see Koop and Korobilis, 2012). Expectations are difficult to measure empirically, and their importance varies over time (e.g. Moretti et al., 2019).

The mainstream literature has dealt with the curse of dimensionality which arises in TVP regressions with many predictors in several ways. Until recently, the two main approaches included principal components or strong Bayesian shrinkage. A comparison of the two approaches can be found in De Mol et al. (2008). Following Raftery et al. (2010), a second stream of research uses model combination to deal with the curse of dimensionality and the fact that models can change over time (e.g. Koop and Korobilis, 2012). Finally, a recent (but expanding) stream of literature forecasts inflation using machine learning techniques (Medeiros et al., 2019). These methods, although useful, suffer from the "black box problem"; while their accuracy compares well with other techniques, they are not able to show how the result is obtained and to offer a simple interpretation. A survey of these techniques is given in Hassani and Silva (2015).

For the reasons above, inflation forecasting is an ideal empirical application in which we can investigate the performance of our methods. An important criterion is the capacity of our approach to generalize over standard TVP models, which are less flexible because they are based on random walk or autoregressive specifications to determine the evolution of the states. A second challenge is the correct detection of well-known structural breaks. In addition, we assess the forecasting performance of our methods relative to alternative approaches.

Following Stock and Watson (1999), we define the target variable as follows

$$y_{t+h} = \ln\left(\frac{P_{t+h}}{P_t}\right) - \ln\left(\frac{P_t}{P_{t-1}}\right),$$

with P_{t+h} denoting the price level (`CPIAUCSL`) in period $t+h$. Using this definition, we estimate a generalized Phillips curve involving 50 covariates that cover different segments of the economy. Further information on the specific variables included and the way they are transformed is provided in Appendix B. The design matrix $\boldsymbol{X}_t$ includes $p = 2$ lags and an intercept and thus features $K = 101$ covariates.

Before we use our model to perform forecasting, we provide some information on computation times, illustrate some in-sample features of our model and briefly discuss selected posterior estimates of key parameters.

Table 2 shows empirical runtimes (in minutes) for estimating the different models for this large dataset. As highlighted in the beginning of Section 6, our approach starts improving upon FFBS-based algorithms in terms of computation time if K exceeds 50, with the improvements increasing non-linearly in K. Hence, it is unsurprising that, for our present application with $K = 101$, our algorithm (without clustering) is more than three times faster than using FFBS. If clustering is added, our approach is still almost twice as fast. The additional computational complexity from using the clustering prior strongly depends on G. If G is close to T (which typically does not occur in practice), then the computation time increases and the advantage of using the SVD is diminished. This arises since estimating the location parameters of the mixtures becomes the bottleneck in our MCMC algorithm. It is worth noting that one key advantage of our approach is that it is relatively easy to implement. In R, sparse algorithms are readily available and manipulating large sparse matrices is relatively cheap. By contrast, efficient implementation of FFBS-based algorithms (which we use throughout the paper) often calls for using low-level computing environments such as C++ or Java.

	SVD (g-prior w. clustering)	SVD (g-prior no clustering)	FFBS	TIV
min.	192	125	377	16

Table 2: Runtime comparison of empirical exercise ($K = 101$; $T = 216$) with $30,000$ draws from the posterior distribution

To further illustrate the properties of the estimated parameters in our SVD approach using the g-prior with clustering we now turn to a small-scale model. In this case, the number of coefficients is relatively small and features such as multipliers with an economic interpretation can be easily plotted. This model is inspired by the New Keynesian Phillips curve (NKPC). The dependent variable is inflation and the right hand side variables include two lags of unemployment and inflation. We set $G = 30$, thus allowing for a relatively large number of clusters.

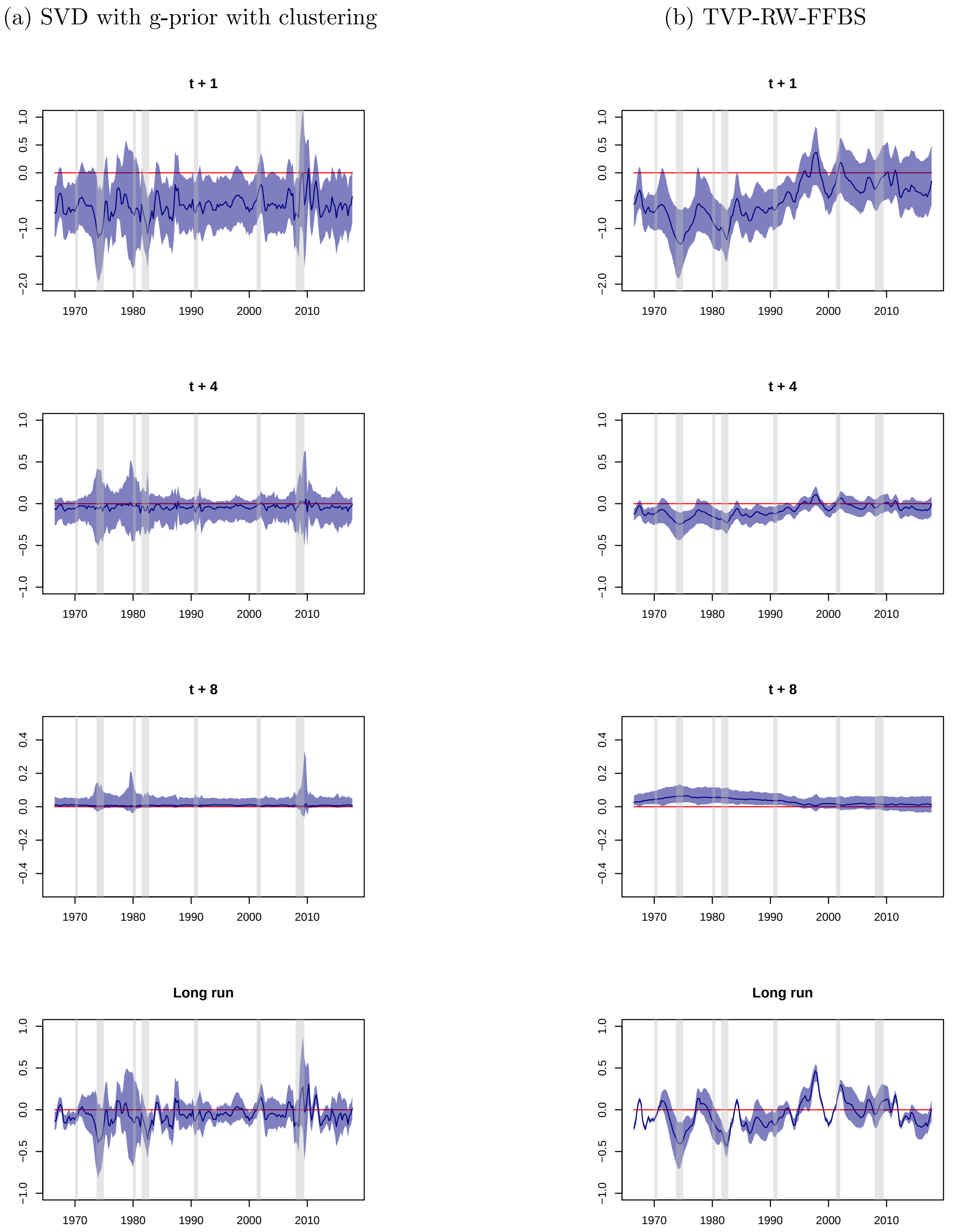
Figure 3 plots multipliers (i.e. the cumulative effect on inflation of a change in unemployment at various horizons). A comparison of SVD to TVP-RW-FFBS shows many similarities. For instance, both models are saying an increase in unemployment has a negative effect on inflation in the very short term for much of the time. This is what the NKPC would lead us to expect. However, for SVD this negative effect remains for most of the time after the financial crisis whereas for TVP-RW-FFBS it vanishes and the NKPC relationship breaks down. Another difference between the two approaches can be seen in many recessions where the estimated effect changes much more abruptly using our approach than with TVP-RW-FFBS. This illustrates the great flexibility of our approach in terms of the types of parameter change allowed for. And this flexibility does not cost us much in terms of estimation precision in the sense that the credible intervals for the two approaches have similar width.

