

# Choosing Prior Hyperparameters: With Applications To Time-Varying Parameter Models

Pooyan Amir-Ahmadi\*

University of Illinois at Urbana-Champaign

and

Christian Matthes

Federal Reserve Bank of Richmond

and

Mu-Chun Wang

University of Hamburg

March 15, 2018

## Abstract

Time-varying parameter models with stochastic volatility are widely used to study macroeconomic and financial data. These models are almost exclusively estimated using Bayesian methods. A common practice is to focus on prior distributions that themselves depend on relatively few hyperparameters such as the scaling factor for the prior covariance matrix of the residuals governing time variation in the parameters. The choice of these hyperparameters is crucial because their influence is sizeable for standard sample sizes. In this paper we treat the hyperparameters as part of a hierarchical model and propose a fast, tractable, easy-to-implement, and fully Bayesian approach to estimate those hyperparameters jointly with all other parameters in the model. We show via Monte Carlo simulations that, in this class of models, our approach can *drastically* improve on using fixed hyperparameters previously proposed in the literature.

*Keywords:* Bayesian inference, Bayesian VAR, Time variation

---

\*We would like to thank Luca Benati, Fabio Canova, Minsu Chan, Todd Clark, Frank Diebold, Luca Gambetti, Thomas Lubik, Frank Schorfheide, Mark Watson, Alexander Wolman, and seminar participants at the University of Pennsylvania as well as conference participants at the EUI workshop on time-varying parameter models and the 9th Rimini Bayesian econometrics workshop for helpful comments. Andrew Owens and Daniel Tracht provided excellent research assistance. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

# 1 Introduction

Multivariate time series models form the backbone of empirical macroeconomics. A common feature of all popular multivariate time series models is that, as researchers include more variables, the number of parameters quickly grows large, a feature that is maybe most evident in VARs that feature time-varying parameters and stochastic volatility (Cogley & Sargent (2005) and Primiceri (2005)), which are the main focus of this paper.

Bayesian inference, via its use of priors, allows researchers to avoid overfitting the observed sample (which would come at the cost of unrealistic out-of-sample behavior). It has thus become the standard approach when estimating multivariate time series models with many parameters. Eliciting priors in such high-dimensional models is a daunting task, though. A common practice is to focus on prior distributions that themselves depend on a substantially smaller number of parameters (which we will call hyperparameters). One prominent example that uses this approach is the 'Minnesota' prior for VARs (Doan et al. (1984)), which is especially useful in applications with many observable variables (Banbura et al. (2010)).

The choice of hyperparameters is crucial because their influence is often sizeable for standard sample sizes. Nonetheless, the choice of those hyperparameters is often ad hoc in the literature. In this paper, we propose a fast, tractable, and easy-to-implement Metropolis step that can easily be added to standard posterior samplers such as the Metropolis-Hastings algorithm or the Gibbs sampler (Gelman et al. (2013)). Researchers can use our approach with minimal changes in their code (and negligible increase in runtime) to estimate these hyperparameters. The estimation algorithm that we present in this paper exploits the hierarchical structure that is automatically present whenever prior hyperparameters are used and thus can be used generally in *any* model with prior hyperparameters. Our approach interprets the structure implied by the interaction of parameters of the model and the associated prior hyperparameters as a hierarchical model, which is a standard model in Bayesian inference (Gelman et al. (2013)).

The Gibbs sampler is already a standard approach to estimate multivariate time series models and thus our approach fits naturally into the estimation approach used for these models.

The importance of hyperparameters for VARs with time-varying parameters and stochastic volatility has been established by Primiceri (2005), who also estimates the hyperparameters (to our knowledge, the only other paper that does so in a Bayesian context for these models). Unfortunately, Primiceri (2005)'s approach to estimating the prior hyperparameters is computationally involved and requires focusing on only a small number of possible values for the hyperparameters. Since the hyperparameters interact with the part of the prior that is set via the use of a training sample (which depends crucially on the specific data sample), it is also not clear that the same discrete grid of possible parameter values that Primiceri (2005) used should be employed for other applications.

Some readers might wonder why the choice of prior hyperparameters is important. Shouldn't the importance of the prior vanish as the data size increases? In this paper, we show that the hyperparameters influence estimation outcomes for the class of models we consider and standard sample sizes available for macroeconomic analysis. This echoes the results in Reusens & Croux (2017), who carry out an extensive Monte Carlo study of prior sensitivity using a VAR with time-varying parameters but no stochastic volatility.

Other papers have addressed related issues in a frequentist framework. Stock & Watson (1996) propose a frequentist approach to estimate scaling parameters in the law of motion for time-varying parameter models. Benati (2015) adapts their approach to a time-varying parameter VAR model without stochastic volatility. Benati's approach is computationally substantially more involved than ours and a mix of Bayesian and frequentist approaches, thus making it harder to interpret in the otherwise Bayesian estimation of these models. Benati focuses on the hyperparameter for the coefficients (since his model does not feature stochastic volatility), while we also estimate the hyperparameters in the law of motion for stochastic volatilities.

Our paper is more generally related to the literature on choosing prior hyperparameters in Bayesian inference. Giannone et al. (2015) estimate prior hyperparameters for time-invariant VARs with conjugate Normal-Inverse Wishart priors by exploiting the fact that in this case the density of the data conditional on the hyperparameters (the marginal likelihood) is known in closed form, which they propose to either maximize with respect to the hyperparameters or to draw from. In the second case they then propose to draw the other

VAR parameters conditional on the hyperparameters. Because the marginal likelihood is known in closed form in their setup, they can first draw the hyperparameters *without* conditioning on the other VAR parameters. As such, their second algorithm is not really a Gibbs sampler, but rather directly generates draws from the joint distribution of hyperparameters and other VAR parameters by first generating a draw from the marginal distribution of the hyperparameters and then generating a draw from the conditional distribution of the other VAR parameters (conditional on the hyperparameters). This is possible *exactly* because the marginal likelihood is known in closed form in their case. Our approach can be applied to any model in which prior hyperparameters are present and thus presents an alternative to the approach in Giannone et al. (2015) for fixed coefficient VARs when the marginal likelihood is not known in closed form (as is the case, for example, if non-conjugate priors are used). In the models with time-varying parameters and stochastic volatility that we focus on in this paper, there is no closed form for the marginal data density. The approach by Giannone et al. (2015) can thus not be easily extended to time-varying parameter models. As highlighted by Giannone et al. (2015), their first approach (which maximizes the marginal likelihood) is an empirical Bayes approach, while our approach and their second approach focus on the hierarchical structure imposed by the use of prior hyperparameters.

In an early attempt to tackle the problem of estimating prior hyperparameters, Lopes et al. (1999) propose an alternative procedure to estimate hyperparameters using sampling importance resampling. Their approach, just as Giannone et al. (2015), requires the calculation of the marginal likelihood conditional on the hyperparameters of interest, i.e. the density of data conditional *only* on the hyperparameters, with all other parameters integrated out. In contrast to Giannone et al. (2015), Lopes et al. (1999) use numerical methods to approximate the marginal likelihood. Computing even one such marginal likelihood is a computationally daunting task in the models we focus on in this paper. The approach in Lopes et al. (1999) would require the computation of such a marginal likelihood for *every* unique draw of the hyperparameters, thus rendering it impractical for the applications we are interested in. Furthermore, in the class of models we study, researchers regularly use loose priors. It is well known (Gelman et al. (2013)) that in the case of loose priors, the exact specification of those priors has a substantial influence on the

value of the marginal likelihood, even though point estimates and error bands are largely unaffected. If a researcher imposes tighter priors so that the inference on marginal likelihoods becomes reliable, our approach does not add any substantial conceptual difficulty to the estimation of the marginal likelihood - the researcher would then have to integrate out the hyperparameters as well as all other parameters of the model when computing the marginal likelihood. Methods such as those presented in Chan & Eisenstat (2017) could then possibly be adapted to estimate posterior odds.

Korobilis (2014) estimates some prior parameters in a VAR with time-varying parameters and stochastic volatility. To be more specific, Korobilis (2014) restricts the prior covariance of the innovations to the parameters to be diagonal. Those diagonal elements are then estimated in a Gibbs sampling step. His approach could be combined with ours since Korobilis (2014) relies on prior hyperparameters for the prior covariance matrix of the innovations to the parameters.

In the next section, we describe the general algorithm before turning to time-varying parameter models in section 3. We then carry out a simulation study in section 4 before showing the effect of estimating prior hyperparameters on two real world applications.

