Data-based priors for vector autoregressions with drifting coefficients

Dimitris Korobilis*
University of Glasgow

January 2014

Abstract

This paper proposes full-Bayes priors for time-varying parameter vector autoregressions (TVP-VARs) which are more robust and objective than existing choices proposed in the literature. We formulate the priors in a way that they allow for straightforward posterior computation, they require minimal input by the user, and they result in shrinkage posterior representations, thus, making them appropriate for models of large dimensions. A comprehensive forecasting exercise involving TVP-VARs of different dimensions establishes the usefulness of the proposed approach.

Keywords: TVP-VAR, shrinkage, data-based prior, forecasting

JEL Classification: C11, C22, C32, C52, C53, C63, E17, E58

*Address: Department of Economics University of Glasgow, Gilbert Scott Building, University Avenue, Glasgow, G12 8QQ, United Kingdom. Tel: +44 (0)141 330 2950, e-mail: Dimitris.Korobilis@glasgow.ac.uk

1 Introduction

During the early stages of the development of vector autoregressive (VAR) models for modelling and forecasting macroeconomic data (Sims, 1989; Littermann, 1979), it has been recognized that such multivariate time series models can be subject to "curse of dimensionality" problems. Coefficients tend to increase exponentially with the number of endogenous variables or the number of lags. Therefore, it is no surprise that the early VAR literature was exclusively Bayesian, since prior distributions can provide a straightforward basis for imposing data-based shrinkage towards zero of irrelevant coefficients. For macroeconomists who typically work with monthly or quarterly data, saving degrees of freedom seems to be recognized as an issue of paramount importance in order to obtain reliable inference and communicate accurate forecasts to policy-makers.

Nevertheless, despite the vast development of Bayesian computational and shrinkage methods since the late 1970s, the econometrics literature has only had a few developments on the shrinkage front. The traditional "Minnesota-prior", an empirical-Bayes prior which is due to Littermann (1979) and co-authors (see, e.g. Doan, Litterman, and Sims, 1984), still dominates many applications of VAR models in economics. With the exception of the recent contribution of Giannone, Lenza and Primiceri (2012), selecting the infamous shrinkage factor of the Minnesota prior, i.e. the prior hyperparameter controlling shrinkage of the VAR coefficients, has been more of an art than exact science.

In this paper we develop a simple algorithm for selecting data-based priors for vector autoregressions with time-varying coefficients and stochastic volatility. We achieve this task by introducing full Bayes (hierarchical) priors which allow prior hyperparameters to be updated by the data using standard posterior expressions. Based on a simple reparametrization for state-space models (e.g. Frühwirth-Schnatter and Wagner, 2010; Belmonte, Koop and Korobilis, 2014) we show that we can develop simple prior structures which add minimal computational complexity to the standard TVP-VAR estimation algorithm used in the macroeconomic literature (e.g. the popular algorithm used in Primiceri, 2005). We show that such data-based priors also have the property of shrinking time-varying VAR coefficients towards zero or towards time-invariance, in which case we can save valuable degrees of freedom for estimation and forecasting.

The recent interest in macroeconomic VARs with drifting coefficients (Primiceri, 2005; Cogley and Sargent, 2005) which can better capture structural changes of interest, such as the Great Moderation or the Great Recession, usually create over-parametrization concerns, as model parameters are allowed to take a different value each and every time period. As mentioned in Korobilis (2013c) a quick review of the TVP-VAR literature reveals that, in light of these overparametrization and computational concerns, many authors fix the number of lags to two or three without implementing any lag/model length selection, thus overlooking one of the most important steps of statistical inference. Additionally, the researchers'

prior opinion about how much time-variation we may expect to find in all TVP-VAR coefficients usually breaks down to the choice of a single hyperparameter; see the detailed discussion in Primiceri (2005, Section 4.1). While such simplification makes prior selection easier and more convenient (only a single hyperparameter to select), there is a "downside risk" of following this approach in the flexible class of TVP-VAR models, since posterior quantities applied economists usually report (e.g. forecasts, impulse responses) can become very sensitive to selection of this hyperparameter[1].

The purpose of this paper is not to criticize the previous empirical macroeconomic literature; rather, its main intent is to propose a simple solution to an issue that can be daunting and confusing to the applied economist. The examples from the literature presented above show vividly that it is essential to establish methods for prior selection in TVP-VARs which allow a more automatic/default and robust determination of the prior structure, instead of relying solely on researchers' beliefs and experience. At the same time, since TVP-VARs are very popular with researchers in central banks and elsewhere, it is quite important to be careful enough to develop a simple posterior estimation framework that can be useful and meaningful to applied economists. That is, estimation should be both robust to subjective prior beliefs and reproducible, helping applied economists to communicate reliable results to their clients, managers, or the public. In that respect, this paper presents very simple tweaks and modifications to the very popular TVP-VAR framework of Primiceri (2005), thus allowing users to shrink coefficients in a data-based manner. Posterior expressions are based on conjugate priors and are, thus, straightforward to derive. Note that computational simplicity is a priority in this paper, so that the Gibbs sampler is preferred compared to other potentially more powerfull and elegant Markov Chain Monte Carlo (MCMC) and Sequential Monte Carlo (SMC) algorithms for prior selection.

The most important aspect of the proposed Bayesian estimator is that it decides (based on information in the likelihood) which coefficients are time-varying or not, as well as which coefficients are zero or not; see Belmonte, Koop and Korobilis (2014), Eisenstat, Chan and Strachan (2013) and Groen, Paap, and Ravazzolo (2012) for similar approaches. In general shrinkage estimators add some more bias (compared to unbiased estimators such as OLS in an unrestricted, time-invariant VAR), in order to massively reduce the variance of the estimated coefficients. We show that in the case of TVP-VARs, this kind of variance-bias tradeoff results in dramatic improvement in model fit, and large gains in forecast accuracy even when we evaluate only mean forecast (and not the whole predictive distribution). In a forecasting exercise using TVP-VARs with four to seven variables[2], we show

[1] Primiceri (2005) chooses this hyperparameter to be $(0.01)^2$, Canova and Gambetti (2009) choose a value of 0.0003, and Cogley and Sargent (2005) choose a value of $3.5e-4$. As with the traditional Minnesota prior, selection of this shrinkage coefficient is an art and not an exact science.

[2] Note that, following the finindings of Clark and Ravazzolo (forthcoming), all TVP-VAR models have stochastic volatility of the sort introduced in Primiceri (2005).

that improvement in forecast accuracy from shrinkage is massive even for the smaller four-variable systems. Additionally, confirming the results of D'Agostino, Gambetti and Giannone (2013) and Korobilis (2013c), forecasting gains from using time-varying instead of time-invariant VARs are substantial for "inherently nonlinear" variables such as GDP and inflation.

The next section serves as an introduction to the properties of time-varying parameter VARs, the overparametrization problem, as well as a formal and intuitive explanation of the source of the numerical (in)stability issues that can occur in this model class. We subsequently present formally the simple modifications explained above, which result in a more stable and robust approach to esitmating TVP-VARs. In Section 3 we begin with an empirical demonstration of the robustness of data-based shrinkage priors, and conclude with a comprehensive forecasting exercise. In Section 4 we summarize the findings of this paper and reflect on the importance of the proposed econometric methods.