Figure 4 displays the posterior of G_0, the number of clusters selected by the algorithm. The posterior is spread over a range of values, although almost all of the posterior probability is associated with a number of clusters between 10 and 20. $G_0 = 1$ implies that $\tilde{\boldsymbol{\beta}}_t$ is centered around a non-zero value that is time-invariant and there is little posterior evidence in this figure indicating support for this. This is the lower bound on the number of clusters. The upper bound on the number of clusters is 30, but the posterior probability lies in a region far away from 30 indicating that the algorithm is successfully finding parsimonious representations for the time variation in parameters. It is worth stressing that these statements hold for the small NKPC model. For the large model with $K = 101$, we find the number of clusters to be even smaller. In this case the posterior mode is eight clusters. This inverse relationship between K and G_0 is to be expected. That is, as model size increases more of the variation over time can be captured by the richer information set in $\boldsymbol{X}_t$, leaving less of a role for time variation in coefficients. Our clustering algorithm automatically adjusts to this effect.

7.2 Forecasting evidence

The forecasting design adopted is recursive. We consider an initial estimation period from 1965Q1 to 1999Q4. The remaining observations (2000Q1 to 2018Q4) are used as a hold-out period to evaluate our forecasting methods. After obtaining $h \in \{1, 4\}$-step-ahead predictive distributions for a given period in the hold-out, we include this period in the estimation sample and repeat this procedure until we reach the end of the sample. In order to compute longer horizon forecasts, we adopt the direct forecasting approach (see e.g. Stock and Watson, 2002). To assess forecasting accuracy, we use root mean square forecast errors (RMSEs) for point forecasts and log predictive likelihoods (LPLs, these are averaged over the hold-out period) for density forecasts. We evaluate the statistical significance of the forecasts relative to random walk (RW) forecasts using the Diebold-Mariano test.

We compare three variants of our SVD approach (i.e. the Minnesota prior and the g-prior with and without clustering) to alternatives which vary in their treatment of parameter change and in the number of explanatory variables. With regards to parameter change, we consider the time-invariant (TIV) model (no change in the regression coefficients which is obtained by setting $\tilde{\boldsymbol{\beta}}_t = \mathbf{0}$ for all t) and the TVP-RW-FFBS approach which has random walk parameter change. With regards to the number of explanatory variables, we consider models with

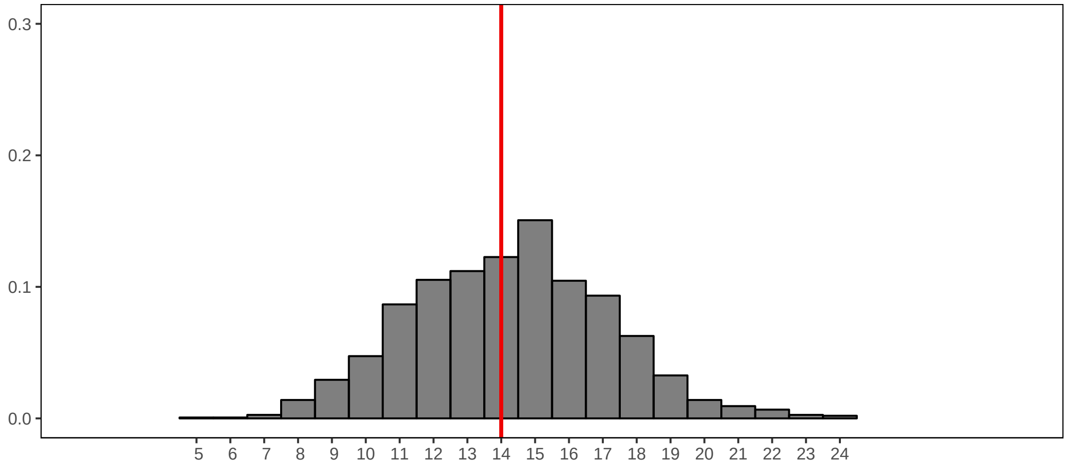


(a) SVD with g-prior with clustering

(b) TVP-RW-FFBS

Notes: Blue shaded areas are 68% credible intervals and gray shaded areas denote NBER recessions.

Figure 3: Posterior means of multipliers



Notes: G_0 refers to the non-empty groups with $G = 30$. The red line denotes the median of G_0.

Figure 4: Posterior distribution of number of non-empty clusters (G_0)

two lags of all 50 of them (labelled FULL in the tables), none of them as well as some specifications which contain a subset of them. To be specific, we present results for all these models using the NKPC specification discussed in the preceding sub-section (labelled NKPC in the tables). We also have versions of the model where the intercept is the only explanatory variable, thus leading to an unobserved components model (labeled UCM in the tables).[9] In addition, we include some simple benchmarks that have been used elsewhere in the literature. These include a constant coefficient AR(2) model, a TVP-AR(2) and an AR(2) augmented with the first three principal components of $\boldsymbol{X}_t$ (this is labeled PCA3). We also present forecasts for a TVP-RW-FFBS model where the explanatory variables are the first three principal components. All models considered include stochastic volatility.

Table 3 contains our main set of forecasting results. Note first that, with some exceptions, the FULL models do best, indicating that there is information in our $K = 50$ variables useful for inflation forecasting. If we focus on results for the FULL models, it can be seen that, for $h = 1$ all of the approaches forecast approximately as well as each other. But for $h = 4$ there are substantial improvements provided by our SVD approaches relative to TVP-RW-FFBS. At this forecast horizon, it is interesting to note that the very parsimonious UCM version of the TVP-RW-FFBS provides point forecasts that are almost as good as those provided by the FULL SVD approaches. However, the density forecasts provided by the UCM are appreciably worse than those provided by the

[9] For the SVD versions of the UCM models, we only present results for the g-prior with clustering as the other priors imply white noise behavior for inflation which is not sensible.