## 2 How to Estimate Prior Hyperparameters

In this section, we derive a Metropolis step to estimate prior hyperparameters. While our focus is on models with time-varying parameters and stochastic volatility, the algorithm is most easily introduced in a general framework, while also showcasing the general applicability of our approach. The model is given by a likelihood function  $p(Y|\theta, K, \kappa)$  where  $Y$  is the matrix of data (note that here the data is not necessarily time series data) and  $\theta$  is the set of all parameters except for the parameter block  $K$  associated with the hyperparameter vector  $\kappa$ . As we will highlight below, the hierarchical nature of hyperparameters implies that  $p(Y|\theta, K, \kappa)$  is actually independent of  $\kappa$ . The prior for  $K$ , which we denote  $p(K|\kappa)$ , depends on the hyperparameter  $\kappa$ . To give a specific example, it might be useful to think of  $\kappa$  as the scaling parameters for the Minnesota prior used in the Bayesian estimation of VARs - then  $K$  would be the intercepts and the coefficients on lagged observables. More detail on the VAR with a Minnesota prior can be found in the appendix. We

assume that  $\theta$  and  $K$  are estimated via Gibbs-sampling or the (possibly multiple-block) Metropolis-Hastings algorithm, as described, for example, in Gelman et al. (2013). The augmented algorithm that includes the estimation of the hyperparameters then alternates between draws from the algorithm for  $\theta$  and  $K$  (both those steps condition on a value for  $\kappa$ ) and the drawing of  $\kappa$  conditional on  $K$  and  $\theta$ , which we describe in this section. The prior beliefs about the hyperparameter  $\kappa$  are encoded in a prior distribution  $p(\kappa)$ . From a conceptual point of view, a researcher could introduce another level of hierarchy and make the prior for  $\kappa$  depend on more hyperparameters as well. Since we are concerned with applications where the dimensionality of  $\kappa$  is already small (such as the time-varying parameter models we describe later), we will not pursue this question further in this paper - our approach could be extended in a straightforward manner if a researcher was interested in introducing additional levels of hierarchy. We focus here on drawing one vector of hyperparameters, but other vectors of hyperparameters could be included in  $\theta$  (which could be high-dimensional, as in our time-varying parameter VAR later). Draws for those other vectors of hyperparameters would then be generated using additional Metropolis steps that have the same structure. If  $J$  vectors of hyperparameters are present, we denote vector  $j$  by  $\kappa_j$  ( $j = 1, \dots, J$ ) and the vector of *all* hyperparameters by  $\tilde{\kappa} = [\kappa'_1 \ \kappa'_2 \ \dots \ \kappa'_J]'$ . When we discuss the algorithm below, we will denote by  $\kappa$  either the only vector of hyperparameters present in the model or one representative vector of hyperparameters  $\kappa_j$ , holding all other hyperparameters fixed (draws for those vectors can then, as mentioned before, be generated from additional Metropolis-steps with the same structure). We assume that the following conditions hold (condition 1 is only necessary if multiple vectors of hyperparameters are present in the model):

**Condition 1** *The different vectors of hyperparameters are a priori independent of each other:  $p(\tilde{\kappa}) = \prod_{j=1}^J p(\kappa_j)$*

**Condition 2** *All parameters of the model except for the parameter block directly linked to a specific hyperparameter are a priori independent of that specific hyperparameter:  $p(\theta, \kappa) = p(\theta)p(\kappa)$*

Neither of these conditions are restrictive. If condition 1 is violated, the dependent vectors of hyperparameters just have to be grouped into one larger vector of hyperparameters.

We later spell out these conditions in more detail for our VAR model. The modifications for the algorithm in this case are straightforward. Violations of the second condition can be handled similarly: The different parameter blocks whose priors depend on the same hyperparameters have to be grouped together in one larger parameter vector, which then depend on the same vector of hyperparameters.

Deriving a Metropolis step for sampling  $\kappa$  amounts to deriving a formula for the acceptance probability in the Metropolis-Hastings step. We draw a realization from the proposal density  $q(\cdot)$ , which will be accepted with probability  $\alpha^i$  at iteration  $i$  of the algorithm. This acceptance probability in the Metropolis-within-Gibbs step at iteration  $i$  is given by

$$\alpha^i = \min \left( 1, \frac{p(\theta, \kappa^{prop}, K|Y)q(\kappa^{prop}|\kappa^{i-1})}{p(\theta, \kappa^{i-1}, K|Y)q(\kappa^{i-1}|\kappa^{prop})} \right) \quad (1)$$

a superscript *prop* denotes a proposed value, a superscript  $i-1$  denotes values from iteration  $i-1$  of the algorithm, and superscripts are dropped for  $K$  and  $\theta$  for convenience. We now simplify  $\alpha^i$  in this general environment.

First, we rewrite  $p(\theta, \kappa, K|Y)$ :

$$p(\theta, \kappa, K|Y) \propto p(Y|\theta, K)p(\theta|\kappa, K)p(K|\kappa)p(\kappa) \quad (2)$$

By the hierarchical nature of the model (the hyperparameters only enter the prior for  $K$ ), the likelihood  $p(Y|\theta, K)$  does not depend on  $\kappa$  since it conditions on  $K$ . Thus,  $p(Y|\theta, \kappa, K)$  cancels out in the numerator and denominator of  $\alpha^i$ . By condition 2 and the hierarchical nature of the hyperparameter structure (and, if necessary, condition 1), the term  $p(\theta|\kappa, K)$  equals  $p(\theta|K)$ , which then also cancels out in the fraction determining  $\alpha^i$ . We are left with

$$\alpha^i = \min \left( 1, \frac{p(K|\kappa^{prop})p(\kappa^{prop})q(\kappa^{prop}|\kappa^{i-1})}{p(K|\kappa^{i-1})p(\kappa^{i-1})q(\kappa^{i-1}|\kappa^{prop})} \right) \quad (3)$$

A key insight to this equation is that all identities that need to be evaluated are either the proposal density  $q(\cdot)$  or prior densities ( $p(\kappa)$  is the prior density for  $\kappa$  while  $p(K|\kappa)$  is the prior density of  $K$ , which depends on the hyperparameter  $\kappa$ ). Generally those densities are known in closed form and thus fast to evaluate, thus making our algorithm computationally efficient.

### 3 The VAR Model and the Estimation of Hyperparameters

This section presents the class of models we focus on in this paper and the necessary additional steps in the Gibbs-sampling algorithm for time-varying parameter VARs to estimate the prior scale parameters. For an introduction to this class of models see Koop & Korobilis (2010). In the appendix, we lay out the estimation algorithm for this class of models in detail.

The observable vector  $y_t$  is modeled as:

$$y_t = \mu_t + \sum_{l=1}^L B_{l,t} y_{t-l} + e_t \quad (4)$$

where the intercepts  $\mu_t$ , the autoregressive matrices  $B_{j,t}$ , and the covariance matrix of  $e_t$  are allowed to vary over time. To be able to parsimoniously describe the dynamics of our model, we define  $X'_t \equiv I \otimes (\mathbf{1}, y'_{t-1}, \dots, y'_{t-L})$ ,  $b_t \equiv \text{vec}(\mu_t B_{1,t} \cdots B'_{L,t})$  and rewrite (4) in the following state space form:

$$y_t = X'_t b_t + e_t \quad (5)$$

$$b_t = b_{t-1} + \omega_{b,t} \quad (6)$$

$I$  denotes a identity matrix of conformable size and  $\mathbf{1}$  denotes a vector of ones of conformable size. The observation equation (5) is a more compact expression for (4). The state equation (6) describes the law of motion for the intercepts and autoregressive matrices. The covariance matrix of the innovations in equation (5) is modeled following Primiceri (2005):

$$e_t = A_t^{-1} \Sigma_t \varepsilon_t \quad (7)$$

$A_t$  is a lower triangular matrix with ones on the main diagonal and representative non fixed element  $a_t^i$ .  $\Sigma_t$  is a diagonal matrix with representative non fixed element  $\sigma_t^j$ . The dynamics of the non fixed elements of  $A_t$  and  $\Sigma_t$  are given by:

$$a_t^i = a_{t-1}^i + \omega_{a,t}^i \quad (8)$$

$$\log \sigma_t^j = \log \sigma_{t-1}^j + \omega_{h,t}^j \quad (9)$$



To conclude the description of our model, we need to make distributional conditions on the innovations  $\varepsilon_t$ ,  $\omega_{b,t}$ ,  $\omega_{h,t}$ , and  $\omega_{a,t}$ , where  $\omega_{h,t}$  and  $\omega_{a,t}$  are vectors of the corresponding scalar innovations in the elements of  $\Sigma_t$  and  $A_t$ . We assume that all these innovations, which govern the time variation for the different parameters in this models, are normally distributed with the following covariance matrix, which we, following Primiceri (2005), restrict as follows:

$$Var \begin{bmatrix} \left( \begin{array}{c} \varepsilon_t \\ \omega_{b,t} \\ \omega_{a,t} \\ \omega_{h,t} \end{array} \right) \end{bmatrix} = \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & \Omega_b & 0 & 0 \\ 0 & 0 & \Omega_a & 0 \\ 0 & 0 & 0 & \Omega_h \end{bmatrix} \quad (10)$$

$\Omega_a$  is further restricted to be block diagonal with  $J$  blocks, which simplifies inference (this is inconsequential for our extension to the standard Gibbs sampler, but we decided to use the standard model in the literature). Note that  $\Omega_h$ , on the other hand, is not restricted, allowing the increments in the stochastic volatility processes to be correlated.

We will now describe the estimation of general prior hyperparameters in this setting before turning to the specific prior hyperparameters used by Primiceri (2005) and the subsequent literature.