2 Methodology

2.1 Understanding time-varying parameters VARs

Let y_t be a vector of n variables of interest for $t = 1, ..., T$. A p-lag time-varying coefficients VAR model for y_t takes the following form

$$y_t = c_t + B_{1,t} y_{t-1} + ... + B_{p,t} y_{t-p} + \varepsilon_t, \tag{1}$$

where c_t is a $n \times 1$ vector of intercepts, $B_{i,t}$, $i = 1, .., p$, are VAR coefficient matrices of dimensions $n \times n$, and $\varepsilon_t \sim N(0, \Omega_t)$ are shocks with Ω_t an $n \times n$ heteroskedastic VAR covariance matrix. The purpose of this paper is to consider shrinkage of the "large" vector of VAR coefficients $B_t = \left(c_t, B_{1,t}, ..., B_{p,t}\right)$. Macroeconomic data are usually highly correlated and subject to abrupt changes in volatility, thus, we assume that there should not be any benefits from shrinking Ω_t towards zero or towards time-invariance; see for instance the discussion in Clark (2009) and Sims and Zha (2006). In light of this assumption we allow Ω_t to follow the typical multivariate stochastic volatility specification of the sort introduced in Primiceri (2005); exact specification details are provided in the appendix.

It is trivial to show that the VAR above can be written as a linear regression of the form

$$\underset{(T \times n)}{Y} = \underset{(T \times Tk)}{X} \underset{(Tk \times n)}{\beta} + \underset{(T \times n)}{\varepsilon} \tag{2}$$

where $Y = (y_1, ..., y_T)'$, $B = (B_1', ..., B_T')'$, $\varepsilon = (\varepsilon_1, ..., \varepsilon_T)'$ and

$$X = \begin{bmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & x_T \end{bmatrix}.$$

where $k = np + 1$ and $x_t = \left[1, y_{t-1}', ..., y_{t-p}'\right]'$. In the formulation above the $Tk \times Tk$ matrix $(X'X)$ is of rank T and is not invertible, meaning that likelihood-based estimation cannot be used here. To deal with this issue, the standard practice in the literature (at least since the time of Cooley, 1971) is to impose more structure in the model by assuming that the VAR coefficients evolve as multivariate random walks. Following the notation of Primiceri (2005) and many others, the full time-varying coefficients VAR takes the form

$$y_t = z_t'\beta_t + \varepsilon_t \quad (3)$$
$$\beta_t = \beta_{t-1} + \eta_t, \quad (4)$$

where $z_t = I_n \otimes x_t$, and $\beta_t = \left(c_t', vec\left(B_{1,t}'\right)', ..., vec\left(B_{1,t}'\right)'\right)'$ subject to the initial condition $\beta_0 \sim N(b_0, P_0)$. In the equation above $\eta_t \sim N(0, Q)$ is a state disturbance term with system covariance matrix Q of dimensions $m \times m$, $m = k \times n$. This is the typical specification of a time-varying parameter VAR which has been embraced so much in macroeconomics; see Doan, Littermann and Sims (1984) for a review. We can show that now the information in X is enough to estimate β_t (which is now centered in β_t) and the covariance matrix Q.

Nevertheless, an important econometric issue arises with such a specification. Using repeated substitution it is easy to show that equation (4) implies that $var(\beta_t) = P_0 + tQ$. This result reveals that even before observing the data, the variance of β_t will tend to explode as $t \to \infty$, which is an obvious implication of the random walk assumption in (4). Thus, macroeconomists tend to choose modest values for P_0, and impose very tight priors on Q, thus regularizing the posterior variance of β_t. A typical implementation of prior selection in such models is to fit a VAR in a training-sample, and use the posterior moments to elicit the priors for the TVP-VAR in the estimation sample (see Primiceri, 2005).

In practical situations, though, application of training sample priors is not optimal. For many short time series (e.g. Euro-Area data) a training or hold-out sample might not be an option. Additionally, whilst a training sample prior is a very informative and subjective prior, the researcher might fail to fully understand its features and its implication for estimating the amount of time variation in the VAR. The size of the training sample or other selected hyperparameters[3] will

[3]See for example the hyperparameters relating to the prior of Q in Primiceri (2005, Section 4.1), and the similar discussion about "business as usual priors" in Cogley and Sargent (2005).

greatly affect posterior inference, so that most probably the researcher will have to set these ex-post (i.e. after observing the posterior) which implies a fair degree of data-minning. Lastly, the information in observed data might not be enough to estimate particular elements of β_t, and the corresponding eigenvalues of the conditional variance $var\left(\beta_t|y^t\right)$. For example, some eignevalues of $var\left(\beta_t|y^t\right)$ can become too high in conflict with prior information or data history (Kulhavý and Zarrop, 1993); or as we add more lags in the VAR and k grows large it is expected that the sample eigenvalues of $var\left(\beta_t|y^t\right)$ will diverge from the population eigenvalues (Stein, 1975).

2.2 A shrinkage representation of the TVP-VAR

In order to be able to specify data-based priors on the TVP-VAR which can provide shrinkage of coefficients, we first follow Belmonte, Koop and Korobilis (2013) and use the following reparametrization

$$\begin{aligned} y_t &= z_t'\alpha + z_t'\alpha_t + \varepsilon_t, \quad (5) \\ \alpha_t &= \alpha_{t-1} + \eta_t, \quad (6) \end{aligned}$$

with initial condition $\alpha_0 \sim N\left(0, 0 \times I_m\right)$, i.e. a point mass at zero. This TVP-VAR is observationally equivalent to the one presented in equations (3)-(4), where it holds that $\alpha \equiv \beta_0$ and $\alpha_t = \beta_t - \beta_0$. The difference is that, roughly speaking, now the VAR is decomposed into a part which is a constant parameter VAR ($z_t'\alpha$) and a part which describes how much additional time-variation we can add to the constant parameter VAR ($z_t'\alpha_t$). Therefore, an immediate advantage is that we can separately focus on the task of saving degrees of freedom by either shrinking the time-varying coefficients α_t, or the constant coefficients α, or both.

The first step in our analysis is to allow the initial condition $\alpha \equiv \beta_0$ to be updated from the data. Therefore, one now can explicitly define full-Bayes priors for α based on the Normal distribution, which are of the form

$$\begin{aligned} \alpha &\sim N\left(0, V\right) \\ V &= diag\left(\tau_1, ..., \tau_m\right) \quad (7) \\ \tau_i &\sim F\left(c_1, c_2\right),\ i = 1, .., m \end{aligned}$$

where V is a diagonal prior variance with element τ_i which has as a prior density F with parameters c_1 and c_2. In the next subsection we follow Korobilis (2013) and explicitly define appropriate forms for the density F. Here it suffices to note that once the τ_i's have their own prior then they will be updated from the data. Unlike standard practice where a common variance parameter (say τ such that $V = \tau \times I_m$) would be selected for all coefficients in the vector α, here we have a dedicated variance parameter τ_i for each coefficient α_i. If the posterior of τ_i is

"large" then α_i will have a fairly uninformative prior variance, however, if the posterior of τ_i is concentrated at zero then the implied prior of α_i is approximately $N(0,0)$ which implies that the posterior of α_i will also be a point mass at zero. Therefore, we have a flexible situation where different elements of α can be shrunk (or not) with a varying degree, depending on the absolute value of each τ_i.

The second step in our analysis is to update in a data-based manner (and also shrink) the time-varying part of the TVP-VAR, that is the coefficient α_t. Notice that if the covariance matrix Q of α_t is shrunk to zero, then the TVP-VAR is shrunk toward a constant parameter VAR, since from (6) we can infer that in this case $\alpha_t = \alpha_{t-1} = ... = \alpha_0$ and $\alpha_0 \sim N(0, 0 \times I_m) \equiv 0$. Therefore, shrinkage of Q towards zero also means shrinkage of a_t towards zero for all t, which in turn results in shrinkage of the TVP-VAR towards a time-invariant VAR. It becomes obvious then that we need to specify a data-based prior for Q in the spirit of the prior presented in equation (7). However, while a conjugate prior for Q is the inverse-Wishart distribution, there are no obvious hierarchical priors that can be used in this case. To be exact, Bouriga and Féron (2013), and references therein, do examine hierarchical shrinkage Wishart priors, nevertheless, their formulations are not conjugate and posterior computation can become demanding in high dimensions[4].