	Specification			**Forecast horizon**	
	TVP/TIV	Type	κ	1-step	4-steps
AR(P)					
	TIV	Benchmark		0.90 (0.08)	0.77*** (0.23***)
	TVP-FFBS-RW	Benchmark		0.90 (0.08)	0.75*** (0.26***)
FULL					
	TIV	Benchmark		0.82* (0.15)	0.61** (0.37)
	TVP-FFBS-RW	Benchmark		0.78* (0.16)	0.92 (0.01)
	TVP-SVD	g-prior	0.1	0.80*** (0.15**)	0.59** (0.42*)
	TVP-SVD	g-prior (clustering)	0.05	0.80*** (0.17**)	0.57*** (0.48***)
	TVP-SVD	Minnesota	0.1	0.82** (0.16*)	0.61** (0.37*)
NKPC					
	TIV	Benchmark		0.91 (0.06)	0.82*** (0.12)
	TVP-FFBS-RW	Benchmark		0.92 (0.07)	0.86 (-0.28*)
	TVP-SVD	g	0.001	0.89 (0.07)	0.79*** (0.13)
	TVP-SVD	g-prior (clustering)	0.001	0.90 (0.07)	0.80*** (0.12)
	TVP-SVD	Minnesota	0.001	0.91 (0.05)	0.81*** (0.13)
PCA3					
	TIV	Benchmark		0.92 (0.06)	0.83*** (0.18***)
	TVP-FFBS-RW	Benchmark		0.88 (0.09)	0.86 (0.05)
UCM					
	TVP-FFBS-RW	Benchmark		0.86*** (0.16)	0.59** (0.16)
	TVP-SVD	g-prior (clustering)	1	0.88* (0.08)	0.71 (0.14)

Table 3: Forecasting Performance of SVD Approaches Relative to Benchmarks

Notes: The table shows RMSEs with LPL's in parentheses below. Asterisks indicate statistical significance for each model relative to a random walk at the 1 (***), 5 (**) and 10 (*) percent significance levels.

SVD approaches. The FULL SVD approaches are also beating PCA approaches, even if we allow for time-variation in the coefficients on the PCAs.

With two different forecast horizons and two different forecast metrics, we have four possible ways of evaluating any approach. For three of these, the FULL SVD approach using the g-prior with clustering performs best. The only exception to this is for MSFEs for $h = 1$, although even here FULL SVD with g-prior is the second best performing approach. The improvements relative to our other SVD approaches which do not involve clustering are small, but are consistently present. This indicates the benefits of the clustering prior.

In general, the TIV approaches do well (for $h = 4$ even better than TVP-RW-FFBS) in terms of point forecasts, but the density forecasts produced by our SVD approaches are slightly better. This suggests there is only a small amount of time-variation in this data set, but that our SVD approach (particularly when we add the hierarchical clustering prior) is effectively picking it out in a manner that the random walk evolution of the TVP-RW-FFBS cannot.

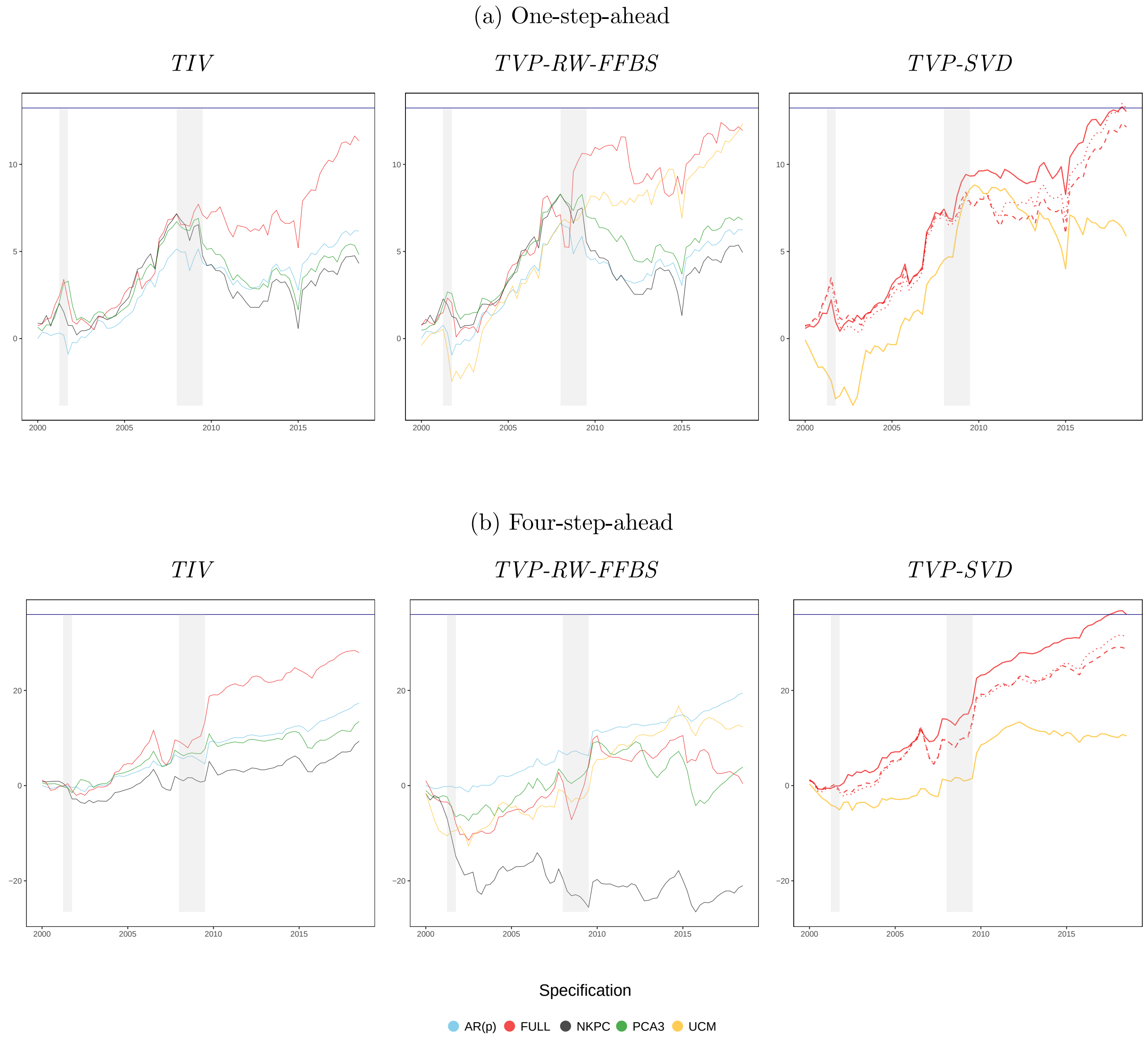
Figure 5 provides evidence of forecast performance over time for the main models used in this forecasting exercise. The lines in this figure are cumulated log predictive Bayes factors relative to a random walk.

One pattern worth noting is that the benefits of using the FULL model increase after the beginning of the financial crisis. This is true not only for our SVD models, but also for the TIV model. However, notice that during the crisis, the slope of the line associated with the FULL SVD approach becomes steeper, indicating that the model strongly outperforms the RW for that specific time period. This potentially arises from the fact that during recessions, we typically face abrupt structural breaks in the regression parameters and our approach is capable of detecting them.