The priors for  $\Omega_b$ ,  $\Omega_a$ , and  $\Omega_h$  are given by:

$$\Omega_b \sim p_{\Omega_b}(\kappa_{\Omega_b}) \quad (11)$$

$$\Omega_h \sim p_{\Omega_h}(\kappa_{\Omega_h}) \quad (12)$$

$$\Omega_{a,j} \sim p_{\Omega_{a,j}}(\kappa_{\Omega_a}) \quad \forall j = 1, \dots, J \quad (13)$$

where  $\kappa_i$ ,  $i \in (\Omega_b, \Omega_h, \Omega_a)$  denotes the vectors of hyperparameters for each set of matrices.  $\Omega_{a,j}$  is the  $j$ -th block of  $\Omega_a$ .

We are interested in estimating the hyperparameters  $\kappa_{\Omega_b}, \kappa_{\Omega_h}, \kappa_{\Omega_a}$ . To do so, we attach priors  $p_X(X)$  to the hyperparameters ( $X = \{\kappa_{\Omega_b}, \kappa_{\Omega_h}, \kappa_{\Omega_a}\}$ ). In our empirical applications, we assume that the prior specification for all other parameters are the same as in Primiceri (2005), but this is inconsequential for our algorithm. We denote by  $\theta$  all parameters to be estimated except for the prior hyperparameters themselves and the associated covariance

matrices  $\Omega_b$ ,  $\Omega_h$ , and  $\{\Omega_{a_j}\}_{j=1}^J$ . Our approach builds on the insight that equations (11) to (13) can be interpreted as a hierarchical model, which in our case is embedded in a larger model, the VAR with time-varying parameters and stochastic volatility.

We now restate conditions 1 and 2 for the specific model at hand:

**Condition 3** *The different vectors of hyperparameters in a TVP-VAR are a priori independent of each other:*

$$p(\kappa_{\Omega_b}, \kappa_{\Omega_h}, \kappa_{\Omega_a}) = p_{\kappa_{\Omega_b}}(\kappa_{\Omega_b})p_{\kappa_{\Omega_h}}(\kappa_{\Omega_h})p_{\kappa_{\Omega_a}}(\kappa_{\Omega_a})$$

**Condition 4** *All parameter blocks of the TVP-VAR model except for the parameter block directly linked to a specific hyperparameter (via one of the equations 11 through 13 in this model) are a priori independent of that specific hyperparameter (e.g.  $\Omega_h$  and  $\Omega_{a,j} \forall j = 1, \dots, J$  are a priori independent of  $\kappa_{\Omega_b}$ ).*

As long as we assume that  $p_{\Omega_b}$ ,  $p_{\Omega_h}$ , and  $p_{\Omega_{a,j}}$  are all inverse Wishart distributions (as is standard in the literature), the drawing of the covariance matrices themselves can be carried out just as in the algorithm described in Del Negro & Primiceri (2015) once we condition on the hyperparameters.

To estimate the hyperparameters, we use a Metropolis-within-Gibbs step (Geweke (2005)) for each vector of hyperparameters. We focus here on the estimation of  $\kappa_{\Omega_b}$  because the other blocks are conceptually the same. The acceptance probability  $\alpha^i$  at iteration  $i$  of the Metropolis-within-Gibbs algorithm is given by:

$$\alpha^i = \min \left( 1, \frac{p(\theta, \kappa_{\Omega_b}^{prop}, \Omega_b, \kappa_{\Omega_a}, \{\Omega_{a_j}\}, \kappa_{\Omega_h}, \Omega_h | y^T) q(\kappa_{\Omega_b}^{prop} | \kappa_{\Omega_b}^{i-1})}{p(\theta, \kappa_{\Omega_b}^{i-1}, \Omega_b, \kappa_{\Omega_a}, \{\Omega_{a_j}\}, \kappa_{\Omega_h}, \Omega_h | y^T) q(\kappa_{\Omega_b}^{i-1} | \kappa_{\Omega_b}^{prop})} \right) \quad (14)$$

where a superscript *prop* denotes the proposed value and a superscript  $i - 1$  the value from the previous iteration (superscripts are dropped for all other parameters for ease of reading).  $y^T$  is the history of observables used for estimation ( $y^T = \{y_t\}_{t=1}^T$ ). Again,  $q(\cdot)$  is the proposal density.

Applying the results from the previous section to this specific model, we find that the acceptance probability simplifies to

$$\alpha^i = \min \left( 1, \frac{p(\Omega_b | \kappa_{\Omega_b}^{prop}) p(\kappa_{\Omega_b}^{prop}) q(\kappa_{\Omega_b}^{prop} | \kappa_{\Omega_b}^{i-1})}{p(\Omega_b | \kappa_{\Omega_b}^{i-1}) p(\kappa_{\Omega_b}^{i-1}) q(\kappa_{\Omega_b}^{i-1} | \kappa_{\Omega_b}^{prop})} \right) \quad (15)$$

$p(\Omega_b|\kappa_{\Omega_b})$  is the prior density for  $\Omega_b$  described above (which is usually an inverse Wishart density) and  $p(\kappa_{\Omega_b})$  is the prior on  $\kappa_{\Omega_b}$ . Once we have fixed a proposal density for  $\kappa_{\Omega_b}$ , evaluating the acceptance probability is thus straightforward. Not only can the same argument be made for the other hyperparameters introduced before, but for any hyperparameter since the logic used for deriving the acceptance probability only hinges on the hierarchical nature of the model with respect to the prior hyperparameters.

Now turning to the exact specification in Primiceri (2005), the priors for  $\Omega_b$ ,  $\Omega_a$  and  $\Omega_h$  are set as follows:

$$\Omega_b \sim IW(\kappa_{\Omega_b}^2 \nu_{\Omega_b} V_{\Omega_b}, \nu_{\Omega_b}) \quad (16)$$

$$\Omega_h \sim IW(\kappa_{\Omega_h}^2 \nu_{\Omega_h} V_{\Omega_h}, \nu_{\Omega_h}) \quad (17)$$

$$\Omega_{a,j} \sim IW(\kappa_{\Omega_a}^2 \nu_{\Omega_{a,j}} V_{\Omega_{a,j}}, \nu_{\Omega_{a,j}}) \quad (18)$$

where  $\nu$  denotes the degrees of freedom,  $IW$  is the inverse Wishart distribution,  $V_X$ ,  $X \in \{\Omega_b, \Omega_h, \Omega_a\}$ , are prior scaling matrices, and  $\kappa_X$  are the scalar hyperparameters we want to estimate. A change in  $\kappa_X^2$  linearly scales the mean and mode of the corresponding inverse Wishart prior distribution while the prior variance is a function of  $\kappa_X^4$ . We follow the literature in having the scaling parameter entering squared in the parameters for the inverse Wishart distribution. This is why the fourth power of  $\kappa_X$  appears in the variance.

In this paper, we focus on the estimation of low-dimensional hyperparameters. In theory, our algorithms could be adapted to estimate the prior scaling matrices  $V_X$ ; however, for most practical applications the  $V_X$  matrices are high-dimensional objects, so we focus instead on picking the  $V_X$  matrices using a training sample, as is standard in the literature. One difference relative to the general algorithm above is that, to be in line with Primiceri (2005) and the subsequent literature, we use the same  $\kappa_a$  for all blocks of  $\Omega_a$ . For the different blocks of  $\Omega_a$ , we use the fact that conditional on  $\kappa_a$  the priors for the different blocks are independent inverse-Wishart densities. Thus, in that case we get

$$P(\Omega_a|\kappa_{\Omega_a}) = \prod_{j=1}^J P(\Omega_{a,j}|\kappa_{\Omega_a}) \quad (19)$$

Some groups of parameters or volatilities might vary at a different rate than other parameters. We now show how to incorporate this idea into our framework. Benati (2015) also estimates different scaling parameters for different equations in his VAR.

We denote by  $\overline{\kappa_x}$  vectors of scaling parameters of dimension  $d_x$ , where matrix  $x$  is of dimension  $d_x$  by  $d_x$ . We then assume the following forms for the priors of the matrices  $\Omega_b$ ,  $\Omega_h$ , and  $\Omega_{a,j}$ :

$$\Omega_b \sim IW(\text{diag}(\overline{\kappa_{\Omega_b}})\nu_{\Omega_b}V_{\Omega_b}\text{diag}(\overline{\kappa_{\Omega_b}}), \nu_{\Omega_b}) \quad (20)$$

$$\Omega_h \sim IW(\text{diag}(\overline{\kappa_{\Omega_h}})\nu_{\Omega_h}V_{\Omega_h}\text{diag}(\overline{\kappa_{\Omega_h}}), \nu_{\Omega_h}) \quad (21)$$

$$\Omega_{a,j} \sim IW(\text{diag}(\overline{\kappa_{\Omega_{a,j}}})\nu_{\Omega_{a,j}}V_{\Omega_{a,j}}\text{diag}(\overline{\kappa_{\Omega_{a,j}}}), \nu_{\Omega_{a,j}}) \quad (22)$$

where *diag* is an operator that turns a  $d \times 1$  dimensional vector into a  $d \times d$  dimensional diagonal matrix with the elements of the vector on the main diagonal. In practice, estimating one  $\kappa$  scaling parameter per coefficient/volatility is not feasible for VARs of the size commonly used in applications because of the large number of coefficients that would have to be estimated. Instead, we propose to group parameters into a relatively small number of groups and use one  $\kappa$  scaling parameter per block of parameters. As mentioned before, natural choices for blocks in the case of the  $b$  coefficients could be intercepts vs. all other parameters or a grouping of  $b$  coefficients by equation. We would then augment our description of the algorithm with a deterministic mapping from the relatively small number of scaling parameters (which we call  $\check{\kappa}_x$ ) to  $\overline{\kappa}_x$ .