In contrast, in this paper we make the assumption that - in the spirit of the Minnesota prior tradition for constant parameter VARs - coefficients which are a-priori shrunk to zero based on the prior (7), should also be shrunk to time-invariance. Alternatively, coefficients which are important and initialized to a value different from zero are the most probable candidates for being time-varying. Additionally, coefficients on higher lags are less probable for being the main source of time-variation in a VAR. Here we consider two versions of this prior belief which - (ab)using terminology from the image processing literature - we term *soft thresholding* and *hard thresholding*, respectively. In soft thresholding we specify the prior

$$Q \sim iW\left(v, k_Q V\right), \tag{8}$$

while in hard thresholding we further restrict Q by assuming it is diagonal, that is, we use the prior

$$Q_{ii} \sim iG\left(v, k_Q V_{ii}\right), \tag{9}$$

for $i = 1, ..., m$. Both priors depend on the diagonal matrix V, which will be estimated by the data. The difference is that in equation (8) if a diagonal element

[4] An alternative way, which is quite attractive, is to use a reparametrization such as the one in Chen and Dunson (2003), Frühwirth-Schnatter and Wagner (2010), Belmonte, Koop and Korobilis (2011), and Eisenstat, Chan and Strachan (2013) who transfer Q in the measuement equation (5), that is, they transform Q to be a "VAR coefficient". This reparametrization has the disadvantage that interpretability of Q as a covariance matrix may be lost, and one has to rely on counterintuitive priors which do not guarantee that the $m \times m$ matrix Q is positive-definite; see Belmonte et al. (2013) for a discussion.

of V happens to be zero, then the respective element of the posterior of Q won't necessarily be exactly zero (but it might become small, i.e. shrunk). In the second prior, given that Q is diagonal, if $V_{ii} \equiv \tau_i$ becomes zero then the posterior of Q_{ii} will also be zero. Therefore, the second prior can result in massive reduction in estimation error (variance of posterior), with the downside risk of a very large bias, while the first prior provides a more balanced tradeoff between bias and variance. For medium-sized TVP-VARs the soft thresholding prior can be a reasonable choice, while for TVP-VARs of very large dimensions (30+ variables) the hard thresholding prior might be needed. Notice that the researcher has to choose hyperparameters v, k_Q[5] for both priors for Q. In Section 3.1 we explain how to choose these hyperparameters, and we show using real data that estimation of the TVP-VAR is not that sensitive to their choice.

To summarize, we have presented a framework for updating the time-varying VAR coefficients using a prior whose hyperparameters are updated from the data. We implement this by using a simple decomposition of the time-varying coefficients into the time-invariant and the time-varying part. Then we apply standard conjugate analysis and simplify selection of the prior hyperparameters into selection of a diagonal prior covariance matrix V. According to equation (7) V has its own prior distribution, reflecting our desire for this matrix to be updated by the data and, thus, admit a parametric posterior distribution of its own. In the next subsection we examine more carefully choices of distributions for V which have desirable properties in the context of VAR models.

2.3 Hyperprior distributions for VARs

In a univariate regression setting, Korobilis (2013a, 2013b) shows several possibilities for specifying hyperprior distributions based on a Normal conjugate prior. In our case we need to consider the fact that in a VAR model coefficients correspond to different equations and to lags of the dependent variables. In particular we use three types of hierarchical priors reflecting a varying degree of subjective beliefs about which coefficients should be shrunk. In the first case, we use a noninformative prior on each element τ_i, $i = 1, ..., m$, of the prior covariance matrix $V = diag(\tau_1, ..., \tau_m)$, in which case the data will update τ_i in a way where coefficients α_i will be a priori equally likely to be shrunk to zero or not. In the second and third cases, we allow the degree of shrinkage to increase as the number of lags increase. That way we can prevent overfitting and preserve degrees of freedom, using this idea from the Minnesota-prior literature; see Litterman (1979). The difference in these two last formulations we propose is that in one we implement a simple variant where the discounting of distant lags is chosen

[5]For the reader familiar with the analysis of Primiceri (2005), k_Q is a crucial hyperparameter causing all the sensitivity in the TVP-VAR. Here we maintain the same notation for this hyperparameter for direct reference to the sensititivity analysis of Primiceri (2005, Section 4).

subjectively from the researcher, while in the other the discounting of distant lags is more data-based.

Noninformative prior

The simplest prior choice for τ_i is to use a Jeffrey's prior of the form

$$\tau_i \propto \frac{1}{\tau_i}. \tag{10}$$

Minnesota-type prior with subjective choice

Here we use the following structure for τ_i

$$\tau_i \sim iG\left(\kappa_1, \kappa_2 \times \frac{1}{r^2}\right), \tag{11}$$

for lag lengths $r = 1, ..., p$. In this case autoregressive coefficients have prior variance τ_i which is updated from an inverse Gamma distribution with prior scale parameter that discounts more distant lags quadratically. This discount factor of $\frac{1}{r^2}$, $r = 1, ..., p$, is similar to the one used in the original Minnesota prior.

Objective Minnesota-type prior

In this case, we set $\tau_i = \lambda^{-1}\xi_r^{-1}$ and we select separate priors for λ and ξ_r where $r = 1, ..., p$ indexes the r-th lag. Following Bhattacharya and Dunson (2011) the prior we define takes the form

$$\begin{aligned} \tau_i &= \lambda^{-1}\xi_r^{-1} \\ \lambda^{-1} &\sim G\left(v/2, v/2\right), \\ \xi_r &= \textstyle\prod_{l=1}^{r} \delta_l, \\ \delta_1 &\sim G\left(\rho_1, 1\right), \\ \delta_l &\sim G\left(\rho_2, 1\right), \quad l \geq 2\ . \end{aligned} \tag{12}$$

In this case the "common" variance for all coefficients is λ. However, coefficients on lag k will have total variance $\lambda^{=1}\xi_r^{-1}$, where $\xi_r = \delta_1\delta_2...\delta_r$ is a multiplicative Gamma process. Under the condition $\rho_1 > 1$ the ξ_r's are stochastically increasing as the number of lags increase, in which case the total prior variance τ_i will be decreasing with the lag length of the VAR. Instead of the fixed geometrical decrease of the prior scale in the traditional Minnesota prior, this prior allows each lag-length to be penalized with different intensity and not necessarily linearly or exponentially. Nevertheless, the prior in (12) can still be considered a Minnesota-type prior since still it is the case that more distant lags will have larger penalty a-priori.

In order to distinguish between the three priors for V, we denote them as noninformative (NI), subjective Minnesota (SM), and objective Minnesota (OM), respectively. The reader should be careful not to confuse this terminology: the Minnesota prior is always a subjective prior; the term "objective Minnesota" is used for the third case in order to denote the fact that shrinkage of more distant lags is done in a more data-based, i.e. objective way.

3 Empirics

3.1 Understanding the effect of shrinkage priors

The first step in our analysis is to delve deeper into the exact effect of the priors presented above on estimation of the time-varying VAR coefficients by means of a real-data example. At this stage the interest does not lie on the exact prior distribution for $V = diag(\tau_1, ..., \tau_m)$, that is we are not interested to assess whether a noninformative, subjective Minnesota, or data-based Minnesota prior is better. This comparison is left as part of the empirical forecasting in the next subsection. Here our aim is to examine different patterns of shrinkage for the VAR coefficients β_t by restricting its state covariance Q through the priors (8) and (9), respectively. For that reason, for the whole analysis presented in this subsection we are using a simple 3-variable TVP-VAR on inflation (growth rate on GDP deflator), unemployment rate, and three-month interest rate estimated over the period 1960Q1-2013Q2. The TVP-VAR has only two lags and an intercept, as well as multivariate stochastic volatility. Therefore, this is the typical New Keynesian system that Primiceri (2005) and others have considered in their analysis.