To examine in more detail forecast performance in the Great Recession, it is worthwhile to keep in mind that inflation was fairly stable through 2008Q3. 2008Q4 and 2009Q1 were the periods associated with a substantial fall in inflation. Subsequently, inflation became more stable again. Accordingly, it is particularly interesting to look at 2008Q4 and 2009Q1 as periods of possible parameter change. We find that the FULL SVD approach performs comparable to a no-change benchmark model. The simple RW model can be expected to handle a one-off structural break well in the sense that it will forecast poorly for the one period where the break occurs and then immediately adjust to the new lower level of the series. Our FULL SVD approach handles the 2008Q4 and 2009Q1 period about as well as the RW. Subseqeuently, its forecasts improve relative to a RW. This improvement occurs in the middle of the Great Recession for $h = 1$ and a bit later for $h = 4$. In contrast, the TVP-RW-FFBS model with the large data set experiences a big drop in forecasting performance at the beginning of the Great Recession and tends not to outperform the random walk after 2010. However, it does well in later 2009. We conjecture that this pattern of performance of the TVP-RW-FFBS reflects two things. First, similarly to our SVD based models, it does allow for structural breaks, but is slow to adjust to them. Second, it overfits the data and, thus, provides a wide predictive distribution. In the latter half of 2009, after the structural break had occured, when there was still uncertainty about the new pattern in inflation, having this wider predictive distribution benefitted forecast performance.

It can also be seen that our SVD approaches tend to perform similarly to one another and never forecast very poorly. This contrasts with the TVP-RW-FFBS models which sometimes forecast well, but sometimes forecast very poorly (see, e.g., results for $h = 4$ using the NKPC data set).

Overall, we are finding the our SVD approaches, and in particular the version that uses the clustering prior, to exhibit the best forecast performance among a set of popular benchmarks. And it is worth stressing that they are computationally efficient and, thus, scaleable. The reason this application uses $K = 100$ explanatory variables as opposed to a much larger number is due to our wish to include the slower TVP-RW-FFBS approach so as to offer a comparison with the most popular TVP regression model. If we were to have omitted this comparison, we could have chosen K to be much larger.


(a) One-step-ahead

(b) Four-step-ahead

Notes: The log predictive Bayes factors are cumulated over the hold-out. The solid line refers to the g-prior with clustering, the dashed line to the Minnesota prior and the dotted line to the g-prior without clustering. The blue lines refer to the maximum Bayes factor at the end of the hold-out sample. The gray shaded areas indicate the NBER recessions in the US.

Figure 5: Evolution of log predictive Bayes factor relative to RW

8 Conclusions

In many empirical applications in macroeconomics, there is strong evidence of parameter change. But there is often uncertainty about the form the parameter change takes. Conventional approaches to TVP regression models have typically made specific assumptions about the form of parameter change (e.g. random walk or structural break). In the specification used in this paper, no restriction is placed on the form that the parameter change can take. However, our very flexible specification poses challenges in terms of computation and surmounting over-parameterization concerns. We have addressed the computational challenge through using the SVD of the high-dimensional set of regressors. We show how this leads to large simplifications since key matrices become diagonal or have banded forms. The over-parameterization worries are overcome through the use of hierarchical priors and, in particular, through the use of a sparse finite mixture representation for the time-varying coefficients. In artificial data, we demonstrate the speed and scaleability of our methods relative to standard approaches. In an inflation forecasting exercise, we show how our methods can uncover different forms of time-variation in parameters than other approaches. Furthermore, they forecast well.

References

Ball L and Mazumder S (2011) Inflation dynamics and the Great Recession. *Brookings Papers on Economic Activity* 42(1 (Spring), 337–405

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

Bitto A and Frühwirth-Schnatter S (2019) Achieving shrinkage in a time-varying parameter model framework. *Journal of Econometrics* 210(1), 75–97

Bitto-Nemling A, Cadonna A, Frühwirth-Schnatter S and Knaus P (2019) Shrinkage in the Time-Varying Parameter Model Framework Using the R Package shrinkTVP. *arXiv preprint arXiv:1907.07065*

Carriero A, Clark T and Marcellino M (2016) Large Vector Autoregressions with stochastic volatility and flexible priors. *Federal Reserve Bank of Cleveland Working Paper, no. 16-17*

Carter C and Kohn R (1994) On Gibbs sampling for state space models. *Biormetrika* 81(3), 541–553

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

Clark T (2011) Real-time density forecasts from BVARs with stochastic volatility. *Journal of Business and Economic Statistics* 29, 327–341

Cogley T and Sargent TJ (2005) Drifts and volatilities: monetary policies and outcomes in the post WWII US. *Review of Economic Dynamics* 8(2), 262 – 302

Coibion O and Gorodnichenko Y (2015) Is the Phillips Curve alive and well after all? Inflation expectations and the missing disinflation. *American Economic Journal: Macroeconomics* 7(1), 197–232

D'Agostino A, Gambetti L and Giannone D (2013) Macroeconomic forecasting and structural change. *Journal of Applied Econometrics* 28(1), 82–101

De Mol C, Giannone D and Reichlin L (2008) Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics* 146(2), 318–328

Doan T, Litterman R and Sims C (1984) Forecasting and conditional projection using realistic prior distributions. *Econometric reviews* 3(1), 1–100

Frühwirth-Schnatter S (1994) Data augmentation and dynamic linear models. *Journal of Time Series Analysis* 15(2), 183–202

Frühwirth-Schnatter S (2001) Markov chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models. *Journal of the American Statistical Association* 96(453), 194–209

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

Giordani P and Kohn R (2008) Efficient Bayesian Inference for Multiple Change-Point and Mixture Innovation Models. *Journal of Business & Economic Statistics* 26(1), 66–77

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

Hassani H and Silva ES (2015) Forecasting with Big Data: A Review. *Annals of Data Science* 2(1), 5–19

Huber F, Koop G and Onorante L (2019) Inducing sparsity and shrinkage in time-varying parameter models. *arXiv preprint arXiv:1905.10787*

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

Kastner G and Frühwirth-Schnatter S (2014) Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models. *Computational Statistics & Data Analysis* 76, 408–423

Kastner G and Huber F (2017) Sparse Bayesian vector autoregressions in huge dimensions. *manuscript*

Koop G and Korobilis D (2012) Forecating inflation using dynamic model averaging. *International Economic Review* 53(3), 867–886

Koop G and Korobilis D (2018) Variational Bayes inference in high dimensional time-varying parameter models. *manuscript*

Koop G, Korobilis D and Pettenuzzo D (2019) Bayesian compressed Vector Autoregressions. *Journal of Econometrics* 210(1), 135–154

Korobilis D (2019) High-dimensional macroeconomic forecasting using message passing algorithms. *Journal of Business and Economic Statistics* (forthcoming), 1–30

Litterman R (1986) Forecasting with Bayesian Vector Autoregressions: Five years of experience. *Journal of Business and Economic Statistics* 4, 25–38

Malsiner-Walli G, Frühwirth-Schnatter S and Grün B (2016) Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing* 26, 303324

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

McCracken MW and Ng S (2016) FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics* 34(4), 574–589

Medeiros MC, Vasconcelos GF, Veiga Á and Zilberman E (2019) Forecasting Inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business and Economic Statistics* , 1–22

Moretti L, Onorante L and Zakipour Saber S (2019) Phillips curves in the euro area. Working Paper Series 2295, European Central Bank

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

Raftery A, Krn M and Ettler P (2010) Online prediction under model uncertainty via Dynamic Model Averaging: Application to a cold rolling mill. *Technometrics* 52(1), 52–66. PMID: 20607102

Rockova V and McAlinn K (2018) Dynamic variable selection with spike-and-slab process priors. *Bayesian Analysis*