In terms of the estimation algorithm, nothing of substance changes: in the proposal step, the proposal density is now multivariate normal and in the calculation of the acceptance probability we have to adjust the evaluation of  $p(\Omega_b|\kappa_{\Omega_b})$  to take into account the updated form of the density (see equation (20)) and the fact that the prior of the hyperparameters is now a density of a multivariate vector. One could use independent priors for each element of  $\check{\kappa}_{\Omega_b}$ , for example. The rest of the Gibbs sampling steps for other parameters are unaffected, with the exception of the step where  $\Omega_b$  is drawn: the scaling matrix for the inverse-Wishart density needs to be updated as described in equation (20).

## 4 Monte Carlo Study

In this section, we use VAR(1) models as data-generating processes to assess the performance of our algorithm vis-a-vis algorithms that use fixed hyperparameters. Using multivariate data-generating processes makes the exercises more realistic, but it comes at a cost: Assuming a random-walk law of motion of the parameters with non-trivial time variation is not straightforward. Either we have to reject many simulated parameters because they yield non-stationary dynamics or our simulated time series are very different from those we actually use in economics (because the simulated series become explosive).

A natural response to this dilemma is that the random walk evolution of parameters was never meant to be a description of the true DGP, but rather a flexible and parsimonious way to approximate a large class of possible patterns of time variation. Furthermore, note that we will confront both our algorithm and the standard algorithm with the same data-generating process, and both implementations assume a random-walk evolution of parameters, so we are not biasing our findings in favor of our approach. In the appendix, we show additional results for both multivariate as well as univariate data-generating processes (including a random walk data-generating process in the univariate case - it is easier to obtain reasonable simulations in a univariate setting without rejecting too many simulations).

For every data-generating process, we simulate 100 samples of 350 observations each. We use a training sample of 40 observations to initialize the prior along the lines of Primiceri (2005). 10000 draws are generated for each sample, of which we use the first 5000 draws to tune the proposal for our Metropolis-Hastings step. In the case of fixed hyperparameters, we use 10000 draws as well.

As data-generating-processes, we use trivariate VARs, but in the appendix we also show results for bivariate VARs to show that our findings are robust to the number of observables and because both bivariate and trivariate VARs feature prominently in the literature.

First, we show a trivariate VAR with a deterministic law of motion by sine and cosine waves (the exact description can be found in the appendix). We compute root mean squared errors of the estimated median parameter paths and root mean squared forecasts errors for each variable in our VAR (computed with the median forecast as point forecast). The exact

Table 1: Monte Carlo forecast results for deterministic and continuous evolution of parameters.

Relative <b>RMSFE</b>			
[Out-of-sample forecast of <i>first</i> variable]			
Horizons	$i\mathcal{G}$	half- $\mathcal{C}$ auchy	Fixed ( $\kappa_{\Omega_b} = 0.1$ )
1	0.969	0.965	0.972
2	0.945	0.946	0.954
3	0.898	0.898	0.910
4	0.889	0.888	0.905
[Out-of-sample forecast of <i>second</i> variable]			
Horizons	$i\mathcal{G}$	half- $\mathcal{C}$ auchy	Fixed ( $\kappa_{\Omega_b} = 0.1$ )
1	0.903	0.903	0.915
2	0.905	0.910	0.920
3	0.939	0.941	0.951
4	0.992	0.989	1.001
[Out-of-sample forecast of <i>third</i> variable]			
Horizons	$i\mathcal{G}$	half- $\mathcal{C}$ auchy	Fixed ( $\kappa_{\Omega_b} = 0.1$ )
1	0.928	0.930	0.919
2	0.948	0.950	0.946
3	0.917	0.916	0.919
4	0.953	0.953	0.946

formula we use can be found in the appendix. We then average over all 100 samples and show the resulting number relative to the value computed for the fixed hyperparameter case with Primiceri (2005)'s values, which were estimated on US data. We use two priors for the hyperparameter: one is an inverse gamma with a standard deviation of 0.1 and a mode of 0.05 (we use the same prior for all hyperparameters), while the other is a half-Cauchy distribution with the same scale parameter as the inverse gamma distribution we use. In the section on empirical examples, we elaborate in more detail on prior choice and argue that the inverse-Gamma prior is a useful benchmark. The half-Cauchy distribution, in contrast with the inverse-Gamma distribution, has the feature that it has positive density

at hyperparameter values of 0. The results in this section are robust to either choice of prior. A natural question to ask is how much our findings depend on the specific values

Table 2: Monte Carlo results for deterministic and continuous evolution of parameters.

Relative <b>RMSE</b>			
[In-sample fit of parameter paths $b_t$ evaluated at posterior median]			
Parameter	iG	half- $\mathcal{C}$ auchy	Fixed ( $\kappa_{\Omega_b} = 0.1$ )
$\mu_1$	0.625	0.620	0.680
$\mu_2$	0.560	0.557	0.615
$\mu_3$	0.448	0.445	0.472
$B_{11}$	0.422	0.419	0.469
$B_{12}$	1.102	1.102	1.073
$B_{13}$	0.926	0.929	0.911
$B_{21}$	0.497	0.497	0.517
$B_{22}$	0.373	0.371	0.417
$B_{23}$	0.727	0.728	1.635
$B_{31}$	0.417	0.415	0.418
$B_{32}$	0.417	0.415	0.442
$B_{33}$	0.270	0.268	0.272

of the hyperparameters we used in the estimation with fixed hyperparameters. While our approach will always have the advantage that no fixed value needs to be chosen for the hyperparameters, for one specific application one could wonder whether a higher value of the hyperparameter can lead to a better performance for the fixed hyperparameter case in a specific application. To check this, we also estimate a version of the model fixed hyperparameters, but the hyperparameter associated with  $b_t$  ( $\kappa_{\Omega_b}$ ) set to 0.1, which is ten times the value estimated by Primiceri (2005). Table 1 shows the forecasting performance of the different variants we consider relative to a VAR estimated using the hyperparameter values considered in Primiceri (2005) ( $\kappa_{\Omega_b} = 0.01 = \kappa_{\Omega_h}, \kappa_{\Omega_{a,j}} = 0.1$ ). Our approach improves forecasts across the board relative to the benchmark of  $\kappa_{\Omega_b} = 0.01$  and is not doing worse than the variant with a larger  $\kappa_{\Omega_b}$ .

Table 2 shows the root mean squared error of the parameter paths for all elements of  $b_t$

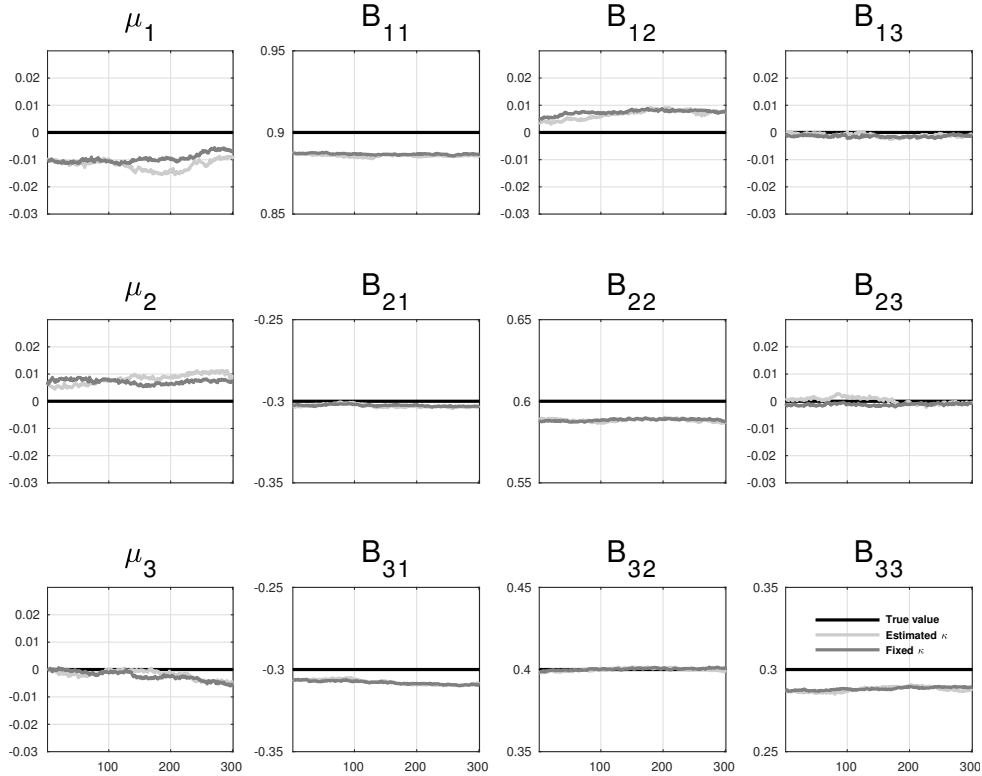


Figure 1: Monte Carlo results - medians across 100 samples

relative to the model with fixed parameters chosen at Princieri’s values. The same picture emerges again: We substantially improve on the benchmark case and are on par with the higher fixed hyperparameter case. In the empirical application we will see that while our approach continues to do well, the success of the higher hyperparameter case in terms of forecasting depends crucially in the data-generating process and is thus not robust. We focus here on the elements of  $b_t$  because volatilities are well estimated across the board for all specifications.