First, we estimate this simple TVP-VAR using a data-based prior for V[6], and the soft thresholding on Q. Note that Q depends on hyperparameters v, k_Q which we need to choose, as well as the scale matrix V which is updated from the data. For the degrees of freedom hyperparameter v we set it to $v = m + 2$, so that the inverse Wishart prior is a well-defined distribution with a finite mean and variance[7]. Therefore, a large part of the prior sensitivity in the TVP-VAR system is expected to originate from selection of the prior scale matrix of $k_Q V$. In the case of Primiceri (2005) V is fixed to the OLS estimate from a constant parameter VAR in a pre-sample (training sample), so that selection of k_Q solely determines prior sensitivity. In this paper, V adapts in each MCMC iteration based on information from the data and is not fixed (it is a random variable with well-defined posterior distribution), therefore, selection of k_Q can be less harmful. Therefore our algorithm allows the prior scale $k_Q V$ in priors (8) and (9) to adapt in such a way that the data will almost always determine their combined optimal value, even though k_Q has to be chosen by the researcher.

In order to demonstrate this case, we plot in Figure 1 the posterior means of $\beta_t = \alpha + \alpha_t$ when using the hard thresholding (left panel) and the soft thresholding (right panel) priors for Q for two substantially different values of the hyperparameter k_Q namely $10,000$ and 1. This is a very extreme example which

[6] In particular we use the noninformative hyper-prior in equation (10), however, we could have come to the same conclusion by using priors (11) or (12).

[7] In the case of the hard thresholding prior, which is based on an inverse gamma distribution and not an inverse wishart, we still use the same formula, that is $v = 1 + 2$, since now Q_{ii} is univariate. This choice ensures that the prior mean and variance of this distribution exists (but we need $v > 3$ for the third and higher moments to exist).

is used to test the numerical stability of our algorithms. Posterior mean estimates based on the hard thresholding prior are strikingly similar, showing that this prior has minimal prior sensitivity. The soft thresholding prior, which assumes that Q is a full matrix with inverse Wishart prior, is considerably more sensitive to selection k_Q, however, it is numerically more stable than what would be expected under traditional priors: the reason why posterior mean estimates using the training sample prior are not plotted in Figure 1 is that estimation collapses in the case $k_Q = 10,000$ (draws of Q become non-positive definite, because Q explodes, leading also β_t to explode).

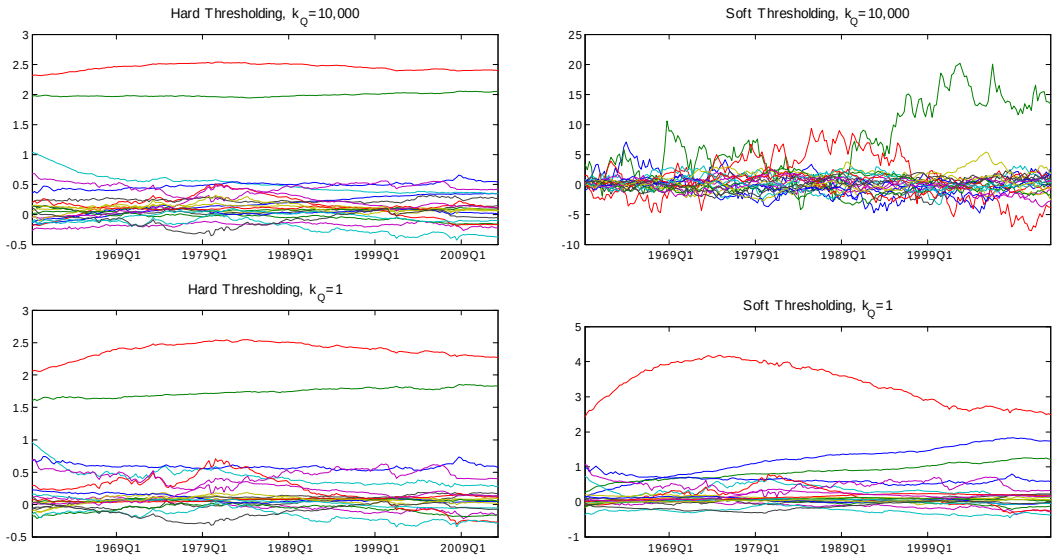

Figure 1. Posterior means of TVP-VAR(4) estimates from a typical three-variable system on prices, output and interest rate. The top panel shows coefficient estimates when $k_Q = 10,000$ for the hard and soft thresholding priors, respectively. The bottom panel shows how time-variation in the estimated coefficients is affected (or not) when we set $k_Q = 1$.

In Figure 2 we present again posterior means of the VAR coefficients for the hard thresholding (left panel) and soft thresholding (right panel) priors, but for tighter values of k_Q which favour more shrinkage toward time-invariance. We specifically compare two values, $k_Q = 0.01$ and $k_Q = 1e-6$. Such values are small enough so that eventually Q will be shrunk to zero, regardless of the posterior values of V, since the prior is so tight that it will dominate the likelihood. We can see clearly that as we move towards lower values, time variation in the posterior mean of the elements of β_t vanishes. Using the notation of the reparametrized TVP-VAR, as $k_Q \to 0$ then $\alpha_t \to 0$ or equivalently $\beta_t = \alpha + \alpha_t \to \alpha$. The hard thresholding prior, which is based on the assumption that Q is diagonal, forces shrinkage much faster than the soft thresholding prior. For the case $k_Q = 1e-6$

the hard thresholding case gives posterior means of VAR coefficients which are indistinguishable from posterior means of a time-invariant VAR.

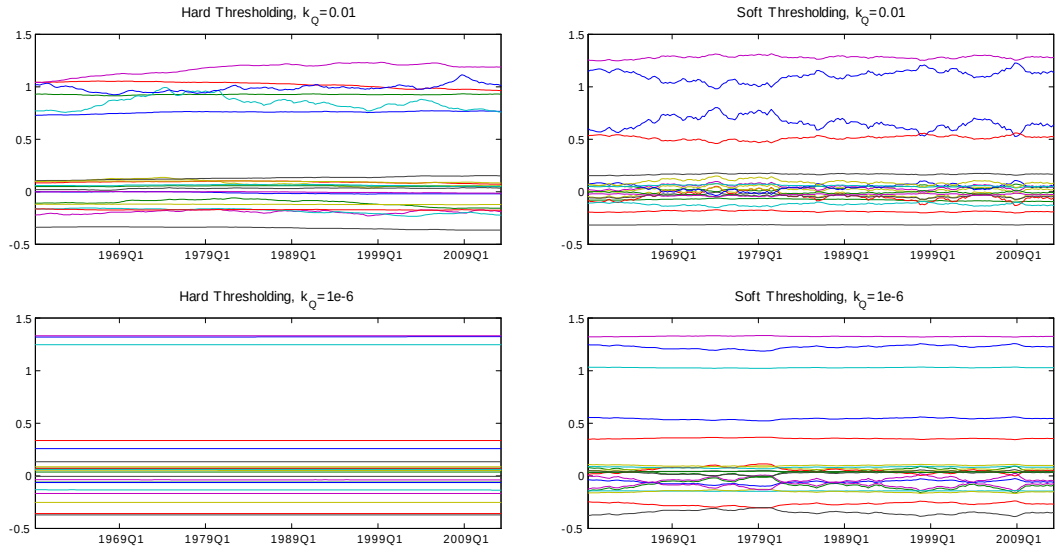

Figure 2. Posterior means of TVP-VAR(4) estimates from a typical three-variable system on prices, output and interest rate. In this case we want to assess the effect of smaller values of k_Q on the posterior mean of the TVP-VAR coefficients β_t.