Stock J and Watson M (1999) Forecasting inflation. *Journal of Monetary Economics* 44(2), 293–335

Stock J and Watson M (2002) Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics* 20(2), 147–162

Stock J and Watson M (2007) Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking* 39(s1), 3–33

Stock J and Watson M (2008) Phillips curve inflation forecasts. NBER Working Papers 14322, National Bureau of Economic Research, Inc

Stock J and Watson M (2011) Dynamic factor models. *Oxford Handbook of Forecasting*

Trippe B, Huggins J, Agrawal R and Broderick T (2019) LR-GLM: High-dimensional Bayesian inference using low-rank data approximations. *arXiv preprint arXiv:1905.07499*

Watson M (2014) Inflation persistence, the NAIRU, and the Great Recession. *American Economic Review* 104(5), 31–36

Zellner A (1986) On Assessing Prior Distributions and Bayesian Regression Analysis with g Prior Distributions. *n Goel, P.; Zellner, A. (eds.). Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. Studies in Bayesian Econometrics and Statistics 6* , 233–243

Appendix A Full conditional posterior distributions

In this section, we provide details on the full conditional posterior distributions of the model described in Section 2. We start by outlining the relevant full conditionals for the time-invariant part of the model. The conditional posterior of the time-invariant coefficients $\boldsymbol{\gamma}$ follows a multivariate Gaussian distribution:

$$\boldsymbol{\gamma}|Data, \tilde{\boldsymbol{\beta}}, \boldsymbol{\sigma}^2, \boldsymbol{\tau}, \psi \sim \mathcal{N}\left(\bar{\boldsymbol{\gamma}}, \boldsymbol{V}_{\gamma}\right), \tag{A.1}$$

with

$$\begin{aligned} \boldsymbol{V}_{\gamma} =& (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} + \boldsymbol{D}_{\gamma}^{-1})^{-1}, \\ \bar{\boldsymbol{\gamma}} =& \boldsymbol{V}_{\gamma}(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}}). \end{aligned}$$

Hereby, we let $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_T^2)'$ denote a T-dimensional vector of volatilities, $\boldsymbol{\tau} = (\tau_1, \dots, \tau_K)'$ stores the K local scaling parameters of the NG prior while the $T \times K$-dimensional matrix $\tilde{\boldsymbol{X}}$ is obtained by stacking the rows of $\boldsymbol{x}_t$ and normalizing by dividing each row by σ_t.

The local scaling parameters follow a generalized inverse Gaussian (GIG) distribution (Griffin and Brown, 2010):

$$\tau_j|\gamma_j, \psi \sim \text{GIG}\left(\vartheta - \frac{1}{2}, \vartheta\psi, \gamma_j^2\right), \text{ for } j = 1, \dots, K. \tag{A.2}$$

The full conditional posterior distribution of the global shrinkage parameter is defined as

$$\psi|\tau_1, \dots, \tau_K \sim \mathcal{G}\left(a_0 + \vartheta K, b_0 + \frac{\vartheta}{2}\sum_{k=1}^{K}\tau_j\right). \tag{A.3}$$

To update $\boldsymbol{\theta}$, irrespective of the prior on $\tilde{\boldsymbol{\beta}}$ adopted, we use a RWMH step. Due to the hierarchical nature of the model, the likelihood $p(\tilde{\boldsymbol{\beta}}|\boldsymbol{\Sigma}, \boldsymbol{b}_0, \boldsymbol{D}_0)$ does not depend on the data and is given by

$$p(\tilde{\boldsymbol{\beta}}|\boldsymbol{\Sigma}, \boldsymbol{b}_0, \boldsymbol{D}_0) = f_{\mathcal{N}}(\tilde{\boldsymbol{\beta}}|\boldsymbol{b}_0, \boldsymbol{\Sigma}\boldsymbol{V}_{\tilde{\beta}}). \tag{A.4}$$

This conditional distribution is then combined with the appropriate Uniform prior and easy to evaluate since $\boldsymbol{\Sigma}\boldsymbol{V}_{\tilde{\beta}}$ is a diagonal matrix. As a proposal distribution for $\boldsymbol{\theta}$, we use a log-Normal distribution:

$$\log \boldsymbol{\theta}^* = \log \boldsymbol{\theta}^{(a)} + \sigma_\theta \boldsymbol{\zeta}, \quad \boldsymbol{\zeta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}).$$

[9] The GIG(a, b, c) is parameterized as $p(x) \propto x^{a-1}\exp\{-(bx + c/x)/2\}$.

Hereby, we let $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^{(a)}$ denote the proposed and previously accepted value of $\boldsymbol{\theta}$, respectively. Moreover, σ_θ is a scaling factor that is specified such that the acceptance rate of the MH algorithm is between 20 and 40%. This is achieved by adjusting σ_θ over the first 25% of the burn-in stage.

With the clustering prior, the algorithm becomes slightly more complicated and the following steps need to be added.

The posterior distribution of the mixture probabilities follows a Dirichlet distribution:

$$\boldsymbol{w}|\boldsymbol{\delta} \sim \mathscr{D}ir(\pi_1, \dots, \pi_G), \tag{A.5}$$

with $\pi_g = \pi + T_g$, where T_g denotes the number of $\tilde{\boldsymbol{\beta}}_t'$s assigned to group g, and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_T)'$.

It can be shown that the regime indicators δ_t $(t = 1, \dots, T)$ follow a Multinomial distribution with

$$\Pr(\delta_t = g|w_g, \boldsymbol{\mu}_g, \sigma_t, \boldsymbol{\Psi}) \propto w_g f_{\mathcal{N}}(\tilde{\boldsymbol{\beta}}_t|\boldsymbol{\mu}_g, \sigma_t\boldsymbol{\Psi}), \quad \text{for } g = 1, \dots, G. \tag{A.6}$$

The full conditional posterior of $\boldsymbol{\mu} = \text{vec}(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_G)$ follows a multivariate Gaussian distribution with diagonal variance-covariance matrix:

$$\boldsymbol{\mu}|\boldsymbol{\Pi}, \boldsymbol{\sigma}, \boldsymbol{\delta}, \boldsymbol{\mu}_0 \sim \mathcal{N}(\overline{\boldsymbol{\mu}}, (\boldsymbol{I}_G \otimes \boldsymbol{\Psi}) \odot \overline{\boldsymbol{V}_\mu}), \tag{A.7}$$

where the posterior variance and mean are given by, respectively:

$$\begin{aligned}\overline{\boldsymbol{V}_\mu} &= \left(\boldsymbol{I}_K \otimes \boldsymbol{Q}'\boldsymbol{Q} + \boldsymbol{I}_G \otimes \boldsymbol{\Pi}^{-1}\right)^{-1}, \\ \overline{\boldsymbol{\mu}} &= \overline{\boldsymbol{V}_\mu}\left(\boldsymbol{I}_K \otimes \boldsymbol{Q}'\tilde{\boldsymbol{\beta}}^* + \boldsymbol{\iota}_G \otimes \boldsymbol{\Pi}^{-1}\boldsymbol{\mu}_0\right).\end{aligned}$$

We let $\boldsymbol{Q}$ denote a $T \times G$ matrix with t^{th} row given by $\boldsymbol{Q}_t = (I(\delta_t = 1)/\sigma_t, \dots, I\,(\delta_t = G)/\sigma_t)$, $\tilde{\boldsymbol{\beta}}^*$ is simply $\tilde{\boldsymbol{\beta}}$ normalized by dividing through σ_t, and $\boldsymbol{\iota}_G$ is a G-dimensional vector of ones. Notice that it is straightforward to sample from (A.7) because all involved quantities are easily vectorized.