A second natural question is whether our approach comes at a cost - if the true coefficients are fixed over time, does our approach do worse than the fixed hyperparameter setup? This is a natural question because, as mentioned before, in many applications the fixed hyperparameter setup finds little to no time variation in many parameters (Cogley & Sargent (2005)), so one might be tempted to think it has an edge when the coefficients are indeed fixed. Furthermore, the inverse gamma prior we use bounds the hyperparameter away from zero, meaning that finding *exactly* zero time variation is not possible. Tables 3 and 4 show that both in terms of parameter estimates and forecasting ability our approach



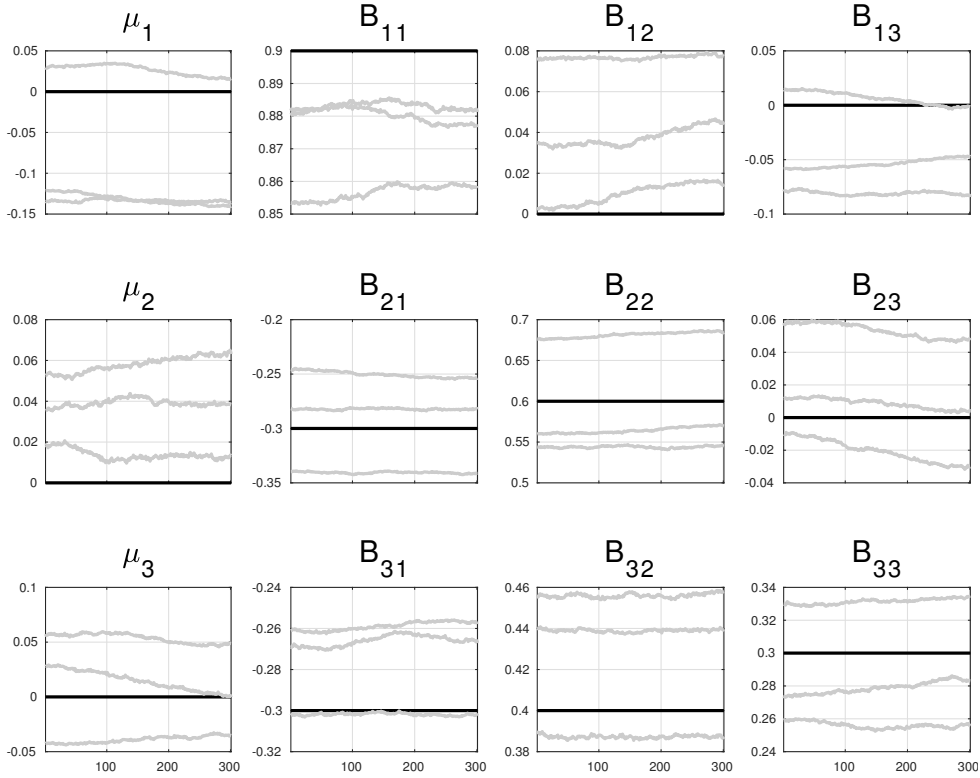


Figure 2: Monte Carlo results - posterior medians for 3 samples

and the fixed hyperparameter approach are very similar in this case. This might leave an interested reader wondering how well TVP-VARs do on an absolute level when approximating fixed coefficient VARs. Figure 1 shows the median across Monte Carlo samples of the estimated posterior median path of all elements of  $b_t$  for the inverse-Gamma prior (results from the Cauchy prior are very similar) and the fixed hyperparameter case (with  $\kappa_{\Omega_b} = 0.01$  - the figure for  $\kappa_{\Omega_b} = 0.1$  looks very similar). The bold straight lines denotes the true values, the light gray lines the results with estimated hyperparameters and an inverse-gamma prior and the dark gray line denotes results for the fixed hyperparameter case. Across simulations we can see that both specifications on average pick up that there is no time variation in the data. Digging deeper, we can also check if for a given sample our approach estimates little to no time variation when there is no time variation present. Figure 2 plots the posterior median paths for three randomly selected samples (out of our 100 simulated samples). Again we see that our algorithm estimates time variation to be small. The main takeaway from this exercise is *not* that the hyperparameters estimated in Primiceri (2005) are 'wrong' in any sense, but rather that, if a researcher is interested

Table 3: Monte Carlo forecast results for fixed parameter VAR.

Relative <b>RMSFE</b>			
[Out-of-sample forecast of <i>first</i> variable]			
Horizons	$i\mathcal{G}$	half- $\mathcal{C}$ auchy	Fixed ( $\kappa_{\Omega_b} = 0.1$ )
1	1.000	1.000	1.022
2	0.999	1.002	1.031
3	0.999	1.002	1.034
4	1.001	1.001	1.049
[Out-of-sample forecast of <i>second</i> variable]			
Horizons	$i\mathcal{G}$	half- $\mathcal{C}$ auchy	Fixed ( $\kappa_{\Omega_b} = 0.1$ )
1	0.997	1.000	0.985
2	1.000	1.002	1.012
3	0.999	1.000	1.009
4	1.000	1.002	1.018
[Out-of-sample forecast of <i>third</i> variable]			
Horizons	$i\mathcal{G}$	half- $\mathcal{C}$ auchy	Fixed ( $\kappa_{\Omega_b} = 0.1$ )
1	1.001	1.002	1.019
2	1.004	0.998	1.026
3	1.001	1.002	1.003
4	1.001	1.002	1.012

in a very different application (including, but certainly not limited to, a different number of observables, a different data frequency, different historical episodes, the different properties of financial versus macroeconomic data etc.), then that researcher should think carefully about the prior hyperparameters. We offer one data-driven and numerically efficient way to take the dependence of the results on the prior hyperparameters into account.

Table 4: Monte Carlo results for fixed parameter VAR.

Relative <b>RMSE</b>			
[In-sample fit of parameter paths $b_t$ evaluated at posterior median]			
Parameter	iG	half- $\mathcal{C}$ auchy	Fixed ( $\kappa_{\Omega_b} = 0.1$ )
$\mu_1$	1.020	0.991	1.738
$\mu_2$	1.007	0.995	1.552
$\mu_3$	1.007	1.008	1.502
$B_{11}$	1.022	0.996	1.598
$B_{12}$	1.010	1.001	1.355
$B_{13}$	1.018	1.011	1.418
$B_{21}$	1.010	0.996	1.363
$B_{22}$	1.003	1.001	1.380
$B_{23}$	1.000	1.000	1.093
$B_{31}$	1.013	1.001	1.369
$B_{32}$	1.003	1.000	1.383
$B_{33}$	1.025	0.990	1.536

## 5 Empirical Application

Going back to the original contributions of Cogley & Sargent (2005) and Primiceri (2005), VARs with time-varying parameters and stochastic volatility have often been used to study questions related to monetary policy and inflation dynamics. Other papers in that vein include Sargent & Surico (2011), D’Agostino & Surico (2012), and Amir-Ahmadi et al. (2016). As the sample size increases, there seems more reason to allow for the possibility of changing parameters and volatilities. These changes can come from various sources - technological progress, changes in institutions, political changes, and international conflicts are just some of the reasons why we might suspect that constant parameter models are ill-suited for longer samples. With samples that are different from the time series used by Primiceri (2005), there is little reason to believe a-priori that the hyperparameters estimated by Primiceri (2005) should reflect a researcher’s view of the amount of time variation present in the data. To assess the importance of estimating the