The analysis above is based on a default prior for V (noninformative), which allows us to evaluate the effect of the soft and hard thresholding priors on Q. Similar analysis can be implemented for prior selection for V, in the case this parameter follows the subjective and objective Minnesota priors in equations (11) and (12), respectively. For instance, tighter choice of hyperparameters on the subjective and objective Minnesota priors for V would not only result in shrinkage of β_t towards time invariance (as it is the case in the bottom panel of Figure 2), but would also force coefficients to be exactly equal to zero and, thus, irrelevant for forecasting. For the sake of brevity such analysis is not implemented here. We rather focus in the next subsection on the ability of the different priors presented in this paper to prevent overparametrization and provide superior TVP-VAR forecasts.

3.2 Forecast evaluation

In this exercise we use quarterly TVP-VARs in order to evaluate the forecasting performance of the priors proposed in this paper. In particular we consider TVP-VARs ranging from four to seven variables, in order to better understand how shrinkage may benefit forecasting as the VAR size increases. Additionally, by

comparing TVP-VARs of different sizes we should be able to understand whether it is more important to consider nonlinearity in the VAR as opposed to adding more information by increasing the number of dependent variables. Table 1 describes all the variables used and how these define our four to seven-variable systems. We estimate all TVP-VARs in this section using data from 1960Q1, however, all variables in Table 1 are collected since 1948Q1 in order to be able to define a training sample 1948Q1-1959Q4 when using the Primiceri (2005) prior. Variables which are not already in rates are transformed into non-annualized quarter-on-quarter growth rates by taking first log-differences multiplied by 100. Monthly variables are transformed to quarterly by taking their average over the quarter.

Table 1. Series used in the forecasting exercise

No	Mnemonic	Tranformation	Description
SERIES USED IN 4-VARIABLE VARS			
1	GDPC1	$\Delta \ln$	real GDP
2	GDPDEFL	$\Delta \ln$	GDP deflator
3	TB3MS	level	3-month interest rate
4	UNRATE	level	unemployment rate
ADDITIONAL SERIES USED IN 5-VARIABLE VARS			
5	PAYEMS	$\Delta \ln$	Total employment
ADDITIONAL SERIES USED IN 6-VARIABLE VARS			
6	PPIACO	$\Delta \ln$	commodity prices
ADDITIONAL SERIES USED IN 7-VARIABLE VARS			
7	DGOERA3Q086SBEA	$\Delta \ln$	energy prices

Note: All 7 series are downloaded from Federal Reserve Economic Data (FRED), and are available at http://research.stlouisfed.org/fred2/.

We estimate several TVP-VARs recursively over the period 1960Q1-2013Q2. All models have four lags and an intercept. In particular, we consider the following prior specifications which define our forecasting TVP-VARs:

1. Training sample prior: We use the original TVP-VAR where $\beta_0 \sim N(\beta_{OLS}, V_{OLS})$, and $Q \sim iW\left(k+2, k_Q^2 \times V_{OLS}\right)$ where β_{OLS} and V_{OLS} are the mean and variance of the OLS estimates from a time-invariant VAR estimated over the sample 1948Q1-1959Q4.

2. Traditional Minnesota prior: We use the original TVP-VAR where $\beta_0 \sim N(\beta_{MIN}, V_{MIN})$, and $Q \sim iW\left(k+2, k_Q^2 \times V_{MIN}\right)$ where β_{MIN} and V_{MIN} are prior hyperparameters of the Minnesota prior (see appendix for exact definition).

3. Soft thresholding: Here we use the reparametrized TVP-VAR and we define $Q \sim iW\left(k+2, k_Q \times V\right)$ where $V = diag(\tau_1, ..., \tau_m)$. We allow for three

different ways to define a prior for V, i.e. noninformative (NI), subjective Minnesota (SM), and objective Minnesota priors (OM); see equations (10) - (12).

4. Hard thresholding: Here we use the reparametrized TVP-VAR and we define $Q_{ii} \sim iG\left(1+2, k_Q \times V_{ii}\right)$, $i = 1, ..., m$, where $V = diag\left(\tau_1, ..., \tau_m\right)$. We allow for three different ways to define a prior for V, i.e. noninformative, subjective Minnesota, and objective Minnesota priors; see equations (10) - (12).

Subsequently, we have in total one TVP-VAR with training sample, one with Minnesota prior, three with hard thresholding, and three with soft thresholding. We also have models with four, five, six, and seven variables. This gives a total of 32 TVP-VARs to be evaluated in this forecasting exercise.

All prior hyperparameters for the time-varying coefficients and covariances used in this forecasting exercise are described in detail in the appendix. Note that selection of k_Q is of paramount importance for all four models in 1-4. In an ideal world selection of k_Q could be based on a grid search, where the optimal value could be selected based on past forecast performance or insample fit (see Koop and Korobilis, 2013, for an illustration). However, when using MCMC such procedures have to be precluded from our analysis due to computational constraints. Therefore, and in order to avoid data-mining we set $k_Q = 0.01$ for all TVP-VAR sizes even though in an ideal world k_Q should decrease as the TVP-VAR size increases; see again Koop and Korobilis (2013). While other choices of k_Q might give quite different forecasting results, given computational restrictions (this is the restriction that any applied economist estimating a TVP-VAR would face), the purpose of this exercise is to evaluate how different priors perform based on an arbitrary choice of prior hyperparameters. The insample results of the previous subsection suggest that the data-based priors proposed in this paper are more robust to suboptimal choices of k_Q, which is something that we should expect to verify out-of-sample.

We estimate the reparametrized TVP-VAR in equations (5) and (6). The original coefficient β_t is recovered using the simple formula $\beta_t = \alpha + \alpha_t$, and forecasting at time t is implemented iteratively using the typical TVP-VAR specification in equations (3)-(4), by assuming for simplicity that the VAR parameters are the ones estimated at the last in-sample observation[8], that is $\beta_{t+h|t} = \beta_t$, $\Omega_{t+h|t} = \Omega_t$. By rearraging the elements of the parameter matrices, we can obtain the VAR in usual companion form as

$$\mathbf{y}_t = \mathbf{c} + \mathbf{B} y_{t-1} + \boldsymbol{\varepsilon}_t$$

[8]This is a simplifying assumption done in the literature with TVP-VAR models, whether they are used for structural analysis or forecasting (see for instance Cogley and Sargent, 2005 and Korobilis, 2013). One can simulate out-of-sample coefficients $\beta_{t+h|t}$ from the random walk evolution equation (4), however, this approach can result in larger forecast errors.

where $\mathbf{y}_t = \left(y_t', ..., y_{t-p+1}'\right)'$, $\boldsymbol{\varepsilon}_t = (\varepsilon_t', 0, ..., 0)'$, $\mathbf{c} = (c_t', 0, ..., 0)'$ and

$$\mathbf{B} = \begin{bmatrix} B_{1,t}...B_{p-1,t} & B_{p,t} \\ I_{n(p-1)} & 0_{n(p-1)\times n} \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_t & 0_{n\times n(p-1)} \\ 0_{n(p-1)\times n(p-1)} \end{bmatrix}.$$

Iterated h-step ahead forecasts can be obtained using the formula

$$E_t\left(\mathbf{y}_{t+h}\right) = \sum_{i=0}^{h-1} \mathbf{B}^i \mathbf{c} + \mathbf{B}^h y_{t-1} \tag{13}$$

The forecasting exercise is performed in pseudo real time. For the daily data, 60% of the observations are used to estimate the model initially, i.e. $T_0 = 0.7T$, and forecasts $\widehat{y}_{T_0+h|T_0}$ are calculated for $h = 1, 4$ and 8 quarters ahead. Then one quarterly observation is added and the sample becomes $T_1 = T_0 + 1$, each model is re-estimated, and the coefficients β_{T_1} are used in order to obtain the forecasts of $\widehat{y}_{T_0+1+h|T_0+1}$. This is done until the whole sample is exhausted, i.e. when $T_n = T$. Notice that evaluating forecasts using only 40% of the sample is not an ideal situation, however, one has to consider the computational demands of estimating TVP-VARs of that size. Additionally, the out-of-sample period includes the turbulent '00s where forecasting episodes, such as the Great Recession, are of extreme interest to policy-makers.