Similarly, the full conditional posterior distribution of the common mean $\boldsymbol{\mu}_0$ follows a Gaussian:

$$\boldsymbol{\mu}_0|\boldsymbol{\Pi}, \boldsymbol{\mu} \sim \mathcal{N}\left(\frac{\sum_{g=1}^G \boldsymbol{\mu}_g}{G}, \frac{1}{G}\boldsymbol{\Pi}\right). \tag{A.8}$$

The full conditional posterior distribution of the elements $\upsilon_1, \dots, \upsilon_K$ of the common variance-covariance is defined as a GIG distribution

$$\upsilon_j|\boldsymbol{R}, \boldsymbol{\mu} \sim \text{GIG}\left(c_0 - \frac{G}{2}, 2c_1, \frac{\sum_{g=1}^G (\mu_{gj} - \mu_{0j})^2}{R_j}\right). \tag{A.9}$$

Here, μ_{gj} $(g = 1, \ldots, G)$ denotes the j^{th} element of group-specific means $\boldsymbol{\mu}_g$, μ_{0j} is the j^{th} element of the common mean $\boldsymbol{\mu}_0$ and R_j corresponding to the j^{th} element of $\boldsymbol{R}$.

Appendix B Data description

FRED.Mnemonic	Description	Trans I(0)
CPIAUCSL	**Consumer Price Index for All Urban Consumers: All Items**	**5**
GDPCTPI	Gross Domestic Product: Chain-type Price Index	5
PCECTPI	Personal Consumption Expenditures: Chain-type Price Index	5
GDPC1	Real Gross Domestic Product	5
PCECC96	Real Personal Consumption Expenditures	5
FPIx	Real private fixed investment	5
INDPRO	IP:Total index Industrial Production Index (Index 2012=100)	5
CUMFNS	Capacity Utilization: Manufacturing (SIC) (Percent of Capacity)	1
PAYEMS	Emp:Nonfarm All Employees: Total nonfarm (Thousands of Persons)	5
CE16OV	Civilian Employment (Thousands of Persons)	5
UNRATE	Civilian Unemployment Rate (Percent)	1
UNRATESTx	Unemployment Rate less than 27 weeks (Percent)	1
UNRATELTx	Unemployment Rate for more than 27 weeks (Percent)	1
LNS14000012	Unemployment Rate - 16 to 19 years (Percent)	1
LNS14000025	Unemployment Rate - 20 years and over, Men (Percent)	1
LNS14000026	Unemployment Rate - 20 years and over, Women (Percent)	1
UEMPLT5	Number of Civilians Unemployed - Less Than 5 Weeks (Thousands of Persons)	5
UEMP5TO14	Number of Civilians Unemployed for 5 to 14 Weeks (Thousands of Persons)	5
UEMP15T26	Number of Civilians Unemployed for 15 to 26 Weeks (Thousands of Persons)	5
UEMP27OV	Number of Civilians Unemployed for 27 Weeks and Over (Thousands of Persons)	5
AWHMAN	Average Weekly Hours of Production and Nonsupervisory Employees: Manufacturing	1
CES0600000007	Average Weekly Hours of Production and Nonsupervisory Employees: Goods-Producing	1
HOUST	Housing Starts: Total: New Privately Owned Housing Units Started	5
PERMIT	New Private Housing Units Authorized by Building Permits	5
IPDBS	Business Sector: Implicit Price Deflator (Index 2012=100)	5
CPILFESL	Consumer Price Index for All Urban Consumers: All Items Less Food & Energy	5
WPSFD49207	Producer Price Index by Commodity for Finished Goods	5
PPIACO	Producer Price Index for All Commodities	5
WPSFD49502	Producer Price Index by Commodity for Finished Consumer Goods	5
WPSFD4111	Producer Price Index by Commodity for Finished Consumer Foods	5
PPIIDC	Producer Price Index by Commodity Industrial Commodities	5
WPSID61	Producer Price Index by Commodity Intermediate Materials: Supplies & Components	5
WPU0561	Producer Price Index by Commodity for Fuels and Related Products and Power	5
OILPRICEx	Real Crude Oil Prices: West Texas Intermediate (WTI) - Cushing, Oklahoma	5
WPSID62	Producer Price Index: Crude Materials for Further Processing	5
PPICMM	Producer Price Index: Commodities: Metals and metal products: Primary nonferrous metals	5
CPIAPPSL	Consumer Price Index for All Urban Consumers: Apparel	5
CPITRNSL	Consumer Price Index for All Urban Consumers: Transportation	5
CPIMEDSL	Consumer Price Index for All Urban Consumers: Medical Care	5
CES2000000008x	Real Average Hourly Earnings of Production and Nonsupervisory Employees: Construction	5
CES3000000008x	Real Average Hourly Earnings of Production and Nonsupervisory Employees: Manufacturing	5
COMPRNFB	Nonfarm Business Sector: Real Compensation Per Hour (Index 2012=100)	5
CES0600000008	Average Hourly Earnings of Production and Nonsupervisory Employees:	5
FEDFUNDS	Effective Federal Funds Rate (Percent)	1
TB3MS	3-Month Treasury Bill: Secondary Market Rate (Percent)	1
GS10	10-Year Treasury Constant Maturity Rate (Percent)	1
GS10TB3Mx	10-Year Treasury Constant Maturity Minus 3-Month Treasury Bill, secondary market	1
M1REAL	Real M1 Money Stock	5
S.P.500	S&P Common Stock Price Index: Composite	5
S.P..indust	S&P Common Stock Price Index: Industrials	5

Table B1: Data is obtained from the FRED data base of the Federal Reserve of St. Louis. The column `Trans I(0)` denotes the transformation applied to each variable so as to make it stationary. The transformation codes are taken from McCracken and Ng (2016) with (1) implying no transformation applied and (5) denoting growth rates, defined as log fist differences $ln\left(\frac{x_t}{x_{t-1}}\right)$. All variables are standardized by substracting the mean and dividing by the standard deviation.

Appendix C Empirical Appendix

Our priors are hierarchical and involve few prior hyperparameters which must be selected by the researcher. For most of these, we can draw upon existing papers such as Malsiner-Walli et al. (2016) to provide suggestions for sensible benchmark values. However, the choice of the prior hyperparameter κ defined in equation (6) does not fall into this category and, hence, it is worth offering some additional discussion of prior robustness relating to κ in this appendix. This hyperparameter determines the upper bound of the grid for the parameters which enter the prior covariance matrix.

We repeat our forecasting exercise using the SVD approach using our three priors and three different model sizes, but allow a range of values for κ. Results are given in Table 3. Note first that for $\kappa = 1$, which is the largest value we considered, no results are given for the FULL or NKPC models. For these cases, $\kappa = 1$ led to severe over-fitting problems and resulting poor forecast performance (e.g. LPLs of minus infinity). Clearly this value does not induce adequate shrinkage in larger models and care must be taken to avoid such regions of the hyperparameter space.