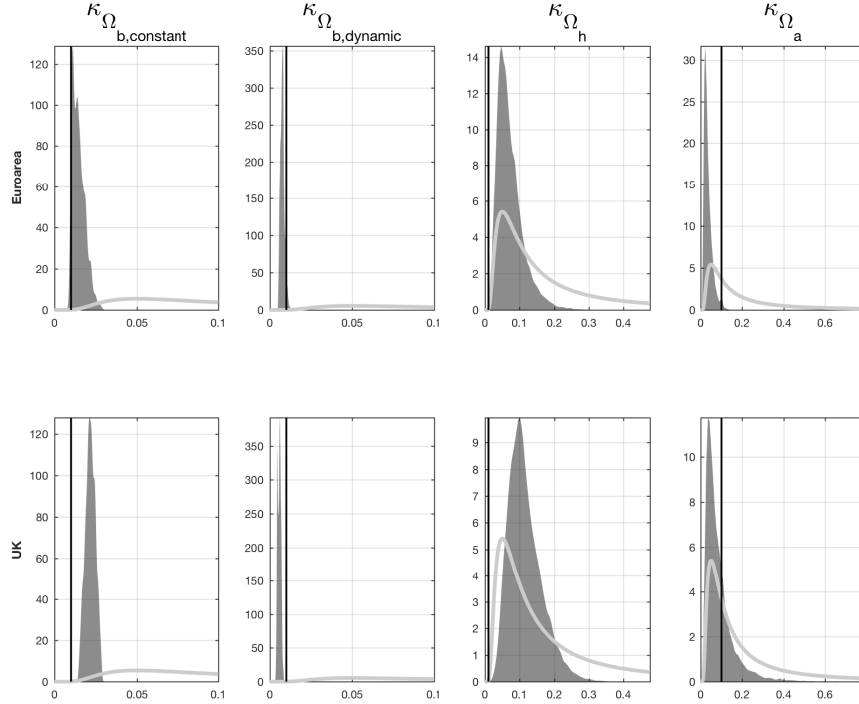


Figure 3: Prior (light gray lines) and Posterior distributions for the hyperparameters

prior hyperparameters, we estimate VARs with time-varying parameters and stochastic volatility for the UK and the Euro Area. The data are annualized quarter-over quarter inflation, annualized quarter-over quarter real GDP growth rate and an annualized short-term nominal rate from 1979 to 2013 for the UK and from 1970 to 2015 for the Euro Area. The prior is set using a training sample of 40 observations along the lines of Primiceri (2005). In this application we use a different hyperparameter for the intercepts (denoted by  $\kappa_{\Omega_{b,constant}}$ ) than for the rest of the elements of  $b_t$  (denoted by  $\kappa_{\Omega_{b,dynamic}}$ ) to allow for more flexibility along the lines of Benati (2015). The data source for the UK is the global VAR database at <https://sites.google.com/site/gvarmodelling/data>. For the Euro Area we use data compiled for the ECB's area wide model available at <http://eabcn.org/page/area-wide-model>. We chose these specific examples because the data series closely resemble those used in many studies of monetary policy that use this class of models, yet the choice of non-US data gives room for hyperparameters to be different from those usually used in the literature to best fit the data. It is useful to remember that the hyperparameters estimated in Primiceri (2005) were estimated on US data. As

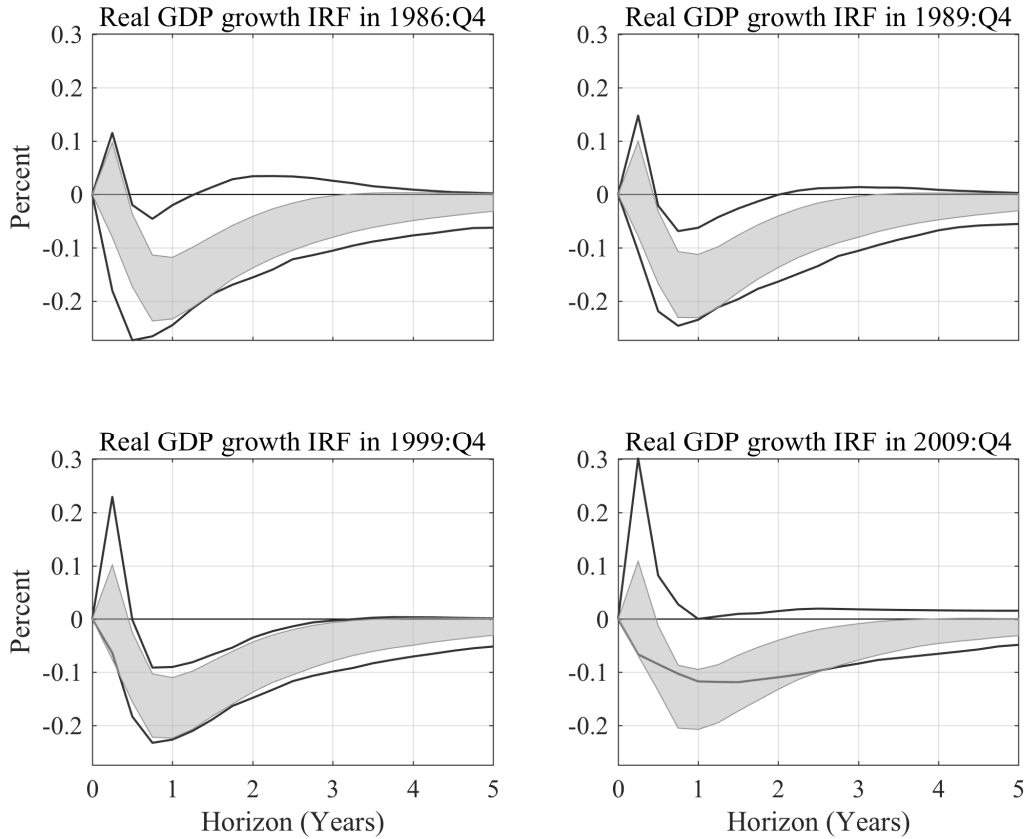


Figure 4: 68 % Posterior Bands for UK impulse responses, estimated hyperparameters in gray

an aside, using our approach on an updated US dataset yields estimates that are broadly in line with those used in Primiceri (2005). In particular the hyperparameters associated with  $b_t$  are very similar, while the hyperparameter associated with  $a_t$  is somewhat smaller and the hyperparameter for  $h_t$  is larger with the updated US dataset. The goal of this section is *not* to represent a comprehensive study of time-varying dynamics and the effect of monetary policy in the UK and the Euro Area, but rather to highlight that estimating the hyperparameters can make a substantial difference when looking at standard model output in this class of models such as impulse responses, estimated parameters, measure of persistence, forecasts, and so on.

To get a sense of whether or not the data calls for hyperparameter values different from those estimated on US data in Primiceri (2005), figure 3 plots the posterior distributions, We see that the posterior distributions peak at values different from those obtained by

Primiceri using US data (Primiceri (2005)'s value is represented by the vertical line in each plot). Furthermore, there is information in the data about the values of the hyperparameters - the marginal posteriors are different from the priors, which are in light gray in each plot.

For simplicity, we assume the same priors across all hyperparameters, which are inverse gamma with a scale parameter of 0.1 and 2 degrees of freedom - this implies a prior mode of 0.05 and an infinite variance. Note that in spite of the mode being higher than Primiceri (2005)'s estimated value and the variance being infinite, the estimated hyperparameters for all elements of  $b_t$  except the intercepts are substantially smaller than Primiceri (2005)'s value. Another takeaway is that, following Benati (2015), the data prefers the scaling parameters to be different across sets of parameters (we use a different scaling parameter for the intercepts). In our forecasting exercise below we show that even when imposing the same hyperparameter within each group of parameters (so that all elements of  $b_t$  use the same hyperparameter), we can still substantially improve the forecasting performance relative to the case of fixed hyperparameters.

A more important question than whether or not these hyperparameters are different from previously used values is whether or not estimating them makes a difference for object economists care about. We first show the estimated impulse responses (obtained using a Cholesky-type recursive ordering just as in Primiceri (2005)) of real GDP growth to an unexpected monetary policy shock that leads to a 1 percent increase in the nominal interest rate. We compare those obtained using our approach with the inverse gamma prior on the hyperparameters with estimates obtained using fixed hyperparameters (fixed at the values from Primiceri (2005)). Since we want to compute some objects that require the companion matrix of the VAR to have eigenvalues that are less than 1 in absolute value at each point in time, we follow Cogley & Sargent (2005) and impose this restriction in exactly the same fashion as in that paper. In the appendix we show that our general findings are robust to not assuming this restriction. Figures 4 and 5 show these impulse responses at 4 dates in the sample. We can see that in the plotted 68% posterior bands there can be substantial differences. In particular, the posterior bands obtained using our approach tend to be narrower than those obtained using the standard approach. These findings carry over