The results are evaluated using the Mean Squared Forecast Error (MSFE)[9]. When forecasting variable i at horizon h, these statistics are defined as

$$MSFE_{i,h} = \frac{1}{n-h} \sum_{t=T_0}^{T-h} \left(\widehat{y}_{i,t+h|t} - y_{i,t+h}^o\right)^2$$

where $y_{i,t+h}^o$ are the observed out-of-sample values of $y_{i,t+h}$. The MSFE statistic is presented relative to a benchmark 4-variable VAR(4) with constant coefficients estimated using least squares, so that values higher (lower) than one signify worse (better) performance of a specific TVP-VAR model compared to the VAR(4). Evaluation of the relative MSFE is based only on the four variables which are common in all TVP-VARs, namely real GDP growth (GDPC1), price inflation (GDPDEFL), short-term interest rate (TB3MS), and the unemployment rate (UNRATE).

[9]It is interesting to note here that according to other metrics which evaluate the whole preditive distribution (e.g. mean predictive likelihood; see Koop and Korobilis, 2013) the models with shrinkage dominate massively the training sample and Minnesota priors for the heavily parametrized TVP-VARs we consider in this exercise. This is no surprise as without sufficient shrinkage, posterior variances of TVP-VAR estimates are quite large and they eventually feed in the posterior predictive distribution resulting in high forecast uncertainty, especially in more distant forecast horizons.

Table 2. Relative MSFE of TVP-VAR models with various priors

	GDPC1			GDPDEFL			TB3MS			UNRATE		
	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8
VAR(4)	0.624	0.768	0.729	0.057	0.126	0.166	0.227	1.916	5.315	0.087	1.378	3.941
						4-variable TVP-VARs:						
	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8
TS	1.06	1.50	4.06	0.77	0.81	1.01	0.89	1.31	1.86	0.75	1.53	4.56
Minnesota	0.97	1.17	1.69	0.76	0.59	0.63	0.90	1.29	1.51	0.70	1.13	1.87
SOFT THRESHOLDING												
NI	0.91	1.33	1.82	0.70	0.95	1.72	0.63	1.01	1.41	0.72	1.10	2.09
SM	0.91	1.04	1.15	0.84	0.85	1.05	0.63	1.06	1.29	0.66	1.07	1.90
OM	0.91	1.02	1.16	0.80	0.75	1.02	0.60	1.05	1.18	0.60	1.03	1.71
HARD THRESHOLDING												
NI	0.87	1.23	1.77	0.72	0.76	0.85	0.85	1.44	1.85	0.66	1.06	2.29
SM	0.87	1.45	2.85	0.73	0.71	0.96	0.86	1.36	1.85	0.59	1.34	3.74
OM	0.88	1.20	1.66	0.67	0.55	0.74	0.87	1.18	1.51	0.48	1.12	2.59
						5-variable TVP-VARs:						
	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8
TS	0.93	1.65	3.88	0.89	0.65	0.69	0.92	1.55	2.07	0.76	1.71	4.62
Minnesota	0.98	1.36	3.08	0.79	0.76	0.77	0.91	1.64	2.34	0.79	1.83	4.95
SOFT THRESHOLDING												
NI	0.96	1.14	1.66	0.70	0.66	0.83	0.68	1.27	1.67	0.58	0.97	2.03
SM	0.82	1.28	1.98	0.83	0.64	0.71	0.79	1.35	1.79	0.46	1.12	2.75
OM	0.76	1.18	1.73	0.76	0.65	0.73	0.79	1.28	1.73	0.42	1.03	2.36
HARD THRESHOLDING												
NI	0.85	1.41	2.26	0.89	0.82	0.94	0.65	1.05	1.46	0.70	0.98	1.82
SM	0.85	1.24	2.46	0.71	0.60	0.85	0.81	1.53	2.36	0.63	1.32	3.95
OM	0.82	1.21	2.33	0.79	0.71	0.87	0.65	1.29	1.73	0.69	1.14	2.87

Table 2. continued

	GDPC1			GDPDEFL			TB3MS			UNRATE		
						6-variable TVP-VARs:						
	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8
TS	1.31	1.34	1.51	1.81	0.89	1.41	1.15	1.32	1.35	1.33	1.02	1.60
Minnesota	1.22	1.17	1.15	1.33	0.62	0.89	1.10	1.26	1.24	1.68	1.23	2.98
SOFT THRESHOLDING												
NI	1.01	1.07	1.35	1.55	0.61	0.82	0.89	1.16	1.44	0.69	0.99	1.57
SM	0.89	0.99	0.97	1.03	0.72	0.74	0.99	1.19	1.37	1.00	0.99	1.59
OM	0.90	0.99	0.85	0.72	0.63	0.76	0.92	1.15	1.40	0.80	0.96	1.58
HARD THRESHOLDING												
NI	0.95	0.98	0.96	0.82	0.55	0.75	0.59	0.96	1.20	0.87	1.13	1.92
SM	1.02	1.37	1.81	1.43	0.85	1.01	1.41	1.89	2.33	0.73	0.83	1.84
OM	0.92	0.59	0.86	0.37	0.68	0.87	0.82	1.41	1.76	0.78	0.95	1.87
						7-variable TVP-VARs:						
	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8
TS	1.27	1.64	1.66	1.07	0.89	1.08	1.55	1.36	1.34	1.38	1.21	2.59
Minnesota	1.23	1.71	1.30	1.79	0.87	0.71	1.09	1.21	1.25	0.67	1.03	3.14
SOFT THRESHOLDING												
NI	1.19	1.36	1.40	1.01	0.55	0.55	0.65	0.97	1.35	0.58	0.60	0.99
SM	0.82	1.05	0.92	1.06	0.81	0.93	1.06	1.08	1.13	0.83	0.89	1.57
OM	0.81	0.98	0.94	0.64	0.61	0.50	1.03	0.97	1.17	0.50	0.69	1.21
HARD THRESHOLDING												
NI	1.03	1.02	1.02	1.13	0.59	0.73	0.95	1.18	1.39	0.82	0.85	1.24
SM	0.91	1.09	0.96	1.20	0.66	1.09	1.36	1.47	1.39	0.86	0.82	1.49
OM	0.90	1.05	0.92	0.86	0.59	0.87	1.08	1.32	1.35	0.80	0.63	1.36

Table entries for the 4-variable VAR(4) model with OLS are absolute MSFE values. For all other models ratios of MSFEs relative to the one for the VAR(4) are reported. Values higher than one signify worse performance compared to the base VAR(4) model. Variable mnemonics are the ones defined in the main text, and follow Federal Reserve Economic Data (FRED).

Table 2 presents the results of this forecasting exercise. There are several stylized facts we can establish from such a table. First, TVP-VAR models improve massively upon simple VARs in forecasts of inflation at all horizons. There is also improvement in short-run forecasts of GDP and unemployment, implying that there are some nonlinearities in the short-run while we are better off in the long-run with the simple linear VAR model. For the short-term interest rate the VAR model seems to be hard to beat at all horizons. These results also comply with the findings of Korobilis (2013c) for UK data.

Second, while in constant parameter VARs we expect larger systems with more information to perform well, when considering nonlinearities in the form of time-varying coefficients and stochastic volatility there is no evidence that larger VAR models are better than smaller VAR models. There are two interpretations to this result. On the one hand, nonlinearities captured by the TVP-VAR might be the main reason for forecast improvement over the simple VAR(4) estimated with OLS. Therefore, larger systems add little to the forecasting ability of a smaller TVP-VAR. On the other hand, one can argue that larger TVP-VARs have the potential of improving upon smaller TVP-VARs conditional on achieving the right amount of shrinkage. In this paper, due to computational constrains we have used "default" prior hyperparameters for all VAR sizes, while careful application of shrinkage would require us to estimate the optimal shrinkage hyperparameter, k_Q, for each VAR size. Koop and Korobilis (2013) implement such an approach, however, estimation of their TVP-VARs is based on fast and efficient approximations and not computationally expensive MCMC methods.