However, provided κ is kept small, we find a high degree of prior robustness. If we consider point forecast performance, we find that as long as the upper bound is specified between 0.01 and 0.1, point forecasts are only slightly affected by the choice of κ. No clear patterns emerge. For one-step-ahead forecasts using the FULL model estimated using the g-prior, relative RMSEs seem to be inversely related to κ. But this does not carry over to the four-steps-ahead horizon. For longer-run forecasts, we observe that accuracy is largest if κ is set equal to 0.05. For the NKPC model, a similar U-shaped pattern arises, indicating that the optimal value of κ should be between 0.005 and 0.05. The only model that strongly profits from using a larger scaling paramater is the UCM. Here, we find that the best forecasting performance is obtained if $\kappa = 1$, a choice that seriously distorts predictive accuracy for larger models. In the case of the Minnesota prior, the specific choice of κ plays a limited role, with point forecasts of the full model being indistinguishable from each other for the one-quarter-ahead horizon and quite similar for the one-year-ahead horizon.

When the full predictive distribution is considered, we also find only limited differences in predictive accuracy for varying values of κ. For the large-scale model and all priors, differences are typically quite small. This suggests that the precise value of κ, as long as it is not specified too large, plays only a minor role in impacting forecasting accuracy. Notice, however, that this does not carry over to smaller models. Interestingly, we find that for the NKPC model, using a smaller κ improves LPLs for both priors and forecast horizons. We conjecture that in this small-dimensional setting, the Uniform prior on the hyperparameters of the Minnesota and the g-prior translate

	Specification		κ					
	Horizon	Information set	0.001	0.005	0.01	0.05	0.1	1
G-PRIOR (NO CLUSTERING)								
	1-step-ahead	FULL	0.81**	0.82**	0.80**	0.79***	0.80***	
			(0.15*)	(0.12)	(0.18**)	(0.13)	(0.15**)	
		NKPC	0.89	0.88	0.89	0.94	1.02	
			(0.07)	(0.06)	(0.00)	(-0.07)	(-0.15**)	
	4-step-ahead	FULL	0.60**	0.59**	0.61**	0.58**	0.59**	
			(0.35)	(0.36)	(0.35)	(0.41*)	(0.42*)	
		NKPC	0.79***	0.78*	0.76*	0.85	0.95	
			(0.13)	(0.11)	(0.10)	(0.03)	(-0.03)	
G-PRIOR (CLUSTERING)								
	1-step-ahead	FULL	0.82**	0.81**	0.80***	0.80***	0.76**	
			(0.16*)	(0.13)	(0.15*)	(0.17**)	(0.13)	
		NKPC	0.90	0.88*	0.88*	0.96	1.31	
			(0.07)	(0.00)	(-0.05)	(-0.14*)	(-0.24***)	
		UCM	1.10**	1.10**	1.11**	1.05	1.01	0.88*
			(-0.26***)	(-0.26***)	(-0.25**)	(-0.19**)	(-0.18*)	(0.08)
	4-step-ahead	FULL	0.60***	0.60***	0.59***	0.57***	0.69**	
			(0.43***)	(0.43***)	(0.46***)	(0.48***)	(0.32***)	
		NKPC	0.80***	0.77*	0.76	0.85	1.00	
			(0.12)	(0.11)	(0.08)	(-0.01)	(-0.01)	
		UCM	1.07*	1.06*	1.06*	1.00	0.96	0.71
			(-0.35**)	(-0.34**)	(-0.34**)	(-0.27*)	(-0.24)	(0.14)
MINNESOTA								
	1-step-ahead	FULL	0.82**	0.82**	0.82**	0.82*	0.82**	
			(0.13)	(0.12)	(0.14)	(0.15)	(0.16*)	
		NKPC	0.91	0.88	0.88*	0.95	1.08	
			(0.05)	(0.05)	(0.03)	(-0.04)	(-0.14)	
	4-step-ahead	FULL	0.61**	0.61**	0.62**	0.61**	0.61**	
			(0.36)	(0.37)	(0.37*)	(0.38*)	(0.37*)	
		NKPC	0.81***	0.78***	0.78**	0.75	0.77	
			(0.13)	(0.12)	(0.12)	(0.06)	(0.03)	

Table C1: : Forecasting Performance for Different Values of κ

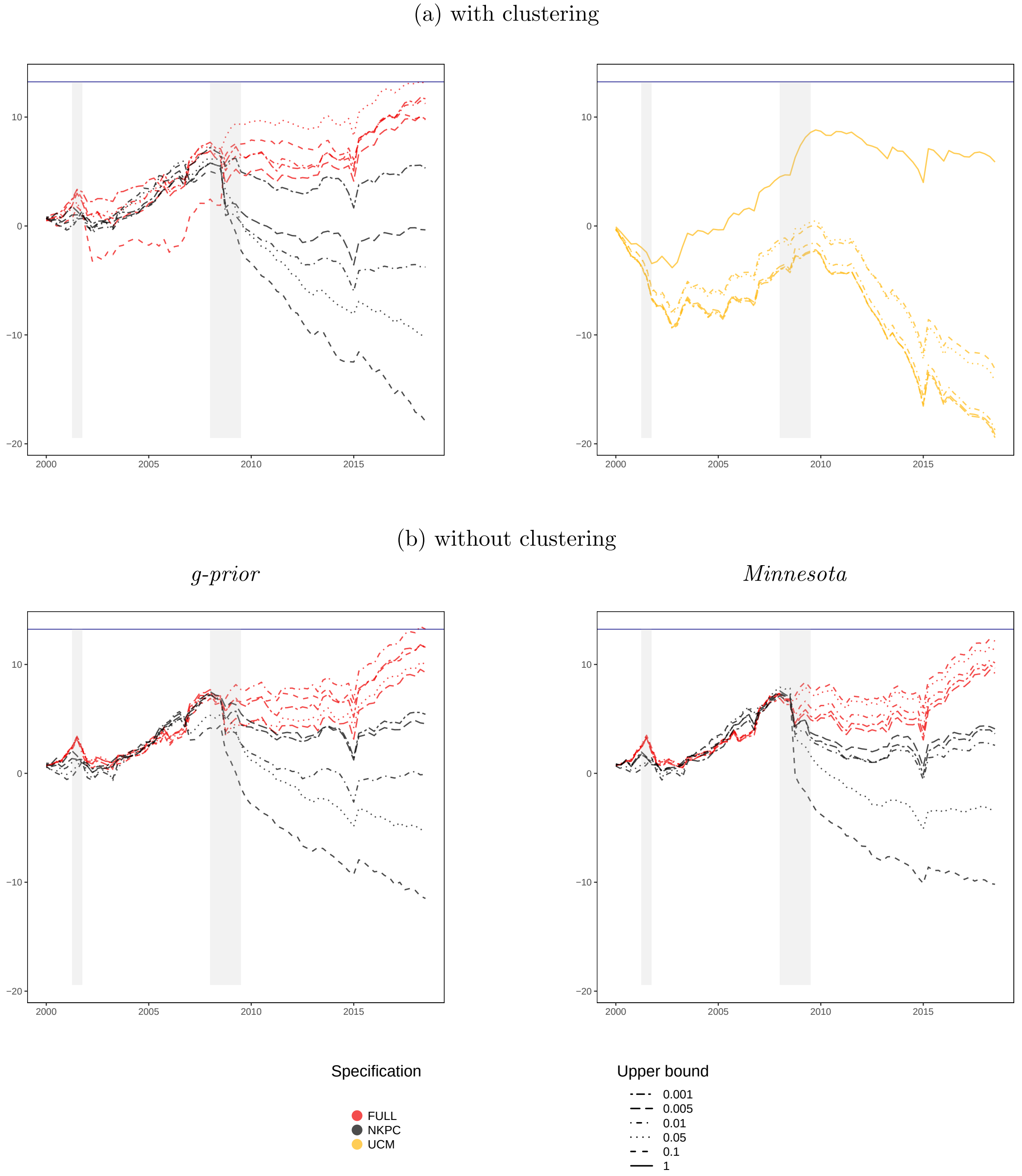
Notes: The table shows RMSEs with LPL's in parentheses below. Asterisks indicate statistical significance for each model relative to a random walk at the 1 (***), 5 (**) and 10 (*) percent significance levels.