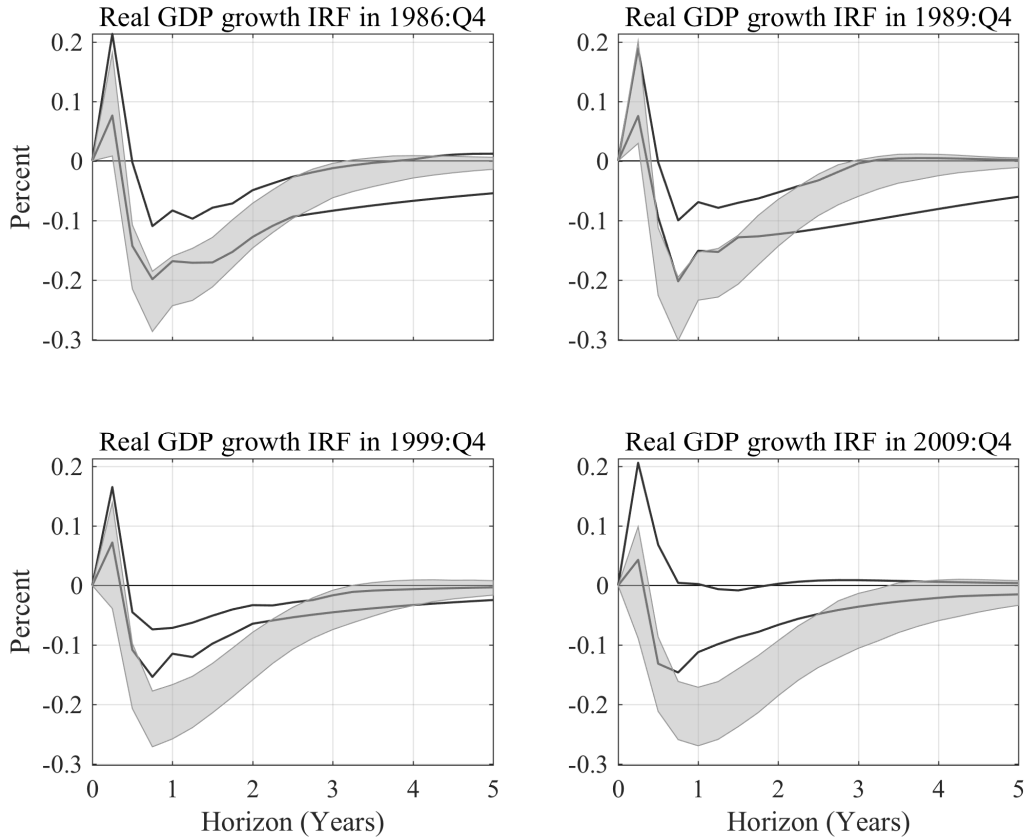


Figure 5: 68 % Posterior Bands for Euro Area impulse responses, estimated hyperparameters in gray

to the responses of inflation and the nominal rate, which can be found in the appendix.

Where do these differences come from? Figure 6 plots the estimated median parameter paths for the fixed and estimated hyperparameter paths for the UK. We can see that the estimated standard deviations of the residuals in the VAR (called  $H_t$  in the graph) are very similar across the specifications, whereas other parameters differ substantially. In particular, the estimated paths of VAR coefficients  $b_t$  are much smoother. This carries over to the Euro Area as well (which we show in the appendix). This is not surprising as the hyperparameter associated with the non-intercept coefficients is estimated to be smaller than the standard value assumed in the literature.

Next, we turn to infinite horizon forecasts, which are often interpreted in these models as trends (to compute infinite horizon forecasts that are guaranteed to be finite, we impose the eigenvalue restriction mentioned before).

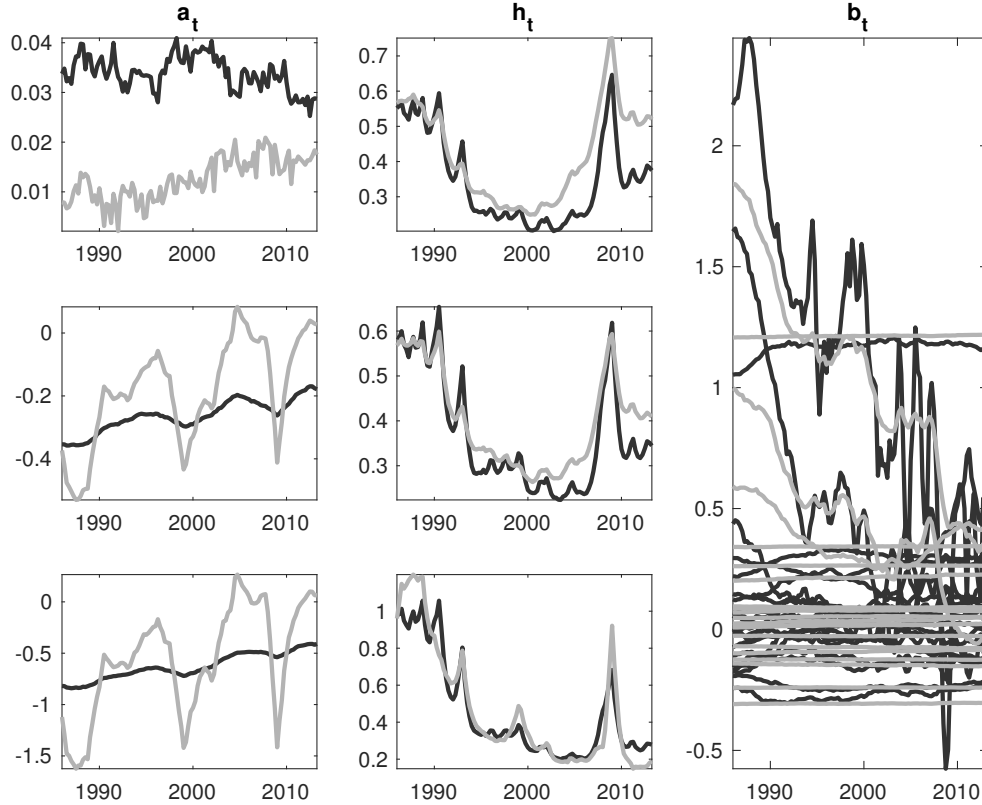


Figure 6: Estimated median parameter paths, fixed hyperparameters in black

Figure 7 shows these infinite horizon forecasts 68 % posterior bands. Since the message is similar, we again relegate the Euro Area figure to the appendix. Our long-run forecasts are much smoother, whereas the fixed coefficient forecast are rather volatile, making it harder to interpret them as trends. A similar message emerges when we look at measure of persistence such as the  $R^2$  measure from Cogley et al. (2010), which we show in the appendix for both the UK and the Euro Area.

## 5.1 Forecasting

D’Agostino et al. (2013) have shown that VAR with time-varying parameters and stochastic volatility can improve upon the forecasting ability of fixed coefficient VARs and other competing models, in particular for inflation. D’Agostino et al. (2013) used fixed hyperparameters. We now ask if our approach can improve forecasting ability even further. We



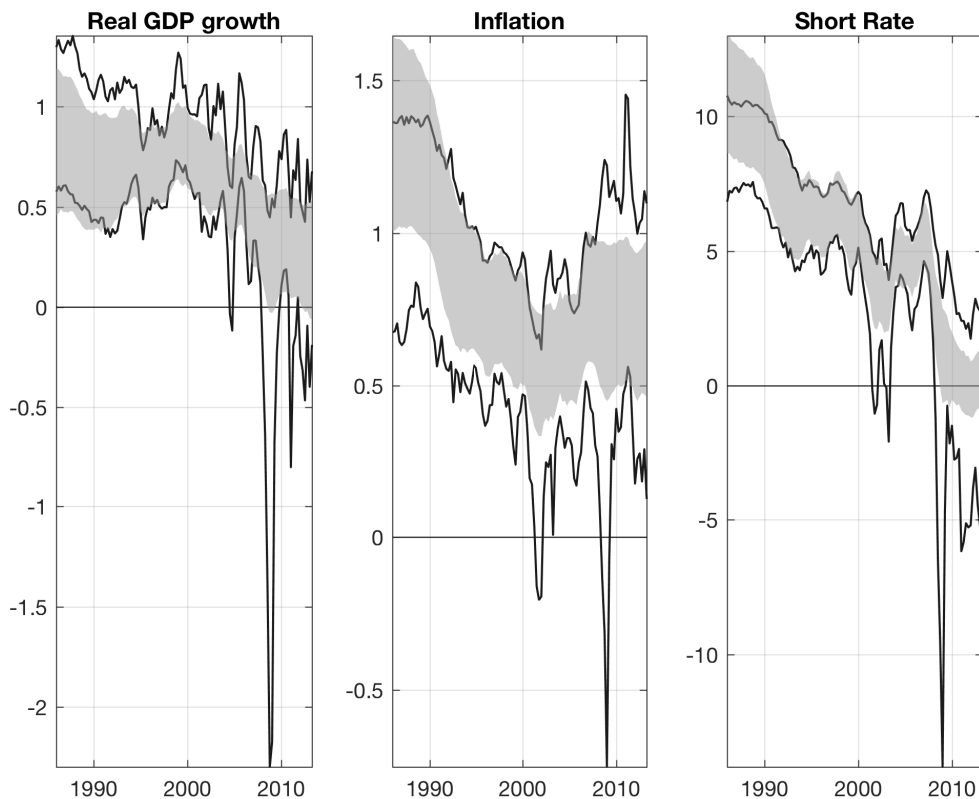


Figure 7: Infinite horizon forecasts, fixed hyperparameters in solid black lines.

are going to tie our hands by using one hyperparameter per parameter block, so that the intercepts share the same hyperparameter with all other coefficients in  $b_t$ . One could in theory use multiple fixed hyperparameters to give the fixed hyperparameter model more flexibility, but it would then not be clear what values to pick for these additional hyperparameters. The advantage of our approach is exactly that we don't have to make such choices. Nonetheless, we choose to be conservative here and use the same number of hyperparameters as in the standard fixed hyperparameter case. The prior we use for the hyperparameters we use here is the aforementioned inverse gamma prior with a prior mode of 0.05 and an infinite variance. All settings ( number of draws etc.) are the same as above. We use the same datasets for the UK and the Euro area as described above. The first sample we estimate our model on for the UK (Euro Area) start in the third quarter of 1979 (second quarter of 1970) and ends in the fourth quarter of 2004 (the same for both datasets). We then compute the posterior median forecast up to 4 quarters ahead. After

that we increase the sample size one quarter at a time and repeat the forecasting exercise until the four quarter ahead forecast is for the second quarter of 2013 (fourth quarter of 2015 for the Euro Area), which is the last data point in our sample. While we focus on point forecasts in this paper, it would be interesting to further study the forecasting performance of models estimated with our approach by looking at, for example, longer horizons or density forecasts. We leave those extensions to future work.