Third, data-based and semi-automatic shrinkage of the sort presented in this paper clearly outperforms the more subjective shrinkage schemes based on training sample and Minnesota priors. While it is hard to establish in detail which specific data-based shrinkage method is the best - different shrinkage priors perform better for different VAR sizes and for different variables - in general all data-based shrinkage priors seem to benefit particularly at longer forecast horizons. The single exception is the impressive performance of the Minnesota prior in forecasting inflation using the 4-variable TVP-VAR. The only pattern we can observe among all six data-based priors is that soft thresholding (regardless of the prior choice for V) is performing better than hard thresholding. This is not surprising as hard thresholding picks up more restrictions and tends to shrink parameters to zero or time-invariance at a much faster rate. This results in faster increase in the bias of posterior mean estimates (but also faster decrease in variance of posterior estimates) and, consequently, increasing mean squared forecast errors. This difference in performance is noticeable in smaller TVP-VARs, however, it almost vanishes as the VAR size increases and the need for faster rates of shrinkage is more imperative.

4 Conclusions

We have presented a framework for dealing efficiently with concerns of overfitting and overparametrization in time-varying parameter VARs. By means of data-based shrinkage priors we have managed to reduce the huge parameter space associated with coefficients which change value each time period, whilst at the same time computation has remained simple and feasible. We have examined several prior specifications which address varying needs for shrinkage in TVP-VAR models. Therefore, this paper has proposed an integrated treatment of TVP-VARs with automatic prior choices that are updated from the likelihood (data). Additionally, the algorithms presented here add minimal computation time to existing MCMC algorithms and result in improved numerical stability (through shrinkage we can prevent time-varying coefficients from exploding, and associated covariance matrices from becoming non-invertible).

Although some of the priors we have introduced seem to be doing better than others in the specific empirical exercise presented in this paper, it is expected that different priors will suit better in different set-ups. In general, use of shrinkage priors is compelling in very large macroeconomic VAR models, or small models with only a few time-series observations (e.g. VARs for Euro-Area data). By introducing some bias compared to unrestricted estimates (e.g. OLS in a constant parameter VAR), shrinkage priors can result in posterior standard deviations of all model parameters which are much smaller. Therefore, as long as TVP-VAR models are becoming more popular over time in central banks for forecasting and monitoring macroeconomic variables, then settings such as the one proposed in this paper have the potential to offer more accurate and robust statistical inference.

References

[1] Belmonte, M. A. G., Koop, G. and Korobilis, D. (2014). Hierarchical shrinkage in time-varying parameter models. Journal of Forecasting, forthcoming.

[2] Bhattacharya, A. and Dunson, D. B. (2011). Sparse Bayesian infinite factor models. Biometrika 98, pp. 291-306.

[3] Bouriga, M. and Féron, O. (2013). Estimation of covariance matrices based on hierarchical inverse-Wishart priors. Journal of Statistical Planning and Inference 143, pp. 795-808.

[4] Chen, Z. and Dunson, D. (2003). Random effects selection in linear mixed models. Biometrics 59 (4), pp. 762-769.

[5] Clark, T. and Ravazzolo, F. (forthcoming). The macroeconomic forecasting performance of autoregressive models with alternative specifications of time-varying volatility. Journal of Applied Econometrics.

[6] Cooley, T. F. (1971). Estimation in the presence of sequential parameter variation. Ph.D Thesis. Department of Economics, University of Pennsylvania.

[7] D'Agostino, A., Gambetti, L. and Giannone, D. (2013). Macroeconomic forecasting and structural change. Journal of Applied Econometrics 28, pp. 82-101.

[8] Doan, T., Litterman, R. and Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. Econometric Reviews, 3, pp. 1-100.

[9] Eisenstat, E., Chan, J. C. C. and Strachan, R. (2013). Stochastic model specification search for time-varying parameter VARs. manuscript.

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

[11] Giannone, D., Lenza, M. and Primiceri, G. E. (2012). Prior selection for vector autoregressions. Working Paper Series 1494, European Central Bank.

[12] Groen, J. J. J., Paap, R. and Ravazzolo, F. (2012). Real-time inflation forecasting in a changing world. Journal of Business and Economic Statistics, 31, pp. 29-44.

[13] Koop, G. and Korobilis, D. (2010). Bayesian multivariate time series methods for empirical macroeconomics. Foundations and Trends in Econometrics 3, pp. 267-358.

[14] Koop, G. and Korobilis, D. (2013). Large Time-Varying Parameter VARs. Journal of Econometrics 177, pp. 185-198.

[15] Korobilis, D. (2013a). Bayesian forecasting with highly correlated predictors. Economics Letters 118, pp. 148-150.

[16] Korobilis, D. (2013b). Hierarchical shrinkage priors for dynamic regressions with many predictors. International Journal of Forecasting 29, pp. 43-59.

[17] Korobilis, D. (2013c). VAR forecasting using Bayesian variable selection. Journal of Applied Econometrics 28, pp. 204-230.

[18] Kulhavý, R. and Zarrop, M. B. (1993). On a general concept of forgetting. International Journal of Control 58, pp. 905-924.

[19] Litterman, R. (1979). Techniques of forecasting using vector autoregressions. Federal Reserve Bank of Minneapolis Working Paper 115.

[20] Sims, C. (1989). A nine variable probabilistic macroeconomic forecasting model. Federal Reserve Bank of Minneapolis Discussion paper no. 14.

[21] Stein, C. (1975). Estimation of a covariance matrix. In: Rietz Lecture, 39th Annual Meeting IMS. Atlanta, Georgia.

A Technical Appendix

A.1 A Gibbs sampler for the TVP-VAR with data-based priors

In this Appendix we describe the hierarchical shrinkage priors for time-varying parameter VARs, and the associated modifications in the Markov Chain Monte Carlo (MCMC) algorithm which are needed in order to sample from the posterior distributions of all parameters. Consider the noncentered parametrization of the VAR

$$y_t = z_t'\alpha + z_t'\alpha_t + \varepsilon_t, \tag{A.1}$$
$$\alpha_t = \alpha_{t-1} + \eta_t, \tag{A.2}$$

subject to the initial condition $\alpha_0 \sim N(0,0) \equiv \delta_\alpha(0)$, with $\delta_\alpha(0)$ denoting the Dirac delta function for coefficient α which is concetrated at zero. In this case the original coefficient β_t can be recovered as $\beta_t = \alpha + \alpha_t = \beta_0 + \alpha_t$, that is we have effectively split the original coefficient into a "initial condition part" and a "time-varying part".

We use the following prior on the $k \times 1$ vector α:

$$\begin{aligned} \alpha &\sim N(0, V), \\ V &= diag(\tau_1, ..., \tau_m), \\ & \quad \tau_i \begin{cases} = 2, & \text{for intercepts} \\ \sim iG\left(\kappa_1, 1 \times \left(\frac{1}{r^2}\right)\right), & \text{for coefficients on } r\text{-th lag} \end{cases}, \end{aligned}$$

for $i = 1, ..., k$. The prior variance of $\alpha = \beta_0$ is determined by the matrix V with diagonal element τ_i, where each diagonal element is defined to have an inverse-gamma prior[10]. The specific inverse-gamma prior is set-up in order to favour more shrinkage as the number of lags increases.