into an overfitting model and this, in turn, hurts forecast accuracy. κ then simply acts as an additional shrinkage parameter that pushes $\boldsymbol{\theta}$ to zero for both priors.

When we consider the performance of the UCM models based on the full predictive density, we find that forecast accuracy appreciably increases with κ. This contrasts with the findings for the other models. This is due to the fact that with small values of κ, it becomes increasingly difficult to control for unobserved heterogeneity and the resulting model approaches a white noise specification.

Figures 6 and 7 plot the cumulative Bayes factors comparing a variety of approaches and values for κ (which is labelled "upper bound" in the figures) against the random walk for $h = 1$ and $h = 4$, respectively. These reinforce the preceding discussion. For the FULL model, containing all the regressors, lines for different values of κ are plotted in red on the figures. Note that all the red lines are similar to one another. Note too, the increasingly good performance of the FULL models after the financial crisis for $h = 4$. For the smaller data sets and, in particular, the UCM model there is much more sensitivity to the choice of κ. For the UCM model, it is interesting to note that the deterioration in forecast performance associated with poor choices of κ occurs largely after the financial crisis. For the NKPC specifications, when $h = 1$ some choices of κ lead to poor forecast performance. But for $h = 4$, with the NKPC specification, we are finding much more robustness.

To sum up this discussion, we find that the precise value of κ plays only a limited role for forecasting accuracy when the large model is adopted, provided we avoid large values of κ which clearly lead to over-fitting problems. In contrast, in smaller models, the precise value of κ has a bigger impact on forecasting accuracy and the researcher needs to carefully select this hyperparameter.



(a) with clustering

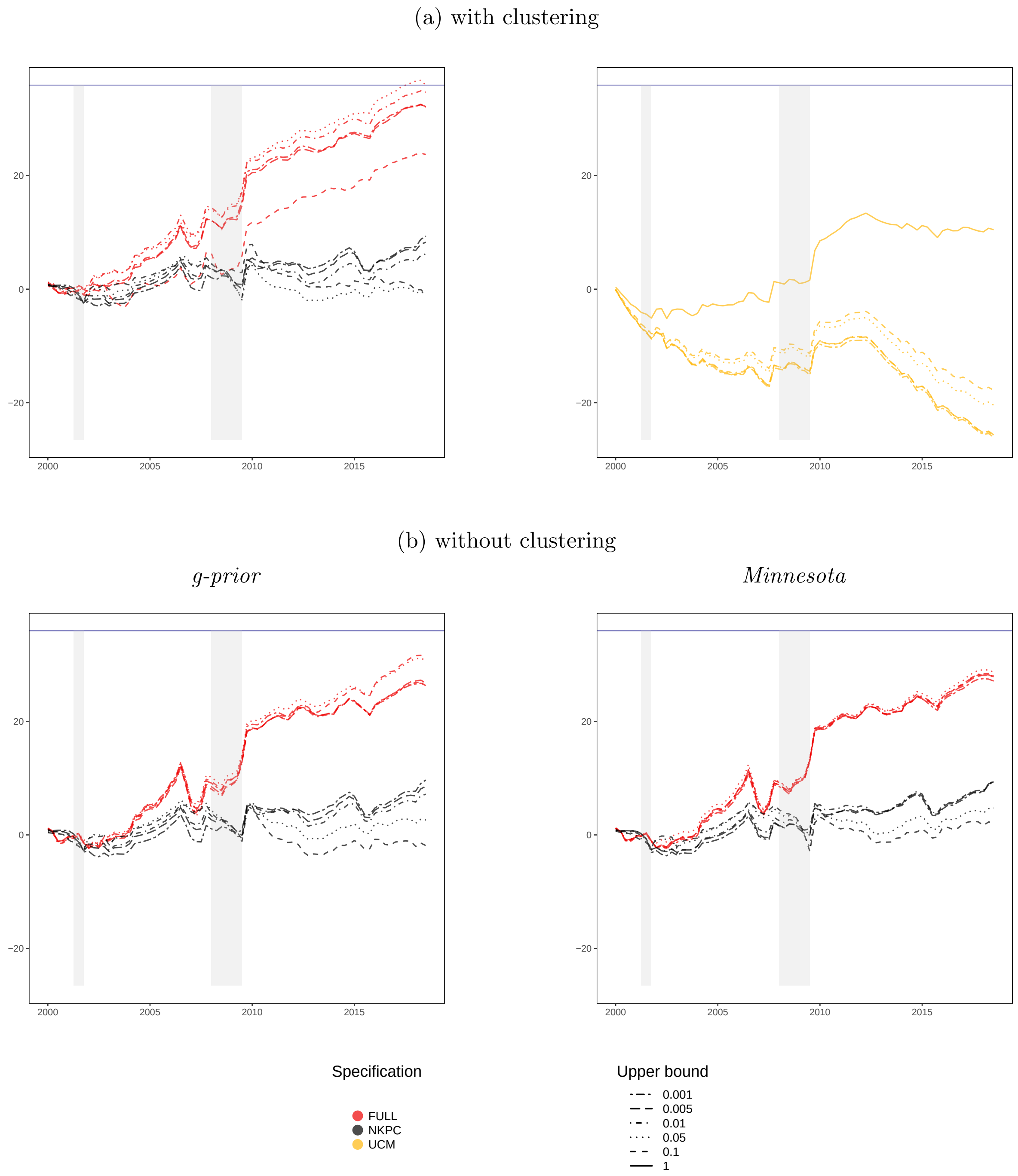
(b) without clustering

g-prior

Minnesota

Notes: The log predictive Bayes factors are cumulated over the hold-out sample. The blue lines refer to the maximum Bayes factor at the end of the hold-out sample. The gray shaded areas indicate the NBER recessions in the US.

Figure 6: Evolution of one-step-ahead log predictive Bayes factors relative to RW


(a) with clustering

(b) without clustering

g-prior

Minnesota

Notes: The log predictive Bayes factors are cumulated over the hold-out sample. The blue lines refer to the maximum Bayes factor at the end of the hold-out sample. The gray shaded areas indicate the NBER recessions in the US.

Figure 7: Evolution of four-step-ahead log predictive Bayes factors relative to RW