We show the root mean squared error of those forecasts relative to the case of hyperparameters fixed at the values used in Primiceri (2005). We also show the forecasting performance for the case the hyperparameter for  $b_t$  is set to ten times the value used in Primiceri (2005). We do not impose any eigenvalue restrictions in the forecasting exercises.

The main takeaway is that our approach improves forecasting performance relative to the standard approach using Primiceri (2005)'s hyperparameters (which were estimated on US data) for the majority of horizons except for the case of the Euro area short-term nominal rate and Euro Area real GDP growth (where our approach does better for two horizons and worse for the two others). Our approach does *substantially* better than the approach where the hyperparameter for  $b_t$  is fixed at a higher value.

One question that a researcher wanting to use our approach is facing is what priors to pick for the hyperparameters. As we showed in the Monte Carlo simulations, even with a prior that bounds the hyperparameters away from 0 (such as our benchmark inverse-Gamma prior) the model can effectively estimate constant coefficient paths if that is what the data calls for. We thus recommend using a prior that puts no mass at a hyperparameter value of 0 to avoid the possibility of a pile-up problem (as described in Stock & Watson (1996)) occurring. The pile-up problems falsely identifies no time variation, when in reality there is some time variation. With, for example, the inverse-Gamma prior, we have ruled out the pile-up problem while still allowing for no estimated time variation, as just discussed. It is worth pointing out though that in our Monte Carlo simulations we have also used priors that have a positive density at a hyperparameter value of 0 and have found no pile-up problem. Still, we recommend guarding against any possibility of this problem occurring. In practice, we have found inverse-Gamma priors with infinite variance to work well (this relieves the user from having to make a choice for the variance so that he/she can focus

Table 5: Forecast results for the Euro Area

Relative <b>RMSFE</b>		
[Out-of-sample forecast for Euro Area real GDP growth]		
Horizons	i $\mathcal{G}$	Fixed ( $\kappa_{\Omega_b} = 0.1$ )
1	0.978	0.594
2	1.059	2.759
3	1.033	2.631
4	0.863	9.420
[Out-of-sample forecast for Euro Area inflation]		
Horizons	i $\mathcal{G}$	Fixed ( $\kappa_{\Omega_b} = 0.1$ )
1	0.708	1.613
2	0.698	2.274
3	0.890	1.168
4	1.062	0.907
[Out-of-sample forecast for Euro Area nominal rate]		
Horizons	i $\mathcal{G}$	Fixed ( $\kappa_{\Omega_b} = 0.1$ )
1	1.233	1.818
2	1.305	1.736
3	1.173	2.261
4	1.142	1.597

on picking the prior mode only) - assuming instead a finite, but large, variance has led to very similar findings. Besides possibly leading to the pile-up problem, we also recommend against using a uniform prior because it is not clear what the natural parametrization of the hyperparameters is that should be used for the uniform prior (a standard problem when using uniform priors for inference). In terms of the details of our algorithm, the user needs to pick proposal densities for the Metropolis-Hastings steps for the hyperparameters. We recommend using as many draws for tuning the proposal density (i.e. adjusting its variance to meet a targeted acceptance probability) as one wants to use for the final inference. This is a conservative choice, but the algorithm is fast enough to make this not costly.

Table 6: Forecast results for the UK

Relative <b>RMSFE</b>		
[Out-of-sample forecast for UK real GDP growth]		
Horizons	i $\mathcal{G}$	Fixed ( $\kappa_{\Omega_b} = 0.1$ )
1	0.644	4.371
2	1.301	2.490
3	0.954	1.568
4	0.674	5.798
[Out-of-sample forecast for UK inflation]		
Horizons	i $\mathcal{G}$	Fixed ( $\kappa_{\Omega_b} = 0.1$ )
1	0.644	4.371
2	1.301	2.490
3	0.954	1.586
4	0.674	5.798
[Out-of-sample forecast for UK nominal rate]		
Horizons	i $\mathcal{G}$	Fixed ( $\kappa_{\Omega_b} = 0.1$ )
1	0.516	5.490
2	0.758	1.882
3	0.657	2.978
4	1.896	11.273

The online appendix contains evidence that the number of observables in the VAR can have a substantial effect on the posterior of the hyperparameters, giving further evidence that choosing the same hyperparameter values irrespective of the specification of the VAR can give misleading results. The appendix also contains additional Monte Carlo and empirical results that confirm the findings in this paper.

## 6 Conclusion

The choice of prior hyperparameters in large multivariate time series models, particularly when time-varying parameters and/or stochastic volatility are present in the models, is a daunting task. Using introspection to obtain a prior is difficult because there are many parameters. Thus, many researchers have turned to automated or semiautomated prior choices that depend only on few hyperparameters. Since those hyperparameters influence the prior distribution of large dimensional objects, their choice can be crucial. The common approach is to fix the hyperparameters at values that have been used before in the literature. We argue that, considering the number of hyperparameters is usually relatively small and considering that many applications use vastly different datasets than the applications from which they borrow the values for their hyperparameters, researchers should instead consider estimating these hyperparameters. This is especially relevant because, as we show in this paper, this estimation can be carried out with only minor changes in existing codes and at negligible computational cost (because the densities that need to be evaluated in the additional estimation step are prior distributions that are usually fast to evaluate).

We show that estimating these hyperparameters can drastically change conclusions about the amount of time variation in parameters. In the online appendix we carry out various additional robustness checks.

## References

- Amir-Ahmadi, P., Matthes, C. & Wang, M.-C. (2016), ‘Drifts and volatilities under measurement error: Assessing monetary policy shocks over the last century’, *Quantitative Economics* **7**(2), 591–611.
- Banbura, M., Giannone, D. & Reichlin, L. (2010), ‘Large Bayesian vector auto regressions’, *Journal of Applied Econometrics* **25**(1), 71–92.
- Benati, L. (2015), How Fast Can Advanced Economies Grow?, Technical report, University of Bern.
- Chan, J. C. & Eisenstat, E. (2017), ‘Bayesian model comparison for time-varying parameter VARs with stochastic volatility’, *Journal of Applied Econometrics* **forthcoming**.
- Cogley, T., Primiceri, G. E. & Sargent, T. J. (2010), ‘Inflation-Gap Persistence in the US’, *American Economic Journal: Macroeconomics* **2**(1), 43–69.
- Cogley, T. & Sargent, T. J. (2005), ‘Drift and volatilities: Monetary policies and outcomes in the post WWII U.S.’, *Review of Economic Dynamics* **8**(2), 262–302.
- D’Agostino, A., Gambetti, L. & Giannone, D. (2013), ‘Macroeconomic forecasting and structural change’, *Journal of Applied Econometrics* **28**(1), 82–101.
- D’Agostino, A. & Surico, P. (2012), ‘A century of inflation forecasts’, **94**(4).
- Del Negro, M. & Primiceri, G. (2015), ‘Time-varying structural vector autoregressions and monetary policy: a corrigendum’, *Review of Economic Studies* **82**(4).
- Doan, T., Litterman, R. B. & Sims, C. A. (1984), ‘Forecasting and conditional projection using realistic prior distribution’, *Econometric Reviews* **31**(3), 1–100.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013), *Bayesian Data Analysis*, Chapman & Hall.
- Geweke, J. (2005), *Contemporary Bayesian Econometrics and Statistics*, Wiley.

- Giannone, D., Lenza, M. & Primiceri, G. E. (2015), ‘Prior Selection for Vector Autoregressions’, *The Review of Economics and Statistics* **97**(2), 436–451.
- Koop, G. & Korobilis, D. (2010), ‘Bayesian Multivariate Time Series Methods for Empirical Macroeconomics’, *Foundations and Trends(R) in Econometrics* **3**(4), 267–358.
- Korobilis, D. (2014), Data-based priors for vector autoregressions with drifting coefficients, MPRA Paper 53772, University Library of Munich, Germany.
- Lopes, H. F., Moreira, A. R. B. & Schmidt, A. M. (1999), ‘Hyperparameter estimation in forecast models’, *Computational Statistics & Data Analysis* **29**(4), 387–410.
- Primiceri, G. (2005), ‘Time varying structural vector autoregressions and monetary policy’, *Review of Economic Studies* **72**(3), 821–852.
- Reusens, P. & Croux, C. (2017), ‘Detecting time variation in the price puzzle: An improved prior choice for time varying parameter VAR models’, *Studies in Nonlinear Dynamics and Econometrics* **forthcoming**.
- Sargent, T. J. & Surico, P. (2011), ‘Two illustrations of the quantity theory of money: Breakdowns and revivals’, *American Economic Review* **101**(1), 109–28.
- Stock, J. H. & Watson, M. W. (1996), ‘Evidence on Structural Instability in Macroeconomic Time Series Relations’, *Journal of Business & Economic Statistics* **14**(1), 11–30.