We define two cases of shrinkage priors for Q (note: Q is the variance of η_t). In the hard thresholding case we use

$$Q_{ii} \sim iG\left(8, \left(\frac{1}{r^2}\right) \times \tau_i^2\right),$$

and in the soft thresholding case we use

$$Q \sim iW(k+1, 1e-3 \times V).$$

For the VAR covariance matrix Ω_t we follow Primiceri (2005) and define the decomposition

$$\Omega_t = L_t^{-1} D_t D_t' L_t^{-1\prime},$$

[10]Note that the noninformative prior in equation (10) can be obtained in the limit using an $iG(0,0)$ prior. The objective Minnesota prior case in equation (12) of the main text, is explained in the following subsection.

where L_t is a lower triangular matrix with ones on the main diagonal, and D_t is a diagonal matrix. Denote by l_t the $n(n-1)/2$ vector of non-zero and non-one elements of L_t, and by d_t the vector consisting of all n diagonal elements in D_t. Then time variation in l_t and d_t is of the form

$$\begin{aligned} l_t &= l_{t-1} + v_t, \\ \log d_t &= \log d_{t-1} + w_t, \end{aligned}$$

where $v_t \sim N(0, S)$ and $w_t \sim N(0, W)$. Given draws of α and α_t, it is easy to show that the posterior conditional distributions for l_t and $\log d_t$ are exactly the ones given in the appendix of Primiceri (2005); see also Koop and Korobilis (2010). Now it suffices to show that for the initial condition on l_t and d_t we set the relatively noninformative values

$$\begin{aligned} l_0 &\sim N\left(0, 10 \times I_{n(n-1)/2}\right), \\ \log d_0 &\sim N(0, 10 \times I_n). \end{aligned}$$

Finally, the state covariances S and W have priors

$$\begin{aligned} S_j &\sim iW\left(j+1, 0.01 \times I_j\right), \\ W_{i,i} &\sim iG(8, 0.1), \end{aligned}$$

for $j = 1, ..., n-1$ and $i = 1, ..., n$, where the reader should note that S is block-diagonal and W is diagonal for estimation reasons; once again the appendix of Primiceri (2005) gives the exact details.

The conditional posteriors[11] of all coefficients are:

1. Sample α from

$$\alpha|- \sim N\left(\overline{\alpha}, \overline{V}_\alpha\right),$$

where $\overline{\alpha} = \overline{V}_\alpha\left(\sum_t z_t \Omega_t^{-1} y_t^*\right)$, $\overline{V}_\alpha = \left(V + \sum_t z_t \Omega_t^{-1} z_t\right)$, $V = diag(\tau_1, ..., \tau_m)$ and $y_t^* = y_t - z_t' \alpha_t$.

2. Sample τ_i (only for those i that do not refer to intercept coefficients) from

$$\tau_i|- \sim iG(\rho_{1i}, \rho_{2i}),$$

where $\rho_{1i} = 0.5 + \kappa_1$, and $\rho_{2i} = 0.5 \times (\alpha_i)^2 + \kappa_2$. When using the Jeffrey's noninformative prior on τ_i then the posterior has the same form with the exception that $\kappa_1 = \kappa_2 = 0$. When using the data-based Minnesota prior then see the next subsection.

[11] Here we use notation where $\theta|-$ is the conditional posterior distribution of θ conditional on the data and draws of all other parameters (from the current or the previous iteration of the Gibbs sampler, depending on the sampling order).

3. Sample $\alpha_t|-$ using Carter and Kohn (1994) with initial condition $\alpha_0 \sim N(0,0)$ and data $(\widetilde{y}_t, z_t)$ where $\widetilde{y}_t = y_t - z_t'\alpha$.

4. Sample $Q|-$ using standard expressions from Inverse Wishart (or inverse Gamma, in the case of hard thresholding) posteriors; see Koop (2003).

5. Sample the covariance matrix conditional on all other parameters as in Primiceri (2005). To see that there are no differences, once we have draws of α_t and α we can obtain a draw of β_t simply by using the identity $\beta_t = \alpha + \alpha_t$, in which case we can estimate Ω_t using the original TVP-VAR in equations (3) and (4).

A.2 Sampling of τ_i under the data-based Minnesota prior

When using the data-based Minnesota prior we just have to sample a few more hyper-parameters we have introduced from their conditional posteriors. We remind that the full prior is of the form

$$\begin{aligned}
\alpha &\sim N(0, V), \\
V &= diag(\tau_1, ..., \tau_m), \\
\tau_i &= \lambda_i^{-1}\xi_{ir}^{-1}, \quad r = 1, ..., p \\
\lambda_i^{-1} &\sim G(v/2, v/2), \\
\xi_r &= \prod_{l=1}^{r} \delta_l, \\
\delta_1 &\sim G(\rho_1, 1), \\
\delta_l &\sim G(\rho_2, 1), \quad l \geq 2 .
\end{aligned}$$

Step 1 of the Gibbs sampler algorithm presented above (sampling of α conditional on τ_i) is not affected by the new hierarchical layers we have added in order to sample τ_i. What we need to adapt is Step 2 which now becomes

2*. (a) Sample λ^{-1} from

$$\lambda_i^{-1}|- \sim Ga\left(\frac{v+1}{2}, \frac{v+\xi_r\alpha_i^2}{2}\right).$$

(b) Sample δ_1 from

$$\delta_1|- \sim G\left(\rho_1 + \frac{m}{2}, 1 + \frac{1}{2}\left[\sum_{r=1}^{p} \xi_r^{(1)} \left(\sum_{j=1}^{n^2} \lambda_{jr}^{-1}\alpha_{jr}^2\right)\right]\right),$$

where $\alpha_r^2 = \left(\alpha_{1r}^2, ..., \alpha_{n^2r}^2\right)$ denotes the n^2 elements of α which correspond to coefficients on the r-th VAR lag, $r = 1, ..., p$, and $\xi_r^{(1)} = \prod_{c=2}^{r} \delta_c$

(c) Sample δ_l for $l \geq 2$ from

$$\delta_l|- \sim G\left(\rho_2 + \frac{n \times n}{2}\left(r - l + 1\right), 1 + \frac{1}{2}\left[\sum_{r=1}^{p} \xi_r^{(l)} \left(\sum_{j=1}^{n^2} \lambda_{jr}^{-1} \alpha_{jr}^2\right)\right]\right),$$

where $\xi_r^{(l)} = \prod_{c=1, c\neq l}^{r} \delta_c$

For more details the reader is referred to Bhattacharya and Dunson (2011), who developed this shrinkage prior for the case of a factor model with unkown number of factors.

A.3 The traditional Minnesota (Litterman, 1979) prior

One of the benchmark priors used in this paper, other than the training sample prior of Primiceri (2005), is the traditional Minnesota prior of Litterman (1979) adapted for the time-varying parameter VAR. This adaption of the Minnesota prior is not new, since very early Doan, Litterman and Sims (1984) have shown how the Minnesota prior can provide shrinkage both in constant parameter and time-varying parameter VARs. We apply the Minnesota prior in the original TVP-VAR[12] of equations (3)-(4). The prior takes the following form

$$\beta \sim N\left(\underline{b}, \underline{V}\right),$$

where $\underline{b}$ is one for coefficients on the first own lag, and zero otherwise, and

$$\underline{V}_{ij} = \begin{cases} 100 s_i^2 & \text{if intercept} \\ \lambda / r^2 & \text{if } i = j \\ \lambda \frac{s_i^2}{r^2 s_l^2} & \text{if } i \neq j \end{cases} \tag{A.3}$$

for $r = 1, ..., p$, $i = 1, ..., n$,and $j = 1, ..., k$ with $k = np + 1$. Here s_i^2 is the residual variance from the unrestricted p-lag univariate autoregression for variable i. The degree of shrinkage depends on a single hyperparameter λ which, for the purpose of computing the forecasts of the empirical section, we fix to 1.

[12]One can also apply the Minnesota prior in the equivalent reparametrized TVP-VAR in equations (5)-(6). In practice it shouldn't matter which specification we use, since this prior provides data-based shrinkage before the model is estimated (i.e. the traditional Minnesota prior is not updated by the likelihood in each iteration).